



## Оглавление

Глава 1. Введение в анализ данных.....	13
1.1. Шкала измерения.....	13
1.2. Табличные данные.....	14
1.2.1. Таблицы 2 x 2.....	14
1.2.1.1. Независимые выборки.....	14
1.2.1.2. Парные выборки.....	14
1.2.2. Двухходовые таблицы типа r x c.....	15
1.2.3. Многоходовые таблицы.....	16
1.3. Проблемы малых и больших выборок.....	17
1.4. Общая методология.....	17
1.4.1. Статистическая популяция.....	17
1.4.2. Статистическая гипотеза.....	17
1.4.3. Р-значение.....	19
1.4.4. Доверительная вероятность.....	20
1.4.5. Мощность критерия.....	21
1.4.6. Сопряженность выборок.....	21
1.4.6.1. Независимые выборки.....	21
1.4.6.2. Сопряженные выборки.....	21
Глава 2. Описательная статистика.....	22
2.1. Введение.....	22
2.2. Теоретическое обоснование.....	22
2.2.1. Численность выборки.....	23
2.2.2. Среднее значение.....	23
2.2.2.1. Общая методика.....	23
2.2.2.2. Оценка среднего на основе теории распределений.....	24
2.2.2.3. Оценка среднего на основе теории множеств.....	25
2.2.2.4. Стандартная ошибка.....	26
2.2.2.5. Дисперсия.....	26
2.2.2.6. Стандартное отклонение.....	28
2.2.2.7. Среднее отклонение.....	29
2.2.2.8. Средняя разность Джини.....	30
2.2.3. Асимметрия.....	30
2.2.4. Эксцесс.....	31
2.2.5. Коэффициент вариации.....	31
2.2.6. Минимум и максимум.....	32
2.2.6.1. Размах выборки.....	32
2.2.7. Медиана.....	32
2.2.7.1. Оценка медианы на основе теории множеств.....	33
2.2.7.2. Псевдомедиана.....	34



2.2.8. Квартили.....	34
2.2.8.1. Межквартильный размах.....	35
2.2.9. Гистограмма.....	35
2.2.9.1. Мода.....	36
2.2.9.2. Оптимальное число классов.....	36
2.2.9.2.1. Метод оптимизации числа классов.....	37
2.2.9.2.2. Метод Шимазаки–Шиномото.....	37
2.2.10. Доля.....	38
2.2.10.1. Ошибка доли.....	39
2.2.10.2. Дисперсия доли.....	40
2.2.11. Показатель точности опыта.....	40
2.2.12. Достаточная численность выборки.....	40
2.2.13. Критерий Аббе.....	41
2.2.14. Формулы для сгруппированных выборок.....	42
Глава 3. Параметрическая статистика.....	43
3.1. Введение.....	43
3.2. Теоретическое обоснование.....	44
3.2.1. Критерий Стьюдента.....	45
3.2.2. Критерий Чен.....	45
3.2.3. Критерий Стьюдента для независимых выборок.....	46
3.2.4. Парный критерий Стьюдента.....	46
3.2.5. Критерий Лорда.....	47
3.2.6. Критерий Уэлча.....	47
3.2.7. Критерий Пагуровой.....	48
3.2.8. Критерий Кокрена–Кокса.....	49
3.2.9. Критерий Крамера.....	49
3.2.10. Критерий Фишера.....	50
3.2.11. Трансгрессия.....	50
3.2.12. График средних значений с доверительными интервалами.....	51
3.2.13. Отношения средних и дисперсий.....	52
Глава 4. Непараметрическая статистика.....	53
4.1. Введение.....	53
4.2. Теоретическое обоснование.....	54
4.2.1. Робастность.....	54
4.2.2. Тестируемые параметры.....	55
4.2.3. Типы критериев.....	56
4.2.3.1. Ранговые критерии.....	56
4.2.3.1.1. Учет связей.....	59
4.2.3.1.2. Учет поправки на непрерывность.....	59
4.2.3.1.3. Критерий Вилкоксона для независимых выборок.....	59
4.2.3.1.4. Критерий Вилкоксона для связанных выборок.....	60



4.2.3.1.5. Критерий Манна–Уитни.....	61
4.2.3.1.6. Критерий Ван дер Вардена.....	62
4.2.3.1.7. Критерий Сэвиджа.....	63
4.2.3.1.8. Критерий Ансари–Бредли.....	63
4.2.3.1.9. Критерий Клотца.....	65
4.2.3.1.10. Критерий Зигеля–Тьюки.....	65
4.2.3.1.11. Критерий Коновера.....	66
4.2.3.1.12. Критерий Муда–Брауна.....	67
4.2.3.2. Критерии на основе сравнения функций распределения.....	68
4.2.3.2.1. Критерий Смирнова.....	68
4.2.3.2.2. Критерий Лемана–Розенблатта.....	69
4.2.3.2.3. Критерий Койпера.....	70
4.2.3.2.4. Критерий Мак–Немара.....	70
4.2.3.2.5. Критерий хи–квадрат.....	71
4.2.3.2.6. Критерий медианы.....	72
4.2.3.3. Критерий серий Вальда–Вольфовица.....	73
4.2.4. Таблицы 2 x 2.....	73
4.2.4.1. Относительный риск.....	74
4.2.4.2. Отношение шансов.....	75
4.2.4.3. Разность долей.....	75
4.2.4.3.1. Разность долей в таблице независимых признаков.....	76
4.2.4.3.2. Разность долей в таблице связанных признаков.....	76
4.2.4.4. Прогностичность.....	77
4.2.4.4.1. Чувствительность.....	78
4.2.4.4.2. Специфичность.....	79
4.2.4.4.3. Распространенность.....	79
4.2.4.4.4. Прогностичность положительного результата.....	80
4.2.4.4.5. Прогностичность отрицательного результата.....	80
4.2.5. График медиан с доверительными интервалами.....	81
4.2.6. График долей с доверительными интервалами.....	82
4.2.7. ROC анализ.....	82
4.2.8. Каппа Козна.....	86
Глава 5. Точные критерии.....	87
5.1. Введение.....	87
5.2. Теоретическое обоснование.....	88
5.2.1. Критерий рандомизации для независимых выборок.....	89
5.2.2. Критерий рандомизации для связанных выборок.....	90
5.2.3. Критерий Вилкоксона для независимых выборок.....	91
5.2.4. Критерий Вилкоксона для связанных выборок.....	91
5.2.5. Точный метод Фишера.....	91
5.2.6. Критерий Барнарда.....	93



5.2.7. Критерий Мак–Немара.....	95
5.2.8. Критерий знаков.....	96
5.2.9. Критерий серий Вальда–Вольфовица.....	96
Глава 6. Кросстабуляция.....	97
6.1. Введение.....	97
6.2. Теоретическое обоснование.....	97
6.2.1. Критерий Кресси–Рида.....	100
6.2.2. Критерий Хеллингера.....	101
6.2.3. Критерий хи–квадрат.....	101
6.2.4. Критерий отношения правдоподобия.....	102
6.2.5. Критерий Зелтермана.....	103
6.2.6. Критерий Фримана–Холтона.....	104
6.2.7. Критерий Стюарта–Максвелла.....	105
6.2.8. Критерий Баукера.....	105
6.2.9. Критерий Бхапкара.....	106
6.2.10. Коэффициент Кендалла.....	106
6.2.11. Коэффициент Крамера.....	107
6.2.12. Коэффициент Сомерса.....	108
6.2.13. Коэффициент сопряженности Пирсона.....	109
6.2.14. Критерий Краскела–Уоллиса.....	109
6.2.15. Диагностика Симонов–Цай.....	110
6.2.16. Диагностика Хабермана.....	111
Глава 7. Проверка нормальности распределения.....	111
7.1. Введение.....	111
7.2. Теоретическое обоснование.....	112
7.2.1. Процедура тестирования.....	113
7.2.2. Типы тестов на нормальность.....	114
7.2.2.1. Простые и сложные гипотезы.....	114
7.2.3. Критерии функций распределения.....	115
7.2.3.1. Критерии типа Колмогорова.....	116
7.2.3.1.1. Критерий Колмогорова.....	117
7.2.3.1.2. Модифицированный критерий Колмогорова.....	117
7.2.3.1.3. Модифицированный критерий Смирнова.....	118
7.2.3.2. Критерии типа омега–квадрат.....	119
7.2.3.2.1. Критерий Крамера–Мизеса.....	120
7.2.3.2.2. Критерий Андерсона–Дарлинга.....	120
7.2.3.2.3. Критерий хи–квадрат Фишера.....	121
7.2.3.3. Критерии типа Эппса–Палли.....	123
7.2.3.3.1. Критерий Эппса–Палли.....	124
7.2.3.3.2. Критерий Хенце–Цирклера.....	124
7.2.4. Критерии, основанные на регрессии.....	125



7.2.4.1. Критерий Шапиро–Уилка.....	125
7.2.4.2. Критерий Шапиро–Франсиса.....	126
7.2.4.3. Критерий Д’Агостино.....	127
7.2.5. Критерии моментов.....	128
7.2.5.1. Критерий коэффициента асимметрии.....	129
7.2.5.2. Критерий эксцесса.....	130
7.2.5.3. Критерий Жарка–Бера.....	131
7.2.5.4. Критерий Гири.....	132
7.2.5.5. Критерий асимметрии Мардиа.....	132
7.2.5.6. Критерий эксцесса Мардиа.....	133
7.2.6. Информационные критерии.....	133
7.2.6.1. Критерий Васичека.....	134
7.2.7. Графические методы.....	134
7.2.7.1. Глазомерный метод.....	135
Глава 8. Дисперсионный анализ.....	135
8.1. Введение.....	135
8.2. Теоретическое обоснование.....	135
8.2.1. Дисперсионный анализ.....	135
8.2.1. Однофакторный дисперсионный анализ.....	137
8.2.1.1. Однофакторный дисперсионный анализ.....	138
8.2.1.1.2. Однофакторный дисперсионный анализ (повторные измерения).....	138
8.2.1.1.4. Критерий Данна.....	139
8.2.1.1.3. Ранговый однофакторный анализ Краскела и Уоллиса.....	140
8.2.1.1.5. Критерий Коновера.....	141
8.2.1.1.6. Критерий Джонкхиера и Терпстра.....	141
8.2.1.1.7. Критерий Бартлетта.....	142
8.2.1.6. Критерий G Кокрена.....	143
8.2.1.1.9. Критерий Шеффе.....	143
8.2.1.1.10. Критерий Дункана.....	144
8.2.1.1.11. Критерий Тьюки.....	145
8.2.1.1.12. Критерий Ливена.....	146
8.2.1.1.13. Критерий Брауна–Форсайта.....	146
8.2.1.1.14. Критерий V Бхапкара.....	147
8.2.1.1.15. Критерий D Дешпанде.....	148
8.2.1.1.16. Критерий L Дешпанде.....	148
8.2.1.2. Многофакторный дисперсионный анализ.....	149
8.2.1.2.1. Двухфакторный дисперсионный анализ.....	149
8.2.1.2.2. Ранговый критерий Фридмана.....	150
8.2.1.2.3. Критерий Квейд.....	151
8.2.1.2.4. Критерий Пэйджа.....	152
8.2.1.2.5. Критерий Q Кокрена.....	152



8.2.1.2.6. Критерий Шеффе для связанных выборок.....	153
8.2.2. Множественные сравнения.....	154
8.2.2.1. Критерий Хотеллинга.....	155
8.2.2.2. Критерий Джеймса–Сю.....	156
8.2.2.3. Критерий Кульбака.....	156
8.2.2.4. Критерий Пури–Сена–Тамура.....	157
8.2.2.5. Критерий Пури–Сена.....	157
8.2.2.6. Критерий Шейрера–Рэя–Хэйра.....	158
8.2.2.7. Критерий Уилкса.....	159
8.2.3. Ковариационный анализ.....	160
8.2.3.1. Однофакторный ковариационный анализ.....	160
Глава 9. Регрессионный анализ.....	164
9.1. Введение.....	164
9.2. Теоретическое обоснование.....	164
9.2.1. Оценка качества аппроксимации.....	164
9.2.2. Регрессионный анализ.....	166
9.2.3. Метод наименьших квадратов.....	167
9.2.4. Полиномиальные модели.....	168
9.2.5. Экспоненциально–степенная аппроксимация.....	169
9.2.6. Логарифмическая функция.....	170
9.2.7. Логистический анализ.....	170
9.2.8. Пользовательская функция.....	171
9.2.8.1. Метод Бройдена–Флетчера–Голдфарба–Шанно.....	171
9.2.8.2. Метод Гаусса– Ньютона.....	172
9.2.9. Кусочно–линейная аппроксимация.....	173
Глава 10. Корреляционный анализ.....	174
10.1. Введение.....	174
10.2. Теоретическое обоснование.....	174
10.2.1. Корреляция количественных признаков.....	174
10.2.1. Коэффициент корреляционного отношения Пирсона.....	175
10.2.1.2. Коэффициент корреляции Фехнера.....	177
10.2.1.3. Ковариация.....	178
10.2.2. Корреляция порядковых признаков.....	179
10.2.2.1. Показатель ранговой корреляции Спирмэна.....	179
10.2.2.2. Коэффициент ранговой корреляции Кендалла.....	180
10.2.3. Корреляция номинальных признаков.....	182
10.2.3.1. Коэффициент Рассела–Рао.....	182
10.2.3.2. Коэффициент сопряженности Бравайса.....	183
10.2.4. Корреляция признаков, измеренных в различных шкалах.....	183
10.2.4.1. Коэффициент Гауэра.....	184
10.2.4.1. Расчет вклада признаков.....	184



10.2.4.2. Точечно–бисериальная корреляция.....	185
10.2.5. Корреляция разнородных признаков.....	186
10.2.6. Канонический корреляционный анализ.....	187
Глава 11. Факторный анализ.....	187
11.1. Введение.....	187
11.2. Теоретическое обоснование.....	187
11.2.1. Метод главных факторов.....	190
11.2.1.1. Компонентный анализ.....	190
11.2.1.2. Факторный анализ методом главных факторов.....	191
11.2.1.3. Проблема общности.....	192
11.2.1.4. Проблема факторов.....	193
11.2.1.5. Измерение факторов.....	193
11.2.2. Метод максимума правдоподобия.....	194
11.2.3. Проблема вращения.....	195
11.2.4. Критерии максимального числа факторов.....	196
11.2.4.1. Адекватность метода главных факторов.....	196
11.2.4.2. Значимость числа факторов метода максимума правдоподобия.....	196
Глава 12. Кластерный анализ.....	197
12.1. Введение.....	197
12.2. Теоретическое обоснование.....	197
12.2.1. Меры различия.....	198
12.2.1.1. Евклидово расстояние.....	199
12.2.1.2. Манхеттенское расстояние.....	199
12.2.1.3. Супремум–норма.....	199
12.2.1.4. Расстояние Махаланобиса.....	200
12.2.1.5. Расстояние Пирсона.....	201
12.2.1.6. Расстояние Спирмэна.....	201
12.2.1.7. Расстояние Кендалла.....	202
12.2.1.8. Расстояние Жаккара.....	202
12.2.1.9. Расстояние Рассела–Рао.....	202
12.2.1.10. Расстояние Бравайса.....	203
12.2.1.11. Расстояние Юла.....	203
12.2.1.12. Расстояние отношений.....	203
12.2.2. Метод средней связи Кинга.....	204
12.2.3. Метод Уорда.....	205
12.2.4. Метод k–средних Мак–Куина.....	205
12.2.5. Модифицированный метод k–средних.....	206
12.2.6. Графическое представление результатов кластерного анализа.....	207
Глава 13. Информационный анализ.....	207
13.1. Введение.....	207
13.2. Теоретическое обоснование.....	207



13.2.1. Число классов.....	208
13.2.2. Число вариант ряда.....	208
13.2.3. Энтропия.....	209
13.2.4. Дисперсия энтропии.....	210
13.2.5. Максимальная энтропия.....	211
13.2.6. Относительная энтропия.....	211
13.2.7. Избыточность.....	212
13.2.8. Организация системы.....	212
13.2.9. Примеры информационного анализа.....	212
13.2.9.1. Разведочный информационный анализ.....	212
13.2.9.2. Исследование структурной перестройки объекта.....	213
13.2.9.2. Сравнение групп по индексам межвидового разнообразия.....	214
Глава 14. Распознавание образов с обучением.....	215
14.1. Введение.....	215
14.2. Теоретическое обоснование.....	215
14.2.1. Оценка качества моделей.....	216
14.2.1. Количественные классификаторы.....	216
14.2.1.2. Бинарные классификаторы.....	216
14.2.2. Оценка значимости модели.....	217
14.2.2.1. Статистика Вальда.....	217
14.2.2.2. Статистика G.....	218
14.2.3. Линейный дискриминантный анализ Фишера.....	218
14.2.4. Канонический дискриминантный анализ.....	219
14.2.5. Линейный дискриминантный анализ.....	220
14.2.6. Линейный множественный регрессионный анализ.....	220
14.2.6.1. Обработка выбросов.....	224
14.2.6.2. Выявление влияющих наблюдений.....	224
14.2.6.3. Автокорреляция остатков.....	225
14.2.7. Логистическая регрессия.....	227
14.2.8. Пробит анализ.....	228
14.2.9. Регрессия Пуассона.....	230
14.2.10. Оценка прогностической ценности параметров.....	231
Глава 15. Многомерное шкалирование.....	232
15.1. Введение.....	232
15.2. Теоретическое обоснование.....	232
15.2.1. Метрики.....	232
15.2.1. Метрика Минковского.....	233
15.2.1.2. Евклидова метрика.....	234
15.2.1.3. Манхеттенское расстояние.....	234
15.2.2. Метрический метод Торгерсона.....	234
15.2.3. Неметрический метод Краскала.....	236





15.2.4. Проблема вращения.....	238
Глава 16. Обработка экспертных оценок.....	239
16.1. Введение.....	239
16.2. Теоретическое обоснование.....	239
16.2.1. Парные сравнения.....	241
16.2.2. Групповое оценивание.....	242
16.2.3. Коэффициент конкордации.....	243
16.2.4. Метод средних рангов.....	243
16.2.5. Медиана Кемени.....	244
16.2.6. Среднее Кемени.....	244
16.2.7. Альфа Кронбаха.....	245
Глава 17. Анализ выживаемости.....	246
17.1. Введение.....	246
17.2. Теоретическое обоснование.....	246
17.2.1. Функция выживания.....	247
17.2.2. Функция риска.....	248
17.2.3. Оценка параметра положения.....	249
17.2.4. Подбор распределения.....	249
17.2.4.1. Общая методика.....	250
17.2.4.2. Логарифмические модели.....	251
17.2.4.2.1. Логнормальное распределение.....	252
17.2.4.2.2. Логлогистическое распределение.....	253
17.2.4.3. Гамма– распределение.....	254
17.2.4.4. Распределение Вейбулла.....	255
17.2.4.5. Экспоненциальное распределение.....	256
17.2.4.6. Распределение Рэлея.....	256
17.2.4.7. Распределение Гомпертца.....	257
17.2.4.8. Оценка качества подгонки модели.....	258
17.2.5. Критерий Кокса.....	259
17.2.6. Критерий Гехана.....	259
17.2.7. Модель пропорциональных рисков Кокса.....	260
Глава 18. Анализ временных рядов и прогнозирование.....	263
18.1. Введение.....	263
18.2. Теоретическое обоснование.....	263
18.2.1. Метод скользящего среднего.....	264
18.2.2. Сезонный разностный оператор.....	265
18.2.3. Сингулярный спектральный анализ.....	266
18.2.3.1. Вложение.....	266
18.2.3.2. Разложение по сингулярным числам.....	266
18.2.3.3. Восстановление.....	267
18.2.4. Гармонический анализ Фурье.....	267



18.2.5. Автокорреляционная функция.....	268
18.2.6. Периодограмма.....	268
Глава 19. Статистический контроль качества.....	269
19.1. Введение.....	269
19.2. Теоретическое обоснование.....	269
19.2.1. Гистограмма качества.....	270
19.2.2. Диаграмма Парето.....	271
19.2.3. Контрольная карта.....	272
19.2.4. Анализ Бланда–Альтмана.....	273
Глава 20. Обработка пропущенных данных.....	275
20.1. Введение.....	275
20.2. Теоретическое обоснование.....	275
20.2.1. Игнорирование пропусков.....	275
20.2.2. Заполнение средним значением.....	276
20.2.3. Заполнение регрессионными значениями.....	276
20.2.4. Заполнение случайными значениями.....	278
Глава 21. Обработка выбросов.....	279
21.1. Введение.....	279
21.2. Теоретическое обоснование.....	279
21.2.1. Критерий Смирнова–Граббса.....	280
21.2.2. Критерий Титъена–Мура.....	281
21.2.3. Правило Томпсона.....	282
21.2.4. Критерий Диксона.....	282
21.2.5. Критерий Дина–Диксона.....	283
21.2.6. Критерий Шовене.....	283
21.2.7. Правило «ящик с усами».....	284
21.2.8. Критерий Кокрена.....	284
Глава 22. Рандомизация и генерация случайных последовательностей.....	285
22.1. Введение.....	285
22.2. Теоретическое обоснование.....	286
22.2.1. Рандомизация в биомедицинских исследованиях.....	286
22.2.2. Генерация случайных последовательностей.....	287
22.2.2.1. Стандартный генератор ANSI.....	287
22.2.2.2. Мультипликативный линейный конгруэнтный датчик.....	288
Глава 23. Преобразования данных.....	288
23.1. Введение.....	288
23.2. Теоретическое обоснование.....	288
23.2.1. Одномерное преобразование.....	289
23.2.1.1. Преобразование Бокса–Кокса.....	290
23.2.1.2. Преобразование Зеллнера–Реванкара.....	290
23.2.1.3. Преобразование гиперболического арксинуса.....	290



23.2.1.4. Преобразование Йео–Джонсона.....	291
23.2.1.5. Преобразование Джона–Дрейпера.....	291
23.2.1.6. Преобразование Манли.....	292
23.2.2. Многомерное преобразование.....	292
23.2.2.1. Многомерное преобразование Бокса–Кокса.....	293
Глава 24. Матричная и линейная алгебра.....	294
24.1. Введение.....	294
24.2. Теоретическое обоснование.....	294
24.2.1. Транспонирование матрицы.....	294
24.2.2. Сложение матриц.....	294
24.2.3. Произведение матриц.....	295
24.2.4. Обратная матрица.....	295
24.2.5. Определитель матрицы.....	296
24.2.6. Умножение матрицы на скаляр.....	296
24.2.7. Псевдообратная матрица.....	297
24.2.8. Решение системы линейных уравнений.....	297
24.2.9. Стандартная проблема собственных значений.....	298
24.2.10. Обобщенная проблема собственных значений.....	298
24.2.11. Разложение Холецкого.....	298
24.2.12. Разложение Краута.....	299
24.2.13. Разложение QR.....	300
24.2.14. Разложение по сингулярным числам.....	300
24.2.15. Мультиколлинеарность.....	301
24.2.15.1. Корреляция между параметрами.....	301
24.2.15.2. Коэффициенты детерминации векторов.....	301
24.2.15.2. Частные коэффициенты корреляции.....	302
24.2.16. Кронекеровское произведение.....	302
Глава 25. Обыкновенные дифференциальные уравнения.....	303
25.1. Введение.....	303
25.2. Теоретическое обоснование.....	303
25.2.1. Математическое моделирование.....	305
25.2.2. Основные предположения.....	307
25.2.3. Устойчивость.....	308
25.2.3.1. Жесткие задачи.....	308
25.2.3.2. Устойчивость решения.....	309
25.2.4. Численное решение дифференциальных уравнений.....	309
25.2.4.1. Одношаговые методы.....	310
25.2.4.1.1. Явные схемы.....	311
25.2.4.1.2. Неявные схемы.....	311
25.2.4.1.3. Метод Рунге–Кутты.....	311
25.2.4.1.4. Методы Мерсона.....	312



25.2.4.1.5. Метод Хаммера–Холлингсуорта.....	313
25.2.4.2. Многошаговые методы.....	314
25.2.4.2.1. Метод Адамса.....	314
25.2.4.2.2. Методы Гира.....	314
Глава 26. Многочлены.....	315
26.1. Введение.....	315
26.2. Теоретическое обоснование.....	315
26.2.1. Многочлены Бернулли.....	315
26.2.2. Многочлены Лагерра.....	316
26.2.3. Многочлены Эрмита.....	316
26.2.4. Многочлены Чебышева.....	316
26.2.5. Многочлены Лежандра.....	317
Приложение. Статистические распределения.....	318
П.1. Биномиальное распределение.....	318
П.2. Гипергеометрическое распределение.....	318
П.3. Нормальное распределение.....	319
П.4. Многомерное нормальное распределение.....	320
П.5. $t$ -распределение.....	320
П.6. $F$ -распределение.....	321
П.7. Бета-распределение.....	321
П.8. Хи-квадрат распределение.....	321
П.9. Нецентральное хи-квадрат распределение.....	321
П.10. Обобщенное гамма-распределение.....	322
П.11. Логнормальное распределение.....	322
П.12. Распределение $S_U$ Джонсона.....	323
П.13. Распределение выборочного размаха.....	323
П.14. Распределение студентизированного размаха.....	324
П.15. Распределение студентизированного максимума модулей.....	324
П.16. Распределение статистики критерия Колмогорова.....	325
П.17. Распределение статистики критерия Койпера.....	325
П.18. Распределения статистик критериев Вилкоксона.....	326
П.19. Распределение статистики критерия Манна–Уитни.....	326
П.20. Распределение статистики критериев типа омега-квадрат.....	327
П.21. Маргинальные распределения.....	327
П.22. Специальные функции.....	327



## Глава 1. Введение в анализ данных

Руководство по методам математической и статистической библиотеки прикладных программ ME.com (Mathematical & Engineering components) содержит теоретическое обоснование<sup>1</sup>, необходимое для понимания, изучения и использования данного программного обеспечения в различных проектах.

### 1.1. Шкала измерения

Принадлежность признака шкале измерения основана на допустимости логических и арифметических операций, которые могут быть проведены над признаком, как показано в таблице.

Шкала	Допустимые операции
Номинальная	Различение
Порядковая	Различение, сравнение
Количественная	Различение, сравнение, сложение, умножение

Классификация включает признаки:

1. Номинальные – качественные признаки с неупорядоченными состояниями. Номинальные признаки могут быть оцифрованы, однако смысла эти цифры, за исключением возможности различать признаки между собой, не имеют. Частным случаем номинальных признаков являются бинарные (дихотомические) признаки, представляющие собой номинальные признаки с двумя градациями.
2. Порядковые – качественные признаки с упорядоченными состояниями. Порядок состояний имеет смысл, признаки могут быть осмысленно оцифрованы и сравниваться между собой, однако расстояния между ними не определены. К порядковой шкале относится также шкала ранжировок.
3. Количественные (численные), подразделяемые на интервальные и относительные признаки. Они различаются положением нулевой отметки на шкале измерения. Численные признаки определяют измеряемые величины и являются истинными количественными, причем могут измеряться как непрерывные, так и целочисленные признаки.
4. Фиктивные (индикаторные) переменные – вспомогательные бинарные переменные, принимающие значения только 1 либо 0, которые применяются для введения в регрессионные модели качественных переменных.

Рассмотрим возможное кодирование качественных переменных фиктивными, что необходимо для обеспечения участия качественных переменных в количественных расчетах наряду с истинно количественными переменными. Если качественная переменная принимает  $S$  фиксированных значений, то теоретически она может быть закодирована  $N$  фиктивными переменными, где минимальное значение  $N$ , очевидно, определяется из целочисленного

<sup>1</sup> Ссылки на использованные и рекомендуемые источники вынесены в отдельный файл.



неравенства  $S \leq 2^N$ . Возможны и иные варианты кодировки. Например, можно качественную переменную с  $S$  фиксированными значениями закодировать  $S$  фиктивными переменными. Шкалы могут приводиться одна к другой: количественная шкала – к порядковой шкале или номинальной, порядковая шкала – к номинальной шкале. Обратные операции считаются некорректными.

См. учебное пособие Борцова.

## 1.2. Табличные данные

### 1.2.1. Таблицы 2 x 2

Двухходовые таблицы сопряженности типа 2 x 2 возникают в результате сопоставления двух бинарных (дихотомических) выборов, т. е. выборов, состоящих из значений 1 и 0, причем под значением 1 понимают наличие признака, под значением 0 понимают отсутствие признака.

Выборки рассматриваемого могут быть представлены в виде таблиц типа 2 x 2 различными способами, в зависимости от того, являются ли выборки независимыми или парными. Ниже представлены способы получения таблиц 2 x 2 и указаны их существенные особенности.

#### 1.2.1.1. Независимые выборки

Порядок построения таблицы из вариант независимых выборов иллюстрируется таблицей:

	Наличие эффекта	
	Да	Нет
Выборка 1	$a$	$b$
Выборка 2	$c$	$d$

При этом в ячейки заносятся:

$a$  – число значений с эффектом первой выборки,

$b$  – число значений без эффекта первой выборки,

$c$  – число значений с эффектом второй выборки,

$d$  – число значений без эффекта второй выборки.

Таблицы данного типа могут применяться при анализе данных типа «опыт – контроль» или сравнении двух независимых методов воздействия типа «группа 1 – группа 2».

#### 1.2.1.2. Парные выборки

Для парных (сопряженных) выборов порядок построения таблицы иллюстрируется таблицей:

		Эффект В	
		Да	Нет
Эффект А	Да	$a$	$b$
	Нет	$c$	$d$



В данном случае анализу подвергается фактически одна двумерная выборка – выборка пар значений, первое значение пары – наличие или отсутствие эффекта  $A$ , второе – наличие или отсутствие эффекта  $B$ . Поэтому в ячейки таблицы заносятся:

$a$  – число пар значений с эффектом  $A$  и с эффектом  $B$ ,

$b$  – число пар значений с эффектом  $A$  и без эффекта  $B$ ,

$c$  – пар число значений без эффекта  $A$  и с эффектом  $B$ ,

$d$  – пар число значений без эффекта  $A$  и без эффекта  $B$ .

Таблицы данного типа могут применяться при анализе данных типа «до — после».

## 1.2.2. Двухходовые таблицы типа $r \times c$

Пусть обозначено:

$r$  – число градаций первого признака,

$c$  – число градаций второго признака,

$n_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – численность вариантов, обладающих одновременно  $i$ -й градацией первого признака и  $j$ -й градацией второго признака.

Тогда таблица сопряженности будет иметь вид:

$n_{11}$	$n_{12}$	$\dots$	$n_{1c}$
$n_{21}$	$n_{22}$	$\dots$	$n_{2c}$
$\dots$	$\dots$	$\dots$	$\dots$
$n_{r1}$	$n_{r2}$	$\dots$	$n_{rc}$

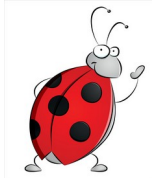
Порядок признаков (столбцы или строки) значения не имеет. При анализе таблицы сопряженности условились количество строк таблицы обозначать символом  $r$  (от английского слова rows), а количество столбцов – символом  $c$  (от английского слова columns), хотя могут встречаться и любые другие обозначения. Двумерная таблица сопряженности будет именоваться  $r \times c$  или  $R \times C$  таблицей. Каждая клетка таблицы сопряженности с индексами  $i$ ,  $i = 1, 2, \dots, r$ , (номер строки) и  $j$ ,  $j = 1, 2, \dots, c$ , (номер столбца) представляет собой количество индивидуумов, обладающих одновременно градацией  $i$  первого признака и градацией  $j$  второго признака. Данное количество называется наблюдаемой (наблюденной) частотой встречаемости признаков. Методами кросстабуляции исследуется зависимость первого номинального признака с числом градаций  $r$  от второго номинального признака с числом градаций  $c$ . Если таблица сопряженности квадратная (числа градаций для первого и второго признаков одинаковы), то часто используется обозначение: таблица типа  $k \times k$ , где  $k$  – число градаций каждого признака.

## 1.2.3. Многоходовые таблицы

Многоходовые таблицы сопряженности возникают, когда число признаков превышает 2.

Сначала для пояснения принципа обозначений рассмотрим трехходовую таблицу, а затем обобщим результаты на таблицы сопряженности произвольной размерности.

Введем новые обозначения. Пусть  $k_i$ ,  $i = 1, 2, 3$  – число градаций  $i$ -го признака,



(.) – обозначение фиксированного уровня 3-го признака. Тогда таблица сопряженности для признаков 1 и 2 при фиксированном  $k_3 = 1$  имеет вид двухвходовой таблицы:

$n_{11}^{(1)}$	$n_{12}^{(1)}$	...	$n_{1k_2}^{(1)}$
$n_{21}^{(1)}$	$n_{22}^{(1)}$	...	$n_{2k_2}^{(1)}$
...	...	...	...
$n_{k_1 1}^{(1)}$	$n_{k_1 2}^{(1)}$	...	$n_{k_1 k_2}^{(1)}$

Действуя аналогично, получаем и все остальные таблицы:

$n_{11}^{(2)}$	$n_{12}^{(2)}$	...	$n_{1k_2}^{(2)}$	$n_{11}^{(\dots)}$	$n_{12}^{(\dots)}$	...	$n_{1k_2}^{(\dots)}$	$n_{11}^{(k_3)}$	$n_{12}^{(k_3)}$	...	$n_{1k_2}^{(k_3)}$
$n_{21}^{(1)}$	$n_{22}^{(1)}$	...	$n_{2k_2}^{(1)}$	$n_{21}^{(\dots)}$	$n_{22}^{(\dots)}$	...	$n_{2k_2}^{(\dots)}$	$n_{21}^{(k_3)}$	$n_{22}^{(k_3)}$	...	$n_{2k_2}^{(k_3)}$
...	...	...	...	...	...	...	...	...	...	...	...
$n_{k_1 1}^{(2)}$	$n_{k_1 2}^{(2)}$	...	$n_{k_1 k_2}^{(2)}$	$n_{k_1 1}^{(\dots)}$	$n_{k_1 2}^{(\dots)}$	...	$n_{k_1 k_2}^{(\dots)}$	$n_{k_1 1}^{(k_3)}$	$n_{k_1 2}^{(k_3)}$	...	$n_{k_1 k_2}^{(k_3)}$

Трехвходовая таблица сопряженности представляет собой своеобразный «куб» со сторонами  $k_1 \times k_2 \times k_3$ . Хорошо заметно, насколько громоздко и неудобно такое представление данных для многовходовых таблиц, поэтому было принято другое представление, фактически отражающее ту же самую сущность, т. е. представления взаимозаменяемы. Данное эквивалентное табличное представление многовходовых таблиц (для рассмотренного примера) представлено ниже. Благодаря тем же самым обозначениям понятен порядок построения табличной формы многовходовой таблицы:

- Первый столбец таблицы – градации признака 1.
- Второй столбец таблицы – градации признака 2.
- Третий столбец таблицы – градации признака 3.
- Последний столбец любой таблицы – это количества индивидуумов (частоты), обладающих одновременно градациями признаков, перечисленных в строке, соответствующей данной частоте.

Таблица представляет собой все возможные сочетания градаций признаков  $k_i$ ,  $i = 1, 2, \dots$ , и

соответствующие им частоты. Поэтому размеры таблицы будут  $\prod_{i=1}^n k_i$  строк на  $n + 1$  столбцов, где  $n$  – количество изучаемых признаков. Рассмотренное представление позволяет изобразить на «плоскости» таблицы сопряженности произвольной размерности.

### 1.3. Проблемы малых и больших выборок

Проблемы малых и больших выборок относятся к основным проблемам, возникающим при практическом применении методов анализа данных. Можно предложить такую классификацию выборок по численности, исходя из требований представленных критериев:

- очень малые выборки – от 5 до 12,
- малые выборки – от 13 до 40,





- выборки средней численности – от 41 до 100,
- большие выборки – от 101 и выше.

Максимальная численность выборки лимитируется повышенной трудоемкостью вычисления статистики критерия, особенно, если в схеме его вычисления применяются комбинаторные алгоритмы. При больших численностях выборок становится оправданным применение менее трудоемких в вычислении тестов, в том числе параметрических.

## 1.4. Общая методология

Под математической статистикой понимают практическое приложение достижений теории вероятностей. Теория вероятности носит всеобщий характер безотносительно к физической природе явления. Поэтому методы математической статистики одинаковы для изучения любой объективной реальности. Статистическое исследование включает шаги:

1. Разработка дизайна.
2. Сбор данных.
3. Математико–статистическая обработка.
4. Выводы, рекомендации и прогноз.

### 1.4.1. Статистическая популяция

Предметом статистических исследований является генеральная совокупность (статистическая популяция), о параметрах которой делается предположение на основании репрезентативной эмпирической выборки (выборочной совокупности) из популяции. Статистической популяцией называется совокупность всех объектов одного класса, различия между которыми определяются только случайными факторами.

### 1.4.2. Статистическая гипотеза

Статистической гипотезой  $H_0$  называется утверждение, в котором предполагается, что истинное распределение вероятностей принадлежит подмножеству семейства возможных вероятностных распределений. Проверяемая гипотеза  $H_0$  называется нулевой гипотезой. Альтернативной (конкурирующей) гипотезой  $H_1$  называется отрицание нулевой гипотезы. Пусть, например, статистический критерий проверяет нулевую гипотезу  $H_0$  о равенстве («нет статистически значимого различия») функций распределения двух выборочных совокупностей  $F(x) = G(x)$ . Альтернативная гипотеза  $H_1$  в данном случае может быть сформулирована одним из трех способов:

1.  $F(x) \neq G(x)$  – «нулевая гипотеза неверна» – это двусторонняя (two-tailed, two-sided) гипотеза;
2.  $F(x) < G(x)$  – это односторонняя (upper-tailed) гипотеза;
3.  $F(x) > G(x)$  – это односторонняя (low-tailed) гипотеза.

Критерий  $T$  проверки статистической гипотезы  $H_0$  есть процедура выработки решения о том, принять или отклонить данную нулевую гипотезу. Критической областью (областью



непринятия нулевой гипотезы)  $U$  является та часть выборочного пространства, которая приводит к отклонению гипотезы  $H_0$ .

Уровнем значимости критерия является вероятность  $\alpha$  того, что этот критерий приведет к отклонению нулевой гипотезы в случае ее истинности  $P(T \in U) = \alpha$ . Если результаты проверки находятся в критической области  $P(T > T_\alpha) < \alpha$ , нулевая гипотеза отклоняется и принимается альтернативная гипотеза. Здесь критическому значению критерия соответствует уровень значимости  $\alpha$ .

Отклонение нулевой гипотезы в случае ее истинности называется ошибкой I (первого) рода.

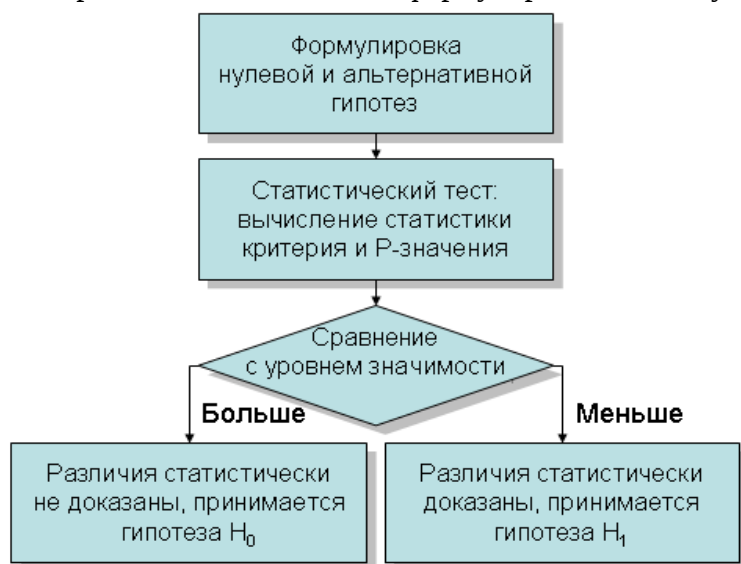
Принятие нулевой гипотезы, когда она не верна, называется ошибкой II (второго) рода.

Вероятность ошибки второго рода обозначается  $\beta$ .

С целью унификации статистических таблиц и стандартизации выводов уровень значимости выбирается из стандартной линейки типа 0,001; 0,005; 0,01; 0,05 ..., либо то же в процентах. Величина уровня значимости зависит от важности предметной области (см. раздел о доверительной вероятности). Чем проводятся исследования более важные (в биомедицине и смежных дисциплинах – более социально значимые), тем меньшим уровнем значимости следует оперировать.

На схеме показан алгоритм действий при практическом решении задачи проверки гипотезы.

Пусть нулевая гипотеза  $H_0$  сформулирована как «нет статистически значимого различия», а альтернативная гипотеза  $H_1$  сформулирована как «нулевая гипотеза неверна».



Результатом статистической проверки является вывод о том, в скольких случаях, например, на каждые 100 проведенных испытаний отклонения можно считать случайными. На заданном стандартном уровне значимости исследователь может остановиться на одной из двух гипотез.



Рассмотрим понятия односторонней (upper-tailed и low-tailed) и двусторонней (two-tailed) гипотез, которым соответствуют односторонний (one-sided) и двусторонний (two-sided) критерии значимости.

Считается, что когда исследователь имеет достаточное количество данных, позволяющих предсказать в альтернативной гипотезе направление различий (например, доля желательных эффектов в опытной группе не просто отличается от доли в контрольной группе, а превышает ее), используется односторонний критерий. В противном случае (доля эффектов в опытной группе просто отличается от доли в контрольной группе) используется двухсторонний критерий. Даже если интересующее различие должно быть в одностороннем направлении, исследователю рекомендуется подстраховаться от неожиданных результатов, выполнив двусторонний тест.

Порядок действий при решении о принятии гипотезы такой.

1. Нулевая гипотеза  $H_0$  (двусторонняя альтернатива) отклоняется, если  $p_2 < \alpha$ .
2. Нулевая гипотеза  $H_0$  (односторонняя upper-tailed альтернатива) отклоняется, если  $(1 - p_U) < \alpha$ .
3. Нулевая гипотеза  $H_0$  (односторонняя low-tailed альтернатива) отклоняется, если  $p_L < \alpha$ .

Здесь обозначено:

$p_2$  – достигнутый уровень значимости двусторонней статистической гипотезы,

$p_U$  и  $p_L$  – достигнутый уровень значимости соответствующей односторонней статистической гипотезы.

При выполнении перечисленных условий соответствующая альтернативная гипотеза  $H_1$  может быть принята.

Если оперировать значением статистики критерия, нулевая гипотеза может быть принята при нахождении вычисленного значения статистики критерия  $T$  в области:

- $T_{1-\alpha} < T \leq T_\alpha$  для двусторонней альтернативы,
- $T_{1-\alpha/2} < T \leq T_{\alpha/2}$  для односторонней альтернативы.

Обсуждение см. в монографиях Тюрина с соавт., Селезнева с соавт., Брандта, Ключина с соавт., Мостеллера (Mosteller) с соавт., книге Глотова с соавт., в статьях Гудмана.

### 1.4.3. P–значение

При подстановке статистики в ее функцию распределения получается величина, имеющая смысл вероятности и интерпретацию, зависящую от решаемой проблемы. Эта вероятность называется фактически достигнутым уровнем значимости, иначе  $P$ –значением.

$P$ –значение дает возможность принимать или отклонять данную гипотезу при любом заранее заданном уровне значимости  $\alpha$  путем простого сравнения вычисленного  $P$ –значения с принятым стандартным уровнем значимости. Возможен иной подход к проверке статистической гипотезы. А именно, сначала вычисляется по выборке статистика  $T$ . Затем вычисляется вероятность  $P$  попадания  $T$  в критическую область.

Рассмотрим, как нужно делать выводы относительно  $P$ –значения статистической гипотезы на основе вычисленного  $P$ –значения статистики критерия в стандартных случаях



статистической гипотезы. Итак, пусть вычислено  $P$ -значение статистики критерия  $p$  путем подстановки статистики критерия в его функцию распределения. Тогда:

1. В случае двусторонней статистической гипотезы ее  $P$ -значение (говорят проще – двустороннее  $P$ -значение) вычисляется как  $p_2 = 2 \cdot \min(p, 1 - p)$ .
2. Если схема вычисления статистического критерия позволяет сразу вычислить два  $P$ -значения стандартных односторонних статистических гипотез (говорят проще – одностороннее  $P$ -значение):  $p_U$  (верхний хвост, upper-tailed) и  $p_L$  (нижний хвост, low-tailed), то двустороннее  $P$ -значение равно  $p_2 = p_U + p_L$ . Если распределение статистики критерия несимметрично, то  $p_U \neq p_L$ . Одностороннее  $P$ -значение вычисляется как  $p_1 = \min(p_U, p_L)$ .
3. Если распределение статистики критерия симметрично, то  $p_U = p_L$  и  $p_2 = 2 \cdot p_U = 2 \cdot p_L$ . Поэтому если вычислено двустороннее  $P$ -значение, а распределение статистики критерия симметричное, одностороннее  $P$ -значение можно получить из двустороннего  $P$ -значения по формуле  $p_1 = p_2 / 2$ .

Рассмотрим пример. Проверяется нулевая гипотеза о равенстве средних значений двух выборок, а также сформулирована двусторонняя альтернатива о том, что средние значения не равны. Зададимся уровнем значимости  $\alpha = 0,05$ . Пусть на основе статистики критерия вычислен достигнутый уровень значимости  $p = 0,988095$ . Тогда двустороннее  $P$ -значение равно  $p_2 = 2 \cdot \min(p, 1 - p) = 2 \cdot \min(0,988095; 0,011905) = 0,023810$ . Очевидно, что  $p_2 < \alpha$ , поэтому нулевая гипотеза отклоняется и принимается альтернативная гипотеза о статистически значимом различии средних значений на уровне значимости  $\alpha = 0,05$ . Данный факт записывают как  $p < 0,05$ .

Обсуждение см. в монографиях Петровича с соавт., Боровкова, Браунли.

#### 1.4.4. Доверительная вероятность

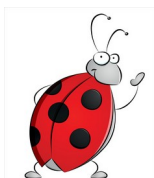
Доверительная вероятность (доверительный уровень, коэффициент доверия) определяется формулой

$$P = 1 - \alpha,$$

где  $\alpha$  – уровень значимости.

Доверительная вероятность требуется для вычисления ряда выборочных статистических показателей. В отличие от ряда других параметров она не вычисляется по выборке, а выбирается исследователем из стандартной линейки (в основном, следуя классификации Плохинского):

- Нулевой порог 0,90 применяется для работы с пониженной ответственностью, при первом ознакомлении с явлением.
- Первый порог 0,95 применяется в большинстве исследований (например, биологические исследования).
- Второй порог 0,99 применяется для работ с повышенной ответственностью (например, медицинские исследования).
- Третий порог 0,999 применяется для работ с высокой ответственностью (например, исследования эффективности лекарств).



Доверительный уровень может быть выражен в долях, например, 0,95, либо в процентах, то же самое, 95%.

## 1.4.5. Мощность критерия

Мощностью называют величину  $1 - \beta$ , где  $\beta$  – вероятность ошибки второго рода статистической гипотезы. Мощность характеризует качество статистического критерия. Мощность – это не число, а функция. Чем эффективнее данная функция стремится к 1, тем более эффективен статистический критерий.

Численное исследование мощности методом Монте–Карло представлено в работах Золотухиной с соавт., Селезнева с соавт., Хассана (Hassan), Лемешко с соавт., монографии Хана с соавт.

## 1.4.6. Сопряженность выборок

Данные, полученные в реальных экспериментах, могут быть представлены независимыми либо сопряженными выборками.

### 1.4.6.1. Независимые выборки

Независимыми будут выборки, отобранные из причинно независимых совокупностей. Не имеет значения, равны ли между собой численности совокупностей.

Примеры независимых выборок:

- параметры двух групп пациентов, к которым применялись различные методики лечения с целью изучения значимости различий между методиками;
- частный случай предыдущей схемы: параметры двух групп пациентов, к одной из которых (опытная группа) применялось воздействие методики, а к другой (контрольной) не применялось, с целью изучения значимости влияния данной методики на результат лечения; данная схема называется «опыт – контроль»;
- частный случай предыдущей схемы: параметры группы пациентов, к которой применяется некоторое лекарственное средство, и контрольной группы пациентов, к которой применяется плацебо, а исследование производится с целью проверки эффективности препарата.

### 1.4.6.2. Сопряженные выборки

Сопряженными будут выборки, отобранные из причинно связанных совокупностей. При анализе сопряженных выборок численности сравниваемых совокупностей всегда равны между собой.

Примеры сопряженных выборок:

- параметры одной и той же испытуемой группы до и после воздействия какого-либо фактора, например, методики лечения; данная схема называется «до и после»;



- параметры одной и той же группы индивидуумов (например, список политических партий, участвующих в парламентских выборах) при воздействии на нее различных факторов (предпочтения электората в различных избирательных округах);
- параметры одного и того же объекта экспериментального исследования, но относящиеся к различным его частям, например состояния двух конечностей в процессе лечения, одна из которых подвергается лечебному воздействию, а другая нет.

## Глава 2. Описательная статистика

### 2.1. Введение

Рассмотрено вычисление основных показателей описательной статистики количественных и качественных выборок. При этом исходные данные могут быть представлены в качестве эмпирической выборки или в сгруппированном виде.

### 2.2. Теоретическое обоснование

Эмпирические (опытные, экспериментальные) выборки (совокупности) состоят из отдельных вариантов (элементов), которые объединены общностью некоторых свойств (признаков, переменных). Источник появления выборок для статистического анализа значения не имеет. Единственное требование к анализируемым выборкам определяется представленными методами расчета. Они применимы только к таким выборкам, варианты которых измерены в соответствующей шкале.

Рассчитываются следующие выборочные статистические показатели описательной статистики:

- численность выборки,
- показатели положения: среднее значение и его стандартная ошибка, медиана, псевдомедиана,
- показатели разброса (рассеяния, масштаба): дисперсия, стандартное отклонение, среднее отклонение, размах, коэффициент вариации, средняя разность Джини, квартили, межквартильный размах,
- показатели формы распределения: коэффициент асимметрии, эксцесс.

Кроме перечисленных показателей, по выборке рассчитываются:

- достаточная численность выборки, из анализа заданных и рассчитанных выборочных показателей,
- оптимальное число классов.
- минимум и максимум.

Для качественных (бинарных) выборок может быть рассчитана доля, ошибка доли и дисперсия доли.

Для всех показателей рассчитываются как точечные, так и интервальные оценки. При этом на распечатке для краткости доверительные интервалы обозначаются аббревиатурой ДИ.





Напомним, что параметры положения и разброса количественной выборки могут оцениваться двумя методами: методом моментов и методом квантилей.

- Использование метода моментов дает в качестве параметрической точечной оценки положения среднее значение, в качестве параметра разброса – дисперсию.
- Использование метода квантилей в качестве непараметрической точечной оценки параметра положения приводит к медиане, в качестве параметра разброса – к межквартильному размаху.

## 2.2.1. Численность выборки

Количество вариант совокупности в источниках называют по-разному. Так, если речь идет об эмпирической выборке, количество ее элементов может называться численностью, величиной или размером. Термин «размерность» употреблять в значении «численность» не следует, т. к. он зарезервирован для описания многомерных совокупностей.

## 2.2.2. Среднее значение

Выборочное среднее значение – статистический показатель, характеризующий середину эмпирической совокупности, оценка параметра положения.

### 2.2.2.1. Общая методика

Среднее значение – это параметрическая оценка параметра положения статистического распределения. Следовательно, для вычисления оценки среднего значения необходимо задаться (или установить на основе эмпирических данных) типом распределения статистической совокупности. Затем следует воспользоваться методом оптимизации с целью вычисления данной оценки. Процесс включает следующие этапы:

- Составление функционала.
- Определение производных функционала по искомым параметрам.
- Приравнивание производных нулю с целью получения системы линейных или нелинейных алгебраических уравнений для вычисления оптимальных (доставляющих экстремум функционалу: минимум – для метода наименьших квадратов, максимум – для метода максимального правдоподобия) значений параметров.
- Решение уравнений. Для некоторых моделей бывает достаточно одного уравнения для данного параметра, в результате преобразования которого получается простая алгебраическая формула. Для других моделей приходится аналитически или численно решать систему уравнений.

### 2.2.2.2. Оценка среднего на основе теории распределений

Пусть имеется количественная выборка, имеющая нормальное распределение с плотностью

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$







где  $\mu$  – параметр положения статистического распределения,  
 $\sigma$  – параметр масштаба.

Вычислим оценку максимального правдоподобия для параметра положения. В рассматриваемом случае функция максимального правдоподобия (ФМП) запишется как

$$L(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}},$$

где  $n$  – численность выборки,

$x_i, i = 1, 2, \dots, n$ , – значения вариант выборки.

Оптимальные значения параметров доставляют максимум ФМП. Вычисления упрощаются, если исследовать не саму ФМП, а ее логарифм, т. к. они достигают максимума при одних и тех же значениях параметров. Логарифмическая ФМП имеет вид

$$\ln L(\mu, \sigma) = -\frac{1}{\sqrt{2\pi}} \left[ n \ln \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right].$$

Максимум логарифмической ФМП достигается при равенстве нулю частных производных по параметрам. Частная производная логарифмической ФМП по параметру  $\mu$  будет

$$\frac{\partial L(\mu, \sigma)}{\partial \mu} = -\frac{1}{\sigma^2 \sqrt{2\pi}} \sum_{i=1}^n (x_i - \mu) = 0,$$

откуда очевидно получается

$$\sum_{i=1}^n (x_i - \mu) = \sum_{i=1}^n x_i - n\mu = 0$$

и, окончательно,

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i.$$

Формула выборочной оценки среднего значения получена в предположении нормального распределения количественной случайной величины. Следовательно, вычисленную по данной формуле оценку допустимо применять только для нормально распределенной количественной величины.

Выборочную оценку среднего значения обозначают символами  $\bar{x}$ ,  $M$  или  $E$ , причем последние символы стандартно принято использовать в смысле оператора над случайной величиной. В некоторых источниках среднее [значение] часто эквивалентно называют средней [величиной].

Доверительный интервал оцениваемого среднего значения вычисляется на заданном доверительном уровне, выражаемом в долях или процентах. Доверительный интервал, вычисленный на доверительном уровне 0,95, означает, что 95% вариант выборочной совокупности попадают в данный интервал. Иначе, истинное значение среднего значения генеральной совокупности (математического ожидания) находится между нижней и верхней значениями доверительного интервала с вероятностью, равной доверительной.





Для вычисления доверительного интервала оцениваемого среднего значения в случае, если эмпирическая выборка распределена нормально, используется формула:

$$I_m = \left( \bar{x} - t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}}; \bar{x} + t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}} \right),$$

где  $\sigma$  – стандартное отклонение,

$t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $n - 1$  и  $(1 + \beta) / 2$ ,

$\beta$  – доверительный уровень, выраженный в долях.

Для вычисления доверительного интервала оцениваемого среднего значения, когда выборка не является нормальной, применяется формула:

$$I_m = \left( \bar{x} - \Psi((1 + \beta) / 2) \frac{\sigma}{\sqrt{n}}; \bar{x} + \Psi((1 + \beta) / 2) \frac{\sigma}{\sqrt{n}} \right),$$

где  $\Psi(.)$  – обратная функция стандартного нормального распределения.

Метод максимума правдоподобия представлен Тiku (Tiku) с соавт. О вычислении доверительных интервалов оцениваемого среднего значения см. книгу Мюллера с соавт.

### 2.2.2.3. Оценка среднего на основе теории множеств

Попытаемся дать универсальное понятие среднего значения, не зависящее ни от шкалы измерения эмпирических данных, ни от их размерности. Пусть имеется эмпирическая случайная выборка  $\{X_1, X_2, \dots, X_n\}$  численностью  $n$ , где  $X_i, i = 1, 2, \dots, n$  – варианты, скалярные или векторные, иначе эмпирические реализации случайной величины. Обозначим через  $d(X, X_i)$  расстояние между произвольной реализацией случайной величины  $X$  и величиной  $X_i, i = 1, 2, \dots, n$ . Основное требование к данному расстоянию состоит в его допустимости в используемой шкале измерения эмпирической выборки.

Средним значением выборки будет случайная величина  $X$ , удовлетворяющая условию

$$M\{X_1, X_2, \dots, X_n\} = \arg \min_{X \in R} \sum_{i=1}^n d(X, X_i),$$

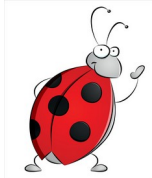
где  $R$  – пространство всех допустимых, с точки зрения шкалы измерений, реализаций случайной величины  $X$ .

Иначе, среднее ищется среди всех возможных, а не только среди полученных в опыте, реализаций  $X$ . Поэтому в общем случае среднее значение не является никакой из вариантов  $X_i, i = 1, 2, \dots, n$ . Это свойство можно считать слабостью в понятии точечной оценки среднего значения, компенсировать которую призваны представленные выше интервальные оценки.

### 2.2.2.4. Стандартная ошибка

Стандартная ошибка среднего значения определяется по формуле:





$$\mu = \frac{\sigma}{\sqrt{n}},$$

где  $\sigma$  – стандартное отклонение,  
 $n$  – численность выборки.

Стандартную ошибку обозначают также символом  $m$ . Запись, характеризующая среднее значение и его стандартную ошибку, приводится в виде  $M \pm m$ .  
См. монографию Тейлора (с. 24 ).

## 2.2.2.5. Дисперсия

Основным статистическим показателем, характеризующим разброс выборки, является выборочная дисперсия. Общая методика оценки и вид функционала для нормальной количественной выборки представлены в разделе «Среднее значение».

Пусть имеется количественная выборка, имеющая нормальное распределение с плотностью

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где  $\mu$  – параметр положения статистического распределения,  
 $\sigma$  – параметр масштаба.

Вычислим оценку максимального правдоподобия для параметра масштаба. В рассматриваемом случае функция максимального правдоподобия (ФМП) запишется как

$$L(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}},$$

где  $n$  – численность выборки,  
 $x_i, i = 1, 2, \dots, n$  – значения вариант выборки.

Оптимальные значения параметров доставляют максимум ФМП. Вычисления упрощаются, если исследовать не саму ФМП, а ее логарифм, т. к. ФМП и логарифм ФМП достигают максимума при одних и тех же значениях параметров. Логарифмическая ФМП имеет вид

$$\ln L(\mu, \sigma) = -\frac{1}{\sqrt{2\pi}} \left[ n \ln \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right].$$

Максимум логарифмической ФМП достигается при равенстве нулю частных производных по параметрам. Частная производная логарифмической ФМП по параметру  $\sigma$  будет

$$\frac{\partial L(\mu, \sigma)}{\partial \sigma} = -\frac{1}{\sigma\sqrt{2\pi}} \left[ n - \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] = 0,$$

откуда очевидно получается

$$n\sigma^2 - \sum_{i=1}^n (x_i - \mu)^2 = 0$$

и, окончательно,





$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

Величину  $\sigma^2$  называют выборочной дисперсией и часто обозначают как  $S$  или  $D$ , причем последний символ стандартно принято использовать в смысле оператора над случайной величиной.

Формула выборочной оценки дисперсии получена в предположении нормального распределения количественной случайной величины. Следовательно, вычисленную по данной формуле оценку допустимо применять только для нормально распределенной величины.

Будет неверным пользоваться полученной выше формулой для дисперсии, если оценка среднего значения совокупности производится также по выборке. Обозначим:

$\xi$  – случайная величина,

$M\xi$  – математическое ожидание,

$D\xi$  – выборочная дисперсия,

$\bar{x}$  – выборочное среднее значение.

Согласно определению и учитывая, что

$$M(\bar{x} - a)^2 = D\bar{x} = \frac{1}{n} D\xi,$$

где  $a$  – известное среднее значение совокупности, можно записать

$$\begin{aligned} nD\bar{x} &= \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n [(x_i - \bar{x}) - (a - \bar{x})]^2 = \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 - 2(a - \bar{x}) \sum_{i=1}^n (x_i - \bar{x}) + n(a - \bar{x})^2. \end{aligned}$$

В последнем выражении сумма во втором члене, очевидно, дает нуль, поэтому перенеся первый член этого выражения в левую часть и сменив знак, получаем

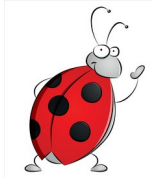
$$\sum_{i=1}^n (x_i - \bar{x})^2 = nD\bar{x} - D\xi = (n - 1)D\bar{x},$$

откуда непосредственно следует, что в случае оценки среднего значения по выборке в качестве оценки дисперсии выборочной совокупности берется величина, определяемая по формуле:

$$D = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Формула вычисляет так называемую несмещенную выборочную оценку дисперсии генеральной совокупности (эмпирическую дисперсию).

Получим эквивалентную формулу для выборочной дисперсии, не содержащую значения выборочного среднего.



$$D = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \sum_{i=1}^n (2x_i\bar{x} - \bar{x}^2) \right] =$$
$$= \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n (2x_i - \bar{x}) \right] = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \bar{x} \left( 2 \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \right) \right].$$

В круглых скобках получилась разность удвоенной суммы вариант выборки и просто суммы вариант, т. к.

$$\sum_{i=1}^n \bar{x} = n\bar{x} = n \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i.$$

Поэтому продолжим, подставив выражение для среднего арифметического значения,

$$D = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n x_i \right] = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right].$$

Для вычисления доверительного интервала оцениваемой дисперсии в случае, если эмпирическая выборка распределена нормально, применяется формула:

$$I_D = (D - t_{(1+\beta)/2} d; D + t_{(1+\beta)/2} d),$$

где  $t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $n-1$  и  $(1+\beta)/2$ ,

$\beta$  – доверительный уровень, выраженный в долях,

$d$  – величина, рассчитанная по формуле

$$d = \sqrt{\frac{1}{n} \left( m_4 - \left( \frac{n-1}{n} \right)^4 D^2 \right)},$$

где  $m_4$  – четвертый центральный выборочный момент, вычисляемый по формуле

$$m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4.$$

Для вычисления доверительного интервала оцениваемой дисперсии, когда выборка не является нормальной, применяется формула:

$$I_D = (D - \Psi((1+\beta)/2) d; D + \Psi((1+\beta)/2) d),$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения.

Метод максимума правдоподобия представлен Тику (Tiku) с соавт. О вычислении доверительных интервалов см. Мюллера с соавт.

## 2.2.2.6. Стандартное отклонение

Стандартным отклонением (средним квадратическим отклонением, средним квадратичным отклонением, стандартом, сигмой) называют корень квадратный из дисперсии. Вычисление стандартного отклонения производится по формуле:

$$\sigma = \sqrt{D},$$





где  $D$  – выборочная дисперсия.

Для вычисления доверительного интервала оцениваемого стандартного отклонения количественной выборки в случае, если эмпирическая выборка распределена нормально, применяется формула:

$$I_{\sigma} = \left[ \sigma \cdot \sqrt{\frac{n-1}{\chi^2_{(1-\beta)/2}}}; \sigma \cdot \sqrt{\frac{n-1}{\chi^2_{(1+\beta)/2}}} \right],$$

где  $n$  – численность выборки,

$\chi^2_{(1-\beta)/2}$  – значение обратной функции  $\chi^2$ -распределения с параметрами  $n-1$  и  $(1-\beta)/2$ ,

$\chi^2_{(1+\beta)/2}$  – значение обратной функции  $\chi^2$ -распределения с параметрами  $n-1$  и  $(1+\beta)/2$ ,

$\beta$  – доверительный уровень, выраженный в долях.

Для вычисления доверительного оцениваемого интервала стандартного отклонения, когда выборка не является нормальной, применяется формула:

$$I_{\sigma} = (\sigma - \Psi((1+\beta)/2)d / (2\sigma); \sigma + \Psi((1+\beta)/2)d / (2\sigma)),$$

где  $\Psi(.)$  – обратная функция стандартного нормального распределения,

$d$  – величина, рассчитанная по формуле

$$d = \sqrt{\frac{1}{n} \left( m_4 - \left( \frac{n-1}{n} \right)^4 D^2 \right)},$$

где  $m_4$  – четвертый центральный выборочный момент, вычисляемый по формуле

$$m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4,$$

где  $x_i, i = 1, 2, \dots, n$  – значения вариант выборки,

$\bar{x}$  – среднее значение.

О вычислении доверительных интервалов см. Мюллера с соавт.

## 2.2.2.7. Среднее отклонение

Выборочное среднее отклонение (выборочная оценка среднего отклонения), подобно стандартному отклонению, характеризует разброс эмпирической выборки относительно среднего значения и вычисляется по формуле

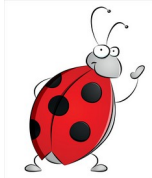
$$\hat{x} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|,$$

где  $n$  – численность выборки,

$x_i, i = 1, 2, \dots, n$  – значения вариант выборки,

$\bar{x}$  – выборочное среднее значение.





Среднее отклонение отражает так называемый модульный подход к вычислению меры отклонения между величинами в противоположность тому, что стандартное отклонение отражает квадратический подход.

### 2.2.2.8. Средняя разность Джини

Средняя разность Джини характеризует разброс значений вариант эмпирической выборки друг относительно друга и не зависит от какого-либо центрального значения, например, от среднего значения или медианы. Вычисление выборочной средней разности Джини производится по формуле

$$g = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n |x_i - x_j|,$$

где  $n$  – численность выборки,  
 $x_i, i = 1, 2, \dots, n$  – значения вариант выборки.

### 2.2.3. Асимметрия

Асимметрия характеризует форму статистического распределения. Если коэффициент асимметрии больше нуля, асимметрия правосторонняя (положительная), форма кривой распределения скошена вправо относительно кривой плотности нормального распределения. Если коэффициент асимметрии меньше нуля, то асимметрия левосторонняя (отрицательная), форма кривой распределения скошена влево относительно кривой плотности нормального распределения. Коэффициент асимметрии выборочной совокупности вычисляется по уточненной формуле:

$$A = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma} \right)^3,$$

где  $n$  – численность выборки,  
 $x_i, i = 1, 2, \dots, n$  – значения вариант выборки,  
 $\bar{x}$  – выборочное среднее значение,  
 $\sigma$  – выборочное стандартное отклонение.

Вычисление доверительного интервала оцениваемого коэффициента асимметрии производится по формуле:

$$I_A = \left( A - \sqrt{\frac{D_A}{\beta}}; A + \sqrt{\frac{D_A}{\beta}} \right),$$

где  $D_A$  – дисперсия коэффициента асимметрии,  
 $\beta$  – доверительный уровень, выраженный в долях.  
Дисперсия коэффициента асимметрии вычисляется по формуле

$$D_A = \frac{6(n-2)}{(n+1)(n+3)}.$$



Асимметрия находит применение, в частности, при исследовании формы распределения выборки. Доверительные интервалы вычислены в книге Иглина.

## 2.2.4. Эксцесс

Эксцесс характеризует форму статистического распределения. Если эксцесс больше нуля, то форма кривой распределения островершинная по сравнению с кривой плотности нормального распределения. Если эксцесс меньше нуля, то форма кривой распределения плосковершинная по сравнению с кривой плотности нормального распределения. Эксцесс выборочной совокупности вычисляется по уточненной формуле:

$$E = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)},$$

где  $n$  – численность выборки,

$x_i, i = 1, 2, \dots, n$  – значения вариант выборки,

$\bar{x}$  – выборочное среднее значение,

$\sigma$  – стандартное отклонение.

Вычисление доверительного интервала оцениваемого эксцесса производится по формуле:

$$I_E = \left[ E - \sqrt{\frac{D_E}{\beta}}; E + \sqrt{\frac{D_E}{\beta}} \right],$$

где  $D_E$  – дисперсия эксцесса,

$\beta$  – доверительный уровень, выраженный в долях.

Дисперсия эксцесса вычисляется по формуле

$$D_E = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Эксцесс находит применение, в частности, при исследовании формы распределения выборки. Доверительные интервалы в книге Иглина.

## 2.2.5. Коэффициент вариации

Коэффициент вариации представляет собой характеристику рассеяния случайной величины.

Он показывает, какой процент составляет стандартное отклонение от среднего значения.

Коэффициент вариации используется для установления степени выравнимости совокупности по тому или иному признаку. Коэффициент вариации вычисляется по формуле:

$$V = \frac{\sigma}{\bar{x}} \text{ в долях или}$$

$$g = \frac{\sigma}{\bar{x}} \cdot 100\% \text{ в процентах,}$$

где  $\sigma$  – стандартное отклонение,

$\bar{x}$  – выборочное среднее значение.





Для вычисления доверительного интервала оцениваемого коэффициента вариации применяется формула:

$$I_V = (V - \Psi((1 + \beta) / 2)d; V + \Psi((1 + \beta) / 2)d),$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях,

$d$  – величина, рассчитанная по формуле

$$d = \sqrt{\frac{1}{n} \left( V^4 - \frac{V^2}{4} + \frac{m_4}{4D\bar{x}^2} - \frac{m_3}{\bar{x}^3} \right)},$$

где  $n$  – численность выборки,

$D$  – выборочная дисперсия,

$m_3$  – третий центральный выборочный момент, вычисляемый по формуле

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3,$$

$m_4$  – четвертый центральный выборочный момент, вычисляемый по формуле

$$m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4,$$

где  $x_i, i = 1, 2, \dots, n$  – значения вариант выборки.

О вычислении доверительных интервалов см. Мюллера с соавт. Коэффициент вариации применяется при проверке репрезентативности (оценке достаточной численности) выборки.

## 2.2.6. Минимум и максимум

Значения минимальной и максимальной вариант выборки:

$x_{\max}$  – значение максимальной варианты выборки,

$x_{\min}$  – значение минимальной варианты выборки.

### 2.2.6.1. Размах выборки

Размах выборки (размах вариации, амплитуда ряда) характеризует степень разброса данных в абсолютных числах. Выборочный размах – это разность между максимумом и минимумом вариант выборки. Вычисление размаха количественной выборки производится по формуле:

$$R = x_{\max} - x_{\min},$$

где  $x_{\max}$  – значение максимальной варианты выборки,

$x_{\min}$  – значение минимальной варианты выборки.

## 2.2.7. Медиана

Существует два типичных определения медианы. Энциклопедия «Вероятность и математическая статистика» определяет медиану случайной величины  $X$  как любое число  $m$  такое, что  $P\{X \geq m\} \geq 1/2$  и  $P\{X \leq m\} \leq 1/2$ . Медиана  $m$  непрерывно распределенной случайной величины  $X$  со строго монотонной функцией распределения  $F(x)$  определяется как единственный корень уравнения  $F(m) = 1/2$ .







Алгоритм определения выборочной медианы количественной выборки определяют следующим образом. Для вычисления медианы эмпирической количественной выборки  $x_i, i = 1, 2, \dots, n$ , численностью  $n$  сначала строится интервальный вариационный ряд  $y_i, i = 1, 2, \dots, n$ , т. е. упорядоченная по возрастанию исходная выборка. Для нечетного  $n = 2k + 1$  медианой будет вариант с номером  $k$ . Для четного  $n = 2k$  медианой будет полусумма вариантов с номерами  $k$  и  $k + 1$ .

Приведенный алгоритм может применяться также и для порядковой выборки нечетной численности. Для порядковой выборки четной численности некоторые авторы рассматривают левую медиану – вариант вариационного ряда с номером  $k$  – и правую медиану – вариант вариационного ряда с номером  $k + 1$  – ввиду того, что для порядковой шкалы измерения операция деления не определена.

Доверительный интервал оцениваемой медианы задается формулой

$$I_m = (y_{c+1}; y_{n-c}),$$

где  $c$  – параметр, вычисляемый по формуле

$$c = [n / 2 - \Psi((1 + \beta) / 2) n^{1/2} / 2],$$

где  $[.]$  – целая часть числа,

$\Psi(.)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

Некоторые исследователи предпочитают медиану среднему значению (для шкалы измерения, в котором данный показатель имеет смысл), считая ее более точной оценкой меры положения выборки.

См. Холлендера с соавт. Уточненная формула дана ГОСТ<sup>2</sup>.

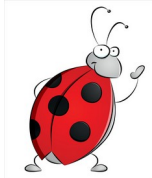
## 2.2.7.1. Оценка медианы на основе теории множеств

Рассмотрим выборочный показатель, представляющий собой вариант выборки, равноудаленную от всех остальных вариантов этой же эмпирической выборки. Данный показатель называется медианой множества (далее – медианой). При этом смысловое наполнение термина «равноудаленная» определяется шкалой измерения выборки.

Попытаемся дать универсальное определение медианы, не зависящее ни от шкалы измерения эмпирических данных, ни от их размерности. Пусть имеется множество реализаций некоторой случайной величины, представляющее собой случайную эмпирическую выборку  $\{X_1, X_2, \dots, X_n\}$ , где  $X_i, i = 1, 2, \dots, n$  – варианты, скалярные или векторные. Обозначим через  $d(X, X_i)$  расстояние между произвольной реализацией случайной величины  $X$  и величиной  $X_i, i = 1, 2, \dots, n$ . Основное требование к данному расстоянию состоит в его допустимости в используемой шкале измерения эмпирической выборки. Определим медиану как решение оптимизационной задачи. Медианой будет случайная величина  $X$ , удовлетворяющая условию

$$\mu\{X_1, X_2, \dots, X_n\} = \arg \min_{X \in D} \sum_{i=1}^n d(X, X_i),$$

2 ГОСТ Р ИСО 16269–7–2004. Статистические методы. Статистическое представление данных. Медиана. Определение точечной оценки и доверительных интервалов. – М.: Издательство стандартов, 2004.



где  $D$  – выборочное пространство реализаций случайной величины  $X$ .

Иначе, медианой множества является одна из вариант  $X_i$ ,  $i = 1, 2, \dots, n$ , удовлетворяющая данному условию.

Поиск медианы множества не вызывает вычислительной сложности в любой шкале измерения и может производиться на основе формального применения представленного определения.

В случае количественной выборки нечетной численности показатель совпадает с обычной медианой. Для порядковой выборки четной численности некоторые авторы рассматривают левую медиану – варианту вариационного ряда с номером  $k$  – и правую медиану – варианту вариационного ряда с номером  $k + 1$  – ввиду того, что для порядковой шкалы измерения операция деления не определена.

См. Холлендера с соавт., Кормена с соавт. (с. 240).

## 2.2.7.2. Псевдомедиана

Пусть вычислено  $m = n(n + 1) / 2$  значений  $w_1 \leq w_2 \leq \dots \leq w_m$  величин  $(x_i + x_j) / 2$ ,  $i \leq j$ ;  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, n$ , где  $x_i$ ,  $x_j$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, n$  – значения вариант исходной количественной выборки. Тогда медиана  $\mu$  полученной выборки  $w_i$ ,  $i = 1, 2, \dots, m$ , называется псевдомедианой (оценкой Ходжеса–Лемана).

Итак, для вычисления медианы полученной выше количественной выборки  $w_i$ ,  $i = 1, 2, \dots, m$ , численностью  $m$  сначала строится интервальный вариационный ряд  $u_i$ ,  $i = 1, 2, \dots, m$ , т. е. упорядоченная по возрастанию выборка. Для нечетного  $m$  медианой является варианта полученного интервального вариационного ряда, имеющая порядковый номер  $(m + 1) / 2$ . Для четного  $m$  медиана равна среднему значению двух средних вариантов. Утверждается, что если распределение симметрично, выборочные оценки медианы и псевдомедианы совпадают. Доверительный интервал оцениваемой псевдомедианы задается формулой

$$I_\mu = (y_{c+1}; y_{m-c})$$

где  $c$  – параметр, вычисляемый по формуле

$$c = \left\lfloor \frac{n(n+1)}{4} - \Psi((1+\beta)/2) \left( \frac{n(n+1)(2n+1)}{24} \right)^{1/2} \right\rfloor,$$

где  $\lfloor \cdot \rfloor$  – целая часть числа,

$n$  – численность выборки,

$\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

## 2.2.8. Квартили

Квартили, а также медиана (50% процентиль), обеспечивают разбиение упорядоченной количественной выборки (в виде вариационного ряда) на 4 подмножества равной численности. Вычисление данных показателей производится по правилам, принятым для вычисления медианы. Верхняя квартиль представляет собой 75% процентиль выборки. Нижняя квартиль представляет собой 25% процентиль выборки.





Укажем на одно полезное употребление квартилей. Тьюки предложил так называемый график «ящик с усами», представляющий собой совокупность следующих элементов:

- точки, обозначающей медиану,
- прямоугольника с верхней и нижней границами (если график изображается вертикально), соответствующими квартилям,
- отрезками, соответствующими максимуму и минимуму выборки.

Иногда изображается график «ящик с усами» для выборки, из которой уже исключены выбросы. В этом случае выбросы накладываются на график в виде точек. О методах исключения выбросов см. главу «Обработка выбросов».

## 2.2.8.1. Межквартильный размах

Как известно, квартили, а также медиана (50% процентиль), обеспечивают разбиение упорядоченной количественной выборки (в виде вариационного ряда) на 4 подмножества равной численности. Вычисление данных показателей производится по правилам, принятым для вычисления медианы.

Межквартильный (интерквартильный) размах выборки характеризует степень разброса данных в абсолютных числах. Выборочный межквартильный размах – это разность между верхней и нижней квартилями выборки, иначе 75% и 25% процентилями выборки.

Вычисление межквартильного размаха упорядоченной по возрастанию количественной выборки производится по формуле:

$$f = f_{3/4} - f_{1/4},$$

где  $f_{3/4}$  – значение верхней квартили выборки,

$f_{1/4}$  – значение нижней квартили выборки.

Точечная оценка стандартного отклонения для нормально распределенной совокупности может быть получена из межквартильного размаха как

$$\sigma = \frac{f}{2\Psi(0,75)} \approx 0,741301f,$$

где  $\Psi(\cdot)$  – функция, обратная функции стандартного нормального распределения, Межквартильный размах находит применение в качестве основы одного из методов выявления аномальных наблюдений (выбросов), применяемых в главе «Обработка выбросов». Величина  $f/2$  используется как характеристика рассеяния и называется семиинтерквартильной шириной.

## 2.2.9. Гистограмма

Гистограмма представляет собой дискретный или интервальный вариационный ряд (ряд распределения), полученный в результате группировки исходной эмпирической выборки, измеренной в порядковой или количественной шкале, по особым образом подобранным классовым интервалам. Данный вариационный ряд служит основой для многих статистических алгоритмов, таких, как глазомерный метод проверки нормальности



распределения, установление типа распределения (как для дискретных, так и для непрерывных распределений), критерии типа хи-квадрат и других.

Имеется два пути практической группировки: задавшись границами классовых интервалов (классов) или задавшись их количеством, а затем вычислить границы. Во втором случае для вариационного ряда число классов равно числу градаций переменной. Число классов дискретного вариационного ряда равно числу градаций вариант выборки, измеренной в порядковой шкале. Для интервального вариационного ряда число классов задается пользователем на основе одного из применяемых правил, рассмотренных ниже.

Критерием правильности выбора количества классов считается верная передача типа распределения эмпирических частот данной выборочной совокупности. Если выбрано слишком мало классов, можно потерять характерную картину эмпирического распределения. При слишком подробном делении можно затушевать реальную картину распределения частот случайными отклонениями.

Инструмент «Гистограмма» позволяет пользователю, при его желании, задать число классов либо делает это автоматически. Выделяются несколько общеупотребительных способов вычисления числа классов для выборок умеренной численности. Правило Стержесса (Sturges rule) основано на формуле

$$k = 1,44 \ln n + 1,$$

где  $k$  – число классов,

$n$  – численность выборочной совокупности.

После решения вопроса о числе классов производится вычисление границ классовых интервалов и разнесение вариант исходной количественной выборки по классовым интервалам.

Обзор способов выбора числа классов см. у Новицкого с соавт., Скотта (Scott).

## 2.2.9.1. Мода

Мода представляет собой значение переменной, при котором функция плотности распределения достигает максимального значения. Визуальным отображением эмпирической функции плотности эмпирического распределения является гистограмма (деленная на численность выборки), поэтому моду удобно рассчитать и вывести в разделе «Гистограмма».

## 2.2.9.2. Оптимальное число классов

Очевидным критерием правильности выбора количества классов считается верная передача типа распределения эмпирических частот данной выборочной совокупности. Если выбрано слишком мало классов, можно потерять характерную картину эмпирического распределения. При слишком подробном делении можно затушевать реальную картину распределения частот случайными отклонениями. Большинство источников ограничиваются данными рекомендациями, предлагая различные эвристические формулы вычисления числа классовых интервалов. Выбор некоторого оптимального количества классов позволит не только верно визуальное передать тип распределения, но и минимизировать существенные потери



информации, содержащейся в исходных данных, которая происходит при фактическом понижении исходной количественной шкалы до шкалы номинальной.

## 2.2.9.2.1. Метод оптимизации числа классов

Предлагается алгоритм, дающий математическое обоснование критерия, с формулировки которого начат данный раздел. Под оптимальным числом классов мы понимаем минимально допустимое, но верно передающее распределение исходной случайной величины. Алгоритм состоит из следующих шагов.

1. Пусть дана количественная эмпирическая выборка  $x_i$ ,  $i = 1, 2, \dots, n$ .
2. Берется минимальное имеющее смысл число классов  $k = 2$ .
3. Производится классификация, в результате которой получается вариационный ряд  $y_j$ ,  $j = 1, 2, \dots, k$ .
4. По вариационному ряду восстанавливается выборка  $z_i$ ,  $i = 1, 2, \dots, n$ , фактически представляющая собой огрубленную до номинальной шкалы с числом градаций, равным  $k$ , исходную выборку.
5. Сравниваются функции распределения исходной выборки  $x_i$ ,  $i = 1, 2, \dots, n$ , и выборки  $z_i$ ,  $i = 1, 2, \dots, n$ . Может использоваться один из тестов, предназначенных для сравнения двух эмпирических функций распределения.
6. Контролируется  $P$ -значение статистики критерия, вычисленного на шаге 5. Первое же значение  $k$ , при котором различия окажутся незначимы, будет оптимальным числом классов – на этом процесс завершается (процесс завершается также при достижении  $k = n$ ). Иначе, при установлении значимости  $p < 0,05$ , значение  $k$  увеличивается на 1 и осуществляется переход к шагу 3.

Значение  $k$ , полученное в результате работы алгоритма, дает необходимую объективную нижнюю оценку числа классовых интервалов равной ширины, при котором тип распределения исходной случайной величины передается верно. В дальнейших расчетах можно уверенно брать любое число классов, равное или немного превышающее данную величину.

Преимуществом предложенного алгоритма является возможность использования для сравнения распределений: исходного и гистограммы – различных метрик, которые зависят от применяемого критерия, и различных уровней значимости. О критериях сравнения функций распределения см. главу «Непараметрическая статистика».

## 2.2.9.2.2. Метод Шимазаки–Шиномото

Метод предложен Шимазаки (Shimazaki) и Шиномото (Shinomoto). Оригинальный метод оптимизирует ширину классового интервала, поэтому мы немного видоизменили схему метода с целью оптимизации числа классов (данные параметры в случае классовых интервалов равной ширины являются однозначно взаимосвязанными).

- Пусть дана количественная эмпирическая выборка  $x_i$ ,  $i = 1, 2, \dots, n$ .
- Берется минимальное имеющее смысл число классов  $k = 2$ .



- Вычисляется соответствующая ширина классового интервала  $\Delta(k)$ .
- Производится классификация, в результате которой получается вариационный ряд  $y_j, j = 1, 2, \dots, k$ . По вариационному ряду вычисляются параметры: среднее значение  $\bar{y} = \frac{1}{k} \sum_{j=1}^k y_j$  и дисперсия  $D = \frac{1}{k} \sum_{j=1}^k (y_j - \bar{y})^2$ .
- Вычисляется функционал («функция стоимости», в терминологии авторов)  $C(\Delta) = \frac{2\bar{y} \cdot D}{\Delta^2}$ .
- Значение  $k$  увеличивается на 1 и осуществляется переход к шагу 3. Процесс повторяется до достижения  $k = n$ .
- Оптимальным числом классов будет то число, которое обеспечивает минимум функционалу  $C(\Delta)$ .

## 2.2.10. Доля

Для бинарной выборки оценка доли (распространенности, binomial proportion), т. е. количества вариантов – «случаев», отнесенное к численности выборки, может быть рассчитана по формуле максимального правдоподобия:

$$\hat{p} = \frac{m}{n},$$

где  $m$  – число случаев,

$n$  – численность выборки.

Доверительные интервалы для оцениваемой доли могут вычисляться различными методами. Стандартно доверительный интервал оцениваемой доли в источниках рассчитывается по классической формуле Вальда (Wald interval)

$$I_{\hat{p}} = (\hat{p} - \Psi((1 + \beta)/2) \sqrt{D_{\hat{p}}}; \hat{p} + \Psi((1 + \beta)/2) \sqrt{D_{\hat{p}}}),$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях,

$D_{\hat{p}}$  – дисперсия доли.

Доверительные интервалы оцениваемой доли могут рассчитываться по «точным» формулам Клоппера–Пирсона (Clopper–Pearson interval). При этом нижняя граница доверительного интервала оцениваемой доли считается как

$$L_p = \left[ 1 + \frac{n - m + 1}{m \cdot F_{2m, 2(n-m+1)}^{-1}(1 - (1 - \beta)/2)} \right]^{-1},$$

где  $F_{m,n}^{-1}(\cdot)$  – обратная функция  $F$ -распределения.

Верхняя граница доверительного интервала оцениваемой доли считается как





$$H_p = \left[ 1 + \frac{n - m}{(m + 1) \cdot F_{2(m+1), 2(n-m)}^{-1}((1 - \beta)/2)} \right]^{-1}.$$

Доверительный интервал оцениваемой доли рассчитывается также по формуле Агрести–Коула (Agresti–Coull interval, иначе называемый уточненным методом Вальда)

$$I_{\tilde{p}} = (\tilde{p} - \Psi((1 + \beta)/2) \sqrt{D_{\tilde{p}}}; \tilde{p} + \Psi((1 + \beta)/2) \sqrt{D_{\tilde{p}}}),$$

где  $\tilde{p}$  – скорректированное значение доли,

$D_{\tilde{p}}$  – значение дисперсии скорректированной доли.

Скорректированное значение доли рассчитывается по формуле

$$\tilde{p} = \frac{m + 2}{n + 4}.$$

Дисперсия скорректированной доли вычисляется по формуле

$$D_{\tilde{p}} = \frac{\tilde{p} \cdot (1 - \tilde{p})}{n}.$$

Доверительный интервал оцениваемой доли рассчитывается также по формуле Вилсона (Wilson interval)

$$I_{\tilde{p}} = (\tilde{p} - \Psi((1 + \beta)/2) \sqrt{D_{\tilde{p}}}; \tilde{p} + \Psi((1 + \beta)/2) \sqrt{D_{\tilde{p}}}),$$

где  $\tilde{p}$  – скорректированное значение доли,

$D_{\tilde{p}}$  – значение дисперсии скорректированной доли.

В дальнейшей записи для простоты обозначим  $k = \Psi((1 + \beta)/2)$ .

Тогда, с учетом введенного обозначения, скорректированное значение доли рассчитывается по формуле

$$\tilde{p} = \frac{m + k^2 / 2}{n + k^2}.$$

Дисперсия скорректированной доли вычисляется по формуле

$$D_{\tilde{p}} = \frac{n}{(n + k^2)^2} \cdot \left( \tilde{p} \cdot (1 - \tilde{p}) + \frac{k^2}{4n} \right).$$

Обзоры методов оценки доли см. в статьях Льюис (Lewis), Льюис с соавт., Брауна (Brown) с соавт. См. оригинальные статьи Агрести (Agresti) с соавт., Клоппера (Clopper) с соавт., а также работы Пирес (Pires) с соавт., Друган (Drugan) с соавт., доклады Пирес, Сауро (Sauro) с соавт., приложение Хромова–Борисова к книге Кайданова, монографии Флейс, Флейс (Fleiss), Флейс с соавт.

## 2.2.10.1. Ошибка доли

Ошибка доли вычисляется по формуле







$$m_{\hat{p}} = \sqrt{D_{\hat{p}}},$$

где  $D_{\hat{p}}$  – дисперсия доли.

Исследователи иногда задаются вопросом, как рассчитать процент и ошибку процента случаев от численности выборки. Идея заключается в том, что вычисления в данном случае производятся по стандартным формулам для доли. Результат же переводится в проценты следующим образом. Процент вычисляется как  $100 \cdot \hat{p}$ , где  $\hat{p}$  – оценка доли. Ошибка процента вычисляется как  $100 \cdot m_{\hat{p}}$ .

## 2.2.10.2. Дисперсия доли

Дисперсия доли может быть вычислена по формуле

$$D_{\hat{p}} = \frac{\hat{p} \cdot (1 - \hat{p})}{n},$$

где  $\hat{p}$  – выборочная оценка доли,  
 $n$  – численность выборки.

## 2.2.11. Показатель точности опыта

Показатель точности опыта, иначе – показатель точности определения среднего значения, выражает величину ошибки среднего значения в процентах от самого среднего. Точность считается удовлетворительной, если величина показателя не превышает 5%, а при значениях, больших 5%, рекомендуется увеличить число наблюдений или повторений. Иногда величину показателя точности можно уменьшить, если повысить точность измерений параметров объектов опыта. Показатель точности опыта вычисляется по формуле:

$$p = \frac{m}{\bar{x}} \text{ в долях или}$$

$$P = \frac{m}{\bar{x}} \cdot 100\% \text{ в процентах,}$$

где  $m$  – стандартная ошибка,  
 $\bar{x}$  – выборочное среднее значение.

## 2.2.12. Достаточная численность выборки

Анализ репрезентативности выборки (иначе – способности выборки адекватно представить всю генеральную совокупность, популяцию) особенно важен на начальном этапе исследований, когда численность генеральной совокупности неизвестна в принципе, но уже известны некоторые параметры опыта, позволяющие оценить репрезентативность.

Достаточная численность выборки может быть рассчитана как для количественных, так и для качественных выборок.

Метод вычисления достаточной численности количественной выборки основан на формуле







$$n = \frac{t_{(1+\beta)/2}^2 \sigma^2}{\Delta^2},$$

где  $t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с числом степеней свободы  $\infty$  и параметром  $(1 + \beta) / 2$ ,

$\beta$  – доверительный уровень, выраженный в долях, к примеру 0,95 (что соответствует 95%),

$\sigma$  – выборочная оценка стандартного отклонения, к примеру, 50 рублей,

$\Delta$  – абсолютная погрешность определения среднего арифметического значения, к примеру, 5 рублей.

Абсолютная погрешность вводится в именованных числах, т. е. в тех же единицах измерения, что и варианты выборки. Например, при подсчете количества неделимых объектов исследования (например, избирательных бюллетеней) абсолютная погрешность может быть установлена равной 1.

В литературе представлена также формула, аналогичная приведенной выше, за исключением того, что используется значение не обратной функции распределения Стьюдента, а обратной функции нормального распределения

$$n = \frac{\Psi^2((1 + \beta) / 2) \sigma^2}{\Delta^2},$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения.

Метод вычисления достаточной численности качественной выборки основан на формуле

$$n = \frac{\Psi^2((1 + \beta) / 2) \cdot \hat{p} \cdot (1 - \hat{p})}{\Delta^2},$$

где  $\hat{p}$  – выборочная оценка доли, к примеру, 0,35,

$\Delta$  – абсолютная погрешность определения доли, к примеру, 0,05.

Если известна численность популяции  $N$ , а вычисленная достаточная численность оказывается 10% и более от численности популяции, то достаточная численность выборки должна быть скорректирована в соответствии с формулой

$$n' = \frac{nN}{N + n - 1}.$$

О вычислении достаточной численности см. Малхотра, Девятко, Голубкова, Лванга (Lwanga) с соавт., Чау (Chow) с соавт., статьи Делл (Dell) с соавт., Кук (Cook) с соавт. Вычисление численности для различных статистических методов и для исходных данных в различных шкалах см. в статьях Кэмпбелл (Campbell) с соавт., Бонетт (Bonett) с соавт., Вальтер (Walter) с соавт.

## 2.2.13. Критерий Аббе

Для проверки, извлечена ли выборка случайно из нормальной генеральной совокупности либо, с другой точки зрения, независимы ли одинаково нормально распределенные случайные величины, можно воспользоваться критерием Аббе. Статистика критерия (отношение фон Ноймана, von Neuman Ratio) может быть подсчитана по формуле:





$$\gamma = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

где  $n$  – численность выборки,

$x_i, i = 1, 2, \dots, n$  – значения вариант выборки,

$\bar{x}$  – выборочное среднее значение.

В литературе под названием статистики Аббе фигурирует величина

$$q = \gamma / 2.$$

При этом  $P$ -значение может быть вычислено с помощью модифицированной статистики

$$T = (q - 1) \sqrt{\frac{2n + 1}{2 - (q - 1)^2}},$$

которая для больших выборок распределена приближенно нормально по закону  $N(0, 1)$ .

Распределение статистики  $\gamma$  изучил фон Нойманн (von Neumann). Аппроксимацию  $P$ -значений предложили Бингхэм (Bingham) с соавт. См. монографии Браунли, Петровича с соавт., Айвазяна с соавт., справочник Большева с соавт., статьи Хэрта (Hart), Лемешко.

## 2.2.14. Формулы для сгруппированных выборок

Группировка выборок может быть как следствием их естественного исходного представления (номинальная либо порядковая шкала измерения), так и результатом понижения количественной шкалы измерения до порядковой или номинальной шкалы. Более подробная информация о шкалах измерения и их преобразовании приводится в «Введение».

Исходные данные в группированном виде могут, к примеру, иметь следующий вид (пусть верхняя строка – оценка за курсовую работу, а нижняя – число студентов, получивших данную оценку):

$b_i, i = 1, 2, \dots, 5$	1	2	3	4	5
$v_i, i = 1, 2, \dots, 5$	0	1	10	19	25

Здесь обозначено:

$b_i, i = 1, 2, \dots, k$  – середины классовых интервалов (для количественных выборок) либо значения для порядковых и номинальных выборок,

$v_i, i = 1, 2, \dots, k$  – частоты наблюдаемых случаев в классах, иначе – численности классов,

$k$  – число классов (групп).

Для вычислений выборочных показателей используются формулы для среднего значения, среднего отклонения и дисперсии (несмещенная оценка), соответственно, в следующей форме:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k b_i v_i,$$

$$\hat{x} = \frac{1}{n} \sum_{i=1}^k |b_i - \bar{x}| v_i,$$





$$D = \frac{1}{n-1} \sum_{i=1}^k (b_i - \bar{x})^2 v_i$$

либо в эквивалентной форме

$$D = \frac{1}{n-1} \left[ \sum_{i=1}^k b_i^2 v_i - \frac{1}{n} \left( \sum_{i=1}^k b_i v_i \right)^2 \right],$$

где  $n$  – общее число наблюдений, вычисляемое по формуле

$$n = \sum_{i=1}^k v_i.$$

Логика вычислений заключается в суммировании по числу классов и домножении каждого выражения под знаком суммы на соответствующую данному классу частоту. На основании данной информации записать эквивалентные формулы для вычисления других статистических показателей не составит труда.

Статистические показатели, в формулы вычислений которых не входят значения вариантов выборки, вычисляются по тем самым формулам и для негруппированных, и для сгруппированных данных.

## Глава 3. Параметрическая статистика

### 3.1. Введение

Все представленные методы применимы только для анализа выборок признаков, измеренных в количественной шкале.

Серьезной проблемой, которая касается представленных методов проверки гипотез, является применимость методов в случае малой численности выборок, что может иметь следствием низкую мощность. Дополнительно о влиянии численности на мощность критериев см. в главе «Введение в практический анализ».

Число наблюдений (численность выборки) для использования параметрических критериев должно быть по возможности большим. Минимальные численности выборок можно установить по таблицам, данным в книге Джонсона с соавт.

Считается, что параметрические методы могут применяться, только если эмпирическое распределение анализируемых выборок не противоречит статистической гипотезе о нормальности распределения. В этой связи необходимо отметить два обстоятельства:

- Данную проверку можно выполнить с помощью статистических тестов главы «Проверка нормальности распределения» (в данной главе содержатся рекомендации, какие именно параметры выборок подлежат проверке). Перед нами – яркий пример того, когда проверка предпосылок применения метода гораздо сложнее самого метода.
- Перед использованием параметрических методов, если данные не показывают нормальности распределения, возможна их нормализация. Методы нормализации представлены в главе «Преобразования данных».





Исследования показывают, что острота проблемы отклонения от нормальности и утверждение, что выборка тем нормальнее, чем многочисленнее, преувеличена. Ряд авторов посвятил свои исследования данной теме.

См. работы Виккерса (Vickers), Бриджа (Bridge) с соавт., Мюллера с соавт., Блэйр (Blair) с соавт.

## 3.2. Теоретическое обоснование

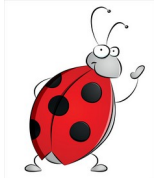
Критерии (тесты), при помощи которых могут быть сравнены статистические совокупности, разделяются на две группы: параметрические и непараметрические. Особенностью параметрических критериев является ряд требований:

- Распределение признака в генеральной совокупности подчиняется некоторому известному, в данном случае нормальному, закону. Нормальность распределения генеральной совокупности может быть статистически установлена на основе проверки эмпирического распределения выборки из данной совокупности до применения любого параметрического теста с помощью одного из методов, представленных в главе «Проверка нормальности распределения». Задача проверки нормальности в целом сложнее задачи проверки гипотезы о математических ожиданиях. Она может быть уверенно решена лишь при больших объемах выборок.
- Для адекватного применения ряда критериев требуется равенство дисперсий сравниваемых выборок. Поэтому многие авторы рекомендуют проверить нулевую гипотезу о равенстве дисперсий сравниваемых совокупностей с помощью критерия Фишера.

Пусть обе выборки извлечены из генеральных совокупностей, имеющих нормальные распределения с равными или неравными между собой неизвестными дисперсиями. Нулевая гипотеза состоит в том, что средние значения совокупностей равны. При анализе выборок из нормальных генеральных совокупностей с неизвестными дисперсиями, равенство которых не предполагается, либо если отношение дисперсий неизвестно, возникает так называемая проблема Беренса–Фишера (Behrens–Fisher problem), решаемая с помощью параметрических методов: критерия Уэлча, критерия Пагуровой или критерия Кокрена–Кокса.

В практических исследованиях решение данной проблемы актуально, т. к. при анализе реальных данных все параметры распределения чаще всего действительно оцениваются по эмпирическим выборкам.

Параметрические критерии в большинстве случаев являются более мощными, чем их непараметрические аналоги. Если существуют предпосылки использования параметрических критериев, но используются непараметрические, увеличивается вероятность ошибки II рода. См. работы Пинто (Pinto), Рейнеке (Reineke).



### 3.2.1. Критерий Стьюдента

Критерий Стьюдента предназначен для проверки нулевой гипотезы о равенстве среднего значения выборочной совокупности заданному математическому ожиданию. Вычисление производится по формуле

$$t = \frac{|\bar{x} - \lambda_0| \sqrt{n}}{s},$$

где  $\bar{x}$  – среднее значение совокупности,

$\lambda_0$  – заданное математическое ожидание,

$n$  – численность совокупности,

$s^2$  – оценка выборочной дисперсии.

Статистика критерия Стьюдента подчиняется  $t$ -распределению с числом степеней свободы  $n - 1$ .

Согласно Мюллеру с соавт. (с. 127, см. также ссылку в источнике), «критерий  $t$  относительно нечувствителен к небольшим отклонениям от распределения генеральной совокупности от нормального (т. е. практически является робастным)».

### 3.2.2. Критерий Чен

Критерий Чен (Chen's test) в качестве обобщения критерия Стьюдента предназначен для проверки нулевой гипотезы о том, что среднее значение выборочной совокупности не превышает заданного математического ожидания

$$\bar{x} \leq \lambda_0,$$

где  $\bar{x}$  – среднее значение совокупности,

$\lambda_0$  – заданное математическое ожидание.

Метод может применяться только при положительном коэффициенте асимметрии.

Вычисление статистики критерия производится по формуле

$$T = t + a(1 + 2t^2) + 4a^2(t + 2t^3),$$

где

$$a = \frac{b}{6\sqrt{n}},$$

$b$  – коэффициент асимметрии,

$n$  – численность совокупности,

$$t = \frac{|\bar{x} - \lambda_0| \sqrt{n}}{s},$$

– статистика критерия Стьюдента,

$s^2$  – оценка выборочной дисперсии.

Статистика критерия подчиняется стандартному нормальному распределению.



### 3.2.3. Критерий Стьюдента для независимых выборок

Критерий Стьюдента для независимых выборок (two-group unpaired  $t$ -test) предназначен для проверки нулевой гипотезы о равенстве средних значений двух нормальных выборочных совокупностей в случае равных неизвестных дисперсий.

Распределение нормальной случайной величины полностью определяется двумя параметрами: математическим ожиданием (его выборочная оценка – среднее значение) и дисперсией. Поэтому в данном случае нулевая гипотеза может быть сформулирована как гипотеза о том, что выборки извлечены из одной статистической популяции.

Вычисление статистики критерия производится по формуле

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s \sqrt{1/n_1 + 1/n_2}},$$

где  $\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей,

$n_1$  и  $n_2$  – численности совокупностей,

$s^2$  – оценка выборочной дисперсии.

Оценка выборочной дисперсии рассчитывается как

$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2},$$

где  $s_1^2$  и  $s_2^2$  – оценки дисперсий, которые считаются по соответствующим выборкам.

Статистика критерия подчиняется  $t$ -распределению с числом степеней свободы  $n_1 + n_2 - 2$ .

Доверительные интервалы для оцениваемой разности средних значений вычислены в статье Сим (Sim) с соавт. Хотя оригинальный критерий изначально предназначен для нормальных количественных выборок, имеется исследование Хирен (Heeren) с соавт. о применении рассмотренного теста к порядковым выборкам.

### 3.2.4. Парный критерий Стьюдента

Критерий Стьюдента для связанных выборок (парный критерий Стьюдента, two-group paired  $t$ -test) предназначен для проверки нулевой гипотезы о равенстве средних значений двух выборочных совокупностей в случае неравных неизвестных дисперсий. В источниках критерий может называться одновыборочным критерием Стьюдента. Это название вызвано тем обстоятельством, что на самом деле, исходя из представленной схемы расчета, анализируется действительно одна совокупность, составленная из попарных разностей вариант исходных связанных выборок. Понятно, что в данном случае проверяется нулевая гипотеза о равенстве среднего значения полученной выборки известному значению, а именно – нулю.

Вычисления производятся по формуле



$$t = \frac{\sum_{i=1}^n \delta_i}{\sqrt{\frac{n \sum_{i=1}^n \delta_i^2 - \left( \sum_{i=1}^n \delta_i \right)^2}{n-1}}},$$

где  $n$  – численность каждой выборки,

$\delta_i = x_i - y_i$ ,  $i = 1, 2, \dots, n$  – попарные разности вариант совокупностей, где

$x_i$ ,  $i = 1, 2, \dots, n$  – варианты первой совокупности,

$y_i$ ,  $i = 1, 2, \dots, n$  – варианты второй совокупности.

Статистика имеет распределение Стьюдента с числом степеней свободы  $n - 1$ .

### 3.2.5. Критерий Лорда

Критерий Лорда (Lord's range test) разработан для проверки нулевой гипотезы о равенстве средних двух совокупностей. Статистика критерия вычисляется по формуле

$$L = \frac{|\bar{x}_1 - \bar{x}_2|}{r_1 + r_2},$$

где  $\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей,

$r_1$  и  $r_2$  – значения размахов. Подробнее о размахе см. главу «Описательная статистика».

Статистику применяют для очень малых выборок. В таблице представлены уровни значимости. Значение  $L$ , равное или большее табличного значения, значимо.

$n_1$	$n_2$	5%	1%
2	2	1,71	3,96
3	3	0,64	1,05
4	4	0,41	0,62

Описание критерия и ссылки даны для полноты информации.

Метод представлен в книге Закса, монографии Лэнгли (Langley). См. также работу Пэтнэйка (Patnaik).

### 3.2.6. Критерий Уэлча

Критерий Уэлча (критерий Велча, критерий Вэлча, критерий Крамера–Уэлча, критерий Саттерзвайта, Satterthwaite's test) предназначен для решения проблемы Беренса–Фишера.

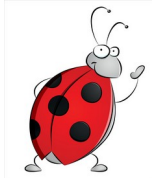
Вычисления производятся по формуле

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

где  $\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей,

$n_1$  и  $n_2$  – численности совокупностей,

$s_1^2$  и  $s_2^2$  – оценки дисперсий, которые считаются по соответствующим выборкам.



Распределение статистики критерия близко к  $t$ -распределению Стюдента при числе степеней свободы, равном

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

Описание критерия см. в книге Закса, Когана с соавт. См. также описание критерия Юена–Уэлча (Yuen–Welch test) в книге Вилкокса (Wilcox).

Уместно указать еще одну модификацию критерия Стюдента, предложенную Хатчесоном (Hutcheson) и предназначенную для сравнения индексов Шеннона двух совокупностей (см. главу «Информационный анализ»):

$$t = \frac{|H_1 - H_2|}{\sqrt{D_{H_1}^2/n_1 + D_{H_2}^2/n_2}},$$

где  $H_1$  и  $H_2$  – индексы Шеннона (энтропии) совокупностей,

$D_{H_1}$  и  $D_{H_2}$  – соответствующие оценки дисперсий индексов Шеннона.

Распределение статистики критерия Хатчесона близко к  $t$ -распределению Стюдента при числе степеней свободы, равном

$$\nu = \frac{(D_1 + D_2)^2}{D_{H_1}^2/n_1 + D_{H_2}^2/n_2}.$$

См. статью Хатчесона, работу Шитикова с соавт.

### 3.2.7. Критерий Пагуровой

Приближенное решение проблемы Беренса–Фишера дано Пагуровой, которая предположила, что распределение статистики критерия существенно зависит от отношения неизвестных дисперсий. Вычисление критерия Пагуровой производится по формуле, аналогичной формуле Уэлча,

$$\nu = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

где  $\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей,

$n_1$  и  $n_2$  – численности совокупностей,

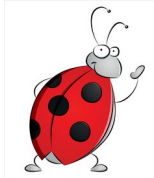
$s_1^2$  и  $s_2^2$  – оценки дисперсий, которые считаются по соответствующим выборкам.

Двустороннее  $P$ -значение вычисляется как решение нелинейного уравнения

$$\nu = t_{n_2-1, p/2} \frac{(\theta - \eta)^2 (1 - \eta)}{\theta^2} + t_{n_1+n_2-1, p/2} \frac{[\theta(1 - \theta) + (2\theta - 1)(\eta - \theta)]\eta(1 - \eta)}{\theta^2(1 - \theta)^2} + t_{n_1-1, p/2} \frac{(\theta - \eta)^2 \eta}{(1 - \theta)^2},$$







где  $t_{\alpha}$  – значение обратной функции  $t$ -распределения,

$$\eta = c - 2c(1 - c) \left( \frac{1 - c}{n_2} - \frac{c}{n_1} \right),$$

$$\theta = \frac{n_1}{n_1 + n_2},$$

$$c = \frac{s_1^2 / n_1}{s_1^2 / n_1 + s_2^2 / n_2}.$$

Уравнение может быть решено одним из методов локальной оптимизации. В простейшем случае используется метод деления отрезка пополам.

Описание критерия приводится в работе Пагуровой.

### 3.2.8. Критерий Кокрена–Кокса

Критерий Кокрена–Кокса (Cochran and Cox test) предназначен для решения проблемы Беренса–Фишера. Вычисления производятся по формуле

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}},$$

где  $\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей,

$n_1$  и  $n_2$  – численности совокупностей,

$s_1^2$  и  $s_2^2$  – оценки дисперсий, которые считаются по соответствующим выборкам.

Распределение статистики критерия близко к  $t$ -распределению Стьюдента при числе степеней свободы, равном

$$\nu = \frac{(s_1^2 / (n_1 - 1) + s_2^2 / (n_2 - 1))^2}{\frac{(s_1^2 / (n_1 - 1))^2}{n_1 + 1} + \frac{(s_2^2 / (n_2 - 1))^2}{n_2 + 1}} - 2.$$

### 3.2.9. Критерий Крамера

Критерий Крамера предназначен для проверки нулевой гипотезы о равенстве средних значений двух выборочных совокупностей в случае равных неизвестных дисперсий.

Вычисление статистики критерия производится по формуле

$$t = \frac{\sqrt{n_1 n_2} |\bar{x}_1 - \bar{x}_2|}{\sqrt{n_2 s_1^2 + n_1 s_2^2}},$$

где  $\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей,

$n_1$  и  $n_2$  – численности совокупностей,





$s_1^2$  и  $s_2^2$  – оценки дисперсий, которые считаются по соответствующим выборкам. Статистика критерия подчиняется стандартному нормальному распределению. См. монографию Крамера.

### 3.2.10. Критерий Фишера

F–критерий Фишера (критерий Фишера–Снедекора) применяют для сравнения дисперсий двух нормальных выборочных совокупностей. Критерий часто называют дисперсионным отношением или просто статистикой Фишера. Вычисление ведется по формуле, предложенной Снедекором:

$$F = \frac{s_1^2}{s_2^2},$$

где в числителе – оценка дисперсии одной выборки, в знаменателе – оценка дисперсии другой выборки. Принято (см. Лакина) брать отношение большего значения дисперсии к меньшему значению, хотя принципиальной разницы нет.

Числа степеней свободы для поиска критического значения по таблице F–распределения (данная таблица – двухвходовая) следует взять  $n_1 - 1$  и  $n_2 - 1$ , где  $n_1$  и  $n_2$  – соответствующие численности совокупностей.

См. книгу Когана с соавт.

### 3.2.11. Трансгрессия

У независимых выборок из различных генеральных совокупностей часть вариантов может оказаться в одних и тех же классах вариационного ряда. Такие ряды называются трансгрессирующими, а их неполное разобщение – трансгрессией. При статистически доказанном различии в средних значениях большая величина трансгрессии (которая может выражаться в долях или в процентах) заставляет предположить, что разделение рядов по анализируемому фактору не является единственным.

В случае нормальных генеральных совокупностей трансгрессия вычисляется по формуле:

$$Tr = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2},$$

где  $n_1$  и  $n_2$  – численности совокупностей.

Остальные величины вычисляются по формулам, соответственно,

$$P_1 = 0,5 + 0,5 \cdot I \left( \frac{\min_2 - \bar{x}_1}{s_1} \right) \quad \text{и} \quad P_2 = 0,5 + 0,5 \cdot I \left( \frac{\max_1 - \bar{x}_2}{s_2} \right),$$

где  $I(.)$  – интеграл вероятностей,

$\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей,

$s_1^2$  и  $s_2^2$  – оценки дисперсий, которые считаются по соответствующим выборкам, а остальные величины вычисляются по формулам, соответственно,

$\min_2 = \bar{x}_2 - 3s_2$  и  $\max_1 = \bar{x}_1 + 3s_1$ ,





Если окажется, что  $\min_2 > \bar{x}_1$  или  $\max_1 < \bar{x}_2$ , то значения величин  $P_1$  и  $P_2$  рассчитываются по формулам, соответственно,

$$P_1 = 0,5 - 0,5 \cdot I \left( \frac{\min_2 - \bar{x}_1}{s_1} \right) \quad \text{и} \quad P_2 = 0,5 - 0,5 \cdot I \left( \frac{\max_1 - \bar{x}_2}{s_2} \right).$$

См. монографию Лакина.

### 3.2.12. График средних значений с доверительными интервалами

Доверительные интервалы оцениваемых средних значений нормальных выборок вычисляются по формуле

$$I_m = \left[ \bar{x} - t_{(1+\beta)/2} \frac{s}{\sqrt{n}}; \bar{x} + t_{(1+\beta)/2} \frac{s}{\sqrt{n}} \right],$$

где  $s$  – выборочная оценка стандартного отклонения,

$t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $n - 1$  и  $(1 + \beta) / 2$ ,

$\beta$  – доверительный уровень, выраженный в долях.

Для вычисления двустороннего доверительных интервалов оцениваемых средних значений, когда выборки не являются нормальными, применяется формула:

$$I_m = \left[ \bar{x} - \Psi((1 + \beta) / 2) \frac{s}{\sqrt{n}}; \bar{x} + \Psi((1 + \beta) / 2) \frac{s}{\sqrt{n}} \right],$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения.

Дополнительно в таблице выводится разность средних анализируемых выборок

$(\bar{x}_1 - \bar{x}_2)$ ,

где  $\bar{x}_1$  и  $\bar{x}_2$  – средние значения совокупностей.

Доверительный интервал оцениваемой разности средних значений (выборки нормальные) вычисляется по формуле

$$I_d = \left[ (\bar{x}_1 - \bar{x}_2) - t_{(1+\beta)/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; (\bar{x}_1 - \bar{x}_2) + t_{(1+\beta)/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right],$$

где  $s_1^2$  и  $s_2^2$  – оценки дисперсий, которые считаются по соответствующим выборкам.

$n_1$  и  $n_2$  – численности совокупностей,

$t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $\nu$  (число степеней свободы) и  $(1 + \beta) / 2$ . При этом число степеней свободы считается как

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$



Доверительный интервал оцениваемой разности средних значений (выборки не являются нормальными) вычисляется по формуле

$$I_d = \left( (\bar{x}_1 - \bar{x}_2) - \Psi((1 + \beta)/2) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; (\bar{x}_1 - \bar{x}_2) + \Psi((1 + \beta)/2) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right).$$

Результаты графического анализа интерпретируются следующим образом. Если 100β% доверительные интервалы оцениваемых средних значений сравниваемых выборок пересекаются, конкурирующая гипотеза (средние не равны) может быть принята на уровне значимости  $p \leq \beta$ . Если 100β% доверительные интервалы оцениваемых средних значений сравниваемых выборок не пересекаются, нулевая гипотеза (средние равны) не отвергается на уровне значимости  $p > \beta$ . Т. к. доверительные интервалы тем шире, чем больше значение β, выбирая различные стандартные значения β, можно получить значение уровня значимости, более точно соответствующее представленным данным.

См. статьи Массон (Masson) с соавт., Вольфе (Wolfe) с соавт., Пэйтон (Payton) с соавт., Остин (Austin) с соавт., Маршалл (Marshall).

### 3.2.13. Отношения средних и дисперсий

Рассматривается возможность вычисления точечных и интервальных оценок отношения средних и отношения дисперсий двух нормальных выборок. Подразумевается, что первая выборка – та, соответствующее значение которой стоит в числителе, вторая выборка – в знаменателе.

Для вычисления доверительного интервала оцениваемого отношения средних значений двух выборок

$$q = \frac{m_1}{m_2},$$

$m_1$  – среднее значение первой выборки,

$m_2$  – среднее значение второй выборки,

сначала вычисляется промежуточная переменная

$$g = \left[ t_{(1+\beta)/2} \frac{\mu_2}{m_2} \right]^2,$$

где  $t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $n_1 + n_2 - 2$  и  $(1 + \beta) / 2$ ,

β – доверительный уровень, выраженный в долях,

$\mu_2$  – стандартная ошибка среднего значения второй выборки.

Дальнейшие вычисления зависят от значения промежуточной переменной, которая является численной характеристикой отношения стандартной ошибки среднего значения второй выборки к самому ее среднему значению.

1. При  $g \geq 1$  искомой интервальной оценки не существует.
2. При малом значении  $g$  стандартная ошибка отношения средних значений вычисляется по формуле





$$SE_q = q \sqrt{\frac{\mu_1}{m_1} + \frac{\mu_2}{m_2}},$$

где  $\mu_1$  – стандартная ошибка среднего значения первой выборки.

При этом доверительный интервал оцениваемого отношения средних значений

$$I_{\frac{m_1}{m_2}} = (q - t_{(1+\beta)/2} SE_q; q + t_{(1+\beta)/2} SE_q).$$

3. При большом значении  $g$  стандартная ошибка отношения средних значений вычисляется по уточненной формуле

$$SE_q = \frac{q}{1-g} \sqrt{(1-g) \left( \frac{\mu_1}{m_1} + \frac{\mu_2}{m_2} \right)},$$

При этом доверительный интервал оцениваемого отношения средних значений

$$I_{\frac{m_1}{m_2}} = \left( \frac{q}{1-g} - t_{(1+\beta)/2} SE_q; \frac{q}{1-g} + t_{(1+\beta)/2} SE_q \right).$$

Доверительный интервал оцениваемого отношения дисперсий вычисляется по формуле

$$I_{\frac{\sigma_1^2}{\sigma_2^2}} = \left( \frac{s_1^2}{s_2^2} F_{(1+\beta)/2}^{-1}(n_1 - 1, n_2 - 1); \frac{s_1^2}{s_2^2} F_{1-(1+\beta)/2}^{-1}(n_1 - 1, n_2 - 1) \right),$$

где  $s_1^2$  – выборочное значение дисперсии 1-й выборки,

$s_2^2$  – выборочное значение дисперсии 2-й выборки,

$n_1$  и  $n_2$  – численности совокупностей,

$F^{-1}(\cdot, \cdot)$  – обратная функция  $F$ -распределения.

Алгоритмы вычислений и поясняющие примеры см. в монографиях Мотульски (Motulsky), Бетеа (Bethea) с соавт. См. также статью Ли (Lee A.F.S.) с соавт.

## Глава 4. Непараметрическая статистика

### 4.1. Введение

Рассматриваются непараметрические методы проверки статистических гипотез и методы анализа качественных (бинарных) данных. Бытует несколько основных соображений относительно полезности непараметрических методов:

- Параметрические методы могут применяться, только если доказана нормальность распределения анализируемых выборок, но эмпирические выборки, полученные в реальных экспериментах, очень часто не являются нормально распределенными.
- Параметрические методы могут применяться для больших выборок. Реальные выборки часто содержат небольшое число вариантов, что тем более делает полезным непараметрические методы.





Исследования показывают, что острота проблемы отклонения от нормальности преувеличена, а утверждение, что выборка тем нормальнее, чем многочисленнее, не имеет основания. Ряд авторов посвятил свои исследования данной теме. См. работы Виккерса (Vickers), Бриджа (Bridge) с соавт., Мюллера с соавт., Блэйр (Blair) с соавт. Серьезной проблемой, которая касается представленных методов проверки гипотез так же, как и параметрических, является применимость методов в случае малой численности выборок, что может иметь следствием низкую мощность критерия (напомним, что мощность – это не число, а монотонная функция численности – чем больше численности выборок, тем выше мощность критерия, к тому же зависящая от альтернативы). Дополнительно о влиянии численности на мощность критериев см. в главе «Введение в практический анализ». Перед применением любого статистического метода необходимо убедиться, что проверяется статистическая значимость различий именно тех параметров выборок, которые интересуют исследователя, а также в том, что метод соответствует шкале измерения исходных данных (признаков). О шкалах измерения см. главу «Введение».

## 4.2. Теоретическое обоснование

Существует большое количество опытных данных, которые не показывают нормальности распределения, поэтому применение параметрических критериев не может быть обоснованным для данных рассматриваемого класса.

Практически ценными явились робастные методы, которые применимы в широком диапазоне условий. Робастные, непараметрические и свободные от распределения процедуры традиционно относят к одному классу, хотя в литературе есть и альтернативные мнения. Сам термин «непараметрическая статистика» был введен в 1942 году Вольфовицем.

### 4.2.1. Робастность

Подробное обсуждение этой темы приводится Хьюбером. Под робастностью мы понимаем слабую чувствительность к отклонениям от стандартных условий (например, эмпирическое распределение может отличаться от теоретического нормального), а методы, применимые в широком диапазоне реальных условий, называем робастными. В этом качестве понятие робастности статистического метода практически совпадает со смыслом данного понятия, которое вкладывается в него в механике и смежных прикладных дисциплинах.

Понятие робастности не тождественно устойчивости статистической процедуры (не путать с численной устойчивостью алгоритма). Как указывает Хьюбер, статистическую процедуру называют устойчивой, если на значение оценки не оказывают влияния малые изменения в выборке (малые изменения всех или большие изменения нескольких значений – см. «Обработка выбросов»). Понятия устойчивости и робастности различны, но иногда их применяют в качестве синонимов.

Непараметрические критерии не требуют предварительных предположений относительно вида исходного распределения и являются более робастными, чем их параметрические аналоги. Их называют также критериями значимости, независимыми от типа распределения. Естественно, непараметрические критерии применимы и для случая нормального



распределения. Однако непараметрические критерии в большинстве случаев являются менее мощными, чем их параметрические аналоги. Если существуют предпосылки использования параметрических критериев, но используются непараметрические, увеличивается вероятность ошибки II рода.

## 4.2.2. Тестируемые параметры

Многие пользователи задают вопрос, почему, к примеру, одним методом между выборками выявляются статистически значимые различия, другим – нет. Дело в том, что все методы предназначены для проверки отсутствия статистических различий в различных параметрах (иногда – в совокупности параметров) выборок. Так, можно себе представить такие выборки, которые имеют одинаковые параметры положения (медианы), но разные параметры рассеяния (дисперсии). В таком гипотетическом случае критерий Ансари–Бредли покажет наличие различий, критерий Вилкоксона – нет. Становится понятным, почему исследователи часто не ограничиваются одним тестом, а пытаются выполнить их совокупность для статистического сравнения всевозможных параметров выборок: средних, медиан, дисперсий, функций распределения.

При формулировании нулевой гипотезы обязательно следует указывать, какие конкретные параметры эмпирических выборок сравниваются с помощью используемого критерия. Данная информация приводится в описании каждого критерия. Нужно указывать это в научной публикации, чтобы читатель имел возможность проверить правильность рассуждений автора. В таблице указаны тестируемые параметры выборок для различных критериев.

Тестируемые параметры	Статистический критерий
Положение (location tests)	Вилкоксона, Манна–Уитни, Ван дер Вардена, Уайта, Фишера–Йейтса–Терри–Гефтинга, Розенбаума, медианы, медианный Муда–Брауна, Гехана, Блома, Тьюки, Мак–Немара, серий Вальда–Вольфовица
Рассеяние/масштаб (scale tests)	Ансари–Бредли, Клотца, Сэвиджа, Коновера, Муда, Дэвида, Зигеля–Тьюки, Мозеса
Функция распределения	Смирнова, Крамера–фон Мизеса, Койпера, Лемана–Розенблатта

В показанной таблице не конкретизировано, какие именно параметры являются параметрами положения, а какие параметрами рассеяния. Уточнение приводится в таблице.

Параметр	Параметрика	Непараметрика
Положение	Среднее значение	Медиана или псевдомедиана (оценка Ходжеса–Лемана)
Рассеяние	Стандартное отклонение <sup>3</sup>	Межквартильный размах или

<sup>3</sup> В качестве параметра рассеяния применяют дисперсию, однако в данном случае удобно взять стандартное отклонение для сопоставления с параметром рассеяния в непараметрическом случае.



семиинтерквартильная широта

Подробнее обо всех перечисленных параметрах см. главу «Описательная статистика».

### 4.2.3. Типы критериев

Все непараметрические критерии проверки гипотез, в зависимости от их конструкции, могут принадлежать к одному из следующих типов:

- ранговые критерии (рангом называют номер варианты в ряду упорядоченных по возрастанию или убыванию вариантов),
- критерии, основанные на сравнении функций распределения,
- точные критерии.

Представленное разделение критериев на типы очень условно и часто относится только к реализации. Лучше говорить о тестируемых параметрах, как это описано в предыдущем разделе. В описании некоторых критериев авторами устанавливаются параллели между ранговыми и перестановочными критериями, ранговыми критериями и критериями на основе функций распределения. Описаны комбинаторные алгоритмы вычисления ранговых критериев. К точным критериям относятся как перестановочные критерии для таблиц сопряженности, являющихся продуктом анализа номинальных признаков, так и критерии первых других типов, для которых известно (и практически применимо) точное распределение статистик.

Многие из представленных критериев имеют многомерные аналоги, представленные в главе «Дисперсионный анализ».

См. монографии Холлендера с соавт., Гаека с соавт., Хеттсманпергера, Коновера (Conover), Руниона, нормативный документ EPA QA/G-9.

#### 4.2.3.1. Ранговые критерии

К ранговым критериям относятся:

- критерий Вилкоксона для независимых выборок,
- критерий Вилкоксона для связанных выборок,
- критерий Манна–Уитни,
- критерий Ван дер Вардена,
- критерий Сэвиджа,
- критерий Ансари–Бредли,
- критерий Клотца,
- критерий Зигеля–Тьюки,
- критерий Коновера,
- медианный критерий Муда–Брауна.

Некоторые из представленных тестов являются эквивалентными. Критерии называются эквивалентными, по определению Холлендера и Вулфа, если для любых возможных выборок





решение, принятое с помощью одного из критериев, согласуется с решением, принятым с помощью другого критерия.

Нетрудно показать эквивалентность ряда критериев. Для упомянутых методов Клотц дает следующую формулу:

$$(T + T' + 1) / 4 = W = (N / 2 + 1) / N / 2 - S,$$

где  $T$  и  $T'$  – статистики Зигеля–Тьюки,

$W$  – статистика Ансари–Бредли,

$S$  – статистика Бартон–Дэвида,

$N$  – численность объединенной выборки.

Также эквиваленты критерии Вилкоксона (для независимых выборок, без учета поправок) и Манна–Уитни. Простая формула их связи имеет вид

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W,$$

где  $U$  – статистика Манна–Уитни,

$W$  – статистика Вилкоксона,

$n_1$  – численность той выборки, для которой вычислялись статистики,

$n_2$  – численность другой выборки.

Данные критерии называются ранговыми (Ф. Вилкоксон, 1945 г.), так как они оперируют не численными значениями вариантов, а их рангами. Сначала производят совместное ранжирование сравниваемых выборок. Данная процедура может быть организована различными способами, однако предпочтительным в смысле простоты понимания процесса и его реализации является объединение двух сравниваемых выборок, их сортировка, ранжирование по требуемой схеме и последующее разнесение рангов на места соответствующих им вариант в обеих выборках. Если имеются совпадающие значения, совпавшим наблюдениям условились назначать средний ранг.

Ранговые критерии могут применяться к признакам, измеренным в количественной или порядковой шкале. Применение ранговых критериев к количественным признакам фактически понижает исходную количественную шкалу до порядковой шкалы (напомним, что ранг – это номер варианты по порядку в ранжированном ряду). Это вызывает опасение некоторых авторов, хотя в литературе показано, что точность выводов снижается гораздо меньше, чем можно было бы себе вообразить.

Схемы вычислений всех ранговых критериев могут быть описаны одними и теми же универсальными соотношениями, отличающимися только способом вычисления ранговых отметок (функций от рангов). Кроме того, перед ранжированием исходные выборки, в зависимости от схемы алгоритма, могут быть подвергнуты преобразованиям.

Обозначим:

$N = n_1 + n_2$  – общее число наблюдений в двух тестируемых выборках, которое может быть скорректировано при наличии совпадающих вариант,

$n_1$  – число наблюдений в одной выборке,

$n_2$  – число наблюдений в другой выборке.



Общая формула вычисления статистики рангового критерия, согласно Хеттсманпергеру, может быть представлена в виде

$$S = \sum_{i=1}^{n_1} a(R_i),$$

где  $R_i, i = 1, 2, \dots, n_1$ , – ранги наблюдений выборки,

$a(R_i), i = 1, 2, \dots, n_1$ , – ранговые метки общего вида.

Для статистик ранговых критериев могут быть известны точные формулы вычисления критических значений, однако вычисления по точным формулам часто трудоемки уже при средних и всегда при больших численностях выборок. Подробнее о вычислениях распределений см. главу «Введение».

Они удобны для построения точных статистических таблиц, однако в практических вычислениях, как показали Гаек и Шидак, может применяться нормальная аппроксимация статистики рангового критерия

$$Z = \frac{S - ES}{\sqrt{DS}},$$

где  $ES$  – математическое ожидание,

$DS$  – дисперсия, которая может быть скорректирована при наличии связей.

Параметры нормального распределения вычисляются по формулам, данным Хеттсманпергером,

$$ES = n_1 \bar{a},$$

$$DS = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^N (a(R_i) - \bar{a})^2,$$

$$\bar{a} = \frac{1}{N} \sum_{i=1}^N a(R_i).$$

где

Обратите внимание, что суммирование в параметрах нормальной аппроксимации производится по обеим выборкам, тогда как статистика критерия вычисляется для одной (любой) из выборок.

Возможно и точное вычисление  $P$ -значений ранговых критериев. Например, методика точного вычисления критериев Вилкоксона полностью совпадает с соответствующими критериями рандомизации компонент, представленными в главе «Точные критерии», с той разницей, что все манипуляции производятся не с вариантами выборок, а с их рангами. По этой причине критерии Вилкоксона могут быть интерпретированы как критерии ранговой рандомизации.



## 4.2.3.1. Учет связей

Связкой (ties) называют совпадающие ранги. При наличии связей статистика критерия (точнее, дисперсия при нормальной аппроксимации статистики критерия) корректируется с помощью особым образом вычисляемой поправки на объединение рангов.

### 4.2.3.1.2. Учет поправки на непрерывность

Поправка на непрерывность (continuity) фактически вводится в формулу вычисления нормальной аппроксимации статистики критерия, т. к. дискретное распределение ранговой статистики аппроксимируется непрерывным нормальным распределением.

См. результаты Пури (Puri), Раджарама (Rajaram).

### 4.2.3.1.3. Критерий Вилкоксона для независимых выборок

W-критерий Вилкоксона (критерий ранговых сумм Вилкоксона, двухвыборочный критерий Вилкоксона, статистика ранговой суммы Уилкоксона, Wilcoxon signed-rank test, Wilcoxon sum-of-ranks test for comparing two unmatched samples) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Вычисление статистики критерия производится по формуле:

$$W = \min \left( \sum_{i=1}^{n_1} R_i, \sum_{i=1}^{n_2} S_i \right),$$

где  $R_i, i = 1, 2, \dots, n_1$  – ранги выборки, имеющей наименьшую сумму рангов,

$S_i, i = 1, 2, \dots, n_2$  – ранги выборки, имеющей наибольшую сумму рангов.

Другой прием – в качестве статистики критерия берется сумма рангов выборки наименьшей численности, хотя принципиальной разницы тут нет.

Вычисленное значение статистики критерия сравнивается с точным критическим значением, однако при большой численности выборок данными формулами пользуются неохотно из-за определенных вычислительных сложностей. Формулы пригодны для построения таблиц, но для практического вычисления значимости критерия применяется подход, учитывающий факт, что статистика

$\frac{W - EW}{\sqrt{DW}}$  распределена по стандартному нормальному закону.

Здесь обозначено:

$EW = n_1(N + 1) / 2$  – математическое ожидание,

дисперсия без связей  $DW = n_1 n_2 (N + 1) / 12$



$$DW = \frac{n_1 n_2}{12} \left[ N + 1 - \frac{b}{N(N-1)} \right],$$

или, при наличии связей,

$N = n_1 + n_2$  – численность объединенной выборки.

$$b = \sum_{j=1}^g t_j (t_j^2 - 1)$$

– поправка на объединение рангов,

где  $t_j, j = 1, 2, \dots, g$  – численность связи,

$g$  – число связей.

При расчете числа связей, при наличии хотя бы одной связи, учитываются также все группы с численностью 1, что, однако, исключает учет данных групп (подобно критерию Ансари–Бредли) из-за особенностей вычисления поправки на объединение рангов. В отчете Хельзель (Helsel) с соавт. со ссылкой на Коновер (Conover) приводится иной способ учета связей.

Если полученное значение статистики превышает 0,02, то в формулу вводится поправка на непрерывность: считается, что новое значение наименьшей суммы рангов равно  $W + 0,5$ .

Критерий рекомендуется для выборок умеренной численности (численность каждой выборки от 12 до 40).

Имеется простая формула связи рассматриваемого критерия с критерием Манна–Уитни, поэтому представленный тест в некоторых источниках носит наименование критерия Вилкоксона–Манна–Уитни.

См. Когана с соавт., Черник (Chernick) с соавт., статью ЛаВанж (LaVange) с соавт. Точное вычисление распределения статистики Вилкоксона см. в работе Лемана (Lehman). Влияние различных поправок в критериях Вилкоксона–Манна–Уитни рассмотрено в работе Бергмана (Bergmann) с соавт. На связь статистики критерия Вилкоксона и площади, отсекаемой ROC кривой (AUC), указано в монографии Власова.

#### 4.2.3.1.4. Критерий Вилкоксона для связанных выборок

$T$ –критерий Вилкоксона (знаковый ранговый критерий Уилкоксона, критерий знаковых рангов Уилкоксона, Wilcoxon signed-ranks test for matched pairs), в отличие от  $W$ –критерия Вилкоксона, применяется для проверки однородности двух совокупностей с попарно сопряженными вариантами. Выборки могут принадлежать порядковой или количественной шкале.

Критерием проверяется статистическая значимость нулевой гипотезы о том, что распределение случайных величин симметрично относительно нуля. Эти случайные величины в рассматриваемом случае представляют собой разности случайных величин, соответствующих двум другим выборкам, поэтому часто критерий называют одновыборочным критерием Вилкоксона. Другое название критерия – критерий Вилкоксона для сопряженных пар,  $T$ –дельта–критерий,  $W$ –критерий Вилкоксона либо просто критерий Вилкоксона.

Методика приближенного вычисления похожа на процедуру вычисления  $W$ –критерия Вилкоксона, однако здесь мы оперируем абсолютными величинами разностей вариант.



Массив разностей ранжируется. Если среди разностей есть нулевые, они отбрасываются (при этом численность сокращается на число отброшенных нулевых разностей). Затем рангам добавляются знаки разностей, и вычисляется наименьшая из сумм положительных  $W^+$  рангов, которая сравнивается с точным критическим значением, однако при большой численности выборок данными формулами пользуются неохотно из-за определенных вычислительных сложностей. Формулы пригодны для построения таблиц, но для практического вычисления значимости критерия применяется подход, учитывающий факт, что статистика

$$\frac{W^+ - EW^+}{\sqrt{DW^+}}$$
 распределена по стандартному нормальному закону.

Здесь обозначено:

где  $EW^+ = N(N + 1) / 4$  – математическое ожидание,

дисперсия без связей  $DW^+ = N(N + 1)(2N + 1) / 24$

$$DW^+ = \frac{1}{24} \left[ N(N + 1)(2N + 1) - \frac{b}{2} \right],$$

или, при наличии связей,

$N$  – численность каждого ряда (после отбрасывания нулевых значений),

$$b = \sum_{j=1}^g t_j(t_j^2 - 1)$$

– поправка на объединение рангов,

где  $t_j, j = 1, 2, \dots, g$  – численность связи,

$g$  – число связей, причем, при наличии хотя бы одной связи, следовало бы учитывать также все группы с численностью 1; однако учет данных групп (подобно критерию Ансари–Бредли) из-за особенностей вычисления поправки на объединение рангов исключен из алгоритма (данные слагаемые – нулевые).

Критерий рекомендуется для выборок умеренной численности (численность каждой выборки от 12 до 40).

Критерий описан практически во всех источниках, посвященных проверке гипотез, непараметрической статистике и ранговым критериям, в частности. Критерий популярен среди биостатистиков. См. например, книгу Черник (Chernick) с соавт.

#### 4.2.3.1.5. Критерий Манна–Уитни

$U$ -критерий Манна–Уитни (Вилкоксона–Манна–Уитни) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности.

Выборки могут принадлежать порядковой или количественной шкале. Наблюдения должны быть независимыми (непарными). Вычисления могут производиться по формулам (в источниках описаны различные схемы, приводящие к аналогичным результатам)

$$U_1 = n_1 n_2 + n_1(n_1 + 1) / 2 - R_1,$$

$$U_2 = n_1 n_2 + n_2(n_2 + 1) / 2 - R_2,$$





$$U = \max(U_1, U_2),$$

где  $R_1$  и  $R_2$  – суммы рангов выборок,

$n_1$  и  $n_2$  – численности соответствующих выборок.

Вычисленное значение статистики критерия сравнивается с точным критическим значением распределения Манна–Уитни, однако при большой численности выборок данными формулами пользуются неохотно из-за определенных вычислительных сложностей.

Формулы пригодны для построения таблиц, но для практического вычисления значимости критерия применяется подход, учитывающий факт, что статистика

$$\frac{U - EU}{\sqrt{DU}},$$

где  $EU = n_1 n_2 / 2$  – математическое ожидание,

$DU = n_1 n_2 (N + 1) / 12$  – дисперсия, которая в случае наличия связей корректируется,

$N = n_1 + n_2$  – численность объединенной выборки.

распределена по стандартному нормальному закону.

Критерий эквивалентен критерию Вилкоксона. Статистические свойства  $U$ -критерия Манна–Уитни и  $W$ -критерия Вилкоксона совпадают. Отметим только, что в критерии Манна–Уитни не используются поправки, разработанные для критерия Вилкоксона, поэтому результаты расчета для одних и тех же данных могут различаться.

См. монографию Уилкса. Точное вычисление распределения статистики Манна–Уитни см. в работе Манна (Mann) с соавт.

#### 4.2.3.1.6. Критерий Ван дер Вардена

Ранговый  $X$ -критерий Ван дер Вардена (Van der Waerden's  $X$ -test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности.

Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$X = \sum_{i=1}^{n_1} \Psi \left( \frac{R_i}{N+1} \right),$$

где  $n_1$  – численность одной выборки,

$n_2$  – численность другой выборки,

$R_i, i = 1, 2, \dots, n_1$  – ранговые метки одной из выборок,

$\Psi(\cdot)$  – функция, обратная функции стандартного нормального распределения,

$N = n_1 + n_2$  – численность объединенной выборки.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{X - EX}{\sqrt{DX}},$$

где  $EX = 0$  – математическое ожидание,





$$DX = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^N \left[ \Psi \left( \frac{i}{N+1} \right) \right]^2 - \text{дисперсия.}$$

распределена по стандартному нормальному закону.

См. также родственный представленному тесту критерий Флигнера–Киллина, описанный Гарретом (Garrett) с соавт.

#### 4.2.3.1.7. Критерий Сэвиджа

Критерий Сэвиджа предназначен для проверки применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Предложено несколько эквивалентных форм записи формулы вычисления статистики критерия. Статистика критерия вычисляется по формуле

$$S = \sum_{i=1}^{n_1} \sum_{j=N+1-R_i}^N \frac{1}{j},$$

где  $R_i$ ,  $i = 1, 2, \dots, n_1$  – ранги выборки с наибольшей численностью,

$N = n_1 + n_2$  – численность объединенной выборки,

$n_1$  – численность выборки с наибольшей численностью,

$n_2$  – численность выборки с наименьшей численностью.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{S - ES}{\sqrt{DS}},$$

где  $ES = n_1$  – математическое ожидание,

$$DS = \frac{n_1 n_2}{N-1} \left( 1 - \frac{1}{N} \sum_{j=1}^N \frac{1}{j} \right) - \text{дисперсия.}$$

распределена по стандартному нормальному закону.

Обобщением критерия Сэвиджа является широко известный критерий Кокса

(логарифмический ранговый критерий), иногда называемый обобщенным критерием

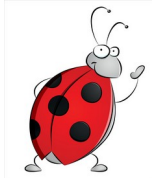
Сэвиджа, предназначенный для анализа цензурированных выборок и представленный в главе «Анализ выживаемости».

См. монографию Скрипника с соавт.

#### 4.2.3.1.8. Критерий Ансари–Бредли

Критерий Ансари–Бредли (Фройнда и Ансари, Freund–Ansari test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности.





Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$W = \sum_{i=1}^{n_1} R_i,$$

где  $R_i, i = 1, 2, \dots, n_1$  – ранги выборки с наибольшей численностью,

$n_1$  – численность одной выборки,

$n_2$  – численность другой выборки,

$N = n_1 + n_2$  – численность объединенной выборки.

Для построения критерия ранжирование производится особым образом. Если  $N$  четно, ранги присваиваются по схеме  $1, 2, 3, \dots, N/2, N/2, \dots, 3, 2, 1$ . Если  $N$  нечетно, ранги присваиваются по схеме  $1, 2, 3, \dots, (N-1)/2, (N+1)/2, \dots, 3, 2, 1$ . При наличии одинаковых наблюдений используются связанные (средние) ранги.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{W - EW}{\sqrt{DW}}$$

распределена по стандартному нормальному закону.

Здесь обозначено:

$EW = n_1(N+2)/4$  – математическое ожидание для четного  $N$ ,

$EW = n_1(N+1)^2/4N$  – математическое ожидание для нечетного  $N$ ,

$$DW = \frac{n_1 n_2 (N+2)(N-2)}{48(N-1)},$$

дисперсия для четного  $N$  без связей

$$DW = \frac{n_1 n_2 [16b - N(N+2)^2]}{48N(N-1)},$$

или, при наличии связей,

$$DW = \frac{n_1 n_2 (N+1)(N^2+3)}{48N^2}$$

дисперсия для нечетного  $N$  без связей

$$DW = \frac{n_1 n_2 [16Nb - (N+1)^4]}{48N^2(N-1)},$$

или, при наличии связей,

$$b = \sum_{j=1}^g t_j r_j^2$$

– поправка на объединение рангов,

где  $t_j, j = 1, 2, \dots, g$  – численность связки,

$r_j, j = 1, 2, \dots, g$  – средний ранг в связке,

$g$  – число связок, причем, при наличии хотя бы одной связки, учитываются также и все группы с численностью 1.

См. Шескин (Sheskin), Джонсон с соавт., Петрович с соавт.





#### 4.2.3.1.9. Критерий Клотца

Критерий Клотца (Klotz test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$K = \sum_{i=1}^{n_1} \left[ \Psi \left( \frac{R_i}{N+1} \right) \right]^2,$$

где  $n_1$  – численность одной выборки,

$n_2$  – численность другой выборки,

$R_i, i = 1, 2, \dots, n_1$  – ранговые метки одной из выборок,

$\Psi(\cdot)$  – функция, обратная функции стандартного нормального распределения,

$N = n_1 + n_2$  – численность объединенной выборки.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{K - EK}{\sqrt{DK}},$$

где  $EK = \frac{n_1}{N} \sum_{i=1}^N \left[ \Psi \left( \frac{i}{N+1} \right) \right]^2$  – математическое ожидание,

$$DK = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^N \left[ \Psi \left( \frac{i}{N+1} \right) \right]^4 - \frac{n_2}{n_1(N-1)} (EK)^2$$

– дисперсия,

распределена по стандартному нормальному закону.

См. Гаек (Најек) с совт., Айвазян с соавт. (1983).

#### 4.2.3.1.10. Критерий Зигеля–Тьюки

Критерий Зигеля–Тьюки (Сиджела–Тьюки, Сайджела–Тьюки, Siegel–Tukey test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале.

Статистика критерия вычисляется по формуле

$$T = \sum_{i=1}^{n_1} R_i,$$

где  $R_i, i = 1, 2, \dots, n_1$  – ранги выборки с наибольшей численностью,

$n_1$  – численность одной выборки,

$n_2$  – численность другой выборки.

Для построения критерия ранжирование производится особым образом. Ранги

присваиваются по схеме  $1, \underline{4}, \underline{5}, \underline{8}, \underline{9}, \dots, \underline{7}, \underline{6}, \underline{3}, \underline{2}$  до исчерпания вариантов объединенной выборки.

При наличии одинаковых наблюдений используются связанные (средние) ранги.





В названии рассмотренного критерия на самом деле объединены два теста – критерий Зигеля и критерий Тьюки. Эти тесты различаются только направлением ранжирования вариантов. Присвоение рангов вариантам в схеме Тьюки начинается не слева направо, как в схеме Зигеля, а справа налево. Построенный таким способом критерий обозначается как  $T'$ . Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{T - ET}{\sqrt{DT}},$$

где  $ET = n_1(N + 1) / 2$  – математическое ожидание,

$DT = n_1 n_2 (N + 1) / 12$  – дисперсия,

$N = n_1 + n_2$  – численность объединенной выборки,

распределена по стандартному нормальному закону.

Представленный критерий эквивалентен критерию Ансари–Бредли.

См. статью Клотца (Klotz), монографии Благовещенского с соавт., Когана с соавт., Шескин (Sheskin).

#### 4.2.3.1.11. Критерий Коновера

Критерий Коновера (Conover's two-sample squared ranks test for equality of variance) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Перед расчетом статистики критерия исходные выборки подвергаются преобразованиям по формулам

$$U_i = |x_i - m_x|, i = 1, 2, \dots, n_1,$$

$$V_i = |y_i - m_y|, i = 1, 2, \dots, n_2,$$

где  $x_i, i = 1, 2, \dots, n_1$  – одна из выборок,

$y_i, i = 1, 2, \dots, n_2$  – другая из выборок,

$n_1$  – численность одной выборки,

$n_2$  – численность другой выборки,

$$m_x = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$$

– среднее значение одной выборки,

$$m_y = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$$

– среднее значение другой выборки.

Статистика критерия вычисляется по формуле

$$K = \sum_{i=1}^{n_1} [R(U_i)]^2,$$

$R_i, i = 1, 2, \dots, n_1$  – ранговые метки одной из выборок.





Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{K - EK}{\sqrt{DK}},$$

где  $EK = n_1 \bar{R}^2$  – математическое ожидание,

$$DK = \frac{n_1 n_2}{N(N-1)} \sum_{i=1}^N [R(U_i)]^4 - \frac{n_1 n_2}{N-1} (\bar{R}^2)^2 \quad \text{– дисперсия,}$$

$$\bar{R}^2 = \frac{1}{N} \sum_{i=1}^N [R(U_i)]^2 \quad \text{– среднее значение суммы квадратов рангов,}$$

$N = n_1 + n_2$  – численность объединенной выборки,  
распределена по стандартному нормальному закону.

Описание метода приводится в монографии Коновера, книге Спрента (Sprent) с соавт., статьях Коновера с соавт., работах Вилкокса (Wilcox), диссертации Бучана (Buchan).

#### 4.2.3.1.12. Критерий Муда–Брауна

Медианный критерий Муда–Брауна (критерий Муда) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Статистика критерия вычисляется по формуле

$$V_+ = \sum_{i=1}^{n_1} \text{sign} \left[ R_i - \frac{N+1}{2} \right],$$

где  $R_i, i = 1, 2, \dots, n_1$  – ранги выборки с наименьшей численностью,

$n_1$  – численность одной выборки,

$n_2$  – численность другой выборки,

$N = n_1 + n_2$  – численность объединенной выборки.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{V_+ - EV_+}{\sqrt{DV_+}},$$

где  $EV_+ = \frac{n_1}{2}$  – математическое ожидание,

$$DV_+ = \frac{n_1 n_2}{4(N-1)} \quad \text{– дисперсия.}$$

распределена по стандартному нормальному закону.



## 4.2.3.2. Критерии на основе сравнения функций распределения

Идея сравнения функций распределения (А.Н. Колмогоров, 1933 г.) оказалась наиболее плодотворной при конструировании критериев согласия. Более подробная информация дана в главе «Проверка нормальности распределения».

Идея стала полезной и при сравнении эмпирических функций распределения эмпирических выборок. Из критериев данного класса нами представлены:

- критерий Смирнова,
- критерий Лемана–Розенблатта,
- критерий Койпера.

Существует группа критериев на основе распределения  $\chi^2$ , предназначенная для анализа таблиц сопряженности, являющихся продуктом сопоставления эмпирических выборок.

Из критериев данного класса нами представлены:

- критерий Мак–Немара (для сопряженных бинарных выборок) в его асимптотическом варианте,
- критерий хи–квадрат (для независимых бинарных выборок),
- критерий медианы (для порядковых или количественных выборок).

Для применения критерия Мак–Немара и критерия хи–квадрат (в представленной форме) анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. (согласно принятому здесь соглашению) состоять только из нулей и единиц, причем ноль означает отсутствие признака, а единица означает наличие признака.

### 4.2.3.2.1. Критерий Смирнова

Критерий Смирнова (критерий Колмогорова–Смирнова, Kolmogorov–Smirnov test, Kolmogorov–Smirnov test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о том, являются ли одинаковыми непрерывные функции распределения генеральных совокупностей, из которых взяты выборки. Иначе, проверяется принадлежность двух выборок одной и той же генеральной совокупности при условии непрерывности ее функции распределения.

Статистика критерия имеет вид

$$D_{m,n} = \sup_{-\infty < x < \infty} |F_n(x') - G_m(x)|,$$

где  $D_{m,n}$  – максимальная разность между частотами рядов  $x'$  и  $x$ ,  
 $m$  и  $n$  – численности вариационных рядов, построенных по эмпирическим выборкам,  
 $G_m(\cdot)$  и  $F_n(\cdot)$  – соответствующие эмпирические функции распределения.

Практически вычисления производятся по формулам:



$$D = \max(D_{m,n}^+, D_{m,n}^-)$$

$$D_{m,n}^+ = \max_{1 \leq r \leq m} \left( \frac{r}{m} - F_n(x_r') \right) = \max_{1 \leq s \leq n} \left( G_m(x_s) - \frac{s-1}{n} \right),$$

$$D_{m,n}^- = \max_{1 \leq r \leq m} \left( F_n(x_r') - \frac{r-1}{m} \right) = \max_{1 \leq s \leq n} \left( \frac{s}{n} - G_m(x_s) \right).$$

Функция распределения модифицированной статистики критерия  $D\sqrt{N}$  (имеются и иные формулы) при  $N = mn / (m + n) \rightarrow \infty$  сходится к функции распределения Колмогорова. Критерий рекомендуется для выборок средней и большой численности (численность каждой выборки от 40 до 100 и выше). При большей численности выборок становится больше теоретических оснований для применения параметрических тестов.

См. учебник Айвазяна с соавт. (критерий однородности Смирнова), статью Лемешко с соавт. Статистика рассматриваемого теста может быть записана как максимум линейных ранговых статистик – модифицированных статистик Муда. Поэтому некоторые авторы рассматривают метод в курсе ранговых критериев. Гудман (Goodman) предложил аппроксимировать статистику критерия распределением  $\chi^2$  (статистика хи-квадрат Гудмана, Goodman approximation of Kolmogorov–Smirnov test).

#### 4.2.3.2. Критерий Лемана–Розенблатта

Критерий Лемана–Розенблатта (Lehmann–Rosenblatt test, Lehmann's two-sample test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о том, являются ли одинаковыми непрерывные функции распределения генеральных совокупностей, из которых взяты выборки. Иначе, проверяется принадлежность двух выборок одной и той же генеральной совокупности при условии непрерывности ее функции распределения.

Статистика критерия вычисляется по формуле

$$T = \frac{1}{m+n} \left[ \frac{1}{m} \sum_{i=1}^n (R_i - i)^2 + \frac{1}{n} \sum_{j=1}^m (S_j - j)^2 - \frac{4mn-1}{6} \right],$$

где  $R_i, i = 1, 2, \dots, n$  – ранги одной выборки,

$S_j, j = 1, 2, \dots, m$  – ранги другой выборки.

$n$  и  $m$  – численности выборок.

Функция распределения статистики критерия при  $m, n \rightarrow \infty$  совпадает с функцией распределения  $a_1$  критериев типа омега-квадрат.

См. таблицы Большева с соавт., книгу Мартынова, статьи Лемана (Lehmann), Розенблатта (Rosenblatt), Сандрама (Sundrum), Вегнера (Wegner), Лемешко (Lemeshko) с соавт., Лемешко с соавт., Фиша (Fisz).



### 4.2.3.2.3. Критерий Койпера

Критерий Койпера (Kuiper test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о том, являются ли одинаковыми непрерывные функции распределения генеральных совокупностей, из которых взяты выборки. Иначе, проверяется принадлежность двух выборок одной и той же генеральной совокупности при условии непрерывности ее функции распределения.

Статистика критерия имеет вид

$$V = D_{m,n}^+ + D_{m,n}^-$$

$$D_{m,n}^+ = \sup_{-\infty < x' < \infty} |F_n(x') - G_m(x)|$$

$$D_{m,n}^- = \sup_{-\infty < x < \infty} |G_m(x) - F_n(x')|$$

где  $D_{m,n}^+$  – максимальная разность  $F_n(x')$  «выше»  $G_m(x)$ ,

$D_{m,n}^-$  – максимальная разность  $F_n(x')$  «ниже»  $G_m(x)$ ,

$F_n(x')$  и  $G_m(x)$  – эмпирические функции распределения вариационных рядов  $x'$  и  $x$ , построенных по эмпирическим выборкам,

$n$  и  $m$  – численности вариационных рядов  $x'$  и  $x$ .

Функция распределения модифицированной статистики критерия  $V\sqrt{N}$  (имеются и иные формулы) при  $N = mn / (m + n) \rightarrow \infty$  сходится к функции распределения Койпера.

Критерий рекомендуется для выборок средней и большой численности (численность каждой выборки от 40 до 100 и выше). При большей численности выборок становится больше теоретических оснований для применения параметрических тестов.

Ряд авторов полагает критерий Койпера предпочтительным относительно критерия Смирнова. Имеются литературные данные о попытках применения критерия Койпера, подобно критериям типа Колмогорова, для проверки согласия распределений (подробнее о проверке нормальности см. в главе «Проверка нормальности распределения»).

### 4.2.3.2.4. Критерий Мак–Немара

Критерий Мак–Немара (McNemar's chi-square test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые попарно сопряженные бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырёхклеточных) таблиц, иначе таблиц типа 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем ноль означает отсутствие признака, а единица означает наличие признака. Перед



применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.

Вычисление статистики критерия производится по формуле:

$$X^2 = \frac{(|b - c| - Y)^2}{b + c},$$

где  $b$  – число пар наблюдений с эффектом  $A$  в первой выборке, но без эффекта  $B$  во второй выборке,

$c$  – число наблюдений без эффекта  $A$  в первой выборке, но с эффектом  $B$  во второй выборке,

$Y = 0$  – поправка на непрерывность (поправка Йейтса), в случае ее неучета,

$Y = 1$  – в случае учета поправки (режим по умолчанию).

Считается, что при величине  $b + c \geq 10$  статистика критерия (двусторонняя гипотеза) удовлетворительно аппроксимируется распределением  $\chi^2$  с числом степеней свободы, равным 1. При  $b + c < 10$  можно использовать точные методы, представленные в главе «Точные критерии», в котором для полноты изложения представлен также критерий Мак–Немара, дополненный его точным вариантом.

О вычислении критерия и точном распределении его статистики см. заметки Беннетта (Bennett) с соавт., материалы компании Cytel. Существует вариант рассмотренного критерия (критерий Стюарта–Максвелла, Stuart–Maxwell test), предназначенный для анализа таблиц типа  $k \times k$ , получающихся из номинальных выборок с числом градаций признаков, равным  $k$ . Аналогичное назначение имеют критерий симметрии Баукера (Bowker's test of symmetry) и критерий Бхапкара (Bhappkar's test). Данные методы представлены в главе «Кросстабуляция».

#### 4.2.3.2.5. Критерий хи–квадрат

Критерий хи–квадрат применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые независимые бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.

Вычисление статистики критерия для данного случая производится по формуле

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(|f_{ij} - \hat{f}_{ij}| - Y)^2}{\hat{f}_{ij}},$$

где  $f_{ij}$ ,  $i, j = 1, 2$  – вычисленные частоты – значения в клетках  $a$ ,  $b$ ,  $c$ ,  $d$  в дальнейшем для наглядности обозначим их этими же литерами,

$\hat{f}_{ij}$ ,  $i, j = 1, 2$ , – соответствующие ожидаемые частоты, вычисляемые по формулам:





$$\hat{f}_{11} = \frac{(a+b)(a+c)}{n},$$

$$\hat{f}_{12} = \frac{(a+b)(b+d)}{n},$$

$$\hat{f}_{21} = \frac{(c+d)(a+c)}{n},$$

$$\hat{f}_{22} = \frac{(c+d)(b+d)}{n},$$

где  $a$  – число наблюдений с эффектом  $A$  в первой выборке,  
 $b$  – число наблюдений без эффекта  $A$  в первой выборке,  
 $c$  – число наблюдений с эффектом  $A$  во второй выборке,  
 $d$  – число наблюдений без эффекта  $A$  во второй выборке,  
 $n = a + b + c + d$  – общая численность всех наблюдений,  
 $Y = 0$  – поправка на непрерывность (поправка Йейтса), в случае ее неучета,  
 $Y = 0,5$  – в случае учета поправки (режим по умолчанию).  
Статистика критерия удовлетворительно аппроксимируется распределением  $\chi^2$  с числом степеней свободы, равным 1.

См. Лванга (Lwanga). Поправки обсуждаются в статье Лузен (Loosen). Существует вариант критерия для анализа таблиц типа  $k \times k$ , получающихся из выборок с числом градаций признаков более 2, представленный в главе «Кросстабуляция».

#### 4.2.3.2.6. Критерий медианы

Критерий медианы (медианный критерий) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Выборки могут принадлежать порядковой или количественной шкале. Этапы вычисления критерия для двух выборок численностями  $n_1$  и  $n_2$  включают:

- Объединение исходных выборок, вычисление медианы объединенной выборки.
- Формирование таблицы типа  $2 \times 2$  по следующему правилу: в ячейку  $A$  заносится число отметок первой выборки, превышающих медиану; в ячейку  $B$  заносится число отметок второй выборки, превышающих медиану; в ячейки  $C$  и  $D$  заносится число отметок, соответственно, первой и второй выборок, не превышающих медиану.

В случае  $n_1 > 15$  и/или  $n_2 > 15$  к полученной таблице применяется критерий хи-квадрат с числом степеней свободы, равным 1.

Существует вариант критерия для анализа таблиц типа  $2 \times k$ , получающихся из  $k$  порядковых выборок с числом вариаций признаков, равным 2.





### 4.2.3.3. Критерий серий Вальда–Вольфовица

Критерий серий Вальда–Вольфовица (Wald–Wolfowitz runs test) применяется для проверки однородности двух независимых совокупностей одинаковой или разной численности. Проверяется нулевая гипотеза о равенстве целого ряда параметров двух сравниваемых выборок, включая медианы и коэффициенты асимметрии. Критерий применяется в случае, если исследователя интересует, имеют ли место любые различия между совокупностями. Выборки могут принадлежать порядковой или количественной шкале. Расчет заключается в объединении выборок с численностями  $n_1$  и  $n_2$  в одну выборку общей численностью  $N = n_1 + n_2$ , ее сортировке по возрастанию или убыванию и подсчете числа серий элементов  $R$ , относящихся к первой и второй выборкам.

Значимость при численности выборок  $n_1 > 20$  и  $n_2 > 20$  может вычисляться посредством нормальной аппроксимации. При этом модифицированная статистика

$$\frac{|R - ER| - 0,5}{\sqrt{DR}},$$

где  $ER = \frac{2n_1n_2}{N} + 1$  – математическое ожидание,

$$DR = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N - 1)} - \text{дисперсия,}$$

0,5 – поправка на непрерывность,  
распределена по стандартному нормальному закону.

Варианты критерия серий и аппроксимации представлены в монографии Браунли. Метод описан в справочнике Руниона. Замечания о применении см. в книге Гаека с соавт., статья Камень с соавт.

### 4.2.4. Таблицы 2 x 2

Рассчитываются следующие продукты анализа таблиц типа 2 x 2

- относительный риск,
- отношение шансов,
- разность долей,
- прогностичность.

Таблицы 2 x 2 возникают в результате сопоставления двух бинарных (дихотомических) выборок, т. е. выборок, состоящих из значений 1 и 0, причем под значением 1 понимают наличие признака, под значением 0 понимают отсутствие признака.

Для расчета пользователь может указать одну из опций расчета таблицы:

- Для независимых выборок.
- Для связанных (парных) выборок.
- Расчет по готовой таблице для независимых выборок.



- Расчет по готовой таблице для связанных выборок.

Важно знать, что таблицы типа 2 x 2 могут быть получены из исходных выборок различными способами, в зависимости от того, являются ли выборки независимыми или связанными. См. монографию Ньюмен (Newman).

#### 4.2.4.1. Относительный риск

Относительный риск (relative risk, RR), или отношение рисков – отношение заболеваемости среди лиц, подвергавшихся и не подвергавшихся воздействию факторов риска.

Относительный риск не несет информации о величине абсолютного риска (заболеваемости). Даже при высоких значениях относительного риска абсолютный риск может быть совсем небольшим, если заболевание редкое. Относительный риск показывает силу связи между воздействием и заболеванием.

Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем ноль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.

Вычисление отношения рисков производится по формуле

$$RR = \frac{n_{11}(n_{21} + n_{22})}{n_{21}(n_{11} + n_{12})},$$

где  $n_{11}$ ,  $n_{12}$ ,  $n_{21}$ ,  $n_{22}$  – ячейки таблицы.

Двусторонний доверительный интервал вычисляется по формуле

$$I_{RR} = (RR - \Psi((1 + \beta) / 2)S_{RR}; RR + \Psi((1 + \beta) / 2)S_{RR}),$$

где  $\Psi(.)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях,

$S_{RR}$  – стандартная ошибка отношения рисков.

Стандартная ошибка логарифма отношения рисков вычисляется по формуле

$$S_{\ln(RR)} = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21}} - \frac{1}{n_{21} + n_{22}}}.$$

Окончательная формула двустороннего доверительного интервала оцениваемого отношения рисков будет:

$$I_{RR} = (\exp(\ln(RR) - \Psi((1 + \beta) / 2)S_{\ln(RR)}); \exp(\ln(RR) + \Psi((1 + \beta) / 2)S_{\ln(RR)})).$$

См. Агрести (Agresti), Хайнес (Haynes) с соавт., статьи Бертелла (Bertell), Гарта (Gart), Барратт (Barratt) с соавт., Подольной с соавт.



## 4.2.4.2. Отношение шансов

Отношение шансов (odds ratio, OR) определяется как отношение шансов события в одной группе к шансам события в другой группе, или как отношение шансов того, что событие произойдет, к шансам того, что событие не произойдет. В исследованиях случай–контроль отношение шансов используется для оценки относительного результата.

Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц 2 x 2. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем ноль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2.

Вычисление отношения шансов производится по формуле

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

где  $n_{11}$ ,  $n_{12}$ ,  $n_{21}$ ,  $n_{22}$  – ячейки таблицы.

Двусторонний доверительный интервал вычисляется по формуле

$$I_{OR} = (OR - \Psi((1 + \beta) / 2)S_{OR}; OR + \Psi((1 + \beta) / 2)S_{OR}),$$

где  $\Psi(.)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях,

$S_{OR}$  – стандартная ошибка отношения шансов.

Стандартная ошибка логарифма отношения шансов вычисляется по формуле

$$S_{\ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Окончательная формула двустороннего доверительного интервала оцениваемого отношения шансов будет

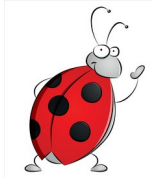
$$I_{OR} = (\exp(\ln(OR) - \Psi((1 + \beta) / 2)S_{\ln(OR)}); \exp(\ln(OR) + \Psi((1 + \beta) / 2)S_{\ln(OR)})).$$

См. Агрести (Agresti), Хайнес (Haynes) с соавт., статьи Бабиш с соавт., Бленда (Bland) с соавт.

## 4.2.4.3. Разность долей

Рассматриваемый метод вычисления разности долей (difference of proportions) предназначен для анализа так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц 2 x 2, возникающих при обработке независимых либо связанных признаков. Анализируемые выборки должны принадлежать дихотомической шкале измерения, т. е. состоять только из нулей и единиц, причем ноль означает отсутствие признака, а единица означает наличие признака. Перед применением метода необходимо ознакомиться с разделом, посвященным описанию таблиц 2 x 2. Предоставляется возможность как ввода исходных массивов, так и готовых таблиц. В последнем случае обязательно необходимо указать, продуктом каких признаков является таблица, ибо формулы их обработки существенно различаются.





#### 4.2.4.3.1. Разность долей в таблице независимых признаков

Вычисление разности долей производится по формуле

$$d = |p_2 - p_1|,$$

где  $p_1 = n_{11} / (n_{11} + n_{12})$  – частота эффекта в первой выборке,

$p_2 = n_{21} / (n_{21} + n_{22})$  – частота эффекта во второй выборке,

$n_{11}, n_{12}, n_{21}, n_{22}$  – ячейки таблицы.

Значимость разности долей тестируется с помощью z-критерия, вычисление статистики которого в данном случае производится по формуле

$$z = \frac{|p_2 - p_1| - Y}{\sqrt{\bar{p}(1 - \bar{p}) \left( \frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21} + n_{22}} \right)}},$$

где  $\bar{p} = (n_{11} + n_{21}) / (n_{11} + n_{12} + n_{21} + n_{22})$ ,

$Y = 0$  – поправка на непрерывность (поправка Йейтса), в случае ее неучета,

$Y = 0,5 \cdot \left( \frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21} + n_{22}} \right)$  – в случае учета поправки (режим по умолчанию).

Квадрат статистики критерия удовлетворительно аппроксимируется распределением  $\chi^2$  с числом степеней свободы, равным 1.

Вычисляется двусторонний доверительный интервал оцениваемой разности долей по формуле Вальда:

$$I_{p_2 - p_1} = (d - \Psi((1 + \beta)/2) S_{p_2 - p_1} - Y; d + \Psi((1 + \beta)/2) S_{p_2 - p_1} + Y),$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

$S_{p_2 - p_1}$  – стандартная ошибка разности долей, вычисляемая по формуле

$$S_{p_2 - p_1} = \sqrt{\frac{p_1(1 - p_1)}{n_{11} + n_{12}} + \frac{p_2(1 - p_2)}{n_{21} + n_{22}}}.$$

#### 4.2.4.3.2. Разность долей в таблице связанных признаков

Вычисление разности долей производится по формуле

$$d = |p_2 - p_1|,$$

где  $p_1 = (n_{11} + n_{12}) / n$  – частота эффекта в первой выборке,

$p_2 = (n_{11} + n_{21}) / n$  – частота эффекта во второй выборке,

$n$  – сумма таблицы, вычисляемая по формуле

$$n = n_{11} + n_{12} + n_{21} + n_{22},$$

$n_{11}, n_{12}, n_{21}, n_{22}$  – ячейки таблицы.

Значимость разности долей тестируется с помощью критерия хи-квадрат, вычисление статистики которого в данном случае производится по формуле





$$\chi^2 = \frac{(|n_{21} - n_{12}| - Y)^2}{n},$$

где  $Y = 0$  – поправка на непрерывность (поправка Йейтса), в случае ее неучета,  
 $Y = 1 / n$  – в случае учета поправки (режим по умолчанию).

Квадрат статистики критерия удовлетворительно аппроксимируется распределением  $\chi^2$  с числом степеней свободы, равным 1.

Вычисляется двусторонний доверительный интервал оцениваемой разности долей по формуле Вальда (Wald interval for difference of proportions):

$$I_{p_2 - p_1} = (d - \Psi((1 + \beta)/2)S_{p_2 - p_1} - Y; d + \Psi((1 + \beta)/2)S_{p_2 - p_1} + Y),$$

где стандартная ошибка разности долей вычисляется по формуле

$$S_{p_2 - p_1} = \frac{1}{n} \sqrt{b + c - \frac{(b - c)^2}{n}}.$$

Вычисляется двусторонний доверительный интервал оцениваемой разности долей по уточненной формуле Вальда (adjusted Wald interval for difference of proportions):

$$I_{p_2 - p_1} = (\hat{p}_2 - \hat{p}_1 - \Psi((1 + \beta)/2)\hat{S}_{p_2 - p_1} - Y; \hat{p}_2 - \hat{p}_1 + \Psi((1 + \beta)/2)\hat{S}_{p_2 - p_1} + Y),$$

где  $|\hat{p}_2 - \hat{p}_1| = |n_{21} - n_{12}| / (n + 2)$ ,

$$\hat{S}_{p_2 - p_1} = \frac{1}{n + 2} \sqrt{b + c + 1 - \frac{(b - c)^2}{n + 2}}.$$

См. Флейс (Fleiss) с соавт., статьи Бурмана (Buhrman), Брауна (Brown) с соавт., Хаука (Hauck) с соавт., Биггерстаффа (Biggerstaff), Чубенко с соавт. Обзор методов вычисления доверительных интервалов оцениваемой разности долей в таблице независимых признаков см. в статье Сантнера (Santner) с соавт. Методы вычисления доверительных интервалов см. в монографиях Агрести (Agresti), Флейс с соавт., статьях Агрести с соавт., Бергер (Berger) с соавт., Хсие (Hsieh), Сюисса (Suissa) с соавт., Ньюскомб (Newcombe), Гарднер (Gardner) с соавт., Танг (Tang) с соавт.

#### 4.2.4.4. Прогностичность

Рассматриваемая опция дает возможность вычислить общепринятые стандартные показатели прогностичности (прогностической ценности) диагностического теста (predictive values). Это следующие показатели:

- чувствительность (Se, sensitivity),
- специфичность (Sp, specificity),
- распространенность (p, преваленс, доля, prevalence),
- прогностичность положительного результата (PPV, positive predictive value),
- прогностичность отрицательного результата (NPV, negative predictive value).



Распространенность – это априорная (претестовая) вероятность наличия болезни до того, как стали известны результаты диагностического теста.

Прогностичность (собственно прогностическая ценность) – это апостериорная (посттестовая) вероятность наличия болезни при известном результате исследования.

Различают прогностичность положительного результата и прогностичность отрицательного результата. Ниже представлены подробные описания данных показателей, включая формулы вычисления их точечных и интервальных оценок.

Рассматриваемые методы предназначены для обработки так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц 2 x 2:

Результат диагностического теста	Положительный Отрицательный	Наличие заболевания	
		Присутствует	Отсутствует
		$n_{11}$	$n_{12}$
		$n_{21}$	$n_{22}$
		$n_1$	$n_0$

Анализируемые выборки должны принадлежать дихотомической шкале измерения и состоять только из нулей и единиц, причем нуль означает отсутствие признака, а единица означает наличие признака. Положительным результатом теста считается такой результат, который показывает наличие заболевания. Отрицательным результатом теста считается такой результат, который показывает отсутствие заболевания. Обозначено:

$n_{11}$  – численность индивидуумов с наличием заболевания, диагностированных тестом как больные,

$n_{21}$  – численность индивидуумов с наличием заболевания, диагностированных тестом как здоровые,

$n_{12}$  – численность индивидуумов без наличия заболевания, диагностированных тестом как больные,

$n_{22}$  – численность индивидуумов без наличием заболевания, диагностированных тестом как здоровые,

$n_1 = n_{11} + n_{21}$  – численность больных,

$n_0 = n_{12} + n_{22}$  – численность здоровых.

Дополнительные пояснения см. в разделе, посвященном ROC-анализу.

#### 4.2.4.4.1. Чувствительность

Чувствительностью называют долю положительных результатов диагностического теста в популяции. Чем чувствительнее тест, тем выше прогностическая ценность его отрицательного результата.

Вычисление оценки чувствительности производится по формуле

$$Se = \frac{n_{11}}{n_1}.$$



Двусторонний доверительный интервал оцениваемой чувствительности вычисляется по формуле

$$I_{Se} = (Se - \Psi((1 + \beta)/2)S_{Se}; Se + \Psi((1 + \beta)/2)S_{Se}),$$

где  $\Psi(.)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях,

$S_{Se}$  – стандартная ошибка чувствительности.

Стандартная ошибка чувствительности вычисляется по формуле

$$S_{Se} = \sqrt{\frac{Se \cdot (1 - Se)}{n_1}}.$$

#### 4.2.4.4.2. Специфичность

Специфичностью называют долю отрицательных результатов диагностического теста в популяции. Чем специфичнее тест, тем выше прогностическая ценность его положительного результата.

Вычисление оценки специфичности производится по формуле

$$Sp = \frac{n_{22}}{n_0}.$$

Двусторонний доверительный интервал оцениваемой специфичности вычисляется по формуле

$$I_{Sp} = (Sp - \Psi((1 + \beta)/2)S_{Sp}; Sp + \Psi((1 + \beta)/2)S_{Sp}),$$

где  $S_{Sp}$  – стандартная ошибка специфичности.

Стандартная ошибка специфичности вычисляется по формуле

$$S_{Sp} = \sqrt{\frac{Sp \cdot (1 - Sp)}{n_0}}.$$

#### 4.2.4.4.3. Распространенность

В литературе встречаются два различных мнения по поводу вычисления распространенности. Распространенность может быть:

- отношением числа выявленных случаев [заболеваний] ко всем обследованным за определенный промежуток времени (например, за год),
- отношением числа выявленных случаев к численности популяции.

Когда распространенность стремится к нулю, прогностическая ценность положительного результата теста стремится к нулю. Когда распространенность стремится к 1, прогностическая ценность отрицательного результата теста стремится к нулю.

Ввод известной из предварительных исследований распространенности относится к Байесовской идеологии, когда те или иные выводы по результатам анализа представленных данных делаются с учетом некоторой априорной (известной до опыта) информации.



Обратите внимание, что целая часть уже показана на форме. Например, для ввода значения распространенности 0,124 следует ввести число 124. Другой пример. Пусть требуется ввести распространенность 23 случая на 1000 обследованных пациентов. В поле вводится значение 023.

В следующем способе (см. Флетчер с соавт.) вычисление точечной оценки распространенности на основе тех же самых представленных для анализа выборочных данных производится по формуле

$$p = n_1 / n,$$

где  $n = n_{11} + n_{12} + n_{21} + n_{22}$  – общая численность.

Доверительный интервал оцениваемой распространенности рассчитываются стандартно по формуле Вальда

$$I_p = (p - \Psi((1 + \beta) / 2) S_p; p + \Psi((1 + \beta) / 2) S_p),$$

где  $S_p$  – стандартная ошибка распространенности.

Стандартная ошибка распространенности может быть вычислена по формуле

$$S_p = \sqrt{\frac{p \cdot (1 - p)}{n}}.$$

#### 4.2.4.4. Прогностичность положительного результата

Вычисление прогностичности положительного результата производится по формуле

$$PPV = \frac{Se \cdot p}{Se \cdot p + (1 - Sp) \cdot (1 - p)}.$$

Двусторонний доверительный интервал вычисляется по формуле

$$I_{PPV} = (PPV - \Psi((1 + \beta) / 2) S_{PPV}; PPV + \Psi((1 + \beta) / 2) S_{PPV}),$$

где  $S_{PPV}$  – стандартная ошибка прогностичности положительного результата.

Стандартная ошибка прогностичности положительного результата вычисляется по формуле

$$S_{PPV} = \sqrt{\frac{\left[ p \cdot (1 - Sp) \cdot (1 - p) \right]^2 \frac{Se \cdot (1 - Se)}{n_1} + \left[ p \cdot Se \cdot (1 - p) \right]^2 \frac{Sp \cdot (1 - Sp)}{n_0}}{\left[ Se \cdot p + (1 - Sp) \cdot (1 - p) \right]^4}}.$$

#### 4.2.4.4.5. Прогностичность отрицательного результата

Вычисление прогностичности отрицательного результата производится по формуле

$$NPV = \frac{Sp \cdot (1 - p)}{(1 - Se) \cdot p + Sp \cdot (1 - p)}.$$

Двусторонний доверительный интервал вычисляется по формуле

$$I_{NPV} = (NPV - \Psi((1 + \beta) / 2) S_{NPV}; NPV + \Psi((1 + \beta) / 2) S_{NPV}),$$

где  $S_{NPV}$  – стандартная ошибка прогностичности отрицательного результата.

Стандартная ошибка прогностичности отрицательного результата вычисляется по формуле





$$S_{NPV} = \sqrt{\frac{\left[ p \cdot Sp \cdot (1 - p) \right]^2 \frac{Se \cdot (1 - Se)}{n_1} + \left[ p \cdot (1 - Se) \cdot (1 - p) \right]^2 \frac{Sp \cdot (1 - Sp)}{n_0}}{\left[ (1 - Se) \cdot p + Sp \cdot (1 - p) \right]^4}}.$$

См. Власова, Флетчер с соавт., Флейс, Флейс (Fleiss), Флейс с соавт., Хайнес (Haynes) с соавт., Халли (Hulley) с соавт., статью Меркалдо (Mercaldo) с соавт., статьи Воробьева, Моссман (Mossman) с соавт., Линн (Linn), Альтман (Altman) с соавт., Сауро (Sauro) с соавт., Агрести (Agresti) с соавт.

## 4.2.5. График медиан с доверительными интервалами

Доверительный интервал оцениваемой медианы задается формулой

$$I_m = (y_c; y_{n+1-c}),$$

где  $c$  – параметр, вычисляемый по формуле

$$c = \left[ n / 2 - \Psi((1 + \beta) / 2) n^{1/2} / 2 \right],$$

где  $[\cdot]$  – целая часть числа,

$\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

Дополнительно в таблице выводится разность медиан анализируемых выборок. Пусть вычислено  $m = n_1 n_2$  разностей значений  $w_1 \leq w_2 \leq \dots \leq w_m$  всех величин  $(x_i - y_j)$ ,  $i = 1, 2, \dots, n_1$ ;  $j = 1, 2, \dots, n_2$ , где  $x_i$ ,  $i = 1, 2, \dots, n_1$  и  $y_j$ ,  $j = 1, 2, \dots, n_2$  – значения вариант исходных количественных выборок. Тогда медиана  $\mu$  полученной выборки  $w_i$ ,  $i = 1, 2, \dots, m$ , будет разностью медиан. Для нечетного  $m$  медианой является варианта полученного интервального вариационного ряда, имеющая порядковый номер  $(m + 1) / 2$ . Для четного  $m$  медиана равна среднему значению двух средних вариантов.

Доверительный интервал оцениваемой разности медиан (интервал Мозеса) задается формулой

$$I_\mu = (z_c; z_{m+1-c}),$$

где  $z_i$ ,  $i = 1, 2, \dots, m$  – интервальный вариационный ряд, представляющий собой упорядоченный по возрастанию ряд разностей  $w_i$ ,  $i = 1, 2, \dots, m$ ,

$c$  – параметр, вычисляемый по формуле

$$c = \left[ \frac{m}{2} - \Psi((1 + \beta) / 2) \left( \frac{n_1 n_2 (m + 1)}{12} \right)^{1/2} \right],$$

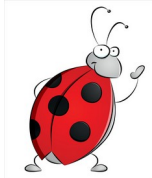
где  $[\cdot]$  – целая часть числа,

$\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

Результаты представленного графического анализа интерпретируются следующим образом.

Если  $100\beta\%$  доверительные интервалы оцениваемых медиан сравниваемых выборок пересекаются, конкурирующая гипотеза (медианы не равны) может быть принята на уровне значимости  $p \leq \beta$ . Если  $100\beta\%$  доверительные интервалы оцениваемых средних значений



сравниваемых выборок не пересекаются, нулевая гипотеза (медианы равны) не отвергается на уровне значимости  $p > \beta$ . т. к. доверительные интервалы тем шире, чем больше значение  $\beta$ , выбирая различные стандартные значения  $\beta$ , можно получить значение уровня значимости, более точно соответствующее представленным данным.

О графическом изображении показателей и интерпретации результатов см. работу Голдстейн (Goldstein) с соавт.

#### 4.2.6. График долей с доверительными интервалами

Границы доверительного интервала доли рассчитываются по «точным» формулам Клоппера–Пирсона (Clopper–Pearson interval). При этом нижняя граница доверительного интервала оцениваемой доли считается как

$$L_p = \left[ 1 + \frac{n - m + 1}{m \cdot F_{2m, 2(n-m+1)}^{-1}(1 - (1 - \beta)/2)} \right]^{-1},$$

где  $m$  – число случаев,

$n$  – численность выборки,

$F_{\dots}^{-1}(\cdot)$  – обратная функция  $F$ -распределения.

$\beta$  – доверительный уровень, выраженный в долях.

Верхняя граница доверительного интервала оцениваемой доли считается как

$$H_p = \left[ 1 + \frac{n - m}{(m + 1) \cdot F_{2(m+1), 2(n-m)}^{-1}((1 - \beta)/2)} \right]^{-1}.$$

Результаты представленного графического анализа интерпретируются следующим образом. Если  $100\beta\%$ , доверительные интервалы оцениваемых долей сравниваемых выборок пересекаются, конкурирующая гипотеза (доли не равны) может быть принята на уровне значимости  $p \leq \beta$ . Если  $100\beta\%$  доверительные интервалы оцениваемых средних значений сравниваемых выборок не пересекаются, нулевая гипотеза (доли равны) не отвергается на уровне значимости  $p > \beta$ . т. к. доверительные интервалы тем шире, чем больше значение  $\beta$ , выбирая различные стандартные значения  $\beta$ , можно получить значение уровня значимости, более точно соответствующее представленным данным.

О графическом изображении показателей и интерпретации результатов см. работу Голдстейн (Goldstein) с соавт.

#### 4.2.7. ROC анализ

ROC (Receiver Operating Characteristic) анализ может иметь различные применения для анализа данных. Дальнейшие обозначения проще всего пояснить с помощью таблицы  $2 \times 2$ .

Исследуемый метод

Стандартный метод





	Положительный исход	Отрицательный исход
Положительный исход	$T_P$	$F_P$
Отрицательный исход	$F_N$	$T_N$

Суть обозначений ясна из первых букв английских терминов:

- True – истинно,
- False – ложно,
- Positive – положительный,
- Negative – отрицательный.

Термины «положительный» и «отрицательный» здесь относятся не к объекту исследования, а, скажем, к способности диагностического теста установить диагноз. Так, при исследовании заболевания положительным исходом будет являться наличие заболевания, отрицательным исходом – отсутствие заболевания.

Термин ROC curve (ROC кривая) в адекватном переводе, заимствованном из радиотехники, означает кривую соотношений правильного и ложного обнаружения сигналов. ROC кривая представляет собой график параметрического типа. При этом абсцисса и ордината кривой являются функциями некоторого параметра, произвольно изменяемого или конкретно измеряемого в эксперименте. В исследовательской практике могут иметь место различные сочетания данных функций, что приводит к различным ROC кривым. Наиболее употребительный тип ROC кривой параметрически отображает величину чувствительности  $Se$  и величину неспецифичности  $1 - Sp$ , где  $Sp$  – специфичность. Порог чувствительности на графике не отображается, однако каждому [в данном случае] заданному значению порога соответствует пара «чувствительность–неспецифичность». На графике данные величины принято изображать в процентах. Показатели определяются следующими формулами.

Чувствительность показывает долю истинно положительных случаев, т. е.

$$Se = \frac{T_P}{T_P + F_N}.$$

Специфичность показывает долю истинно отрицательных случаев, т. е.

$$Sp = \frac{T_N}{T_N + F_P}.$$

Некоторые авторы величину  $Sp$  называют частотой истинно отрицательных результатов (true negative rate), а величину  $1 - Sp$  называют ценой метода либо частотой ложно положительных результатов (false positive rate, FPR). По аналогии величину  $Se$  иногда называют частотой истинно положительных результатов (true positive rate, TPR). Некоторые авторы полагают, что в таких терминах ROC кривая более понятна для чтения. Также условились для построения ROC кривой использовать показатели в процентах.

Сочетание значений чувствительности и специфичности в дальнейшем анализе может быть выбрано различным в зависимости от требований исследователя. При этом соответствующее значение диагностического параметра называют порогом отсечения. Используется критерий



Юдена (Йоден, Youden), максимизирующий сумму чувствительности и специфичности. О порогах отсечения дополнительно см. главу «Распознавание образов с обучением». Рассмотрим алгоритм построения ROC кривой. Пусть даны исследуемая выборка численностью  $n$  и стандартная выборка численностью  $m$ .

Алгоритм ROC анализа предлагается сформулировать следующим образом:

1. Задаться интервалом изменения параметра. Удобнее всего данный интервал получить, объединив представленные выборки в массив диагностических параметров численностью  $n + m$ , а затем отсортировав данный массив по убыванию.
2. Используя варианты полученного в предыдущем пункте алгоритма массива диагностических параметров в качестве порогов отсечения, составить на основе исходных выборок для каждой варианты данного массива таблицу  $2 \times 2$ . При этом решающее правило имеет вид «параметр  $\geq$  порога».
3. Подсчитать для каждой составленной в предыдущем пункте алгоритма таблицы чувствительность и неспецифичность. Массив чувствительностей численностью  $n + m$  будет массивом абсцисс ROC кривой. Массив неспецифичностей численностью  $n + m$  будет массивом ординат ROC кривой.
4. Построить график ROC кривой по парам точек «абсцисса–ордината», полученным в предыдущем пункте алгоритма.
5. Подсчитать площадь, отсекаемую ROC кривой.

Позиции 2, 3, 4 и 5 представленного алгоритма выгоднее выполнять в цикле по всем  $n + m$  вариантам массива диагностических параметров.

Объективную оценку качества диагностического метода может показать площадь под ROC кривой, в литературе кратко называемая AUC (Area Under Curve). Оценка данной площади подсчитывается по формуле трапеций:

$$\hat{A} = \frac{1}{2} \sum_{j=1}^{n+m-1} (Se_i + Se_{i+1})(Sp_i - Sp_{i+1}).$$

При расчете оценки площади условились использовать показатели в долях. Чем выше AUC, тем большую прогностическую ценность имеют представленные данные (представленный метод). Максимальное значение AUC равно 1. При значении AUC, равном 0,5, прогностическая ценность отсутствует. Возможна такая конфигурация исходных данных, что кривая ROC окажется ниже диагонали, а AUC окажется, соответственно, в интервале от 0 до 0,5. В этом случае следует изменить решающее правило (позиция 2 алгоритма) на противоположное: «параметр  $\leq$  порога» – и выполнить алгоритм заново.

Стандартная ошибка оценки AUC подсчитывается по формуле, представленной Хэнли (Hanley) с соавт. (1982),

$$SE(\hat{A}) = \sqrt{\frac{\hat{A}(1 - \hat{A}) + (n - 1)(Q_1 - \hat{A}^2) + (m - 1)(Q_2 - \hat{A}^2)}{n \cdot m}},$$

где для краткости записи обозначено:

$$Q_1 = \hat{A}/(2 - \hat{A}),$$





$$Q_2 = 2\hat{A}^2 / (1 + \hat{A}).$$

Хэнли с соавт. предложили метод сравнения двух ROC кривых по отсекаемым ими AUC. Для этого в простейшем случае используется статистика

$$Z = \frac{|\hat{A}_1 - \hat{A}_2|}{\sqrt{SE(\hat{A}_1)^2 + SE(\hat{A}_2)^2}},$$

распределенная асимптотически нормально.

Удобно использовать статистику  $Z$  при сравнении оценки AUC для данной ROC кривой с величиной AUC, равной 0,5 (случай «бесполезной» классификации). Статистика  $Z$  позволяет объективно судить о статистической значимости полученной классификации. При этом  $SE(0,5)$  вычисляется по показанной выше формуле.

Для вычисления двустороннего доверительного интервала оцениваемой AUC применяется формула:

$$I_{AUC} = (\hat{A} - \Psi((1 + \beta)/2) \cdot SE(\hat{A}); \hat{A} + \Psi((1 + \beta)/2) \cdot SE(\hat{A}))$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

Для ROC анализа удобно реализовать возможности:

- Ввод данных типа «выборка – выборка». При этом в качестве первой выборки указывается выборка с одним из значений классификатора (например, общепринятое значение 1, или наличие симптома болезни). В качестве второй выборки указывается выборка с другим значением классификатора (например, общепринятое значение 0, или отсутствие симптома болезни). При этом сами значения классификатора не вводятся.
- Ввод данных типа «выборка – классификатор». При этом в качестве первой выборки вводится весь массив исходных данных (для всех значений классификатора). В качестве второй выборки вводится [соответствующий массиву исходных данных] массив классификатора, состоящий из единиц и нулей. Кодировка классификатора аналогична предыдущему случаю.

Рассматривая возможность ROC анализа для исходных данных, представленных в виде таблицы 2 x 2, необходимо сделать вывод, что такой анализ сделать нельзя. Дело в том, что ROC кривая, как уже упоминалось выше – это не просто изображенная на графике зависимость  $Se$  от  $1 - Sp$ . ROC кривая представляет собой особый математический объект, называемый параметрической кривой. Параметрическая кривая возникает, когда две величины, участвующие в построении графика, на самом деле зависят не одна от другой, а от третьего параметра (на графике не изображаемого) – в данном случае от порога отсека. Аргументом является именно порог отсека, а изображаются на графике произведенные от него чувствительность и неспецифичность. Вот порога–то отсека как раз и нет в представленной таблице. Формально, конечно, можно для таблицы посчитать чувствительность и неспецифичность (в %), добавить еще две точки – (0;0) и (100;100) и нарисовать некий график, даже посчитать площадь под таким объектом, но это будет не ROC



анализ. При необходимости данные формальные построения пользователь выполнит самостоятельно.

В некоторых публикациях бытует ошибочное изображение ROC в виде кривой гладкой. Это демонстрирует непонимание авторами публикаций самой сути ROC анализа как графического отображения результатов бинарной классификации. ROC кривая – не график зависимости одной непрерывной величины от другой непрерывной величины. ROC кривая может изображаться только в виде лесенки (в этом смысле название «кривая» – curve не является корректным). Она дискретна по своей природе, меняя значения абсциссы и ординаты скачками даже при непрерывном изменении порога отсечения, не может быть гладкой, поэтому ее нельзя аппроксимировать гладкой кривой.

Популярное введение в ROC см. в статье Сюэтс (Swets) с соавт. В дополнение к упомянутым источникам по ROC анализу см. монографии Флетчер с соавт., Жоу (Zhou) с соавт., Хайнес (Haynes) с соавт., статьи Метц (Metz), Обучовски (Obuchowski), Дэвис (Davis) с соавт., Фараджи (Faraggi) с соавт., Парк (Park) с соавт., Шистерман (Schisterman) с соавт., Цвайг (Zweig) с соавт., Альтман (Altman) с соавт., Ланглотц (Langlotz), Клотше (Klotsche) с соавт., статьи и отчет Фосетт (Fawcett). Тема упомянута в книгах Петри с соавт., ван Бель (van Belle) с соавт. О порогах отсечения см. также статью Флуцс (Fluss) с соавт. О статистическом сравнении ROC кривых см. статьи Вергара (Vergara) с соавт., Хэнли с соавт. (1983), Метц с соавт., ДеЛонг (DeLong) с соавт. На связь AUC и статистики непараметрического критерия Вилкоксона указано в работе Фосетт (Fawcett), монографии Власова.

#### 4.2.8. Каппа Козна

Для оценки согласия двух классификаций применяется показатель – каппа Козна (Cohen's Kappa). Интерпретация каппы поясняется в следующей таблице.

Значение каппы	Уровень согласия
< 0,00	Плохое согласие (poor)
0,00 – 0,20	Небольшое согласие (slight)
0,21 – 0,40	Удовлетворительное согласие (fair)
0,41 – 0,60	Среднее согласие (moderate)
0,61 – 0,80	Существенное согласие (substantial)
0,81 – 1,00	Почти прекрасное согласие (almost perfect)

Вычисление выборочной оценки каппы производится по формуле

$$\hat{K} = \frac{p_0 - p_e}{1 - p_e},$$

где  $p_0$  – доля случаев, относительно которых существует согласие,

$p_e$  – доля случаев, относительно которых ожидается согласие.

Упомянутые доли вычисляются по формулам, соответственно,





$$p_0 = \frac{n_{11}}{n} \cdot \frac{n_{22}}{n}$$

$$p_e = \frac{r_1}{n} \cdot \frac{c_1}{n} + \frac{r_2}{n} \cdot \frac{c_2}{n},$$

где  $r_1 = n_{11} + n_{12}$  – численность первой строки таблицы,

$c_1 = n_{11} + n_{21}$  – численность первого столбца таблицы,

$r_2 = n_{21} + n_{22}$  – численность второй строки таблицы,

$c_2 = n_{12} + n_{22}$  – численность второго столбца таблицы,

$n = n_{11} + n_{21} + n_{12} + n_{22}$  – численность таблицы,

$n_{11}, n_{21}, n_{12}, n_{22}$  – ячейки таблицы.

Стандартная ошибка каппы вычисляется по формуле

$$SE(\hat{\kappa}) = \sqrt{\frac{p_0(1 - p_0)}{n(1 - p_e)^2}}.$$

Двусторонний доверительный интервал оцениваемой каппы вычисляется по формуле

$$I_{\kappa} = (\hat{\kappa} - \Psi((1 + \beta)/2) \cdot SE(\hat{\kappa}); \hat{\kappa} + \Psi((1 + \beta)/2) \cdot SE(\hat{\kappa})),$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

См. монографию Флейс (Fleiss) с соавт., статьи Коэн (Cohen), Крюсон (Crewson), Флейс, Брайнгтон (Bryington) с соавт., Костина, Заславского с соавт., Виера (Viera) с соавт., Ли (Lee) с соавт., Кундел (Kundel) с соавт. Взвешенную каппу также рассмотрели Коэн, Флейс с соавт. Расчет доверительных интервалов оцениваемой каппы см. также в статьях Блэкман (Blackman) с соавт., Гарнер (Garner), Гарнер с соавт.

## Глава 5. Точные критерии

### 5.1. Введение

Точные (exact) методы проверки статистических гипотез иначе известны как комбинаторные (перестановочные, permutational). Отметим, что точность здесь понимается в смысле решения задачи с установленными ограничениями и принятыми допущениями используемой статистической модели. К данной группе критериев также принято относить методы, для которых известны точные распределения статистик критериев.

Имеется несколько соображений относительно полезности точных непараметрических методов:

- параметрические методы могут применяться, только если доказана нормальность распределения (см. главу «Проверка нормальности распределения») анализируемых выборок, но эмпирические выборки, полученные в реальных экспериментах, очень часто не являются нормально распределенными.







- параметрические методы могут применяться для больших выборок. Реальные выборки часто содержат небольшое число вариантов.

Перед применением любого статистического метода необходимо убедиться, что проверяется статистическая значимость различий именно тех параметров выборок, которые интересуют исследователя, а также в том, что метод соответствует шкале измерения исходных данных (признаков). При выборе метода, неадекватного шкале измерения представленных данных, полученный числовой результат расчета может оказаться лишен какого-либо смысла.

## 5.2. Теоретическое обоснование

Все точные критерии базируются на возможности точного вычисления  $P$ -значения. Основной группой точных критериев являются методы, основанные на перестановках. Алгоритмы методов позволяют вычислить точное  $P$ -значение, зная число благоприятных исходов и общее число исходов опыта, представляющее собой все мыслимые варианты исхода. Следовательно, при вычислении критериев не избежать применения комбинаторных алгоритмов и вычисления дискретных функций распределения, которые могут быть очень трудоемкими в реализации, особенно для больших выборок.

К точным критериям (exact tests) относятся:

- критерий рандомизации для независимых выборок,
- критерий рандомизации для связанных выборок,
- критерий Вилкоксона для независимых выборок,
- критерий Вилкоксона для связанных выборок,
- точный метод Фишера,
- критерий Барнарда,
- критерий Мак-Немара,
- критерий знаков,
- критерий серий Вальда-Вольфовица.

При формулировании нулевой гипотезы обязательно следует указывать, какие конкретные параметры эмпирических выборок сравниваются с помощью используемого критерия. Данная информация приводится в описании каждого критерия. Нужно указывать это в научной публикации, чтобы читатель имел возможность проверить правильность рассуждений автора. В таблице указаны тестируемые параметры выборок для различных критериев.

Тестируемые параметры

Положение: среднее и/или медиана (location tests)

Функция распределения

Точный статистический критерий

Рандомизации, точный метод Фишера, критерий Барнарда, серий Вальда-Вольфовица, критерий Мак-Немара, Вилкоксона  
Знаков







Для пользователей рекомендуются простые и практические источники, например, переведенные на русский язык монография Кендалла с соавт. и книга Руниона. Сравнительный обзор критериев для проверки однородности таблиц сопряженности приводится в статье Мехротра (Mehrotra) с соавт. Обзор подходов Фишера и Барнарда к анализу таблиц сопряженности см. в работе Мартина Андреса (Martin Andres) с соавт.

## 5.2.1. Критерий рандомизации для независимых выборок

Критерий рандомизации компонент Фишера (критерий рандомизации Фишера–Питмана) для независимых выборок применяется для проверки нулевой гипотезы о том, отобраны ли две независимые выборки из совокупностей с одинаковыми средними значениями. Выборки должны принадлежать количественной шкале.

Критерий рандомизации называется также критерием перестановок, выборочное распределение которого при каждом вычислении должно быть получено заново перебором всех возможных исходов.

Методика теста базируется на идее перебора всех комбинаций наблюдаемых отметок. Пусть даны две выборки:  $x_i, i = 1, 2, \dots, n_x$  и  $y_i, i = 1, 2, \dots, n_y$ , где  $n_x, n_y$  – численности выборок. Сумма, меньшая из двух наблюдаемых, будет

$$S = \min \left( \sum_{i=1}^{n_x} x_i, \sum_{i=1}^{n_y} y_i \right).$$

Число благоприятных исходов вычисляется по формуле

$$N = 2 \sum_{i=1}^{C_n^m} n_i, n_i = \begin{cases} 0, s_i < S, \\ 1, s_i \geq S, \end{cases}$$

где  $n_i$  – оценка  $i$ -го исхода,

$C_n^m$  – общее число исходов – число сочетаний из  $n$  по  $m$ ,

$n = n_x + n_y$  – численность объединенной выборки,

$m$  – численность выборки, соответствующей минимальной сумме

$$s_i = \sum_{j=1}^m z_j, i = 1, \dots, C_n^m,$$

где  $z_j, j = 1, 2, \dots, m$  – массив сочетаний из объединенной выборки.

Двустороннее  $P$ -значение вычисляется по формуле

$$p = \frac{N}{C_n^m}$$

и сравнивается с заданным уровнем значимости.

При больших численностях выборок вместо описанного здесь критерия рекомендуется применять  $W$ -критерий Вилкоксона (см. главу «Непараметрическая статистика»), являющийся критерием ранговой рандомизации.



См. также описание и пример критерия Питмана–Уэлча в монографии Файнштейн (Feinstein). См. справочник Руниона, статьи Питмана (Pitman), Кайзера (Kaiser), монографии Фишера (Fisher), Зигеля (Siegel) с соавт.

## 5.2.2. Критерий рандомизации для связанных выборок

Критерий рандомизации компонент Фишера (критерий рандомизации Фишера–Питмана) для связанных выборок применяется для проверки нулевой гипотезы о равенстве средних значений двух связанных совокупностей. Выборки должны принадлежать количественной шкале.

Критерий рандомизации называется также критерием перестановок, выборочное распределение которого при каждом вычислении должно быть получено заново перебором всех возможных исходов.

Основным моментом в реализации критерия является перебор возможных исходов, построенных из разностных отметок. Пусть даны две выборки:  $x_i, y_i, i = 1, 2, \dots, n$ , где  $n$  – число пар экспериментальных значений. Тогда сумма массива разностных отметок будет

$$S = \sum_{i=1}^n s_i.$$

Определим значения разностных отметок:

$$s_i = \sum_{j=1}^n a_{ij} (x_j - y_j), i = 1, \dots, 2^n,$$

где  $a_{ij}, i = 1, 2, \dots, 2^n; j = 1, 2, \dots, n$  – элементы матрицы возможных исходов.

Отметим, что в некоторых источниках разность вариантов в показанной выше формуле берется по модулю. Однако анализ формулы показывает, что в процессе перебора операция взятия модуля в данном случае значения не имеет.

Систематизацию перебора всех возможных исходов удобно провести в соответствии с ортогональным планом эксперимента первого порядка. Размер полного ортогонального плана составляет  $2n$  строк на  $n$  столбцов, причем  $j$ -й столбец размером  $2n$  представляет собой чередующиеся с шагом  $2^{j-1}$  величины  $+1$  и  $-1, j = 1, 2, \dots, n$ .

Число благоприятных исходов вычисляется по формуле:

$$N = \sum_{i=1}^{2^n} n_i, n_i = \begin{cases} 0, s_i < S, \\ 1, s_i \geq S. \end{cases}$$

Двустороннее  $P$ -значение, вычисляемое по формуле

$$p = \frac{N}{2^n},$$

сравнивается с заданным уровнем значимости.

При больших численностях выборок рекомендуется применять  $T$ -критерий Вилкоксона (см. главу «Непараметрическая статистика»), являющийся критерием ранговой рандомизации.



См. справочник Руниона, статьи Питмана (Pitman), Кайзера (Kaiser), монографии Фишера (Fisher), Зигеля (Siegel) с соавт. Об ортогональных планах см. монографии Шеффлера (Scheffler), Корикова, Монтгомери.

### 5.2.3. Критерий Вилкоксона для независимых выборок

Критерий Вилкоксона для независимых выборок является аналогом критерия рандомизации для независимых выборок с той разницей, что все операции производятся не над вариантами выборок, а над их рангами.

Метод имеет те же ограничения, что и критерий рандомизации, уступает ему в мощности, но не уступает в трудоемкости вычислений и поэтому находит ограниченное применение.

Для больших выборок следует использовать асимптотический  $W$ -критерий Вилкоксона (см. главу «Непараметрическая статистика»).

О точном вычислении критерия Вилкоксона для независимых выборок см. Браунли, Уилкса.

### 5.2.4. Критерий Вилкоксона для связанных выборок

Критерий Вилкоксона для связанных выборок является аналогом критерия рандомизации для связанных выборок с той разницей, что все операции производятся не над вариантами выборок, а над их рангами.

Метод имеет те же ограничения, что и критерий рандомизации, уступает ему в мощности, но не уступает в трудоемкости вычислений и поэтому находит ограниченное применение.

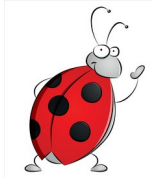
Для больших выборок следует использовать асимптотический  $T$ -критерий Вилкоксона, реализованный в главе «Непараметрическая статистика».

О точном вычислении критерия Вилкоксона для связанных выборок см. Браунли.

### 5.2.5. Точный метод Фишера

Точный метод Фишера (критерий Фишера, точный метод Фишера–Ирвина, критерий Фишера–Ирвина, Fisher's exact test, Fisher–Irwin test, Fisher–Yates–Irwin exact test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц типа  $2 \times 2$ . Для применения критерия анализируемые выборки должны принадлежать дихотомической шкале измерения. Принято, что исходные выборки должны состоять только из нулей и единиц, причем нуль означает отсутствие признака (эффекта), а единица означает наличие признака (эффекта).

	Наличие эффекта А		
	Да	Нет	
Выборка (группа) 1	$a$	$b$	$M_1 = a + b$
Выборка (группа) 2	$c$	$d$	$M_2 = c + d$



Сумма

$$N_1 = a + c$$

$$N_2 = b + d$$

$$n = a + b + c + d$$

Вычисление односторонних достигнутых уровней значимости критерия производится путем суммирования вероятностей всех вариантов  $p(X)$  заполнения таблицы сопряженности:

$$P_U = \sum_{\substack{T(X) > T(X_0) \\ ad < bc}} p(X),$$

$$P_L = \sum_{\substack{T(X) > T(X_0) \\ ad \geq bc}} p(X),$$

где  $T(X)$  – статистика Вальда для текущего варианта заполнения таблицы,

$T(X_0)$  – статистика Вальда исходной таблицы сопряженности.

Двусторонний достигнутый уровень значимости критерия равен

$$P_F = P_U + P_L.$$

Статистика Вальда в данном случае вычисляется по формуле.

$$T(X) = \frac{n \left| \frac{a}{a+c} - \frac{b}{b+d} \right|}{\sqrt{(a+b)(c+d) \left( \frac{1}{a+c} + \frac{1}{b+d} \right)}},$$

где  $a$  – число наблюдений с эффектом  $A$  в первой выборке,

$b$  – число наблюдений без эффекта  $A$  в первой выборке,

$c$  – число наблюдений с эффектом  $A$  во второй выборке,

$d$  – число наблюдений без эффекта  $A$  во второй выборке,

$n = a + b + c + d$  – численность таблицы сопряженности.

Статистика Вальда выводится в качестве критериальной статистики.

Варианты заполнения таблицы сопряженности планируются при условии сохранения всех маргинальных сумм. Это означает, что для всех вариантов таблицы маргинальные суммы  $N_1$ ,  $N_2$ ,  $M_1$ ,  $M_2$  должны быть одинаковыми.

Некоторыми авторами приводится эквивалентная (и гораздо более быстрая в вычислении) формула вычисления критерия. Отличие заключается в замене статистики Вальда текущего варианта заполнения таблицы и статистики Вальда исходной таблицы на, соответственно, вероятность  $p(X)$  и вероятность исходной таблицы  $p(X_0)$ .

Вместо указанной точной условной вероятности биномиального распределения Фишер предложил использовать вероятность гипергеометрического распределения

$$p(X) = \frac{C_{a+b}^a C_{c+d}^c}{C_n^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}.$$

По этой причине описываемый критерий и все методы, основанные на данной идее, называются условными критериями (conditional tests). Подробные соображения по данному вопросу изложены в т. 1 справочника под ред. Ллойда с соавт. В этом же источнике описаны методики получения таблиц сопряженности.



См. монографии ван Бель (van Belle) с соавт., Черник (Chernick) с соавт., Ле (Le), диссертацию Бучана (Buchan). Сравнительный обзор приводится в статье Мехротра (Mehrotra) с соавт. Описание критерия см. в книгах Лемана, Руниона, Флейса, Кендалла с соавт., Бергера (Berger) с соавт. На основе идеи Фишера для обработки таблиц сопряженности типа  $r \times c$  Фриман (Freeman) и Холтон (Halton) разработали расширенный тест, представленный в главе «Кросстабуляция».

## 5.2.6. Критерий Барнарда

Критерий Барнарда (Barnard's test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц типа  $2 \times 2$ . Для применения критерия анализируемые выборки должны принадлежать дихотомической шкале измерения. Принято, что исходные выборки должны состоять только из нулей и единиц, причем нуль означает отсутствие признака (эффекта), а единица означает наличие признака (эффекта).

	Наличие эффекта А		
	Да	Нет	
Выборка (группа) 1	$a$	$b$	$a + b$
Выборка (группа) 2	$c$	$d$	$c + d$
Сумма	$N_1 = a + c$	$N_2 = b + d$	$n = a + b + c + d$

Точный двусторонний достигнутый уровень значимости критерия  $P_B$  определяется как

$$P_B = \sup_{0 < \pi < 1} \left\{ \sum_{T(X) > T(X_0)} p(X, \pi) \right\},$$

где  $\pi$  – параметр распределения,

$p(X, \pi)$  – вероятность варианта заполнения таблицы,

$T(X)$  – статистика Вальда варианта заполнения таблицы сопряженности,

$T(X_0)$  – статистика Вальда исходной таблицы сопряженности.

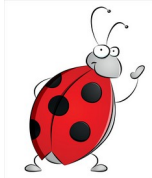
Максимум целевой функции может быть найден с помощью методов оптимизации.

Вероятность таблицы сопряженности вычисляется по формуле вероятности биномиального распределения

$$p(X, \pi) = C_{a+c}^a C_{b+d}^b \pi^{a+b} (1 - \pi)^{c+d}.$$

Это первое отличие критерия Барнарда от критерия Фишера (см. точный метод Фишера). По этой причине и в противоположность условным критериям описываемый критерий и все методы, основанные на данной идее, называются безусловными (unconditional tests).

Статистика Вальда в рассматриваемом случае имеет вид



$$T(X) = \frac{n \left| \frac{a}{a+c} - \frac{b}{b+d} \right|}{\sqrt{(a+b)(c+d) \left( \frac{1}{a+c} + \frac{1}{b+d} \right)}},$$

где  $a$  – число наблюдений с эффектом  $A$  в первой выборке,  
 $b$  – число наблюдений без эффекта  $A$  в первой выборке,  
 $c$  – число наблюдений с эффектом  $A$  во второй выборке,  
 $d$  – число наблюдений без эффекта  $A$  во второй выборке,  
 $n = a + b + c + d$  – численность таблицы сопряженности.

Вычисляется «оптимальное» значение параметра  $\pi$ . В качестве критериальной статистики выводится статистика Вальда и  $P$ -значение.

Варианты заполнения таблицы сопряженности планируются при условии сохранения маргинальных сумм  $N_1$  и  $N_2$ . Это означает, что для всех вариантов таблицы значения маргинальные суммы  $N_1$  и  $N_2$  должны быть одинаковыми. Это второе отличие критерия Барнарда от критерия Фишера. По сути, эти два отличия отражают все основные подходы к обработке таблиц сопряженности (как  $2 \times 2$ , так  $r \times c$ ) и определяют их два основных направления развития:

- Подход Фишера: гипергеометрическое распределение и все фиксированные маргинальные суммы.
- Подход Барнарда: биномиальное распределение (с вычислением оптимального параметра) и фиксированные суммы столбцов.

Критерий Барнарда более трудоемок в вычислении, чем точный метод Фишера. Это вызвано необходимостью решать задачу поиска оптимального значения параметра. Задача упрощается тем, что зависимость целевой функции от данного параметра, как показывают исследования, симметрична относительно  $\pi = 0,5$  и имеет форму либо «шляпы», либо «сомбреро», в зависимости от соотношения частот таблицы сопряженности. Стратегия решения – стандартная. На начальном (глобальном) этапе простым методом перебора производится поиск интервала локализации параметра распределения, который затем уточняется до нужной точности с помощью быстродействующего локального метода.

При работе нужно учитывать, что критерий трудоемок в вычислении, причем снять задачу с выполнения можно, не дожидаясь ее нормального окончания, только средствами операционной системы.

Описание критерия см. в оригинальных статьях Барнарда (Barnard), работах Мехта (Mehta) с соавт., статье Мартин Андрес (Martin Andres) с соавт. В сравнительном обзоре Мехротра (Mehrotra) с соавт. дано современное состояние вопроса и представлены дальнейшие развития идеи Барнарда. Методы локальной оптимизации см. в пособии Вержбицкого.



### 5.2.7. Критерий Мак–Немара

Критерий Мак–Немара (McNemar's test) применяется для проверки нулевой гипотезы о том, отобраны ли две исследуемые парные бинарные выборки из генеральных совокупностей с одинаковой частотой встречаемости изучаемого эффекта. Рассматриваемый метод предназначен для обработки так называемых четырехпольных (четырёхклеточных) таблиц, или таблиц типа 2 x 2:

Эффект A		Эффект B	
		Да	Нет
Да		<i>a</i>	<i>b</i>
Нет		<i>c</i>	<i>d</i>

Метод идеально подходит для анализа данных типа «до и после».

Вычисление статистики критерия производится по формуле:

$$\chi^2 = \frac{(|b - c| - Y)^2}{b + c},$$

где *b* – число индивидуумов с наличием эффекта A и отсутствием эффекта B,

*c* – число индивидуумов с отсутствием эффекта A и наличием эффекта B,

*Y* = 0 – если не используется поправка на непрерывность (поправка Йейтса),

*Y* = 1 – если используется поправка на непрерывность.

Представлены 3 варианта критерия:

1. Асимптотика хи–квадрат.
2. Асимптотика хи–квадрат с поправкой Йейтса.
3. Точный вариант критерия.

По поводу вычисления *P*–значений в первых двух вариантах критерия см. главу «Непараметрическая статистика».

Точный двусторонний достигнутый уровень значимости критерия *P<sub>x</sub>* определяется суммированием вероятностей всех вариантов заполнения таблицы, при условии сохранения суммы ячеек *b* и *c* исходной таблицы, как

$$P_x = \sum_{p(X) \geq p(X_0)} p(X),$$

где *p(X)* – вероятность варианта заполнения таблицы,

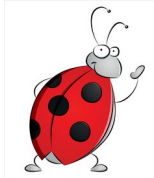
*p(X<sub>0</sub>)* – вероятность исходной таблицы.

Вероятность заполнения таблицы вычисляется по формуле биномиального распределения (адаптированной к данному случаю)

$$p(X) = \frac{0,5^{b+c} (b+c)!}{b!c!}.$$

Описание критерия см. в статьях Беннетта (Bennett) с соавт., Дуайера (Dwyer), Лиделла (Liddell).





## 5.2.8. Критерий знаков

Критерий знаков Фишера предназначен для проверки гипотезы об однородности распределения совокупности, что эквивалентно проверке гипотезы о равенстве функций распределения. Критерий часто используется при сравнении эффективности двух различных способов воздействия на  $n$  объектов. Он может применяться и для связанных выборок. Выборки могут принадлежать порядковой или количественной шкале. Требованием является равная численность сравниваемых выборок, в том числе и независимых выборок. Статистика критерия вычисляется как число положительных разностей вариант выборок:

$$B = \sum_{i=1}^n s(x_i, y_i),$$

$$s(x_i, y_i) = \begin{cases} 1, & x_i > y_i, \\ 0, & x_i < y_i, \end{cases}$$

где

$x_i, y_i, i = 1, 2, \dots, n$  – варианты выборок,

$n$  – численность каждой выборки.

Если среди значений вариант есть совпадающие, т. е.  $x_i = y_i, i = 1, 2, \dots, n$ , то данные пары значений отбрасываются и, соответственно, на число отброшенных значений сокращается численность  $n$ .

Точное критическое значение критерия знаков для любой численности вычисляется на основе функции биномиального распределения с параметрами  $(B; n; 0,5)$ .

Для больших выборок (на самом деле аппроксимация хорошо работает уже при численности, равной 25 вариант в каждой выборке)  $P$ -значение может также вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{B - EB}{\sqrt{DB}},$$

где  $EB = n / 2$  – математическое ожидание,

$DB = n / 4$  – дисперсия,

распределена по стандартному нормальному закону.

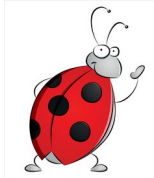
См. книги Браунли, Лемана.

## 5.2.9. Критерий серий Вальда–Вольфовица

Критерий серий Вальда–Вольфовица (Wald–Wolfowitz runs test) предназначен для проверки нулевой гипотезы о равенстве целого ряда параметров двух сравниваемых выборок, включая медианы и коэффициенты асимметрии. Критерий применяется в случае, если исследователя интересует, имеют ли место любые различия между совокупностями. Выборки могут принадлежать порядковой или количественной шкале. Расчет заключается в объединении выборок с численностями  $n_1$  и  $n_2$  в одну выборку общей численностью  $n_1 + n_2$ , ее сортировке по возрастанию или убыванию и подсчете числа серий элементов  $R$ , относящихся к первой и второй выборкам.







Точное одностороннее  $P$ -значение статистики критерия вычисляется как

$$P(r \leq R) = \frac{1}{C_{n_1+n_2}^{n_1}} \sum_{i=2}^R F_i,$$

где величины под знаком суммы вычисляются так:

для четных индексов  $F_i = 2C_{n_1-1}^{k-1}C_{n_2-1}^{k-1}$ , где  $k = i / 2$ ,

для нечетных индексов  $F_i = C_{n_1-1}^{k-2}C_{n_2-1}^{k-1} + C_{n_1-1}^{k-1}C_{n_2-1}^{k-2}$ , где  $k = (i + 1) / 2$ .

Отметим, что в формуле для вероятности не имеет значения, стоит в знаменателе число сочетаний из  $n_1 + n_2$  по  $n_1$  или по  $n_2$ , т. к. показано (Браунли), что

$$C_{n_1+n_2}^{n_1} = C_{n_1+n_2}^{n_2} = \frac{(n_1 + n_2)!}{n_1!n_2!}.$$

В асимптотической версии критерия (см. главу «Непараметрическая статистика») для вычисления  $P$ -значения используется нормальная аппроксимация.

Метод описан в монографии Браунли. Замечания о применении см. в книге Гаека с соавт., статье Камень с соавт.

## Глава 6. Кросстабуляция

### 6.1. Введение

Предлагаются методы анализа однородности и сопряженности (связи типа корреляции) в таблицах сопряженности, полученных на основе выборок, измеренных в номинальной шкале.

### 6.2. Теоретическое обоснование

Кросстабуляцией (cross-tabulation, analysis of cross-tabulated data) называют анализ двухвходовых (двумерных) таблиц сопряженности (таблиц смежности, contingency tables). Таблицы сопряженности возникают при анализе признаков, измеренных в номинальной шкале либо в более высоких шкалах, преобразованных к номинальной шкале.

Двухвходовые таблицы могут быть проанализированы методами:

анализ однородности (согласия) в таблицах типа  $r \times c$ :

- критерий Кресси–Рида,
- критерий Хеллингера,
- критерий хи–квадрат,
- критерий отношения правдоподобия,
- критерий Зелтермана,
- критерий Фримана–Холтона.



анализ однородности (согласия) и симметрии в таблицах типа  $k \times k$ :

- критерий Стюарта–Максвелла,
- критерий Баукера,
- критерий Бхапкара.

Для исследования сопряженности признаков (связи типа корреляции, не путать с корреляцией, которая для номинальных признаков, отражением которых являются таблицы сопряженности, не определена), предназначены специальные методы, как-то:

- коэффициент Кендалла,
- коэффициент Крамера,
- коэффициент Сомерса,
- коэффициент сопряженности Пирсона.

Дисперсионный анализ выборок, представленных таблицами типа  $r \times c$ , может быть выполнен методами:

- критерий Краскела–Уоллиса.

С другой стороны, данные методы отражают два подхода к решению проблемы:

- Критерии первой группы включают в себя, наряду с классическим критерием хи-квадрат, основанные на хи-квадрат методы: коэффициенты Крамера и сопряженности Пирсона.
- Непараметрические ранговые методы включают: коэффициенты Кендалла и Сомерса, критерий Краскела–Уоллиса.

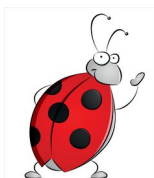
Критерии, основанные на хи-квадрат, с одной стороны, и коэффициенты, на хи-квадрат не основанные (например, коэффициент Кендалла), могут при вычислении давать различные результаты. Это вызвано тем, что критерии, основанные на хи-квадрат, нечувствительны к упорядочению строк и столбцов таблицы сопряженности.

Для исследования сопряженности признаков также предназначены не рассмотренные здесь мера  $\tau_c$  Стюарта, коэффициент ранговой корреляции  $r_s$  Спирмэна, другие методы. Все эти методы представлены в монографии Афифи с соавт., включая нормальные аппроксимации, позволяющие использовать данные методы для проверки значимости связи.

Укажем на особенность тестов, основанных на распределении хи-квадрат. Распределения статистик данных критериев лишь приблизительно соответствуют хи-квадрат:

- Согласно Кокрену (см. Сергиенко с соавт., с. 79), если для таблицы  $2 \times 2$  сумма таблицы  $< 20$  или сумма таблицы от 20 до 40, но при этом в одной из ячеек ожидаемая частота  $< 5$ , то следует использовать не критерий хи-квадрат, а точный метод Фишера (см. главу «Точные критерии»).
- Согласно Аптону (глава 3), приближение работает достаточно хорошо, пока ожидаемые частоты в ячейках таблицы сопряженности не опустятся примерно до трех.

Объективными критериями допустимости аппроксимации хи-квадрат являются так называемые диагностики: Симонов–Цай, Хабермана, Мудхолкара–Хадсона и другие.



В практике иногда возникает необходимость проверки однородности данных, представленных в виде строк таблицы сопряженности, методами дисперсионного анализа. Как известно, построение таблицы сопряженности из количественных данных понижает шкалу. При этом восстановление исходных данных по имеющейся таблице невозможно. Однако в случае, если исходные данные были порядковыми, понижения шкалы не происходит (хотя исходные данные также восстановить нельзя).

Для проведения дисперсионного анализа исходных порядковых данных и данных, восстановленных по таблице сопряженности, могут применяться одни и те же методы непараметрического дисперсионного анализа. Результат непараметрического дисперсионного анализа восстановленных с точностью до коэффициентов данных будет совпадать с результатами анализа исходных порядковых данных. Данная возможность обеспечивается процедурой ранжирования, применяемых в данных методах.

На следующем примере показана возможность восстановления порядковой выборки из строки таблицы сопряженности. Пусть имеется таблица результатов лечения для группы пациентов.

	Плохо	Результат лечения Удовлетворительн	Хорошо	
		0		
Группа 1	2	5	10	17
Группа 2	5	4	4	13

Не имеет значения для непараметрического дисперсионного анализа, какие величины имели те или иные варианты до построения таблицы сопряженности, однако их соотношение должно соблюдаться. Поэтому для определенности можно выбрать кодировку: плохо – 1, удовлетворительно – 2, хорошо – 3. В показанном примере можно восстановить исходные порядковые выборки:

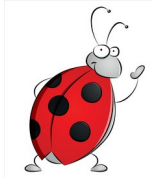
1. Группа 1 (численность 17): 1 1 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3,
2. Группа 2 (численность 13): 1 1 1 1 1 2 2 2 2 3 3 3 3.

Данные выборки могут быть проанализированы любыми непараметрическими методами дисперсионного анализа. При восстановлении исходных порядковых данных из таблиц сопряженности неизбежно появление связей, поэтому применяемые методы, как вариант метода Краскела–Уоллиса, используются только с учетом связей.

Представленные параметры сопряженности (связи типа корреляции для номинальных признаков) могут применяться в качестве показателей статистической значимости связи между признаками. Данную связь допустимо интерпретировать как корреляционную, но нельзя называть корреляцией, т. к. для номинальных признаков корреляция не определена. Поэтому лучше использовать термины «сопряженность» или «связь типа корреляции».

Подробнее о корреляции см. «Корреляционный анализ».

Афифи с соавт. приводят общую формулу расчета показателей, для которых неизвестно или затруднительно вычисление критических значений. Используется тот факт, что статистика



$$z = \frac{X}{\sqrt{DX}},$$

где  $X$  – статистика критерия,

$DX$  – дисперсия,

асимптотически имеет стандартное нормальное распределение  $N(0,1)$ .

## 6.2.1. Критерий Кресси–Рида

Критерий Кресси–Рида (power-divergence family Cressie–Read) является наиболее общим методом анализа однородности таблиц сопряженности. Вычисление критерия производится по формуле

$$CR(\lambda) = \sum_{i=1}^r \sum_{j=1}^c \frac{2}{\lambda(1+\lambda)} A_{ij} \left[ \left( \frac{A_{ij}}{E_{ij}} \right)^{\lambda} - 1 \right],$$

где  $A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$E_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – ожидаемые частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности,

$\lambda$  – параметр, равный  $2/3$ .

По условиям вычисления статистики критерия, при  $A_{ij} = 0$ , во избежание численных проблем, условились считать, что  $A_{ij}[\dots] = 0$ .

Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \dots, r; j = 1, 2, \dots, c,$$

$$n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r,$$

где – суммы строк таблицы сопряженности,

$$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$$

– суммы столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl}$$

– общее число наблюдений.

Статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $(r-1)(c-1)$ .

Представленный критерий является основой семейства тестов, которые получаются в результате того или иного выбора параметра  $\lambda$ . Например, можно получить при значениях параметра:

- $\lambda = 1$  – критерий хи-квадрат,
- $\lambda = 0$  – критерий отношения правдоподобия.

Находят применение и другие значения параметра, в том числе отрицательные.

См. работы фон Давье (Von Davier), Браво (Bravo), Базу (Basu) с соавт.





## 6.2.2. Критерий Хеллингера

Критерий Хеллингера (blended weight Hellinger) является методом анализа однородности в таблицах сопряженности. Вычисление критерия производится по формуле

$$BWH(\alpha) = \sum_{i=1}^r \sum_{j=1}^c \left( \frac{A_{ij} - E_{ij}}{\alpha \sqrt{A_{ij}} + (1-\alpha) \sqrt{E_{ij}}} \right)^2,$$

где  $A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$E_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – ожидаемые частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности,

$\alpha$  – параметр, равный 1/2 или 1/9.

Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \dots, r; j = 1, 2, \dots, c,$$

$$n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r,$$

где – суммы строк таблицы сопряженности,

$$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$$

– суммы столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl} \quad \text{– общее число наблюдений.}$$

Статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $(r-1)(c-1)$ .

См. работы Бхаттачарья (Bhattacharya) с соавт., Базу (Basu) и Рэй (Ray) с соавт.

## 6.2.3. Критерий хи-квадрат

Классический критерий хи-квадрат (критерий хи-квадрат Пирсона, Pearson chi-square test, Pearson's  $X^2$  test) является стандартным для анализа таблиц сопряженности. Вычисление критерия производится по формуле

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

где  $A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$E_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – ожидаемые частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности.

Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \dots, r; j = 1, 2, \dots, c,$$





$$n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r,$$
 где – суммы строк таблицы сопряженности,

$$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$$
 – суммы столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl}$$
 – общее число наблюдений.

Для больших выборок статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $(r-1)(c-1)$ .

См. работы Мехта (Mehta) с соавт., Стаффорда (Stafford). В работе Мудхолкара (Mudholkar) и Хатсона (Hutson) проведен анализ возможности аппроксимации распределения статистики критерия, введены т. н. диагностики, которые позволяют судить о правомерности данной процедуры.

Модификации критерия хи-квадрат для анализа многовходовых таблиц сопряженности см. в монографии Аптона, статьях Кастенбаума (Kastenbaum) с соавт., Гудмана (Goodman).

Модификация критерия хи-квадрат для анализа таблиц сопряженности типа  $2 \times k$  носит наименование критерия тренда Кокрена-Армитеджа (Cochran–Armitage test for trend) и представлена Агрести (Agresti, 2002). Особенностью критерия является введение т. н. весовых функций, позволяющих формулировать различные нулевые гипотезы в рамках представленной таблицы.

#### 6.2.4. Критерий отношения правдоподобия

Классический критерий отношения правдоподобия (likelihood ratio test,  $G^2$  test) является стандартным методом исследования однородности таблиц сопряженности. Вычисление критерия производится по формуле

$$G^2 = 2 \sum_{i=1}^r \sum_{j=1}^c A_{ij} \log \frac{A_{ij}}{E_{ij}},$$

где  $A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$E_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – ожидаемые частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности.

По условиям вычисления статистики критерия, при  $A_{ij} = 0$ , во избежание численных проблем, условились считать, что  $A_{ij} \log \dots = 0$ .

Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \dots, r; j = 1, 2, \dots, c,$$



$$n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r,$$
 где – суммы строк таблицы сопряженности,

$$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$$
 – суммы столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl}$$
 – общее число наблюдений.

Статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $(r-1)(c-1)$ .

Описание см. в работах Мехта (Mehta) с соавт.

## 6.2.5. Критерий Зелтермана

Статистика критерия Зелтермана (Zelterman's statistic) для исследования однородности таблиц сопряженности. Вычисление критерия производится по формуле

$$D_z^2 = X^2 - \sum_{i=1}^r \sum_{j=1}^c \frac{A_{ij}}{E_{ij}} + rc,$$

где  $X^2$  – статистика критерия хи-квадрат,

$A_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$E_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, c$  – ожидаемые частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности.

Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \dots, r; j = 1, 2, \dots, c,$$

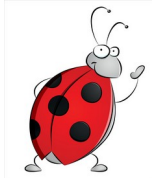
$$n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r,$$
 где – суммы строк таблицы сопряженности,

$$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$$
 – суммы столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl}$$
 – общее число наблюдений.

Статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $(r-1)(c-1)$ .

См. материалы Лаззаротто (Lazzarotto) с соавт.



### 6.2.6. Критерий Фримана–Холтона

Критерий Фримана–Холтона (Фишера–Фримана–Холтона, Fisher–Freeman–Halton test) предназначен для проверки однородности таблицы сопряженности. Критерий является расширением точного метода Фишера.

Пусть  $X_0$  – заданная таблица сопряженности, а  $X$  – вариант заполнения таблицы сопряженности при условии сохранения сумм строк и сумм столбцов заданной таблицы (маргинальных сумм). Тогда достигнутый уровень значимости критерия Фримана–Холтона будет вычисляться как сумма вероятностей всех таблиц  $X$ , таких, что  $P(X) < P(X_0)$ , иначе

$$p = \sum_{P(X) < P(X_0)} P(X).$$

Вероятности таблиц сопряженности  $P(X)$  вычисляются по формуле вероятности гипергеометрического распределения

$$P(X) = \frac{1}{N!} \prod_{i=1}^r R_i! \prod_{j=1}^c C_j! / \prod_{i,j=1}^r x_{ij}!,$$

где  $N$  – численность заданной таблицы сопряженности,

$R_i, i = 1, 2, \dots, r$  – суммы строк заданной таблицы,

$C_j, j = 1, 2, \dots, c$  – суммы столбцов заданной таблицы,

$x_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, c$  – частоты, составляющие таблицу сопряженности,

$r$  – число строк таблицы,

$c$  – число столбцов таблицы.

Оценка требуемого числа генерируемых таблиц позволяет прогнозировать время, затраченное на расчет. Для этого применяется алгоритм, разработанный Гейлом (Gail) и Мантелем (Mantel), согласно которому оценка числа генерируемых таблиц производится по приближенной формуле

$$n \approx \left( \frac{c-1}{2\pi\sigma^2 c} \right)^{(c-1)/2} \sqrt{c} e^{-Q/2} \prod_{i=1}^r C_{R_i+c-1}^{c-1},$$

$$\sigma^2 = \frac{c-1}{(c+1)c^2} \sum_{i=1}^r R_i(R_i+c)$$

где – дисперсия,

$$Q = \frac{c-1}{\sigma^2 c} \left( \sum_{j=1}^c C_j^2 - \frac{N^2}{c} \right)$$

– параметр.

По условиям алгоритма, если  $c > r$ , для вычислений оценки числа таблиц исходная таблица сопряженности [автоматически] транспонируется.

Расчет критерия выполняется методом Монте–Карло (генерируется заданное число таблиц). Генерация таблиц осуществляется по алгоритму Пэйтфилда (Patefield). Результатом расчета является приближенное  $P$ -значение, получающееся как отношение числа таблиц, удовлетворяющего показанному выше условию, к общему числу сгенерированных таблиц. По умолчанию число генерируемых таблиц равно 1 миллиону. Этого достаточно для многих





задач и не является трудоемким в предложенной реализации (рассчитывается практически мгновенно). Если это число окажется равным или меньшим числа, примерно оцениваемого по алгоритму Гейла–Мантеля, следует увеличить число генерируемых таблиц как минимум до оцениваемого по алгоритму Гейла–Мантеля, затем повторить расчет.

Алгоритм Пэйтфилда был модифицирован с тем, чтобы использовать более качественные псевдослучайные числа. Использован алгоритм, представленный в работах Лекюйе (L'Ecuyer), Фокс (Fox), Брэтли (Bratley) с соавт. Описание критерия см. в работах Мехта (Mehta) с соавт. Методы решения предложены Халворсеном (Halvorsen), Борковым (Borkowf), Сандерсом (Saunders), Бойеттом (Boyett).

## 6.2.7. Критерий Стюарта–Максвелла

Критерий однородности Стюарта–Максвелла (Stuart–Maxwell test) является расширением критерия Мак–Немара (см. главу «Непараметрическая статистика») для анализа таблиц сопряженности типа  $k \times k$ . Вычисление критерия производится по формуле

$$\chi^2 = D'S^{-1}D,$$

где  $D$  – вектор–столбец, составленный из величин  $d_i = n_i - n_{.i}$ ,  $i = 1, 2, \dots, k-1$ ,

$S$  – квадратная матрица порядка  $k-1$ , составленная из величин

$$s_{ij} = \begin{cases} -(A_{ij} + A_{ji}), & i \neq j, \\ n_{.i} + n_{.i} - 2A_{ii}, & i = j, \end{cases}$$

где  $k$  – число строк и столбцов таблицы сопряженности,

$A_{ij}$ ,  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, k$  – заданные частоты таблицы сопряженности,

$$n_{.i} = \sum_{j=1}^k A_{ij}, i = 1, 2, \dots, k,$$

– суммы строк таблицы сопряженности,

$$n_{.i} = \sum_{j=1}^k A_{ji}, i = 1, 2, \dots, k,$$

– суммы столбцов таблицы сопряженности.

Статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $k-1$ .

См. статьи Максвелла (Maxwell), Стюарта (Stuart).

## 6.2.8. Критерий Баукера

Критерий симметрии Баукера (Bowker test) является расширением критерия Мак–Немара (см. главу «Непараметрическая статистика») для анализа таблиц сопряженности типа  $k \times k$ . Вычисление критерия производится по формуле

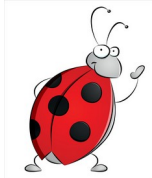
$$\chi^2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^k \frac{(A_{ij} - A_{ji})^2}{A_{ij} + A_{ji}},$$

где  $k$  – число строк и столбцов таблицы сопряженности,

$A_{ij}$ ,  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, k$  – заданные частоты таблицы сопряженности.

Статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $k(k-1)/2$ .





См. статью Баукера (Bowker), отчет Льюиса (Lewis) с соавт.

## 6.2.9. Критерий Бхапкара

Критерий однородности Бхапкара (Bhapkar's test) предназначен для анализа таблиц сопряженности типа  $k \times k$ . Вычисление критерия производится по формуле

$$W = nD'S^{-1}D,$$

$$n = \sum_{i=1}^k \sum_{j=1}^k A_{ij}$$

где  $A_{ij}$ ,  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, k$  – заданные частоты таблицы сопряженности,

$k$  – число строк и столбцов таблицы сопряженности,

$D$  – вектор–столбец, составленный из величин  $d_i = n_{i.} - n_{.i}$ ,  $i = 1, 2, \dots, k - 1$ ,

$S$  – квадратная матрица порядка  $k - 1$ , составленная из величин

$$s_{ij} = \begin{cases} - (n_{ij} + n_{ji}) - (n_{i.} - n_{.i})(n_{.j} - n_{.j}), i \neq j, \\ n_{i.} + n_{.i} - 2n_{ii} - (n_{i.} - n_{.i})^2, i = j, \end{cases}$$

где  $n_{ij} = A_{ij} / n$ ,  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, k$  – частоты,

$$n_{i.} = \sum_{j=1}^k n_{ij}, i = 1, 2, \dots, k, \quad \text{– суммы строк таблицы частостей,}$$

$$n_{.i} = \sum_{j=1}^k n_{ji}, i = 1, 2, \dots, k, \quad \text{– суммы столбцов таблицы частостей.}$$

Статистика критерия подчиняется распределению  $\chi^2$  с числом степеней свободы  $k - 1$ .

См. статьи Бхапкара (Bhapkar), Бхапкара с соавт., отчет Льюиса (Lewis) с соавт.

## 6.2.10. Коэффициент Кендалла

Коэффициент  $\tau_b$  Кендалла (коэффициент Кендэла, Kendall's  $\tau_b$ ) вычисляется по формуле, подробно рассмотренной Аффиси с соавт. и удобной для численных расчетов:

$$\tau_b = S / \sqrt{\left[ \frac{1}{2} n(n-1) - T_1 \right] \left[ \frac{1}{2} n(n-1) - T_2 \right]},$$

где  $S = P - Q$ ,

$$P = \sum_{i=1}^r \sum_{j=1}^c A_{ij} \left( \sum_{k>i} \sum_{l>j} A_{kl} \right) \quad \text{– число пар объектов с взаимно возрастающими переменными,}$$

$$Q = \sum_{i=1}^r \sum_{j=1}^c A_{ij} \left( \sum_{k>i} \sum_{l<j} A_{kl} \right) \quad \text{– число пар объектов с взаимно убывающими переменными,}$$

$A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,





$r$  – число строк таблицы сопряженности,  
 $c$  – число столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl} \quad \text{– общее число наблюдений,}$$

$$T_1 = \frac{1}{2} \sum_{i=1}^r n_{i.} (n_{i.} - 1) \quad \text{– число пар объектов с взаимно равными значениями по одной переменной,}$$

$$n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r, \quad \text{– суммы строк таблицы сопряженности,}$$

$$T_2 = \frac{1}{2} \sum_{j=1}^c n_{.j} (n_{.j} - 1) \quad \text{– число пар объектов с взаимно равными значениями по другой переменной,}$$

$$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c, \quad \text{– суммы столбцов таблицы сопряженности.}$$

Вычисление значимости связи основано на том факте, что статистика

$$\frac{\tau_b}{\sqrt{D\tau_b}},$$

где  $D\tau_b = (4n + 10) / (9(n^2 - n))$  – дисперсия,

асимптотически имеет стандартное нормальное распределение  $N(0,1)$ .

Вариантами рассмотренного коэффициента являются коэффициенты  $\tau_a$  и  $\tau_c$  Кендалла, которые подробно описаны в монографии Кендалла (Кендэла), посвященной ранговым корреляциям. В данной монографии приведены также точные рекуррентные формулы распределений для малых выборок.

## 6.2.11. Коэффициент Крамера

Коэффициент Крамера (мера связанности Крамера) рассчитывается по формуле

$$V = \sqrt{\frac{\chi^2}{n \min(r - 1, c - 1)}},$$

где  $\chi^2$  – статистика критерия хи-квадрат ,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl} \quad \text{– общее число наблюдений,}$$

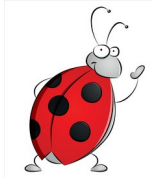
$A_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности.

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности.

Вычисление значимости связи основано на том факте, что статистика





$$\frac{V}{\sqrt{DV}},$$

$$DV = \frac{1}{n(q-1)}$$

где  $n(q-1)$  – дисперсия,

асимптотически имеет стандартное нормальное распределение  $N(0,1)$ .

См. источники: Крамер, Аптон, Кендалл с соавт.

## 6.2.12. Коэффициент Сомерса

Коэффициент Сомерса (Somers' D) является одной из разновидностей семейства мер Гудмана–Краскела. Он аналогичен коэффициенту Кендалла с той разницей, что при его вычислении производится дифференциальный учет пар с равными значениями переменных, учитывающих равенство первой и второй переменной. Коэффициент вычисляется по формулам

$$D_x = S / \left[ \frac{1}{2} n(n-1) - T_1 \right] \quad \text{– статистика «для строк»,}$$

$$D_y = S / \left[ \frac{1}{2} n(n-1) - T_2 \right] \quad \text{– статистика «для столбцов»,}$$

где  $S = P - Q$ ,

$$P = \sum_{i=1}^r \sum_{j=1}^c A_{ij} \left( \sum_{k>i} \sum_{l>j} A_{kl} \right) \quad \text{– число пар объектов с взаимно возрастающими переменными,}$$

$$Q = \sum_{i=1}^r \sum_{j=1}^c A_{ij} \left( \sum_{k>i} \sum_{l<j} A_{kl} \right) \quad \text{– число пар объектов с взаимно убывающими переменными,}$$

$A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl} \quad \text{– общее число наблюдений,}$$

$$T_1 = \frac{1}{2} \sum_{i=1}^r n_i (n_i - 1) \quad \text{– число пар объектов с взаимно равными значениями по одной переменной,}$$

$$n_i = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r, \quad \text{– суммы строк таблицы сопряженности,}$$





$$T_2 = \frac{1}{2} \sum_{j=1}^c n_j (n_j - 1)$$

– число пар объектов с взаимно равными значениями по другой переменной,

$$n_j = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$$

– суммы столбцов таблицы сопряженности.

Асимптотические распределения статистик  $D_x$  и  $D_y$  вычисляются наподобие асимптотического распределения меры  $\tau_c$  Стьюарта, приводятся в ряде источников.

### 6.2.13. Коэффициент сопряженности Пирсона

Коэффициент сопряженности Пирсона (Pearson's contingency coefficient) рассчитывается по формуле

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

где  $\chi^2$  – статистика критерия хи-квадрат,

$$n = \sum_{i=1}^r \sum_{j=1}^c A_{ij}$$

– общее число наблюдений,

где  $A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности.

Значимость статистики критерия может быть оценена, ориентируясь на значимость статистики хи-квадрат, которая подчиняется распределению  $\chi^2$  с числом степеней свободы  $(r - 1)(c - 1)$ .

### 6.2.14. Критерий Краскела–Уоллиса

Критерий Краскела–Уоллиса (ранговый однофакторный анализ Краскела–Уоллиса) является непараметрическим аналогом однофакторного дисперсионного анализа и предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Нулевая гипотеза заключается в том, что все совокупности одинаково распределены. Вычисление критерия производится по формуле

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1),$$

где  $R_i$ ,  $i = 1, 2, \dots, k$  – сумма рангов наблюдений  $i$ -ой выборки,

$$N = \sum_{i=1}^k n_i$$

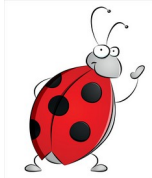
– общая численность,

$n_i$ ,  $i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$k$  – количество столбцов (групп).

Введена поправка на объединение рангов





$$b = 1 - \frac{1}{N(N^2 - 1)} \sum_{j=1}^g t_j(t_j^2 - 1),$$

где  $t_j, j = 1, 2, \dots, g$  – численность связки,  
 $g$  – число связей.

Тогда модифицированная статистика будет записана как

$$H' = H / b.$$

Статистика критерия (равно и модифицированная статистика) имеет  $\chi^2$ -распределение с параметром  $k - 1$ .

См. работы Бикела с соавт., Петровича с соавт., Холлендера с соавт. Точное вычисление критерия Краскала–Уоллиса см. в работе Клотца (Klotz) с соавт.

### 6.2.15. Диагностика Симонов–Цай

Диагностика Симонов–Цай (Simonoff–Tsai diagnostic) применяется для решения вопроса, допустима ли аппроксимация хи-квадрат в решении задачи кросстабуляции для конкретной таблицы сопряженности. Вычисление диагностики производится по формуле

$$S = \frac{(\chi^2(v, \alpha))^{1/2}}{3(X^2)^{3/2}} \sum_{i=1}^r \sum_{j=1}^c \frac{|(A_{ij} - E_{ij})|^3}{E_{ij}^2},$$

где  $\chi^2(v, \alpha)$  – значение обратной функции распределения  $\chi^2$  для  $v = (r - 1)(c - 1)$  степеней свободы и доверительного уровня  $\alpha$  (обычно берется 0,95),

$X^2$  – статистика критерия хи-квадрат,

$A_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности,

$E_{ij}, i = 1, 2, \dots, r; j = 1, 2, \dots, c$  – ожидаемые частоты таблицы сопряженности,

$r$  – число строк таблицы сопряженности,

$c$  – число столбцов таблицы сопряженности.

Ожидаемые частоты вычисляются по формуле

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \dots, r; j = 1, 2, \dots, c,$$

где  $n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r,$  – суммы строк таблицы сопряженности,

$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$  – суммы столбцов таблицы сопряженности,

$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl}$  – общее число наблюдений.

Если значение диагностики превышает значение 0,25, то это указывает на потенциальные проблемы с аппроксимацией  $\chi^2$ .



См. материалы Хромова (Khromov) с соавт., Лаззаротто (Lazzarotto) с соавт.

## 6.2.16. Диагностика Хабермана

Диагностика Хабермана (Haberma diagnostic) применяется для решения вопроса, допустима ли аппроксимация хи-квадрат в решении задачи кросстабуляции для конкретной таблицы сопряженности. Вычисление диагностики производится по формуле

$$S = \frac{1}{\sqrt{32(rc - 1)}} \sum_{i=1}^r \sum_{j=1}^c \left( \frac{1}{E_{ij}} - \frac{rc}{n} \right),$$

где  $r$  – число строк таблицы сопряженности,  
 $c$  – число столбцов таблицы сопряженности,

$$E_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \dots, r; j = 1, 2, \dots, c,$$

– ожидаемые частоты таблицы сопряженности,

$$n_{i.} = \sum_{k=1}^c A_{ik}, i = 1, 2, \dots, r,$$

– суммы строк таблицы сопряженности,

$$n_{.j} = \sum_{k=1}^r A_{kj}, j = 1, 2, \dots, c,$$

– суммы столбцов таблицы сопряженности,

$$n = \sum_{k=1}^r \sum_{l=1}^c A_{kl}$$

– общее число наблюдений,

$A_{ij}$ ,  $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, c$  – заданные частоты таблицы сопряженности.

Если значение диагностики превышает значение 0,1, то это указывает на возможные проблемы с аппроксимацией  $\chi^2$ . Значение диагностики более 1 указывает на серьезные проблемы с аппроксимацией.

См. материалы Хромова (Khromov) с соавт., Лаззаротто (Lazzarotto) с соавт.

## Глава 7. Проверка нормальности распределения

### 7.1. Введение

Проверка типа распределения эмпирической выборки, частная задача которого – проверка нормальности, имеет важнейшее значение в прикладной статистике. Перечислим некоторые из задач, которые решаются с использованием данных методов:

- Для принятия решения, применять тот или иной метод статистической обработки данных, часто необходимо установить, является ли нормальным распределение количественной эмпирической выборки.





- Важной задачей анализа согласия распределения является тестирование датчиков случайных чисел, применяемых в моделировании методом Монте–Карло в различных областях науки и техники.
- По типу статистического распределения параметров технологического процесса можно сделать определенные выводы о качестве этого процесса и вовремя скорректировать процесс.

Можно указать и другие задачи, в которых необходима проверка типа распределения. Применяются разнообразные методы, предназначенные для тестирования различных параметров распределения, в той или иной степени позволяющих исследовать его нормальность. Выводов бывает достаточно для принятия решения о выборе методов дальнейшего прикладного анализа, в частности, параметрической или непараметрической статистики.

Для проверки нормальности распределения реализации случайной одномерной величины, представленной в виде эмпирической выборки, предлагаются различные критерии. Также представлены методы проверки согласия эмпирического многомерного распределения с нормальным теоретическим многомерным распределением. При проверке согласия многомерного распределения размерность эмпирической выборки может быть произвольной. Отметим, что в данном случае размерностью выборки называют число измерений, которым представлена каждая варианта многомерной выборки. Удобна геометрическая интерпретация данного параметра. Фактически каждая варианта (элемент) такой выборки представлена точкой в многомерном пространстве, размерность которого и есть размерность выборки вариант. Размерность не следует путать с численностью выборки, представляющей собой количество вариантов.

Методы охватывают выборки практически любой численности. Однако показано (см. статью Селезнева с соавт., ссылки и другие работы), что для малых выборок (при численности выборки менее 50) и уровня значимости  $\leq 0,05$  все критерии проверки нормальности «работают» плохо вследствие малой мощности при малой численности выборки.

Дополнительно о влиянии численности на мощность критериев см. главу «Введение». Тем не менее, критерий Шапиро–Уилка показывает для таких выборок лучшие результаты, чем другие тесты.

## 7.2. Теоретическое обоснование

Если тип распределения некоторой случайной величины нам неизвестен, располагая случайной эмпирической выборкой (реализацией случайной величины), мы можем захотеть проверить, совпадает ли эмпирическая функция распределения случайной величины с некоторой заданной или вычисленной по выборочным параметрам теоретической функцией эмпирического распределения. При такой постановке говорят о проверке статистической гипотезы согласия.

Частным случаем данной задачи является установление нормальности распределения (соответствия эмпирической функции распределения непрерывной количественной





случайной величины и нормальной функции распределения). Парадоксальным эпиграфом к данной главе могли быть слова Фишера: «Отклонения от нормальной формы распределения, если только они не представляются явными без всякой оценки, могут быть обнаружены только в случае большой выборки; при малых же выборках оказывается невозможным определение сколько-нибудь надежных статистических критериев для этих отклонений». К счастью, за полвека, прошедшие со времени данной публикации, были выполнены определенные исследования.

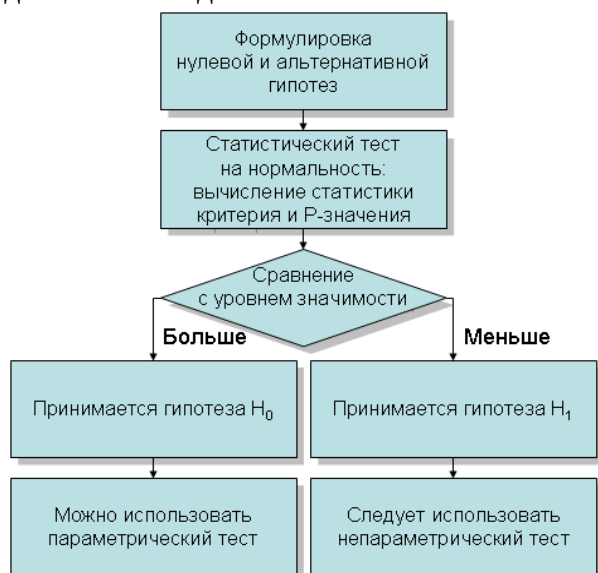
Все критерии проверки типа распределения (и, в частном случае, проверки нормальности) часто называют критериями согласия, хотя, по нашему мнению, критериями согласия справедливо называть только критерии, основанные на функциях распределения, названные так на основе термина «согласие распределений».

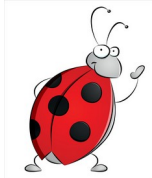
## 7.2.1. Процедура тестирования

Нормальность – основная предпосылка применения параметрических тестов. Поэтому часто исследователя интересует вопрос, соответствует ли распределение эмпирической выборки, измеренной в количественной шкале, нормальному распределению. На схеме показан алгоритм действий при проверке нормальности распределения.

Анализ выполняется стандартно. Сначала формулируют нулевую гипотезу и задаются уровнем значимости. Нулевая гипотеза  $H_0$ : «нет статистически значимого различия».

Альтернативная двусторонняя гипотеза  $H_1$ : «нулевая гипотеза не верна». С учетом, является ли гипотеза простой или сложной, проверяется согласие эмпирического распределения либо иных характеристик нормальному распределению, после чего по результатам проверки делается вывод.





## 7.2.2. Типы тестов на нормальность

Важно представлять, для какой цели производится проверка нормальности распределения. К примеру, соответствие асимметрии или эксцесса эмпирического распределения тем же параметрам нормального теоретического распределения совсем не тождественно согласию эмпирической и теоретической функций эмпирического распределения. Авторами показано, что в ряде задач достаточно проверить лишь некоторые параметры распределения.

Считается, что для выборок, немного отличающихся от нормальных, результаты применения критерия Стьюдента (см. главу «Параметрическая статистика») будут близки к верным результатам, если эксцесс и коэффициент асимметрии анализируемых выборок, как у нормальных выборок.

Более подробную информацию по данному вопросу можно найти в работе Рейнеке (Reineke) с соавт. и в статье Д'Агостино (D'Agostino) с соавт. (1990 г.). Ряд критериев предназначен для тестирования нескольких параметров одновременно. Эти критерии называют омнибусными (в отечественных источниках принято наименование – составные). Название «омнибусный» заимствовано из социологии. В социологии омнибусным исследованием принято называть исследование, проводимое одновременно для нескольких клиентов и по нескольким темам. Такая организация исследования дает возможность каждому из клиентов за меньшие деньги и в более короткий срок получить оперативную информацию по интересующим вопросам, что позволяет снизить затраты на проведение самостоятельных исследований в несколько раз. Отметим, что омнибусные критерии относятся не к отдельному типу исследования согласия распределений, а к способу организации такого исследования. Поэтому омнибусные критерии могут иметь место не только в категории «Критерии моментов», но и в других категориях.

Проверка нормальности распределения может быть выполнена с помощью специальных статистических критериев, в зависимости от анализируемых характеристик эмпирической выборки. Современными авторами выделяются критерии следующих типов:

- критерии функций распределения,
- критерии, основанные на регрессии,
- критерии моментов, включая составные тесты,
- информационные критерии,
- графические методы.

Сводка основных идей проверки типа распределения (в т. ч. различные подходы к проверке нормальности) представлена Кобзарем. Подробный обзор типов критериев дан в диссертации Ли (Lee). В специальной литературе предложены и другие идеи по поводу проверки нормальности статистического распределения. См. работы Деклерка (Declercq) и Дюво (Duvaut), Лианг (Liang) и Бентлера (Bentler).

### 7.2.2.1. Простые и сложные гипотезы

При проверке согласия эмпирического и некоторого теоретического распределения различают простые и сложные гипотезы:





- простой гипотеза будет в том случае, если теоретическое распределение задано всеми своими параметрами;
- сложной гипотеза будет, если все или некоторые параметры теоретического распределения неизвестны и оцениваются по выборке.

Иначе, если распределение имеет  $l$  параметров и гипотеза утверждает, что  $k$  из них имеют заданные значения, то гипотеза будет:

- простой, если  $k = l$ ,
- сложной, если  $k < l$ .

Разность  $l - k$  называется числом степеней свободы гипотезы, а  $k$  будет числом ограничений, наложенных гипотезой.

В случае нормального распределения по выборке могут оцениваться математическое ожидание (его оценка – среднее значение) и дисперсия (для других типов распределения число оцениваемых по эмпирической выборке параметров может быть другим). Поэтому для нормального распределения сложная гипотеза может быть трех видов:

- по выборке оценивается математическое ожидание (его оценка – среднее значение), дисперсия задана,
- по выборке оценивается дисперсия, математическое ожидание задано,
- по выборке оцениваются и математическое ожидание, и дисперсия.

Хотя статистика критерия вычисляется во всех случаях по одним и тем же алгоритмам, необходимо наличие статистических таблиц или, лучше, формул вычисления критических значений либо  $P$ -значений, особых как для каждого типа распределения, так и для каждого случая сложной гипотезы. В литературе опубликованы формулы или таблицы для многих критериев и для различных гипотез.

Некоторые критерии согласия изначально, по замыслу алгоритмов своего вычисления, представленному их авторами, не предполагают различие простой и сложной гипотез. Все параметры оцениваются по эмпирической выборке, поэтому данные критерии предназначены только для сложных гипотез.

### 7.2.3. Критерии функций распределения

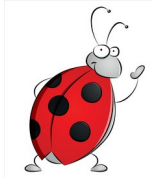
Критерии, построенные на основе функций распределения, в зависимости от метрики подразделяются на следующие типы:

- критерии типа Колмогорова,
- критерии типа омега-квадрат,
- критерии типа Эппса–Палли.

Данные критерии являются эффективными методами проверки согласия распределений.

Рассматриваются критерии нормальности, построенные на непосредственном сравнении эмпирической и теоретической функций эмпирического распределения и различающиеся метриками.

Эмпирической функцией распределения (empirical distribution function, EDF) называют такую функцию  $F_n(x)$  от вариант упорядоченной в порядке возрастания выборки, что



$$F_n(x_i) = \frac{i}{n}, i = 1, \dots, n,$$

где  $x_i, i = 1, 2, \dots, n$  – варианты упорядоченной выборки,  
 $n$  – численность выборки.

Эмпирическая функция распределения от каждой варианты выборки показывает, сколько вариантов выборки меньше данной варианты. График функции  $F_n(x)$  является равномерным по оси ординат ступенчатым графиком с шагом ступеньки, в точности равным  $i/n$ . Весь график заключен в полосу, ограниченной сверху и снизу ординатами с численными значениями 0 и 1. По оси абсцисс график в общем случае равномерным не является.

Как показывает приведенная выше формула, эмпирическая функция распределения может строиться непосредственно по заданной эмпирической выборке, минуя какие-либо промежуточные вычисления.

В случае простой гипотезы теоретическая функция распределения полностью определена заданными параметрами. В случае сложной гипотезы, когда все параметры распределения оцениваются по выборке, теоретическая функция – это функция нормального распределения, определенная параметрами, вычисленными по эмпирической выборке.

Эмпирическая характеристическая функция (empirical characteristic function, ECF) распределения имеет вид

$$\psi_n(t) = \frac{1}{n} \sum_{j=1}^n e^{it(X_j - \bar{X})S^{-1}},$$

$i$  – мнимая единица,

$t$  – нормированное отклонение,

$n$  – численность выборки,

$X_j, j = 1, 2, \dots, n$  – исходная выборка, в общем случае многомерная,

$\bar{X}$  – вектор средних значений,

$S$  – матрица дисперсий–ковариаций.

Сравнительный обзор критериев для различных гипотез дал Стефенс (Stephens, 1974).  
Методика моделирования методом Монте–Карло представлена Стефенсом (1970).

### 7.2.3.1. Критерии типа Колмогорова

Представлены следующие критерии рассматриваемого типа:

- Критерий Колмогорова (классический, для простой гипотезы).
- Модифицированный критерий Колмогорова.
- Модифицированный критерий Смирнова.

Критерии типа Колмогорова предназначены для проверки согласия эмпирической и теоретической функций распределения и построены на модульной метрике. Статистика задается формулой



$$D_n = \sup_{|x| < \infty} |F_n(x) - F(x)|,$$

где  $F_n(\cdot)$  – эмпирическая функция распределения, построенная тем или иным способом по исходной эмпирической выборке,  
 $F(\cdot)$  – теоретическая функция распределения.

Модифицированный критерий Колмогорова известен также под наименованием точного критерия Дарбина (Durbin's exact test). См. также работу Дайера (Dyer). Интересным вариантом рассматриваемого теста можно считать критерий, предложенный Ляо (Liao) и Шимокава (Shimokawa) и описанный в аналитическом обзоре Хассана (Hassan), изученный для некоторых специальных типов распределений.

### 7.2.3.1. Критерий Колмогорова

Статистика критерия Колмогорова представляет собой результат сравнения эмпирической и заданной теоретической функций распределения в модульной метрике

$$D_n = \sup_{|x| < \infty} |F_n(x) - F(x)|,$$

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|,$$

что эквивалентно

где  $x$  – случайная величина,

$F_n(\cdot)$  – эмпирическая функция распределения.

$F(\cdot)$  – теоретическая функция распределения.

Предполагается, что теоретическая функция распределения полностью задана своими параметрами. Иначе, рассматривается простая гипотеза. Это означает, что параметры распределения не могут быть вычислены по эмпирической выборке.

Статистика критерия Колмогорова обладает тем интересным свойством, что для любой

непрерывной теоретической функции распределения распределение статистики  $\sqrt{n}D_n$  при  $n \rightarrow \infty$  подчиняется  $\lambda$ -распределению (распределению Колмогорова):

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n > x) = K(x),$$

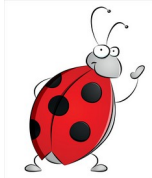
где  $K(x)$  – функция распределения Колмогорова.

Критерий Колмогорова представлен в оригинальной работе 1933 г. См. также статью Каца (Kac). Пример ошибочного вычисления см. в монографии Руниона. Для понимания принципа расчета может помочь графическая интерпретация в статьях Мэйджа (Mage), Аймэна (Iman), работах Шора с соавт., Мюллера с соавт.

### 7.2.3.1.2. Модифицированный критерий Колмогорова

Практическое вычисление статистики модифицированного критерия Колмогорова (критерия типа Колмогорова для сложной гипотезы) производится по формуле





$$D_n = \max(D_n^+, D_n^-),$$

где

$$D_n^+ = \max_{1 \leq m \leq n} \left( \frac{m}{n} - F(\eta_m) \right),$$

$$D_n^- = \max_{1 \leq m \leq n} \left( F(\eta_m) - \frac{m-1}{n} \right),$$

$\eta_m, m = 1, 2, \dots, n$  – эмпирическая выборка, отсортированная в порядке возрастания значений вариантов,

$n$  – численность выборки,

$F(\cdot)$  – теоретическая функция распределения.

Все или некоторые параметры теоретической функции непрерывного распределения (в данном случае функции нормального распределения) для случая сложной гипотезы оцениваются по эмпирической выборке.

В Рекомендациях<sup>4</sup> предложено вычислять модифицированную статистику

$$S_k = \frac{6nD_n + 1}{6\sqrt{n}},$$

хотя можно использовать значение статистики  $D_n$ .

Распределение статистики критерия не обладает свойством независимости от типа распределения, характерным для критерия Колмогорова. Поэтому для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В упомянутых Рекомендациях рассматриваются различные варианты критерия.  $P$ -значения статистики  $S_k$  для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке, могут быть аппроксимированы обобщенной функцией гамма-распределения с параметрами (4,9014; 0,0691; 0,2951).

Пример у Шора с соавт. Иные применяемые аппроксимации описаны в монографиях Тюрина и Тюрина с соавт., статье Лиллиефорса (в зарубежных источниках представленный критерий может называться Kolmogorov–Smirnov test with Lilliefors critical values или Kolmogorov–Smirnov test with Lilliefors correction) и других источниках.

### 7.2.3.1.3. Модифицированный критерий Смирнова

Вычисление статистики критерия типа Смирнова (модифицированного критерия Смирнова) производится по формуле

$$D_n^+ = \max_{1 \leq m \leq n} \left( \frac{m}{n} - F(\eta_m) \right),$$

<sup>4</sup> Рекомендации по стандартизации Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. – М.: Издательство стандартов, 2002.



где  $\eta_m$ ,  $m = 1, 2, \dots, n$  – эмпирическая выборка, отсортированная в порядке возрастания значений вариантов,

$n$  – численность выборки,

$F(\cdot)$  – теоретическая функция распределения.

В Рекомендациях<sup>5</sup> предложено вычислять модифицированную статистику

$$S_M = \frac{(6nD_n^+ + 1)^2}{9n}.$$

хотя можно использовать и статистику  $D_n^+$ .

Для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В упомянутых Рекомендациях рассматриваются различные варианты критерия. В данных Рекомендациях установлено, что  $P$ -значения статистики  $S_M$  для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке, могут быть аппроксимированы функцией логнормального распределения с параметрами (0,1164; 0,5436).

О методе см. справочник Большева с соавт., также Руководство по пакету прикладных программ SSJ (Stochastic Simulation in Java), составленному Лекюйе (L'Ecuyer), и указанные в нем источники, в том числе относительно точного вычисления распределения статистики Смирнова.

### 7.2.3.2. Критерии типа омега–квадрат

Представлены следующие критерии рассматриваемого типа:

- Критерий Крамера–Мизеса.
- Критерий Андерсона–Дарлинга.
- Критерий хи–квадрат Фишера.

Критерии типа омега–квадрат основаны на идее сравнения эмпирической и теоретической функций распределения в квадратичной метрике

$$\omega^2 = \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \psi[F(x)] dF(x),$$

где  $F_n(\cdot)$  – эмпирическая функция распределения, построенная тем или иным способом по исходной эмпирической выборке,

$F(\cdot)$  – теоретическая функция распределения,

$\psi[\cdot]$  – некоторая весовая функция.

Таблицы для определения критических значений критериев будут различаться для простой гипотезы и для каждого случая сложной гипотезы при оценке согласия эмпирического распределения с конкретным типом теоретического распределения.

<sup>5</sup> Рекомендации по стандартизации Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. – М.: Издательство стандартов, 2002.





Применение критериев типа омега–квадрат для проверки согласия различных распределений исследовано Г.В. Мартыновым. См. также Рекомендации<sup>6</sup>.

### 7.2.3.2.1. Критерий Крамера–Мизеса

При выборе весовой функции в критерии типа омега–квадрат в виде  $\psi(t) = 1$  получается критерий Крамера–Мизеса (Мизеса, Крамера–Фон Мизеса, Крамера–Мизеса–Смирнова и др.). Как и в алгоритме вычисления критерия Колмогорова, функция распределения может строиться непосредственно по эмпирической выборке, без разнесения вариантов по классам, поэтому практическое вычисление статистики критерия Крамера–Мизеса удобно производить по формуле

$$S_{\omega} = n\omega^2 = \frac{1}{12n} + \sum_{j=1}^n \left[ F(\eta_j) - \frac{2j-1}{2n} \right]^2,$$

где  $\eta_m$ ,  $m = 1, 2, \dots, n$  – эмпирическая выборка, отсортированная в порядке возрастания значений вариантов,

$n$  – численность выборки,

$F(\cdot)$  – теоретическая функция распределения.

Для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В Рекомендациях по стандартизации Р.50.1.037–2002 рассматриваются различные варианты критерия. Там же установлено, что  $P$ –значения критерия для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке, могут быть аппроксимированы функцией логнормального распределения с параметрами  $(-2,9794; 0,5330)$ .

Подробное исследование критерия см. в монографии Мартынова. Близок к рассматриваемому тесту критерий  $U^2$  Уотсона (Watson), описанный в ряде зарубежных источников.

### 7.2.3.2.2. Критерий Андерсона–Дарлингга

При выборе весовой функции в критерии типа омега–квадрат в виде

$$\psi(t) = \frac{1}{t(1-t)}$$

получается критерий Андерсона–Дарлингга ( $A^2$  критерий Андерсона–Дарлингга).

Практическое вычисление статистики критерия производится по формуле

<sup>6</sup> Рекомендации по стандартизации Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. – М.: Издательство стандартов, 2002.





$$A^2 = n\Omega^2 = -n - 2 \sum_{j=1}^n \left\{ \frac{2j-1}{2n} \ln F(\eta_j) + \left( 1 - \frac{2j-1}{2n} \right) \ln [1 - F(\eta_j)] \right\},$$

где  $\eta_m$ ,  $m = 1, 2, \dots, n$  – эмпирическая выборка, отсортированная в порядке возрастания значений вариантов,

$n$  – численность выборки,

$F(\cdot)$  – теоретическая функция распределения.

Для каждого тестируемого теоретического распределения и каждого случая сложной гипотезы распределение статистики критерия будет отличаться. В Рекомендациях по стандартизации Р.50.1.037–2002 рассматриваются различные варианты критерия. Там же установлено, что  $P$ -значения критерия для проверки нормальности в случае сложной гипотезы, когда оба параметра распределения оцениваются по эмпирической выборке, могут быть аппроксимированы функцией распределения  $S_U$  Джонсона с параметрами  $(-2,7057; 1,7154; 0,0925; 0,1043)$ . Имеется незначительное различие обозначений (следуя Хану с соавт., см. главу «Введение») в упомянутых Рекомендациях: последний и предпоследний параметры аппроксимации функцией распределения  $S_U$  Джонсона в Рекомендациях по неизвестным нам причинам поменяны местами.

Распределение статистики критерия для простой гипотезы теоретически исследовано Мартыновым. Описание дано в справочнике Степнова.

### 7.2.3.2.3. Критерий хи-квадрат Фишера

Критерий хи-квадрат Фишера (Пирсона–Фишера) является одним из старейших и самых популярных среди исследователей критериев согласия, применяемых для анализа выборок большой численности.

Критерий хи-квадрат Фишера предназначен для проверки сложных гипотез и является модификацией критерия хи-квадрат Пирсона, предназначенного для проверки простых гипотез. Вычисление статистики критерия хи-квадрат Фишера в случае проверки согласия непрерывного эмпирического распределения и непрерывного теоретического распределения производится по формуле

$$\chi^2 = \sum_{i=1}^k \frac{(v_i - nd_i p_i)^2}{nd_i p_i},$$

где  $v_i$ ,  $i = 1, 2, \dots, k$  – частоты наблюдаемых случаев в  $k$  классах,

$nd_i p_i$ ,  $i = 1, 2, \dots, k$  – соответствующие ожидаемые частоты,

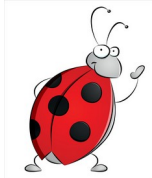
$p_i$ ,  $i = 1, 2, \dots, k$  – теоретические вероятности, вычисленные по формуле плотности распределения (в данном частном случае – нормального),

$k$  – число классов распределения,

$n$  – общее число наблюдений, вычисляемое по формуле

$$n = \sum_{i=1}^k v_i,$$





$d_i, i = 1, 2, \dots, k$  – величина классового интервала (разность соседних значений интервала); умножение на данную величину необходимо для непрерывных распределений, к которым принадлежит распределение нормальное.

При появлении интервалов с ожидаемыми частотами менее 5, по условным предпосылкам применения алгоритма, их рекомендуется объединять с соседними интервалами. Величины классовых интервалов при этом подлежат пересчету. Афифи с соавт. указывают, что некоторые ожидаемые частоты могут быть  $\geq 2$  (часто они располагаются на концах интервала), но при этом остальные обязательно должны быть  $\geq 5$ .

Статистика критерия хи-квадрат Фишера распределена как  $\chi^2$  с числом степеней свободы  $k - s - 1$ , где  $s$  – число оцениваемых параметров распределения. В рассматриваемом случае при проверке нормальности распределения, когда по выборке оцениваются среднее значение и дисперсия,  $s = 0$ , а число степеней свободы будет  $k - 3$ . Здесь нужно отметить, что параметры нормального распределения для расчета теоретических вероятностей, используемых при расчете статистики рассматриваемого критерия, должны быть вычислены по эмпирическим частотам, а не по исходным выборкам. Поэтому для вычислений данных выборочных показателей используются формулы для среднего значения и дисперсии (смещенная оценка), соответственно, в следующей форме:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k b_i \nu_i,$$

$$\sigma^2 = \frac{1}{n} \left[ \sum_{i=1}^k b_i^2 \nu_i - \frac{1}{n} \left( \sum_{i=1}^k b_i \nu_i \right)^2 \right],$$

где  $b_i, i = 1, 2, \dots, k$  – середины классовых интервалов.

Несколько нерешенных проблем, свойственных рассматриваемому критерию:

- Число классовых интервалов вычисляется по правилу Стержесса (см. главу «Описательная статистика»). От выбора числа классовых интервалов существенно зависит результат анализа рассматриваемым критерием, но нельзя сказать, что проблема выбора оптимального числа классов решена. По этой причине многие исследователи полагают, что использовать критерии типа хи-квадрат для обработки количественных данных нецелесообразно.
- На другую проблему указал проф. Воинов (цитируется по личной переписке): «... параметры должны быть оценены по эмпирическим частотам, а не по исходной выборке. Это условие необходимо, но не достаточно!!! Достаточным условием того, что критерий будет в пределе хи-квадрат с  $k - s - 1$  степенью свободы и не зависеть от параметров, является то, что предельная ковариационная матрица стандартизованных частот будет такая же, как и в случае оценок, полученных по методу минимума хи-квадрат. Я не уверен, что это условие выполняется для выборочных среднего и дисперсии по группированным данным ...». Данное утверждение может быть проверено с помощью методов, представленных в главах «Параметрическая статистика», «Непараметрическая статистика» и «Дисперсионный анализ».



- В руководствах по прикладной статистике указывается, что числа классов должно быть достаточно для верной передачи характеристик эмпирической функции распределения. При этом рекомендаций о проверке данного утверждения не приводится. Оно может быть проверено с помощью методов, представленных в главе «Непараметрическая статистика».

Выдача результатов включает дополнительные параметры:

- число классов,
- классовый интервал,
- середины классовых интервалов,
- численности классов,
- теоретические частоты.

Критерий представлен в книге Тюрина с соавт., работах Лемешко, Кобзаря, Рекомендациях<sup>7</sup>. Критерий  $J$  Ястремского, основанный на хи-квадрат, статистика которого имеет нормальное распределение, описывает Лакин. Обзор методов выбора числа классов дан в книге Новицкого с соавт. См. также работу Карлис (Karlis) с соавт. Вклад в развитие теории критериев типа хи-квадрат внесли Никулин, Мирвалиев, Воинов, Пя. Из важнейших результатов данных авторов нужно отметить группу критериев типа хи-квадрат, свободных от метода разбиения на классовые интервалы и от способа оценки неизвестных параметров распределения.

### 7.2.3.3. Критерии типа Эппса–Палли

В разделе рассмотрены:

- Критерий Эппса–Палли.
- Критерий Хенце–Цирклера.

Критерии типа Эппса–Палли (Epps–Pulley test) основаны на измерении расстояния эмпирической характеристической функции и модельной (теоретической) функции распределения

$$T_n = n \int_{-\infty}^{\infty} \left| \psi_n(t) - e^{-t^2/2} \right|^2 \varphi(t) dt,$$

где  $\psi_n(t)$  – эмпирическая характеристическая функция,

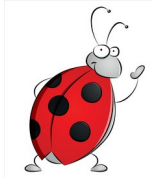
$t$  – нормированное отклонение,

$n$  – численность выборки,

$|\cdot|$  означает модуль комплексного выражения.

Обзор критериев рассматриваемого типа, включая аппроксимации и результаты компьютерного моделирования, представлен Эппсом (Epps).

<sup>7</sup> Рекомендации по стандартизации Р 50.1.033–2001. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа хи-квадрат. – М.: Издательство стандартов, 2002.



### 7.2.3.3.1. Критерий Эппса–Палли

Представив эмпирическую характеристическую функцию (обозначения выбраны так, чтобы они совпадали с аналогичными обозначениями критерия Хенце–Цирклера)

$$\psi_n(t) = \frac{1}{n} \sum_{j=1}^n e^{it(X_j - \bar{X})/S},$$

где  $i$  – мнимая единица,

$S$  – дисперсия,

$\bar{X}$  – среднее значение выборки  $X_j, j = 1, 2, \dots, n$ ,

в тригонометрической форме и взяв выражение  $\varphi(t)$  в виде плотности стандартного нормального распределения, несложно получить удобную формулу для вычисления статистики критерия Эппса–Палли

$$T_n = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n} \sum_{j=2}^n \sum_{k=1}^{j-1} e^{-\frac{1}{2}(X_j - X_k)^2/S^2} - \sqrt{2} \sum_{j=1}^n e^{-\frac{1}{4}(X_j - \bar{X})^2/S^2}.$$

Согласно Хенце (Henze),  $P$ -значение для малых выборок берется по таблице, а для выборок численностью от 10 и выше вычисляется по формуле

$$P = \Phi(z),$$

где  $\Phi(z)$  – функция стандартного нормального распределения.

Величина  $z = z(T_n^*)$  рассчитывается как

$$z = \gamma + \delta \log((T_n^* - \xi)/(\xi + \lambda - T_n^*)),$$

где  $T_n^* = (T_n - 0,365/n + 1,34/n^2)(1 + 1,3/n)$ ,

а греческими буквами обозначены константы.

Минимальная численность выборки, анализируемой критерием Эппса–Палли, равна 4.

Максимальная численность равна 200.

См. статьи Эппса, Рекомендации<sup>8</sup>, статью Хенце (Henze). Многомерный аналог критерия Эппса–Палли представлен критерием Хенце–Цирклера.

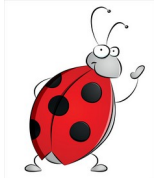
### 7.2.3.3.2. Критерий Хенце–Цирклера

Существует аналог критерия Эппса–Палли, предназначенный для проверки нормальности многомерного распределения. Вычисление критерия Хенце–Цирклера (инвариантного теста Хенце–Цирклера, Henze–Zirkler test) производится по формуле

$$D_{n,\beta} = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n e^{-\frac{\beta^2}{2}|y_j - y_k|^2} - 2(1 + \beta^2)^{-d/2} \frac{1}{n} \sum_{j=1}^n e^{-\frac{\beta^2}{2(1+\beta^2)}|y_j|^2} + (1 + 2\beta^2)^{d/2},$$

где  $\beta$  – вычисляемый особым образом или задаваемый параметр,

<sup>8</sup> Рекомендации по стандартизации Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. – М.: Издательство стандартов, 2002.



$d$  – размерность многомерной ( $d$ -мерной) выборки  $X_j, j = 1, 2, \dots, n$ ,

$n$  – число вариантов  $d$ -мерной выборки.

Остальные входящие в формулу параметры вычисляются как

$$|Y_j - Y_k|^2 = (X_j - X_k)' S^{-1} (X_j - X_k),$$

$$|Y_j|^2 = (X_j - \bar{X})' S^{-1} (X_j - \bar{X}),$$

где  $S^{-1}$  – матрица, обратная дисперсионно-ковариационной матрице,

$\bar{X}$  –  $d$ -мерный вектор среднего значения, вычисленный по  $d$ -мерной выборке,

штрих означает операцию транспонирования.

$P$ -значения критерия вычислены путем нормальной аппроксимации.

См. работу Свантессон (Svantesson) с соавт.

## 7.2.4. Критерии, основанные на регрессии

К тестам, основанным на регрессии и корреляции (иногда их называют критериями, основанными на регрессии порядковых статистик), относятся группа критериев типа Шапиро–Уилка и  $D$  критерий Д’Агостино.

Представляют интерес исследования критериев типа Шапиро–Уилка, выполненные Райан (Ryan) и Джойнером (Joiner), Чен (Chen) и Шапиро. Обзор см. в статьях Баи (Bai) с соавт., Веррилл (Verrill) с соавт.

### 7.2.4.1. Критерий Шапиро–Уилка

В ряде опытов, особенно в экспериментальных и клинических биомедицинских исследованиях, часто возникает ситуация, когда численность выборки мала. Специально для проверки нормальности распределения малых, численностью от 3 до 50 вариантов, выборок Шапиро (Shapiro) и Уилк (Wilk) разработали критерий. На основе формул оригинальной статьи критерий в принципе можно применять для любых по численности выборок, однако авторы табулировали константы, необходимые для вычисления статистики критерия и аппроксимации  $P$ -значения, только до 50 вариантов.

Статистика критерия имеет вид

$$W = \frac{\left( \sum_{i=1}^n a_i x_i \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

где  $x_i, i = 1, 2, \dots, n$  – отсортированная в порядке возрастания выборка,

$n$  – численность выборки,

$a_i, i = 1, 2, \dots, n$  – константы.

В матричной форме формула вычисления констант имеет вид

$$a = (m' V^{-1} V^{-1} m)^{-1/2} m' V^{-1},$$





где  $m$  и  $V$  – соответственно, вектор математических ожиданий и дисперсионно–ковариационная матрица массива упорядоченных сгенерированных выборок численностью  $n$ , распределенных по стандартному нормальному закону. Вычисление данных величин сопряжено с большими вычислительными сложностями, вызванными требованиями к объему (используется от 2000 до 8000 выборок, и, если математические ожидания можно просто накапливать, для получения дисперсионно–ковариационной матрицы все выборки необходимо хранить) и адресации памяти, быстродействию. Методика вычислений также приводится в более поздних публикациях Ройстона (J.P. Royston) и Ройстона (P. Royston). Поэтому практически вычисление статистики оригинального критерия производится по формуле, пригодной для быстрых вычислений,

$$W = \frac{\left( \sum_{i=1}^k a_{n-i+1} (x_{n-i+1} - x_i) \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

где  $k = n / 2$ , если  $n$  – четное,

$k = (n - 1) / 2$ , если  $n$  – нечетное,

$a_{n-i+1}; i = 1, 2, \dots, k; n = 3, 4, \dots, 50$ , – табулированные константы.

Для вычисления  $P$ –значений критерия применяется нормальная аппроксимация. Величина

$$Z = y_n + \eta_n \ln \frac{W - \varepsilon_n}{1 - W},$$

где  $y_n, \eta_n, \varepsilon_n$  – табулированные константы для соответствующих значений  $n$ , распределена нормально как  $N(0, 1)$ .

Другие аппроксимации, действительные для численности выборок до 5000, получены в работе Ройстона (P. Royston, 1993). Критерий реализован на основе монографии Хана с соавт. (Hahn et al., имеется русский перевод). См. также справочник Степнова. Ройстон (J.P. Royston) в 1983 году представил критерий  $H$  – многомерный аналог критерия Шапиро–Уилка. О критерии  $H$  Ройстона см. также работу Свантессон (Svantesson) с соавт. Очень простое многомерное обобщение критерия Шапиро–Уилка под наименованием маргинального алгоритма (marginals algorithm) предложили Петерсон (Peterson) с соавт.

## 7.2.4.2. Критерий Шапиро–Франсиа

Шапиро (Shapiro) и Франсиа (Francis) предположили, что для больших выборок статистика критерия  $W$  может быть вычислена менее трудоемко, чем это сделано в критерии Шапиро–Уилка. Она имеет другое обозначение, но похожую запись

$$W' = \frac{\left( \sum_{i=1}^n b_i x_i \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$



где  $x_i, i = 1, 2, \dots, n$  – отсортированная в порядке возрастания выборка,

$n$  – численность выборки,

$b_i, i = 1, 2, \dots, n$  – константы.

В матричной форме формула вычисления констант имеет совсем простой вид

$$b = (m'm)^{-1/2}m,$$

где  $m$  – вектор математических ожиданий, вычисленный на основе упорядоченных сгенерированных выборок численностью  $n$ , распределенных по стандартному нормальному закону. Определение данной величины сопряжено с большими вычислительными сложностями, вызванными требованиями к быстродействию компьютера. Поэтому авторы теста воспользовались тем, что ранее Блом (Blom, см. Дэйвида) записал простую в вычислении оценку компонент вектора математических ожиданий

$$\tilde{m}_i = \Psi\left[(i - 3/8)/(n + 1/4)\right], i = 1, 2, \dots, n,$$

где  $\Psi(\cdot)$  – функция, обратная функции стандартного нормального распределения.

Статистика критерия не относится к какому-либо стандартному типу распределения,

поэтому Ройстон (J.P. Royston, 1983) для практических вычислений предложил ее

трансформацию с последующей аппроксимацией по стандартному нормальному закону.

Другие аппроксимации, также действительные для численности выборок до 5000, даны в работе Ройстона (P. Royston, 1993).

### 7.2.4.3. Критерий Д'Агостино

$D$  критерий Д'Агостино (D'Agostino's  $D$  test) построен, как и критерий Шапиро–Уилка, на порядковых статистиках. Вычисление статистики критерия производится по формуле

$$D = \frac{\sum_{i=1}^n \left( i - \frac{n+1}{2} \right) x_i}{n^2 s},$$

где  $x_i, i = 1, 2, \dots, n$  – отсортированная в порядке возрастания выборка,

$n$  – численность выборки,

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

– смещенная оценка дисперсии,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

где – выборочное среднее значение.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{D - ED}{\sqrt{DD}},$$

$$ED = \frac{(n-1)}{2\sqrt{(2n\pi)}} \frac{\Gamma(n/2 - 1/2)}{\Gamma(n/2)} \approx (2\sqrt{\pi})^{-1} \approx 0,28209479$$

где – математическое ожидание,







$$\sqrt{DD} = \left( \frac{12\sqrt{3} - 27 + 2\pi}{24n\pi} \right)^{1/2} \approx 0,02998598/\sqrt{n}$$

– стандартное отклонение.

распределена по стандартному нормальному закону. По этой причине  $D$  критерий Д’Агостино полагается более удобным в вычислении, чем критерий Шапиро–Уилка, требующий для своего вычисления либо таблиц, либо довольно сложных трудоемких аппроксимаций, связанных с объемными вычислениями.

Формулы взяты из источников: Д’Агостино (D’Agostino, 1971), Донг (Dong) с соавт. и Уайт (White) с соавт. Во втором источнике асимптотические формулы для математического ожидания и стандартного отклонения записаны неправильно, причем опечатка в формуле для стандартного отклонения идет из оригинальной работы.

## 7.2.5. Критерии моментов

Существует группа критериев, которые позволяют оценить отклонение некоторых параметров эмпирического распределения (коэффициент асимметрии, эксцесс или оба параметра одновременно) от тех же параметров нормального распределения. Подробнее о данных параметрах эмпирической выборки см. главу «Описательная статистика».

Рассматриваемые критерии принадлежат к группе критериев, основанных на обычных и абсолютных моментах распределения. По результатам применения данных критериев нельзя делать заключение о соответствии тестируемой выборки нормальному распределению. Данными критериями можно лишь проверить, что тестируемые параметры эмпирической выборки принимают определенные значения, соответствующие нормальному распределению.

Наиболее распространены следующие критерии, основанные на моментах распределения:

- критерий коэффициента асимметрии (третий нормированный центральный момент),
- критерий эксцесса (четвертый нормированный центральный момент),
- критерий Жарка–Бера, построенный на идее одновременного анализа коэффициента асимметрии и эксцесса,
- критерий Гири (первый нормированный центральный абсолютный момент),
- многомерный критерий асимметрии Мардиа,
- многомерный критерий эксцесса Мардиа.

Напомним, что центральные выборочные моменты определяются формулами

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, k = 1, 2, \dots,$$

где  $x_i, i = 1, 2, \dots, n$  – эмпирическая выборка,

$n$  – численность выборки,

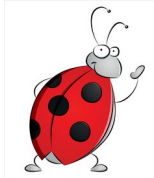
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

– выборочное среднее значение,

$k$  – порядок момента.







Центральные абсолютные выборочные моменты определяются формулами

$$\beta_k = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|^k, k = 1, 2, \dots$$

Абсолютные моменты четных порядков совпадают с обычными моментами. Первый центральный абсолютный момент называется средним арифметическим отклонением.

Наряду со средним квадратическим отклонением, данный показатель может применяться в качестве характеристики рассеяния.

Хорошо проработаны многомерные аналоги критериев коэффициента асимметрии и эксцесса – критерии Мардиа. Многомерный критерий Мардиа–Фостера идейно близок к составным тестам типа Жарка–Бера (одновременно тестируются асимметрия и эксцесс).

Применение критерия коэффициента асимметрии и критерия эксцесса рекомендуется для проверки отклонения от нормальности, например, при решении вопроса о применении критерия Стьюдента, представленного в главе «Параметрическая статистика».

## 7.2.5.1. Критерий коэффициента асимметрии

Коэффициент асимметрии (skewness) характеризует несимметричность распределения случайной величины. Для нормального распределения коэффициент асимметрии равен нулю.

Коэффициент асимметрии – величина, не зависящая от выбора начала отсчета и от единиц измерения случайной величины. Выборочный коэффициент асимметрии (sample skewness) может вычисляться по формуле выборочных моментных отношений

$$b_1 = \frac{m_3}{S^3},$$

где  $S^2$  – оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где  $x_i, i = 1, 2, \dots, n$  – варианты эмпирической выборки,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

– выборочное среднее значение.

$n$  – численность выборки,

$$m_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n (x_i - \bar{x})^3$$

– выборочная оценка 3-го центрального момента.

Запишем модифицированную статистику

$$B_1 = \frac{\sqrt{n(n-1)}}{n-2} b_1.$$

Тогда статистика  $B_1$  для большой численности выборки распределена асимптотически нормально с нулевым средним и дисперсией



$$DB_1 = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}.$$

Минимальная численность выборки, анализируемой критерием коэффициента асимметрии, равна 3.

Описание см. у Крамера, Ван дер Вардена. В литературе имеются и другие формы записи статистики критерия, а также ее аппроксимации. См. Большева с соавт., Степнова, Стенгоса (Stengos) с соавт. Исследователями предложены следующие варианты критерия коэффициента асимметрии:

- критерий асимметрии Д'Агостино (D'Agostino's test for skewness), подробно представленный в статье Д'Агостино с соавт. (1990),
- критерий  $g_1$  Фишера (Fisher  $g$  statistics for skewness), описанный там же.

## 7.2.5.2. Критерий эксцесса

Эксцесс (kurtosis, excess) характеризует степень выраженности хвостов распределения – частоту появления удаленных от среднего значений. Для нормального распределения эксцесс равен трем, поэтому при вычислении эксцесса от полученного значения часто отнимают число три, чтобы показать, насколько эксцесс эмпирической выборки отличается от эксцесса нормального распределения. Эксцесс – величина, не зависящая от выбора начала отсчета и от единиц измерения случайной величины. Выборочный эксцесс (sample kurtosis) может вычисляться по формуле выборочных моментных отношений

$$b_2 = \frac{m_4}{S^4},$$

где  $S^2$  – оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где  $x_i, i = 1, 2, \dots, n$  – варианты эмпирической выборки,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

– выборочное среднее значение.

$n$  – численность выборки,

$$m_4 = \frac{n(n+1) \sum_{i=1}^n (x_i - \bar{x})^4 - 3(n-1) \left\{ \sum_{i=1}^n (x_i - \bar{x})^2 \right\}^2}{(n-1)(n-2)(n-3)}$$

– выборочная оценка 4-го центрального

момента.

Запишем модифицированную статистику

$$B_2 = \frac{n-1}{(n-2)(n-3)} [(n+1)b_2 + 6].$$



Тогда статистика  $B_2$  для большой численности выборки распределена асимптотически нормально с нулевым средним и дисперсией

$$DB_2 = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}.$$

Минимальная численность выборки, анализируемой критерием эксцесса, равна 4. Описание см. у Крамера, Ван дер Вардена. В литературе имеются и другие аппроксимации статистики критерия. См. Большева с соавт., Степнова.

### 7.2.5.3. Критерий Жарка–Бера

Известным представителем составных тестов является широко применяемый (и широко критикуемый) критерий Жарка–Бера (Jarque–Bera test, он же Bowman–Shenton  $K^2$  test). С помощью данного критерия производится одновременный анализ коэффициента асимметрии и эксцесса. Статистика критерия вычисляется по формуле

$$J = n \left( \frac{b_1^2}{6} + \frac{(b_2 - 3)^2}{24} \right),$$

где  $b_1$  – коэффициент асимметрии,

$b_2$  – эксцесс,

$n$  – численность выборки.

В соответствии с требованиями алгоритма, коэффициент асимметрии вычисляется по формуле

$$b_1 = k_3 / S^3,$$

где  $S^2$  – оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где  $x_i, i = 1, 2, \dots, n$  – варианты эмпирической выборки,

$\bar{x}$  – выборочное среднее значение.

$$k_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3.$$

Эксцесс вычисляется по формуле

$$b_2 = \frac{k_4}{S^4},$$

$$k_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4.$$

где

Статистика критерия для большой численности выборки распределена асимптотически как  $\chi^2$  с параметром, равным 2.

Критерии описаны во множестве оригинальных источников. Гел и Гаствирт представили робастный вариант критерия (The Gel–Gastwirth robust Jarque–Bera test). См. также обзор



Дурник (Doornik) с соавт., Ромао (Romaо) с соавт., книгу Селезнева с соавт., справочник Степнова. Составной критерий Д'Агостино–Пирсона (D'Agostino–Pearson test) представлен в статье Д'Агостино с соавт. (1990 г.), но в настоящее время дезавуирован из-за обнаруженных теоретических проблем. ГОСТ<sup>9</sup> определяет так называемый составной критерий, представляющий собой совокупность двух тестов, одним из которых является вариант критерия Гири.

#### 7.2.5.4. Критерий Гири

Гири предложил серию критериев, построенных на соотношениях для центральных абсолютных моментов. Вместо критерия эксцесса может применяться критерий Гири (Geary's kurtosis test), построенный на соотношении первого центрального абсолютного момента:

$$d = \frac{1}{nS} \sum_{i=1}^n |x_i - \bar{x}|,$$

где  $S^2$  – смещенная оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где  $x_i, i = 1, 2, \dots, n$  – варианты эмпирической выборки,

$\bar{x}$  – выборочное среднее значение.

$n$  – численность выборки.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом статистика  $d$  распределена нормально с

математическим ожиданием  $\sqrt{2/\pi}$  и дисперсией  $(1 - 3/\pi) / n$ .

Критерий изучен Большевым с соавт., Д'Агостино (D'Agostino) и Розман (Rosman), Чо (Cho) с соавт., Уолпоул (Walpole) с соавт. Родственным описанному тесту является критерий Бонетта–Сайера (Bonett–Seier test).

#### 7.2.5.5. Критерий асимметрии Мардиа

Многомерный аналог критерия коэффициента асимметрии предложен Мардиа. Статистика критерия вычисляется по формуле

$$b_{1,d} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n [(X_i - \bar{X})' S^{-1} (X_j - \bar{X})]^3,$$

где  $d$  – размерность многомерной ( $d$ -мерной) выборки  $X_j, j = 1, 2, \dots, n$ ,

$n$  – число вариантов  $d$ -мерной выборки,

$S^{-1}$  – матрица, обратная дисперсионно-ковариационной матрице,

$\bar{X}$  –  $d$ -мерный вектор среднего значения, вычисленный по  $d$ -мерной выборке,

9 ГОСТ Р ИСО 5479–2002. Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения. – М.: Издательство стандартов, 2002.



штрих означает операцию транспонирования.

Для практического исследователя–расчетчика многомерность эмпирической выборки означает, что она представлена таблицей чисел, строки которой являются вариантами (в данном случае – векторными)  $d$ -мерной выборки, число строк равно численности выборки, а число столбцов равно размерности («числу измерений»).

Статистика  $\frac{n}{6}b_{1,d}$  распределена асимптотически как  $\chi^2$  с параметром  $d(d+1)(d+2)/6$ .

О критериях Мардиа см. оригинальные работы Mardia, а также статью и библиографию Канкайнена (Kankainen) с соавт., справочник Родионова с соавт.

## 7.2.5.6. Критерий эксцесса Мардиа

Многомерный аналог критерия эксцесса предложен Мардиа. Статистика критерия вычисляется по формуле

$$b_{2,d} = \frac{1}{n} \sum_{i=1}^n [(X_i - \bar{X})' S^{-1} (X_i - \bar{X})]^2,$$

где  $d$  – размерность многомерной ( $d$ -мерной) выборки  $X_j, j = 1, 2, \dots, n$ ,

$n$  – число вариант  $d$ -мерной выборки,

$S^{-1}$  – матрица, обратная дисперсионно–ковариационной матрице,

$\bar{X}$  –  $d$ -мерный вектор среднего значения, вычисленный по  $d$ -мерной выборке,

штрих означает операцию транспонирования.

Для исследователя многомерность эмпирической выборки означает, что она представлена таблицей чисел, строки которой являются вариантами (в данном случае – векторными)  $d$ -мерной выборки, число строк равно численности выборки, а число столбцов равно размерности («числу измерений»).

Статистика  $b_{2,d}$  распределена асимптотически нормально со средним  $d(d+2)$  и дисперсией  $8d(d+2)/n$ .

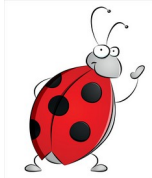
О критериях Мардиа см. оригинальные работы (Mardia), а также статью и библиографию Канкайнена (Kankainen) с соавт., справочник Родионова с соавт.

## 7.2.6. Информационные критерии

Информационные критерии согласия основаны на информационной мере – энтропии (см. «Информационный анализ»). Они основаны на том научном факте, что энтропия непрерывного распределения максимальна, если распределение нормальное.

Наиболее известным является критерий Васичека (Vasicek's test).

Помимо оригинальных работ, все данные критерии описаны в обзоре Эстебана (Esteban) с соавт.



## 7.2.6.1. Критерий Васичека

Статистика критерия Васичека (Vasicek's test) вычисляется по формуле

$$K_{mn} = \frac{n}{2mS} \left\{ \prod_{i=1}^n (x_{i+m} - x_{i-m}) \right\}^{1/n},$$

где  $x_i$ ,  $i = 1, 2, \dots, n$  – варианты упорядоченной (от меньшего значения к большему значению) эмпирической выборки, причем условились, что при индексе варианты в данной формуле  $(i + m) > n$  индекс берется  $n$ , при индексе  $(i - m) < 1$  индекс берется 1,  $m$  – ширина окна, положительное наименьшее целое значение из интервала от 1 до  $(n - 1) / 2$ ,  $n$  – численность выборки,

$S^2$  – смещенная оценка выборочной дисперсии, вычисляемая по формуле

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

где – выборочное среднее значение.

Гипотеза о нормальности распределения не отклоняется на заданном уровне значимости при выполнении условия

$$K_{mn} \geq K^*,$$

где  $K^*$  – критическое значение, взятое из таблицы, вычисленной методом компьютерного моделирования.

Таблица критических значений в оригинальном источнике отличается лаконичностью, поэтому мы используем более подробную таблицу, вычисленную Эстебаном (Esteban) с соавт. Как и в оригинальной статье Васичека, вычисления выполнены для  $n = 1, 2, \dots, 50$ , поэтому при большей численности выборки критерий не применяется. Кроме того, таблицы получены для значений  $1 \leq m \leq 9$  с учетом представленного выше правила выбора ширины окна и только для уровня значимости 0,05. См. также книги Кобзаря, Тику (Tiku) с соавт., статью Мудхолкара (Mudholkar) и Тиань (Tian). В последнем источнике разъясняется роль такого важного параметра алгоритма критерия Васичека, как ширина окна  $m$ . При конкретных альтернативных распределениях эмпирической выборки и фиксированной ее численности максимальная мощность критерия (см. главу «Введение») достигается при определенной ширине окна.

## 7.2.7. Графические методы

Простейшим из графических методов является глазомерный метод, когда визуально сравниваются график функции распределения или плотности распределения эмпирической и график наложенной на нее теоретической. В практической реализации графических методов может оказаться полезным использование инструмента «Гистограмма», представленного в главе «Описательная статистика».



О чтении гистограмм, в числе огромного числа источников, см. монографию под ред. Кумэ. Кроме того, некоторые методы имеют очевидную графическую интерпретацию. См., например, статьи Мэйджа (Mage), Аймэна (Iman).

## 7.2.7.1. Глазомерный метод

Простейшим из графических методов является так называемый глазомерный метод, когда визуально сравниваются график плотности распределения эмпирической и график наложенной на нее соответствующей теоретической функции. Сравнение производится пользователем, который играет в данном случае роль эксперта.

Результаты анализа включают параметры:

- число классов,
- номера классов,
- численности классов,
- теоретические частоты нормального распределения,
- диаграмму, на которой гистограмма представляет собой отображение эмпирического распределения, а точечная диаграмма со значениями, соединенными сглаживающими линиями, отображает теоретическое нормальное распределение.

## Глава 8. Дисперсионный анализ

### 8.1. Введение

Назначение представленных в данной главе дисперсионного анализа, множественных сравнений и ковариационного анализа подробно разъясняется в соответствующих теоретических разделах.

Методы дисперсионного анализа и множественных сравнений могут быть предназначены для нормально распределенных совокупностей (т. е. будут многомерными аналогами параметрических тестов) и для выборок, свободных от предположения о типе распределения (т. е. будут многомерными аналогами непараметрических тестов). Методы ковариационного анализа предполагают нормальность распределения ошибок (относительно линейной регрессии). Нормальность распределения произвольных по численности и «числу измерений» выборок может быть проверена с помощью методов главы «Проверка нормальности распределения».

### 8.2. Теоретическое обоснование

#### 8.2.1. Дисперсионный анализ

Дисперсионным анализом называют совокупность статистических методов, предназначенных для обработки данных экспериментов, целью которых являлось не



установление каких-то свойств и параметров, а сравнение эффектов различных воздействий на каком-либо экспериментальном материале. Методы дисперсионного анализа используются для проверки гипотез о наличии связи между результативным признаком и исследуемыми факторами, а также для установления силы влияния факторов и их взаимодействий.

Из представленных критериев одна многочисленная группа тестов является параметрическими и требуют нормальности распределения исходных выборок. Данные методы предназначены только для нормально распределенных количественных данных. Исследованию свойств некоторых параметрических методов при нарушении предположений о нормальности посвящены работы Лемешко с соавт. Проверить нормальность распределения, включая многомерный случай, можно с помощью методов главы «Проверка нормальности распределения». Другие методы являются непараметрическими и не требуют предположений относительно вида исходного распределения.

Критерий Q Кокрена предназначен для бинарных (дихотомических) данных.

Для параметрических и непараметрических методов проверки гипотез (см.

«Параметрическая статистика» и «Непараметрическая статистика») существуют многомерные аналоги в дисперсионном анализе, как показано в таблице.

Метод проверки гипотезы для двух выборок «Функциональный аналог» из дисперсионного анализа

Параметрические тесты

Критерий Стьюдента

Однофакторный дисперсионный анализ

Критерий Шеффе

Критерий Пейджа

Критерий Дункана

Критерий Тьюки

Критерий Стьюдента парный

Однофакторный дисперсионный анализ с повторными измерениями

Многофакторный дисперсионный анализ

Критерий Шеффе для связанных выборок

F-критерий

Критерий Бартлетта

Критерий G Кокрена

Критерий Ливена

Непараметрические тесты

Критерий Вилкоксона

Критерий Джонкхиера–Терпстра

Критерий Краскела и Уоллиса

Критерий Данна

Критерий Коновера

Критерий Кьюзика

Критерий Вилкоксона парный

Ранговый критерий Фридмана

Критерий Квейд

Точный метод Фишера

Критерий Q Кокрена







Критерий  $V$  Бхапкара  
Критерий  $D$  Дешпанде  
Критерий  $L$  Дешпанде  
Критерий Брауна–Форсайта

Методы дисперсионного анализа следует использовать, когда число выборок больше двух. Нельзя применять критерии, предназначенные для сравнения выборок попарно, а затем делать какие-либо выводы относительно всей совокупности.

В дисперсионном анализе, как и в других областях анализа данных, сложилась определенная терминология. Фактором называют величину, определяющую свойства исследуемого объекта или системы, иначе – причину, влияющую на конечный результат. Конкретную реализацию фактора называют уровнем фактора или способом обработки. Значение измеряемого признака называют откликом.

См. книги Браунли, Кобзаря, Холлендера с соавт., нормативный документ EPA QA/G–9.

## 8.2.1. Однофакторный дисперсионный анализ

Исходные данные для однофакторного дисперсионного анализа представлены в виде таблицы (прямоугольной матрицы), причем число столбцов (выборок) соответствует числу уровней фактора (уровней обработки), число строк равно числу наблюдений. При этом выборки могут иметь как одинаковое число вариантов (равные объемы), так и различное, в зависимости от требований применяемого метода.

Предлагаются методы однофакторного дисперсионного анализа:

- Однофакторный дисперсионный анализ (ANOVA).
- Однофакторный дисперсионный анализ с повторными измерениями.
- Ранговый однофакторный анализ Краскела–Уоллиса.
- Критерий Данна.
- Критерий Коновера.
- Критерий Джонкхиера–Терпстра.
- Критерий Бартлетта.
- $G$ –критерий Кокрена.
- Критерий Шеффе.
- Критерий Дункана.
- Критерий Тьюки.
- Критерий Ливена.
- Критерий Брауна–Форсайта.
- Критерий  $V$  Бхапкара.
- Критерий  $D$  Дешпанде.
- Критерий  $L$  Дешпанде.



## 8.2.1.1. Однофакторный дисперсионный анализ

При однофакторном дисперсионном анализе (дисперсионном анализе по одному признаку, analysis of variance, ANOVA) предполагается, что результаты наблюдений для разных уровней представляют собой выборки из нормально распределенных генеральных совокупностей. Эти совокупности имеют свои средние и дисперсии, которые полагаются одинаковыми. Задачей анализа является проверка нулевой гипотезы о равенстве средних рассматриваемых совокупностей. Вычисление критерия производится по формуле

$$t = \frac{N - k}{k - 1} \frac{\sum_{i=1}^k n_i (\bar{x}_i - \bar{x}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2},$$

где  $N = \sum_{i=1}^k n_i$  – общая численность,

$n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2, \dots, k$  – среднее значение  $i$ -й выборки,

$\bar{x}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}$  – общее среднее значение,

$x_{ij}, i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$  – варианты выборки,

$k$  – число столбцов (выборок).

Сумма, стоящая в числителе формулы вычисления критерия, служит приближенной мерой вариации между анализируемыми выборками, а двойная сумма, стоящая в знаменателе, служит мерой вариации внутри выборок.

Статистика критерия имеет  $F$ -распределение с параметрами  $k - 1$  и  $N - k$ .

См. монографию Шеффе.

## 8.2.1.1.2. Однофакторный дисперсионный анализ (повторные измерения)

При однофакторном дисперсионном анализе с повторными измерениями (repeated measurements ANOVA) предполагается, что результаты наблюдений одного и того же процесса для разных временных уровней представляют собой выборки из нормально распределенных генеральных совокупностей. Эти совокупности имеют свои средние и дисперсии, которые полагаются одинаковыми. Задачей анализа является проверка нулевой гипотезы о равенстве средних рассматриваемых совокупностей.

Вычисления производятся по формулам:





$$t = \frac{D_{col}}{D},$$

где  $D_{col} = \frac{SS_{col}}{c - 1}$  – дисперсия, объясняемая столбцами,

$$D = \frac{SS}{(r - 1)(c - 1)} \text{ – остаточная дисперсия,}$$

$$SS_{col} = r \sum_{j=1}^c (T_{.j} - T_{..})^2 \text{ – средний квадрат столбцов,}$$

$$SS = \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - T_{i.} - T_{.j} + T_{..})^2 \text{ – средний квадрат погрешности,}$$

$$T_{i.} = \frac{1}{c} \sum_{j=1}^c x_{ij}, i = 1, 2, \dots, r \text{ – средние суммы строк,}$$

$$T_{.j} = \frac{1}{r} \sum_{i=1}^r x_{ij}, j = 1, 2, \dots, c \text{ – средние суммы столбцов,}$$

$$T_{..} = \frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c x_{ij} \text{ – общее среднее,}$$

$c$  – число столбцов (выборок),

$r$  – число строк (параметров).

Статистика критерия имеет  $F$ -распределение с параметрами  $r - 1$  и  $(r - 1)(c - 1)$ .

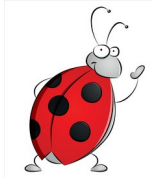
Результаты расчета совпадают с эффектом столбцов в двухфакторном дисперсионном анализе.

Описание см. в монографии Дэйвиса (Davis).

#### 8.2.1.1.4. Критерий Данна

Ранговый однофакторный анализ Краскела и Уоллиса может показать, что параметры положения совокупностей различаются. Однако данный критерий не позволяет узнать, параметры каких совокупностей действительно различаются между собой. Для решения проблемы применяется непараметрический критерий Данна (Bonferroni–Dunn post hoc test, Dunn's multiple comparison post-test). Критерий применим для независимых групп как равной, так и различной численности. Вычисление критерия производится по формуле

$$Q_{ij} = \frac{|\bar{R}_i - \bar{R}_j|}{\sqrt{\frac{N(N+1)}{12} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}, i = 1, 2, \dots, k; j = i + 1, \dots, k,$$



$$\bar{R}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} R_{il}, i=1,2,\dots,k$$

где  $\bar{R}_i$  – средний ранг  $i$ -й выборки,  
 $R_{il}, i=1,2,\dots,k$  – ранги наблюдений  $i$ -ой выборки,

$$N = \sum_{i=1}^k n_i$$

– общая численность,

$n_i, i=1,2,\dots,k$  – численность  $i$ -й выборки,

$k$  – количество столбцов (групп).

$P$ -значения критерия  $p_{ij}, i=1,2,\dots,k; j=i+1,\dots,k$ , являются решениями нелинейных уравнений

$$Q_{ij} = \Psi\left(\frac{p_{ij}}{k(k-1)}\right), i=1,2,\dots,k; j=i+1,\dots,k,$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения.

Уравнения могут быть решены одним из методов локальной оптимизации. В простейшем случае используется метод деления отрезка пополам.

Описание см. у Гланца, Даниэла (Daniel), Зигеля (Siegel) с соавт., Холлендера с соавт.

### 8.2.1.1.3. Ранговый однофакторный анализ Краскела и Уоллиса

Критерий Краскела–Уоллиса (ранговый однофакторный анализ Краскела–Уоллиса) является непараметрическим аналогом однофакторного дисперсионного анализа и предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Нулевая гипотеза заключается в том, что все совокупности одинаково распределены. Вычисление критерия производится по формуле

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1),$$

где  $R_i, i=1,2,\dots,k$  – сумма рангов наблюдений  $i$ -ой выборки,

$$N = \sum_{i=1}^k n_i$$

– общая численность,

$n_i, i=1,2,\dots,k$  – численность  $i$ -й выборки,

$k$  – количество столбцов (групп).

Поправка на объединение рангов

$$b = 1 - \frac{1}{N(N^2-1)} \sum_{j=1}^g t_j(t_j^2-1),$$

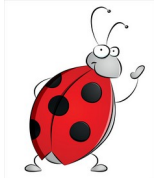
где  $t_j, j=1,2,\dots,g$  – численность связки,

$g$  – число связок.

Тогда модифицированная статистика будет записана как

$$H' = H / b.$$





Статистика критерия (равно и модифицированная статистика) имеет  $\chi^2$ -распределение с параметром  $k - 1$ .

См. книги Бикела с соавт., Петровича с соавт., Холлендера с соавт. Точное вычисление критерия см. в работе Клотца (Klotz) с соавт.

### 8.2.1.1.5. Критерий Коновера

Ранговый однофакторный анализ Краскела и Уоллиса может показать, что параметры положения совокупностей различаются. Однако данный критерий не позволяет узнать, параметры каких совокупностей действительно различаются между собой. Для решения проблемы применяется непараметрический критерий Коновера (Conover post hoc test). Критерий применим для независимых групп как равной, так и различной численности. Вычисление критерия производится по формуле

$$C_{ij} = \frac{|\bar{R}_i - \bar{R}_j|}{\sqrt{\frac{N(N+1)}{12} \cdot \frac{N-1-H}{N-k} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}, i=1,2,\dots,k; j=i+1,\dots,k,$$

$$\bar{R}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} R_{il}, i=1,2,\dots,k$$

где  $\bar{R}_i$  – средний ранг  $i$ -й выборки,

$R_{il}, i=1,2,\dots,k$  – ранги наблюдений  $i$ -ой выборки,

$$N = \sum_{i=1}^k n_i$$

– общая численность,

$n_i, i=1,2,\dots,k$  – численность  $i$ -й выборки,

$H$  – статистика критерия Краскела–Уоллиса,

$k$  – количество столбцов (групп).

$P$ –значения критерия  $p_{ij}, i=1,2,\dots,k; j=i+1,\dots,k$ , подчиняются  $t$ -распределению с параметром  $N - k$ .

Описание см. у Бортца (Bortz) с соавт.

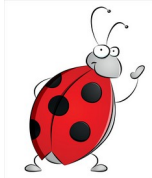
### 8.2.1.1.6. Критерий Джонкхиера и Терпстра

Критерий Джонкхиера–Терпстра (критерий Джонкхиера) представляет собой многомерное обобщение критерия Манна–Уитни (см. главу «Непараметрическая статистика») и предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Вычисление критерия производится по формуле

$$J = \sum_{i=1}^{k-1} \sum_{j=i+1}^k U_{ij},$$

где  $U_{ij}, i=1,2,\dots,k-1; j=2,3,\dots,k$  – статистика критерия Манна–Уитни для выборок с номерами  $i$  и  $j$ ,





$k$  – число столбцов (выборок).

Для больших выборок распределение преобразованной статистики

$$\frac{J - MJ}{\sqrt{DJ}}$$

является приближенно нормальным. Здесь математическое ожидание и дисперсия рассчитываются по формулам, соответственно:

$$MJ = \frac{1}{4} \left( N^2 - \sum_{i=1}^k n_i^2 \right),$$

$$DJ = \frac{1}{72} \left( N^2(2N + 3) - \sum_{i=1}^k n_i^2(2n_i + 3) \right),$$

$$N = \sum_{i=1}^k n_i$$

где  $N$  – общая численность,

$n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки.

Описание см. в книгах Тюринга с соавт., Холлендера с соавт., в работе Кьюзика (Cuzick).

### 8.2.1.1.7. Критерий Бартлетта

Критерий Бартлетта ( $M$ -критерий Бартлетта) служит для проверки нулевой гипотезы о равенстве дисперсий нормальных генеральных совокупностей. Вычисления статистики критерия производится по формуле

$$M = \left[ 1 + \frac{1}{3(k-1)} \left( \sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{N - k} \right) \right]^{-1} \left[ (N - k) \ln s^2 - \sum_{i=1}^k (n_i - 1) \ln s_i^2 \right],$$

где  $n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$$s^2 = \frac{\sum_{i=1}^k (n_i - 1) s_i^2}{N - k},$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, i = 1, 2, \dots, k$$

– выборочная дисперсия  $i$ -й выборки,

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2, \dots, k$$

– среднее значение  $i$ -й выборки,

$$N = \sum_{i=1}^k n_i$$

– суммарная численность всех выборок,

$k$  – число столбцов (выборок).

Статистика критерия имеет  $\chi^2$ -распределение с параметром  $k - 1$ .

Описание см. у Браунли, Когана с соавт.





## 8.2.1.6. Критерий G Кокрена

Критерий G Кокрена используется для проверки нулевой гипотезы о равенстве дисперсий нормальных генеральных совокупностей по независимым выборкам с одинаковыми численностями. Вычисление статистики критерия производится по формуле

$$G = \frac{\max_{1 \leq i \leq k} \sigma_i^2}{\sum_{i=1}^k \sigma_i^2},$$

где  $\sigma_i^2, i = 1, 2, \dots, k$  – выборочные дисперсии совокупностей,  
 $k$  – число выборочных совокупностей.

$P$ –значение модифицированной статистики

$$G' = \frac{G(k-1)}{1-G}$$

является решением нелинейного уравнения

$$G' = F_{(n-1)(n-1)(k-1)}^{-1} \left( \frac{p}{k} \right),$$

где  $F_{\dots}^{-1}(\cdot)$  – обратная функция  $F$ –распределения,  
 $n$  – численность каждой совокупности.

Уравнение может быть решено одним из методов локальной оптимизации. В простейшем случае используется метод деления отрезка пополам.

Описание см. в монографиях Мюллера с соавт., Налимова, Зигеля (Siegel) с соавт.

## 8.2.1.1.9. Критерий Шеффе

Однофакторный анализ может показать, что средние значения совокупностей различаются. Однако он не позволяет узнать, средние значения каких совокупностей действительно различаются между собой. Для решения проблемы применяется метод множественного сравнения Шеффе (критерий Шеффе). Критерий Шеффе предназначен для проверки так называемой гипотезы о линейном контрасте. Линейный контраст

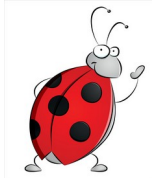
$$L = \sum_{i=1}^k c_i \mu_i$$

представляет собой линейную функцию от средних значений  $\mu_i, i = 1, 2, \dots, k, k$  независимых нормальных выборок с неизвестными, но равными дисперсиями, и известных констант  $c_i, i = 1, 2, \dots, k$ , удовлетворяющих условию

$$\sum_{i=1}^k c_i = 0.$$

В частном случае проверяется серия гипотез о простых линейных контрастах вида

$$L_0 = \mu_i - \mu_j, i = 1, 2, \dots, k-1; j = i+1, \dots, k.$$



Вычисление критерия производится по формуле

$$t = \frac{\sum_{i=1}^k c_i \bar{x}_i}{\sqrt{M \sum_{i=1}^k \frac{c_i^2}{n_i}}},$$

$$M = \frac{1}{N - k} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

где – средний квадратичный остаток,

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2, \dots, k$$

– среднее значение  $i$ -й выборки,

$$N = \sum_{i=1}^k n_i$$

– общая численность,

$n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$k$  – число столбцов (выборок).

Статистика критерия имеет  $F$ -распределение с параметрами  $k - 1$  и  $N - k$ .

Обсуждение см. в книгах Полларда, Шеффе, Мюллера с соавт., Ликеша с соавт., Бикела с соавт., Браунли.

### 8.2.1.1.10. Критерий Дункана

Однофакторный анализ может показать, что средние значения совокупностей различаются.

Однако он не позволяет узнать, средние значения каких совокупностей действительно различаются между собой. Для решения проблемы применяется критерий Дункана (Duncan's test). Вычисление критерия производится по формуле

$$d = \frac{|\bar{x}_i - \bar{x}_j|}{\sqrt{\frac{M}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}}, i = 1, 2, \dots, k; j = i + 1, \dots, k,$$

$$M = \frac{1}{N - k} \sum_{i=1}^k \sum_{l=1}^{n_i} (x_{il} - \bar{x}_i)^2$$

где – средний квадратичный остаток,

$$\bar{x}_i = \frac{1}{n_i} \sum_{l=1}^{n_i} x_{il}, i = 1, 2, \dots, k$$

– среднее значение  $i$ -й выборки,

$$N = \sum_{i=1}^k n_i$$

– общая численность,

$n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$k$  – число столбцов (выборок).







$P$ -значение критерия  $p$  является решением нелинейного уравнения

$$P_{r+2, N-k}(d) = (1-p)^{r+1},$$

где  $P_{\dots}(\cdot)$  – функция распределения студентизированного размаха,

$r$  – количество средних значений, расположенных между  $\bar{x}_i$  и  $\bar{x}_j$  в упорядоченном по возрастанию ряду  $k$  средних.

Обратная функцию распределения рассматриваемого критерия

$$p = 1 - \exp \frac{\ln P_{r+2, N-k}(d)}{r+1}.$$

Описание см. в сборниках таблиц Оуэна, Мюллера с соавт.

### 8.2.1.1.11. Критерий Тьюки

Если независимые выборки имеют равные численности, гипотезы о простых линейных контрастах могут быть проверены с помощью критерия Тьюки (метода Тьюки). Критерий Тьюки имеет аналогичные критерию Шеффе предпосылки для своего применения.

Линейный контраст

$$L = \sum_{i=1}^k c_i \mu_i$$

представляет собой линейную функцию от средних значений  $\mu_i$ ,  $i = 1, 2, \dots, k$ ,  $k$  независимых нормальных выборок с неизвестными, но равными дисперсиями, и известных констант  $c_i$ ,  $i = 1, 2, \dots, k$ , удовлетворяющих условию

$$\sum_{i=1}^k c_i = 0.$$

В частном случае проверяется серия гипотез о простых линейных контрастах вида

$$L = \mu_i - \mu_j, i = 1, 2, \dots, k-1; j = i+1, \dots, k.$$

Вычисление критерия при проверке нулевой гипотезы  $L = L_0$  производится по формуле

$$t = \frac{\sum_{i=1}^k c_i \bar{x}_i - L_0}{\frac{1}{2} \sqrt{M} \sum_{i=1}^k |c_i|} \sqrt{m},$$

$$M = \frac{1}{k(m-1)} \sum_{i=1}^k \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2$$

где – средний квадратичный остаток,

$$\bar{x}_i = \frac{1}{m} \sum_{j=1}^m x_{ij}, i = 1, 2, \dots, k$$

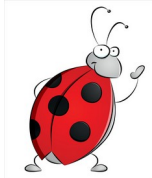
– среднее значение  $i$ -й выборки,

$m$  – численность каждой выборки,

$k$  – число столбцов (выборок).

Статистика критерия Тьюки подчиняется распределению студентизированного размаха с параметрами  $k$  и  $k(m-1)$ .





Обсуждение см. в книгах Мюллера с соавт., Ликеша с соавт., Бикела с соавт., Гланца, Аффифи с соавт.

## 8.2.1.1.12. Критерий Ливена

Критерий Ливена (Levene's test for equality of variance) является аналогом критерия Бартлетта. Перед вычислением статистики критерия выполняется преобразование исходных данных по формуле

$$z_{ij} = |x_{ij} - \bar{x}_i|, i = 1, 2, \dots, k; j = 1, 2, \dots, n_i,$$

$k$  – число столбцов (выборок),

$n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2, \dots, k$$

– среднее значение  $i$ -й выборки (групповое среднее значение).

Вычисление статистики критерия производится по формуле, аналогичной статистике однофакторного дисперсионного анализа,

$$W = \frac{N - k}{k - 1} \frac{\sum_{i=1}^k n_i (\bar{z}_i - \bar{z}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2},$$

$$N = \sum_{i=1}^k n_i$$

где – общая численность,

$$\bar{z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}, i = 1, 2, \dots, k$$

– среднее значение  $i$ -й выборки,

$$\bar{z}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} z_{ij}$$

– общее среднее значение.

Статистика критерия имеет  $F$ -распределение с параметрами  $k - 1$  и  $N - k$ .

См. Шукри (Shoukri) с соавт.

## 8.2.1.1.13. Критерий Брауна–Форсайта

Критерий Брауна–Форсайта (Brown–Forsythe test for equality of group variances) является вариантом критерия Ливена. Перед вычислением статистики критерия выполняется преобразование исходных данных по формуле

$$z_{ij} = |x_{ij} - \tilde{x}_i|, i = 1, 2, \dots, k; j = 1, 2, \dots, n_i,$$

$k$  – число столбцов (выборок),

$n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$\tilde{x}_i, i = 1, 2, \dots, k$  – медиана  $i$ -й выборки (групповая медиана).





Вычисление статистики критерия производится по формуле, аналогичной статистике однофакторного дисперсионного анализа,

$$W = \frac{N - k}{k - 1} \frac{\sum_{i=1}^k n_i (\bar{z}_{i.} - \bar{z}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})^2},$$

где  $N = \sum_{i=1}^k n_i$  – общая численность,

$\bar{z}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}, i = 1, 2, \dots, k$  – среднее значение  $i$ -й выборки,

$\bar{z}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} z_{ij}$  – общее среднее значение.

Статистика критерия имеет  $F$ -распределение с параметрами  $k - 1$  и  $N - k$ .  
См. руководство NIST/SEMATECH.

#### 8.2.1.1.14. Критерий V Бхапкара

Критерий V Бхапкара предназначен для проверки нулевой гипотезы о равенстве параметров положения (сдвига в средних) и масштаба (сдвига в дисперсиях). Вычисление статистики критерия производится по формуле

$$V = (2k - 1) \prod_{i=1}^k n_i \left\{ \frac{1}{\prod_{i=1}^k n_i} \sum_{i=1}^k n_i \left( u_i - \frac{1}{k} \right)^2 - \left[ \frac{1}{\prod_{i=1}^k n_i} \sum_{i=1}^k \left( u_i - \frac{1}{k} \right) \right]^2 \right\},$$

где  $n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

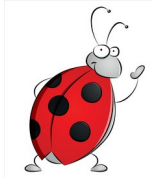
$k$  – число столбцов (выборок),

$u_i, i = 1, 2, \dots, k$  – количества подвыборок в сгенерированных выборках,  $i$ -ая варианта в

которых меньше остальных  $k - 1$  вариантов; при этом  $\prod_{i=1}^k n_i$  подвыборок генерируются из исходных выборок, чтобы в каждой подвыборке была представлена одна варианта каждой из  $k$  выборок.

Статистика критерия имеет  $\chi^2$ -распределение с параметром  $k - 1$ .

См. монографию Кобзаря.

**8.2.1.15. Критерий D Дешпанде**

Критерий  $D$  Дешпанде ( $D$ -критерий Дешпанде, Дюфора и Люнга) предназначен для проверки нулевой гипотезы о равенстве параметров масштаба (сдвига в дисперсиях). Вычисление статистики критерия производится по формуле

$$D = \frac{(2k-1)(k-1)^2 C_{2(k-1)}^{k-1} \prod_{i=1}^k n_i}{2[k^2 + (k^2 + 4k + 2)C_{2(k-1)}^{k-1}]} \left\{ \frac{1}{\prod_{i=1}^k n_i} \sum_{i=1}^k n_i (u_i + v_i)^2 - \left[ \frac{1}{\prod_{i=1}^k n_i} \sum_{i=1}^k (u_i + v_i) \right]^2 \right\},$$

где  $n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$k$  – число столбцов (выборок),

$u_i, i = 1, 2, \dots, k$  – количества подвыборок в сгенерированных выборках,  $i$ -ая варианта в

которых меньше остальных  $k-1$  вариантов; при этом  $\prod_{i=1}^k n_i$  подвыборок генерируются из исходных выборок, чтобы в каждой подвыборке была представлена одна варианта каждой из  $k$  выборок,

$v_i, i = 1, 2, \dots, k$  – количества подвыборок в сгенерированных выборках,  $i$ -ая варианта в которых больше остальных  $k-1$  вариант.

Статистика критерия имеет  $\chi^2$ -распределение с параметром  $k-1$ .

См. монографию Кобзаря.

**8.2.1.16. Критерий L Дешпанде**

Критерий  $L$  Дешпанде ( $L$ -критерий Дешпанде, Дюфора и Люнга) предназначен для проверки нулевой гипотезы о равенстве параметров положения (сдвига в средних). Вычисление статистики критерия производится по формуле

$$L = \frac{(2k-1)(k-1)^2 C_{2(k-1)}^{k-1} \prod_{i=1}^k n_i}{2k^2 [C_{2(k-1)}^{k-1} - 1]} \left\{ \frac{1}{\prod_{i=1}^k n_i} \sum_{i=1}^k n_i (-u_i + v_i)^2 - \left[ \frac{1}{\prod_{i=1}^k n_i} \sum_{i=1}^k (-u_i + v_i) \right]^2 \right\},$$

где  $n_i, i = 1, 2, \dots, k$  – численность  $i$ -й выборки,

$k$  – число столбцов (выборок),

$u_i, i = 1, 2, \dots, k$  – количества подвыборок в сгенерированных выборках,  $i$ -ая варианта в которых

меньше остальных  $k-1$  вариантов; при этом  $\prod_{i=1}^k n_i$  подвыборок генерируются из исходных выборок, чтобы в каждой подвыборке была представлена одна варианта каждой из  $k$  выборок,



$v_i, i = 1, 2, \dots, k$  – количества подвыборок в сгенерированных выборках,  $i$ –я варианта в которых больше остальных  $k - 1$  вариант.

Статистика критерия имеет  $\chi^2$ –распределение с параметром  $k - 1$ .

См. монографию Кобзаря.

## 8.2.1.2. Многофакторный дисперсионный анализ

Исходные данные для двухфакторного дисперсионного анализа представлены в виде таблицы (прямоугольной матрицы), причем число столбцов соответствует числу уровней первого фактора (уровней обработки), число строк равно числу уровней второго фактора (уровней обработки).  $n$  строк – блоков наблюдений параметров объектов – расположены в  $k$  столбцах, соответствующих видам обработки (видам воздействия на объекты). При этом каждый блок может быть результатом измерений параметров как на одном объекте, так и на группе объектов, например в виде среднего значения какого–либо параметра, вычисленного по всем объектам исследуемой группы при определенном виде воздействия на группу. Следующий блок будет средним значением другого параметра по всем объектам группы при том же виде воздействия.

Предлагаются методы двухфакторного дисперсионного анализа:

- двухфакторный дисперсионный анализ (MANOVA),
- ранговый двухфакторный анализ Фридмана,
- критерий Квейда,
- критерий Пэйджа,
- Q–критерий Кокрена,
- критерий Шеффе для связанных выборок.

### 8.2.1.2.1. Двухфакторный дисперсионный анализ

Результаты опытов никогда в точности не соответствуют степени влияния на них того или иного признака. Происходит это потому, что на результаты оказывают влияние и неучтенные в условиях эксперимента факторы. При включении в дисперсионный анализ двух и более факторов имеет место многофакторный дисперсионный анализ (MANOVA).

Двухфакторный дисперсионный анализ, иначе называемый дисперсионным анализом по двум признакам (двухфакторный дисперсионный анализ без повторений), применяется для зависимых нормально распределенных выборок. Нулевая гипотеза состоит в утверждении о равенстве эффектов строк между собой и равенстве эффектов столбцов между собой.

Вычисления производятся по формулам:

$$t_{row} = \frac{D_{row}}{D},$$

эффект строк

$$t_{col} = \frac{D_{col}}{D},$$

эффект столбцов



где  $D_{row} = \frac{1}{r-1} SS_{row}$  – дисперсия, объясняемая строками,

$D_{col} = \frac{1}{c-1} SS_{col}$  – дисперсия, объясняемая столбцами,

$D = \frac{SS - SS_{row} - SS_{col}}{(r-1)(c-1)}$  – остаточная дисперсия,

$SS_{row} = \frac{1}{c} \sum_{i=1}^r T_i^2 - \frac{T_{..}^2}{rc}$  – средний квадрат строк,

$SS_{col} = \frac{1}{r} \sum_{j=1}^c T_j^2 - \frac{T_{..}^2}{rc}$  – средний квадрат столбцов,

$SS = \sum_{i=1}^r \sum_{j=1}^c x_{ij}^2 - \frac{T_{..}^2}{rc}$  – средний квадрат погрешности,

$T_i = \sum_{j=1}^c x_{ij}, i = 1, 2, \dots, r$  – суммы строк,

$T_j = \sum_{i=1}^r x_{ij}, j = 1, 2, \dots, c$  – суммы столбцов,

$T_{..} = \sum_{i=1}^r \sum_{j=1}^c x_{ij}$  – общая сумма,

$c$  – число столбцов (выборок),

$r$  – число строк (параметров).

Статистика критерия имеет  $F$ -распределение с параметрами  $r-1$  и  $(r-1)(c-1)$  в случае исследования эффекта строк и с параметрами  $c-1$  и  $(r-1)(c-1)$  в случае исследования эффекта столбцов.

## 8.2.1.2.2. Ранговый критерий Фридмана

Если не выполнены предположения, позволяющие провести двухфакторный дисперсионный анализ, применяется свободный от типа распределения непараметрический критерий Фридмана. Ранговый двухфакторный анализ Фридмана (ранговый критерий Фридмана, Кендалла и Бэбингтона Смита) применяется для проверки нулевой гипотезы о том, что различные методы обработки или иных воздействий на изучаемый объект (процесс) дают одинаковые результаты. Нулевая гипотеза состоит в отсутствии эффектов столбцов (эффектов обработки). Критерий может также применяться в качестве непараметрического аналога однофакторного дисперсионного анализа с повторными измерениями. Вычисление статистики критерия производится по формуле



$$S = \frac{12 \sum_{j=1}^k (R_j - nR_{..})^2}{nk(k+1) - \frac{1}{k-1} \sum_{i=1}^n \left( \sum_{j=1}^{g_i} t_{ij}^3 - k \right)},$$

где  $R_j, j = 1, 2, \dots, k$  – соответствующие суммы рангов в строках,  
 $n$  – численность каждой совокупности,  
 $k$  – число эффектов обработки (воздействий, уровней фактора),

$$R_{..} = \frac{k+1}{2},$$

$g_i, i = 1, 2, \dots, n$  – число связей в блоке,

$t_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, g_i$  – численность соответствующей связи, равная 1 при отсутствии связей в блоке.

Суммы рангов вычисляются по формуле

$$R_j = \sum_{i=1}^n r_{ij}, j = 1, 2, \dots, k,$$

где  $r_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, k$  – ранги, причем ранжирование производится по каждой строке отдельно.

Статистика критерия имеет  $\chi^2$ -распределение с параметром  $k - 1$ .

Описание см. в книгах Хотеллинг с соавт., Браунли, Петровича с соавт., в справочнике Оуэна.

### 8.2.1.2.3. Критерий Квейд

Если не выполнены предположения, позволяющие провести двухфакторный дисперсионный анализ, применяется свободный от типа распределения непараметрический ранговый критерий Квейд (Quade's test). Нулевая гипотеза состоит в отсутствии эффектов столбцов.

Вычисление статистики критерия производится по формуле

$$S = \frac{n-1}{n} \sum_{j=1}^k T_j^2 \left[ \sum_{i=1}^n \sum_{j=1}^k R_{ij}^2 - \frac{1}{n} \sum_{j=1}^k T_j^2 \right]^{-1},$$

где  $R_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, k$  – скорректированные ранги,

$T_j, j = 1, 2, \dots, k$  – суммы столбцов матрицы скорректированных рангов,

$n$  – численность каждой совокупности,

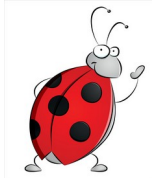
$k$  – число эффектов обработки (воздействий, уровней фактора).

Скорректированные ранги рассчитываются по формуле

$$R_{ij} = Q_i \cdot \left( r_{ij} - \frac{k+1}{2} \right), i = 1, 2, \dots, n; j = 1, 2, \dots, k,$$

где  $r_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, k$  – ранги, причем ранжирование производится отдельно по каждой строке матрицы исходных данных,





$Q_j, j = 1, 2, \dots, n$  – ранги размахов строк матрицы исходных данных.

Статистика критерия имеет  $F$ -распределение с параметрами  $k - 1$  и  $(k - 1)(n - 1)$ .

См. книгу Петровича с соавт., работы Понтеса (Pontes, 2000), Кемпбелл (Campbell).

#### 8.2.1.2.4. Критерий Пэйджа

Критерий Пэйджа (критерий L Пэйджа, дисперсионный анализ Пэйджа) предназначен для проверки нулевой гипотезы о равенстве эффектов обработки (воздействия) на выборки с неизвестными, но равными средними. Нулевая гипотеза состоит в утверждении о равенстве эффектов строк между собой и равенстве эффектов столбцов между собой. Статистика критерия вычисляется по формуле

$$L = \sum_{i=1}^k iR_i,$$

где  $R_i, i = 1, 2, \dots, k$  – упорядоченные по возрастанию суммы рангов блоков,

$k$  – число эффектов обработки (воздействий, уровней фактора).

$$\frac{L - ML}{\sqrt{DL}}$$

Для больших выборок распределение преобразованной статистики является приближенно нормальным. Здесь математическое ожидание и дисперсия рассчитываются по формулам, соответственно:

$$ML = \frac{1}{4}nk(k+1)^2,$$

$$DL = \frac{n(k^3 - k)}{144(k-1)},$$

где  $n$  – численность каждой совокупности.

См. источники: Лисенков, Тюрин с соавт., Холлендер с соавт.

#### 8.2.1.2.5. Критерий Q Кокрена

Критерий Q Кокрена используется в случае, если группы однородных субъектов подвергаются более чем двум экспериментальным воздействиям, и их ответы носят двухвариантный (бинарный, дихотомический) характер. Предполагается, что 0 означает отрицательный ответ, 1 – положительный. Каждая выборка представляет собой измерения одного условия по всем группам. Варианты выборки – это измерения в рассматриваемых группах по данному условию. Нулевая гипотеза состоит в том, что в генеральной совокупности доли всех экспериментальных условий равны. Вычисление производится по формуле





$$Q = \frac{(c-1) \left( c \sum_{j=1}^c T_{.j}^2 - \left( \sum_{j=1}^c T_{.j} \right)^2 \right)}{c \sum_{i=1}^r T_{i.} - \sum_{i=1}^r T_{i.}^2},$$

где  $T_{.j} = \sum_{i=1}^r x_{ij}, j = 1, 2, \dots, c$  – суммы столбцов,

$T_{i.} = \sum_{j=1}^c x_{ij}, i = 1, 2, \dots, r$  – суммы строк,

$c$  – число столбцов (выборок),

$r$  – число строк (параметров).

Статистика критерия имеет  $\chi^2$ -распределение с параметром  $c - 1$ .

Описание см. у Браунли.

### 8.2.1.2.6. Критерий Шеффе для связанных выборок

Двухфакторный позволяет обнаружить существование эффектов столбцов (эффектов обработки) в таблице дисперсионного анализа. Однако он не дает возможности точно указать столбцы, которые обладают нулевыми эффектами. Для решения проблемы применяется метод множественного сравнения Шеффе для связанных выборок (парный критерий Шеффе). Критерий Шеффе для связанных выборок предназначен для проверки так называемой гипотезы о линейном контрасте. Линейный контраст

$$L = \sum_{i=1}^k c_i \mu_i$$

представляет собой линейную функцию от средних значений  $\mu_i, i = 1, 2, \dots, k$ ,  $k$  независимых нормальных выборок с неизвестными, но равными дисперсиями, и известных констант  $c_i, i = 1, 2, \dots, k$ , удовлетворяющих условию

$$\sum_{i=1}^k c_i = 0.$$

В частном случае проверяется серия гипотез о простых линейных контрастах вида

$$L_0 = \mu_i - \mu_j, i = 1, 2, \dots, k-1; j = i+1, \dots, k.$$

Вычисление статистики критерия производится по формуле

$$t = \frac{\left( \sum_{i=1}^r c_i \bar{x}_i \right)^2}{(r-1) S \sum_{i=1}^r c_i^2},$$



$$S = \sum_{i=1}^r \sum_{j=1}^c x_{ij}^2 - \frac{T_{..}^2}{rc}$$
 – остаточный средний квадрат,

$$T_{..} = \sum_{i=1}^r \sum_{j=1}^c x_{ij}$$
 – общая сумма,

$$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2, \dots, k$$
 – среднее значение  $i$ -й выборки,

$c$  – число столбцов (выборок),

$r$  – число строк (параметров).

Статистика критерия имеет  $F$ -распределение с параметрами  $r - 1$  и  $(r - 1)(c - 1)$  в случае исследования эффекта строк и с параметрами  $c - 1$  и  $(r - 1)(c - 1)$  в случае исследования эффекта столбцов.

См. справочник Полларда.

## 8.2.2. Множественные сравнения

Методы множественного сравнения применяются, если исходные данные представлены многомерными выборками. В разделе предлагаются несколько популярных методов множественных сравнений, представляющих собой обобщения методов проверки гипотез (в т.ч. дисперсионного анализа) на многомерные выборки.

Для параметрических и непараметрических методов проверки гипотез (см. главы «Параметрическая статистика» и «Непараметрическая статистика») и дисперсионного анализа существуют многомерные аналоги в множественном сравнении, как показано в таблице.

Метод проверки гипотезы для двух выборок и Многомерный «функциональный аналог» из дисперсионного анализа

Параметрические тесты

Критерий Стьюдента

Критерий Уэлча

$F$ -критерий

Критерий Бартлетта

Непараметрические тесты

Критерий Вилкоксона

Критерий Муда

Критерий Краскела–Уоллиса

множественных сравнений

Критерий Хотеллинга

Критерий Джеймса–Сю

Критерий Кульбака (2 многомерные выборки)

Критерий Кульбака ( $k > 2$  выборок)

Критерий Уилкса

Критерий Пури–Сена–Тамура

Критерий Пури–Сена

Критерий Шейрера–Рэя–Хэйра (2 многомерные выборки)

Представлены:

- критерий Хотеллинга,
- критерий Джеймса–Сю,





- критерий Кульбака,
- критерий Пури–Сена–Тамура,
- критерий Пури–Сена.
- критерий Шейрера–Рэя–Хэйра.

В главе для полноты информации описан также критерий Уилкса и даны рекомендации по его самостоятельному вычислению.

Исходные данные для множественных сравнений представлены в виде таблиц (прямоугольных матриц). Каждой выборке соответствует одна матрица, причем число столбцов каждой матрицы соответствует размерности многомерной выборки, число строк равно числу наблюдений. При этом выборки могут иметь как одинаковое число вариантов (равные объемы), так и различное, в зависимости от требований применяемого метода. Размерности сравниваемых многомерных выборок должны быть одинаковы. Обзор см. в диссертации Понтеса (Pontes, 2005). См. также источники: Родионов с соавт., Коган с соавт., Пури (Puri) с соавт., Сен (Sen) с соавт., Тамура (Tamura).

### 8.2.2.1. Критерий Хотеллинга

Критерий  $T^2$  (критерий следа, критерий Лоули и Хотеллинга), для случая двух многомерных выборок предложенный Хотеллингом, применяется в задаче статистической проверки гипотезы о равенстве векторов средних двух многомерных совокупностей. Предполагается, что многомерные выборки извлечены из нормальных многомерных распределений с равными между собой ковариационными матрицами. Статистика критерия Хотеллинга вычисляется по формуле

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2),$$

где  $n_1$  – количество многомерных вариантов первой многомерной выборки,

$n_2$  – количество многомерных вариантов второй многомерной выборки,

$\bar{x}_1$  и  $\bar{x}_2$  – векторы средних двух многомерных совокупностей,

$S$  – дисперсионно–ковариационная матрица совокупности.

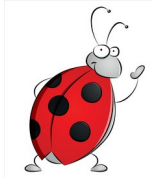
Если дисперсионно–ковариационная матрица совокупности неизвестна, она вычисляется через выборочные дисперсионно–ковариационные матрицы совокупностей по формуле:

$$S = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)S_1 + (n_2 - 1)S_2],$$

где  $S_1$  и  $S_2$  – выборочные дисперсионно–ковариационные матрицы многомерных совокупностей.

$$\frac{n_1 + n_2 - m - 1}{m(n_1 + n_2 - 2)} T^2$$

Модифицированная статистика критерия имеет  $F$ -распределение с параметрами  $m$  и  $n_1 + n_2 - m - 1$ , где  $m$  – размерность каждой выборки.



Описание см. у Андерсона, Афффи с соавт., Джонсона с соавт., Кульбака, Мэйндоналда, Хальда, в справочнике под редакцией Ллойда с соавт. Связь с расстоянием Махаланобиса выведена Уилксом.

## 8.2.2.2. Критерий Джеймса–Сю

Критерий Джеймса–Сю предназначен для проверки гипотезы о равенстве векторов средних двух многомерных совокупностей. Предполагается, что многомерные выборки извлечены из нормальных многомерных распределений с неизвестными или неравными между собой ковариационными матрицами. Критерий является решением многомерной проблемы Беренса–Фишера. Статистика критерия вычисляется по формуле

$$2I = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2),$$

где  $\bar{x}_1$  и  $\bar{x}_2$  – векторы средних двух многомерных совокупностей,

$S$  – дисперсионно–ковариационная матрица совокупности.

Дисперсионно–ковариационная матрица совокупности вычисляется через выборочные дисперсионно–ковариационные матрицы совокупностей по формуле:

$$S = S_1 / n_1 + S_2 / n_2,$$

где  $n_1$  – количество многомерных вариантов первой многомерной выборки,

$n_2$  – количество многомерных вариантов второй многомерной выборки,

$S_1$  и  $S_2$  – выборочные дисперсионно–ковариационные матрицы многомерных совокупностей.

Статистика критерия  $2I$  подчиняется асимптотически распределению  $\chi^2$  с  $m$  степенями свободы.

Критерий описан Родионовым с соавт.

## 8.2.2.3. Критерий Кульбака

Критерий Кульбака предназначен для проверки равенства ковариационных матриц двух или более многомерных совокупностей. Предполагается, что многомерные выборки извлечены из совокупностей, подчиняющихся нормальному многомерному распределению. Для двух выборок статистика критерия может вычисляться по формуле

$$2I_0 = \sum_{i=1}^2 (n_i - 1) \ln \frac{|S|}{|S_i|},$$

где  $n_1$  и  $n_2$  – количества многомерных вариантов сравниваемых совокупностей,

$S_1$  и  $S_2$  – выборочные дисперсионно–ковариационные матрицы многомерных совокупностей,

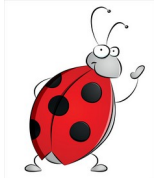
$|\cdot|$  – определитель матрицы.

Статистика критерия  $2I_0$  подчиняется асимптотически  $B$ –распределению Фишера, иначе нецентральному распределению  $\chi^2$  с параметром нецентральности

$$\lambda = \frac{(2m^3 + 2m^2 - m)}{12} \left( \frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2} \right) \text{ и } m(m + 1) / 2 \text{ степенями свободы, где } m -$$

размерность каждой выборки.





Критерий представил Кульбак, подробно описали Родионов с соавт.

#### 8.2.2.4. Критерий Пури–Сена–Тамура

Ранговый непараметрический критерий Пури–Сена–Тамура предназначен для проверки гипотезы о равенстве векторов средних двух многомерных совокупностей. Статистика критерия вычисляется по формуле

$$\Lambda = \sum_{i=1}^2 (\bar{r}_i - \bar{r})' S^{-1} (\bar{r}_i - \bar{r}),$$

где  $\bar{r}_1$  и  $\bar{r}_2$  – векторы средних рангов двух многомерных совокупностей,

$\bar{r}$  – вектор средних рангов объединенной совокупности,

$S$  – дисперсионно–ковариационная матрица рангов объединенной совокупности.

Статистика критерия подчиняется асимптотически распределению  $\chi^2$  с  $m$  степенями свободы, где  $m$  – размерность каждой выборки.

Критерий описан Родионовым с соавт., где рассмотрены также случаи использования иных ранговых отметок.

#### 8.2.2.5. Критерий Пури–Сена

Ранговый непараметрический критерий Пури–Сена предназначен для проверки равенства ковариационных матриц двух многомерных совокупностей. Статистика критерия вычисляется по формуле

$$\Lambda = \sum_{i=1}^2 (\bar{e}_i - \bar{e})' S^{-1} (\bar{e}_i - \bar{e}),$$

где  $\bar{e}_1$  и  $\bar{e}_2$  – векторы средних ранговых отметок двух многомерных совокупностей,

$\bar{e}$  – вектор средних ранговых отметок объединенной совокупности,

$S$  – дисперсионно–ковариационная матрица ранговых отметок объединенной совокупности.

При этом ранговые отметки вычисляются как

$$E_{ij} = \left( \frac{R_{ij}}{N+1} - 0,5 \right)^2, i=1,2,\dots,N; j=1,2,\dots,m,$$

где  $R_{ij}$ ,  $i=1,2,\dots,N$ ;  $j=1,2,\dots,m$  – ранги соответствующей выборки,

$N$  – численность соответствующей выборки:  $n_1$  или  $n_2$  – многомерных вариант сравниваемых совокупностей,  $n_1 + n_2$  – объединенной совокупности,

$m$  – размерность каждой выборки.

Статистика критерия подчиняется асимптотически распределению  $\chi^2$  с  $m$  степенями свободы.

Критерий описан Родионовым с соавт., где рассмотрены также случаи использования иных ранговых отметок.



### 8.2.2.6. Критерий Шейрера–Рэя–Хэйра

Критерий Шейрера–Рэя–Хэйра представляет собой многомерное расширение критерия Краскела–Уоллиса. Критерий парный, т. е. формально представленные для анализа выборки должны иметь равные количества строк и столбцов. При этом предполагается, что по  $n$  строкам располагаются значения  $k$ -мерных выборочных значений.

Представлен вариант критерия для анализа двух многомерных выборок. Алгоритм может быть обобщен на произвольное количество многомерных выборок. Пусть даны две многомерных выборки:  $X_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, k$ , и  $Y_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, k$ . Многомерные выборки совместно ранжируются по убыванию. При этом совпадающим значениям присваиваются средние (по связке) ранги. В результате ранжирования получаются массивы рангов, соответственно,  $R_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, k$ , и  $S_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, k$ .

Затем составляется таблица размером  $2 \times k$ . В ячейки таблицы записываются суммы рангов, вычисленные по формулам, соответственно,

$$T_{1j} = \sum_{i=1}^n R_{ij}, j = 1, 2, \dots, k,$$

$$T_{2j} = \sum_{i=1}^n S_{ij}, j = 1, 2, \dots, k.$$

Вычисления статистик критерия производятся по формулам:

$$H_c = \frac{RSS_c}{RMS_{total}},$$

эффект столбцов

$$H_r = \frac{RSS_r}{RMS_{total}},$$

эффект строк

$$H_{rc} = \frac{RSS_{rc}}{RMS_{total}},$$

эффект взаимодействия строк и столбцов

где квадратичные остатки вычисляются по формулам:

$$RSS_c = \frac{1}{2k} \sum_{j=1}^k \left( \sum_{i=1}^2 T_{ij} \right)^2 - \frac{N(N+1)^2}{4},$$

$$RSS_r = \frac{1}{nk} \sum_{i=1}^2 \left( \sum_{j=1}^k T_{ij} \right)^2 - \frac{N(N+1)^2}{4},$$

$$RSS_{rc} = \frac{1}{n} \sum_{i=1}^2 \sum_{j=1}^k T_{ij}^2 - \frac{N(N+1)^2}{4} - RSS_c - RSS_r,$$

$$RMS_{total} = \frac{N(N+1)}{12},$$

где  $N = 2nk$  – общая численность представленных выборок.





Статистика  $H_c$  имеет  $\chi^2$ -распределение с параметром  $k - 1$ . Статистика  $H_r$  имеет  $\chi^2$ -распределение с параметром 1. Статистика  $H_{rc}$  (в случае двух выборок) имеет  $\chi^2$ -распределение с параметром  $k - 1$ .

См. монографию Сокала (Sokal) с соавт., статью Шайрера (Scheirer) с соавт.

## 8.2.2.7. Критерий Уилкса

Критерий  $\lambda$  Уилкса предназначен для выполнения однофакторного многомерного дисперсионного анализа. Его можно считать обобщением множественного критерия Хотеллинга на случай  $k > 2$  многомерных выборок. Предполагается, что многомерные выборки извлечены из нормальных многомерных распределений с равными между собой ковариационными матрицами. Статистика критерия вычисляется по формуле

$$\lambda = \frac{|W|}{|W + B|},$$

где  $W$  – общая матрица внутригруппового разброса,

$B$  – матрица межгруппового разброса,

$|\cdot|$  – операция вычисления определителя.

Элемент  $w_{ij}$  матрицы  $W$  вычисляется как

$$w_{ij} = \sum_{r=1}^k s_{ij}^{(r)}, i = 1, 2, \dots, m; j = 1, 2, \dots, m,$$

где  $s_{ij}^{(r)}, r = 1, 2, \dots, k$ , – элемент т.н. Матрицы  $S^{(r)}$  остаточных сумм квадратов и произведений выборки  $r$ ,

$k$  – количество многомерных выборок,

$m$  – число переменных (размерность) каждой многомерной выборки,

$r, r = 1, 2, \dots, k$  – верхний индекс, означающий номер многомерной выборки.

Элемент  $s_{ij}^{(r)}$  матрицы  $S^{(r)}$  вычисляется как

$$s_{ij}^{(r)} = \frac{1}{n - k} \sum_{l=1}^{n_r} (x_{il}^{(r)} - \bar{x}_i^{(r)})^T (x_{jl}^{(r)} - \bar{x}_j^{(r)}), i = 1, 2, \dots, m; j = 1, 2, \dots, m,$$

где  $x_{il}^{(r)}, i = 1, 2, \dots, n_r$ , – значение варианты переменной  $i$ ,

$x_{jl}^{(r)}, j = 1, 2, \dots, n_r$ , – значение варианты переменной  $j$ ,

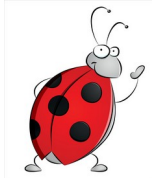
$\bar{x}_i^{(r)}, i = 1, 2, \dots, m$ , – среднее значение переменной  $i$ ,

$\bar{x}_j^{(r)}, j = 1, 2, \dots, m$ , – среднее значение переменной  $j$ ,

$n_r, r = 1, 2, \dots, k$  – численность выборки  $r$  (число  $m$ -мерных вариантов в каждой многомерной выборке),

$n = \sum_{r=1}^k n_r$  – общее количество многомерных выборок.





Элемент  $b_{ij}$  матрицы  $B$  вычисляется как

$$b_{ij} = \sum_{r=1}^k n_r \bar{x}_i^{(r)} \bar{x}_j^{(r)} - n \bar{x}_i \bar{x}_j, j = 1, 2, \dots, m; j = 1, 2, \dots, m,$$

где  $\bar{x}_i, i = 1, 2, \dots, m$ , – среднее значение переменной  $i$  по всем  $k$  выборкам,

$\bar{x}_j, j = 1, 2, \dots, m$ , – среднее значение переменной  $j$  по всем  $k$  выборкам.

Модифицированная статистика

$$\hat{\lambda} = - \left( n - 1 - \frac{m + k}{2} \right) \ln \lambda$$

подчиняется  $\chi^2$ -распределению с  $m(k - 1)$  степенями свободы (аппроксимация Бартлетта).

Метод представлен Андерсоном, Афифи с соавт., Петровичем с соавт. Андерсон указал точное распределение статистики критерия, а в последних двух источниках представлена также аппроксимация статистики  $F$ -распределением, предложенная Рао. Кульбак представил свой тест также и для  $k > 2$  многомерных выборок.

### 8.2.3. Ковариационный анализ

Однофакторный ковариационный анализ использует концепции однофакторного дисперсионного анализа и линейной регрессии. Предполагается, что исходные данные представляют собой совокупность регрессий («предиктор–зависимая переменная» или, в смысле построения графика регрессии, «абсцисса–ордината»), соответствующих различным уровням значения качественного признака. При этом сами значения качественного признака не вводятся.

Метод позволяет протестировать ряд статистических гипотез, как указано в соответствующем разделе.

Предпосылки применения ковариационного анализа:

- нормальность распределения ошибок (относительно линейной регрессии),
- однородность дисперсии ошибок,
- зависимость отклика от количественного предиктора линейна,
- равный наклон регрессий на уровнях качественного фактора.

Если данные не удовлетворяют представленным требованиям, они могут быть преобразованы соответствующими методами.

#### 8.2.3.1. Однофакторный ковариационный анализ

Однофакторный ковариационный анализ (one-way ANCOVA) использует концепции однофакторного дисперсионного анализа, линейной регрессии и множественных сравнений. Представление исходных данных для расчета имеет свою особенность. Массив вводится в виде совокупности  $k$  регрессий (иначе – уровней, соответствующих градациям качественного фактора), как показано на следующей иллюстрации, причем численности  $n_i, i = 1, 2, \dots, k$ , пар





предикторов  $x_{ij}$ ,  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, n_i$ , и зависимых переменных  $y_{ij}$ ,  $i = 1, 2, \dots, k$ ;  $j = 1, 2, \dots, n_i$ , могут полагаться как равными, так и различными:

Уровни качественного фактора (не вводятся)

Уровень 1		Уровень 2		...		Уровень $k$	
$x_{11}$	$y_{11}$	$x_{21}$	$y_{21}$	...	...	$x_{k1}$	$y_{k1}$
$x_{12}$	$y_{12}$	$x_{22}$	$y_{22}$	...	...	$x_{k2}$	$y_{k2}$
...	...	...	...	...	...	...	...
$y_{1n_1}$	$y_{1n_1}$	$x_{2n_2}$	$y_{2n_2}$	...	...	$x_{kn_k}$	$y_{kn_k}$

Обозначим:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, i = 1, 2, \dots, k,$$

– среднее значение предиктора на уровне  $i$ ,  $i = 1, 2, \dots, k$ ,

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, i = 1, 2, \dots, k,$$

– среднее значение зависимой переменной на уровне  $i$ ,  $i = 1, 2, \dots, k$ ,

$$\bar{x}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}$$

– общее среднее значение предикторов на всех уровнях,

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$$

– общее среднее значение зависимых переменных на всех уровнях.

где  $k$  – количество уровней качественного фактора,

$N$  – общая численность пар предикторов и зависимых переменных, вычисляемая как

$$N = \sum_{i=1}^k n_i$$

С учетом введенных обозначений суммы квадратов и смешанные произведения между уровнями вычисляются как:

$$M_{xx} = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x}_{..})^2 \quad M_{xy} = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x}_{..})(\bar{y}_i - \bar{y}_{..}) \quad M_{yy} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y}_{..})^2$$

и

Соответственно, суммы квадратов и смешанные произведения внутри всех уровней находятся как:

$$E_{xx} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad E_{xy} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) \quad E_{yy} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

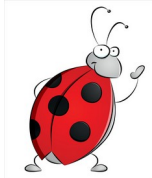
и

Полные суммы квадратов и смешанные произведения будут:

$$T_{xx} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 \quad T_{xy} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..}) \quad T_{yy} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

и

С учетом введенных обозначений, некоторые средние квадраты, необходимые для дальнейшего конструирования статистик, вычисляются как:



$$MS_M = \frac{1}{k-1} \left( M_{yy} - \frac{T_{xy}^2}{T_{xx}} + \frac{E_{xy}^2}{E_{xx}} \right), \quad MS_Z = \frac{E_{xy}^2}{E_{xx}}, \quad MS_E = \frac{1}{N-k-1} \left( E_{yy} - \frac{E_{xy}^2}{E_{xx}} \right),$$
$$MS_B = \frac{1}{k-1} \left( B_M - \frac{E_{xy}^2}{E_{xx}} \right), \quad MS_R = \frac{1}{N-2k} (E_{yy} - B_M)$$

где  $B_M$  – сумма средних квадратов внутри уровней, вычисляемая по формуле

$$B_M = \sum_{i=1}^k b_i$$

где  $b_i, i = 1, 2, \dots, k$  – групповые коэффициенты регрессии.

Групповые коэффициенты регрессии вычисляются как

$$b_i = \frac{E_{xy(i)}^2}{E_{xx(i)}}, i = 1, 2, \dots, k,$$

где  $E_{xy(i)}, i = 1, 2, \dots, k$  – смешанные произведения внутри уровней,

$E_{xx(i)}, i = 1, 2, \dots, k$  – суммы квадратов внутри уровней.

Данные параметры вычисляются, соответственно, по формулам:

$$E_{xy(i)} = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.}), i = 1, 2, \dots, k, \quad E_{xx(i)} = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2, i = 1, 2, \dots, k.$$

Стандартно выводятся вычисленные ранее средние значения на уровнях, групповые коэффициенты регрессии, а также скорректированные (adjusted) групповые средние значения на уровнях, вычисляемые по формуле

$$\bar{y}_{i(adj)} = \bar{y}_{i.} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}), i = 1, 2, \dots, k,$$

где  $\hat{\beta}$  – оценка коэффициента регрессии, вычисляемая по формуле

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}}.$$

Соответствующие стандартные ошибки скорректированных групповых средних значений вычисляются по формуле

$$S\bar{y}_{i(adj)} = \sqrt{MS_E \left( \frac{1}{n_i} + \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{E_{xx}} \right)}, i = 1, 2, \dots, k.$$

Доверительные интервалы оцениваемых скорректированных групповых средних значений вычисляются по формуле

$$I_{\bar{y}_{i(adj)}} = (\bar{y}_{i(adj)} - \sqrt{0,5 | m_{p,k',f} | S\bar{y}_{i(adj)}}, \bar{y}_{i(adj)} + \sqrt{0,5 | m_{p,k',f} | S\bar{y}_{i(adj)}}), i = 1, 2, \dots, k,$$

где  $|m_{p,k',f}|$  – значение обратной функции распределения студентизированного максимума модулей,

$p = 0,95$  – строится 95% доверительный интервал,



$$k' = k(k - 1) / 2,$$

$$f = N - k - 1.$$

Далее остановимся на регрессиях, которые могут быть построены по данным ковариационного анализа. Групповые регрессии задаются уравнениями

$$\hat{y} = \bar{y}_i + b_i(x - \bar{x}_i), i = 1, 2, \dots, k.$$

Можно построить групповые регрессии, используя оценку коэффициента регрессии  $\hat{\beta}$  (при этом мы получим параллельные групповые регрессии):

$$\hat{y} = \bar{y}_i + \hat{\beta}(x - \bar{x}_i), i = 1, 2, \dots, k.$$

Возможно построение регрессии средних значений

$$\hat{y} = \bar{y}_.. + \hat{\beta}_M(x - \bar{x}_..), i = 1, 2, \dots, k,$$

где  $\hat{\beta}_M$  – оценка коэффициента регрессии средних значений, вычисляемая по формуле

$$\hat{\beta}_M = \frac{M_{xy}}{M_{xx}}.$$

Возможно также построение полной регрессии

$$\hat{y} = \bar{y}_.. + \hat{\beta}_T(x - \bar{x}_..), i = 1, 2, \dots, k,$$

где  $\hat{\beta}_T$  – оценка полного коэффициента регрессии, вычисляемая по формуле

$$\hat{\beta}_T = \frac{T_{xy}}{T_{xx}}.$$

Для всех групповых регрессий, регрессии средних значений и полной регрессии выводится свободный член уравнения соответствующей регрессии в стандартной форме

$$y = a + bx,$$

где  $a$  – свободный член уравнения,

$b$  – коэффициент регрессии (здесь и далее подставить соответствующее значение из показанных выше уравнений регрессий),

$x$  – значение предиктора,

$y$  – значение зависимой переменной.

Свободный член вычисляется как

$$a = \bar{y} - b\bar{x},$$

где  $\bar{x}$  – среднее значение предиктора,

$\bar{y}$  – среднее значение зависимой переменной.

Проверяются следующие статистические гипотезы.

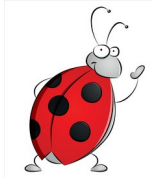
Гипотеза о равенстве скорректированных групповых средних значений: статистика

$$S_m = \frac{MS_M}{MS_E}$$

подчиняется  $F$ -распределению со степенями свободы  $k - 1$  и  $N - k - 1$ .

Гипотеза о равенстве наклона регрессии средних значений нулю: статистика





$$S_g = \frac{MS_Z}{MS_E}$$

подчиняется  $F$ -распределению со степенями свободы 1 и  $N - k - 1$ .

Гипотеза о равенстве наклонов групповых регрессий: статистика

$$S_b = \frac{MS_B}{MS_R}$$

подчиняется  $F$ -распределению со степенями свободы  $k - 1$  и  $N - 2k$ .

См. монографии Аффифи с соавт., Вайлдта (Wildt) с соавт., Милликена (Milliken) с соавт., Сокала (Sokal) с соавт., Вестфалла (Westfall) с соавт., Сахаи (Sahai) с соавт., статью и монографию Хсу (Hsu), монографию под ред. Эдвардс (Edwards).

## Глава 9. Регрессионный анализ

### 9.1. Введение

Регрессионный анализ предназначен для вычисления параметров аппроксимирующих функций различными методами и их статистических оценок.

Номенклатура методов насчитывает несколько различных аппроксимирующих функций.

Если пользователь не найдет необходимую функцию в предлагаемом перечне, можно воспользоваться универсальным методом – пользовательской функцией. Выбор параметров для данного метода не очевиден, поэтому порядок работы с данным методом представлен в виде подробного примера.

### 9.2. Теоретическое обоснование

Аппроксимацией называется замена одних математических объектов другими, в том или ином смысле близкими исходным. В более узком смысле аппроксимация – вычисление (подбор) неизвестных параметров алгебраических уравнений, в том числе приближение одних функций другими, причем аналитическое выражение для аппроксимируемой функции может быть известно или неизвестно.

#### 9.2.1. Оценка качества аппроксимации

Качество аппроксимации оценивается коэффициентом детерминации, вычисляемым по формуле

$$R^2 = 1 - \frac{\sigma_E^2}{\sigma_Y^2},$$

где  $\sigma_E^2$  – дисперсия остатков,

$\sigma_Y^2$  – дисперсия функции выхода эксперимента (далее – функции).

Дисперсия остатков вычисляется как





$$\sigma_E^2 = \frac{1}{N} \sum_{i=1}^N (e_i - \bar{e})^2,$$

где  $N$  – число экспериментальных значений (пар аргумент–функция),

$e_i, i = 1, 2, \dots, N$  – остатки,

$\bar{e}$  – среднее значение остатков.

Остатки рассчитываются по формуле

$$e_i = \hat{y}_i - y_i, i = 1, 2, \dots, N,$$

где  $\hat{y}_i, i = 1, 2, \dots, N$ , – рассчитанные значения модели (модельной оценки),

$y_i, i = 1, 2, \dots, N$  – заданные значения функции.

Среднее значение остатков вычисляется по формуле

$$\bar{e} = \frac{1}{N} \sum_{i=1}^N e_i.$$

Дисперсия функции вычисляется как

$$\sigma_Y^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2,$$

где  $\bar{y}$  – среднее значение функции.

Среднее значение функции рассчитывается по формуле

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i.$$

В источниках при записи формулы коэффициента детерминации часто опускают множители  $1 / N$  в выражениях для дисперсий, поэтому в числителе и знаменателе второго слагаемого помещают только соответствующие квадраты.

Исправленный коэффициент детерминации

$$R_{(adj)}^2 = 1 - \frac{\sigma_{E(adj)}^2}{\sigma_{Y(adj)}^2},$$

где  $\sigma_{E(adj)}^2$  – исправленная дисперсия остатков,

$\sigma_{Y(adj)}^2$  – исправленная дисперсия функции.

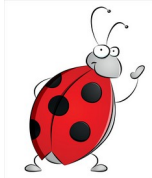
Исправленная дисперсия остатков вычисляется как

$$\sigma_{E(adj)}^2 = \frac{1}{(N - k)} \sum_{i=1}^N (e_i - \bar{e})^2,$$

где  $k$  – количество оцениваемых параметров модели.

Исправленная дисперсия функции вычисляется как

$$\sigma_{Y(adj)}^2 = \frac{1}{(N - 1)} \sum_{i=1}^N (y_i - \bar{y})^2.$$



Чем ближе вычисленное значение коэффициента детерминации или исправленного коэффициента детерминации к 1, тем лучше модель аппроксимирует представленные экспериментальные данные. И наоборот, чем меньше 1 вычисленное значение, тем хуже аппроксимирующая функция соответствует представленным данным.

Для проверки значимости коэффициента детерминации рассчитывается статистика

$$F = \frac{R^2 / k}{(1 - R^2) / (N - k - 1)},$$

где  $k$  – число оцениваемых параметров модели.

Статистика подчиняется  $F$ -распределению с параметрами  $k$  и  $N - k - 1$ .

Для проверки автокорреляции остатков применяется статистика Дарбина–Уотсона

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}.$$

В грубом приближении автокорреляция отсутствует при  $1,5 \leq d \leq 2,5$ .

## 9.2.2. Регрессионный анализ

Основные задачи регрессионного анализа:

1. Выбор наилучшей регрессионной модели (функции) для заданного набора экспериментальных данных (независимой переменной – выхода эксперимента).
2. Вычисление оптимальных параметров модели. Наиболее популярным для рассматриваемого типа задач является метод наименьших квадратов.
3. Оценка значимости и вычисление доверительных интервалов параметров модели.
4. Оценка значимости и вычисление доверительных интервалов выхода модели.

Решение первой задачи иногда находится вне вычислительных методов. К формулировке данной задачи могут привести математическое моделирование или просто интуиция исследователя. Также можно предположить вид аппроксимирующей функции, исходя из вида графика, построенного по результатам эксперимента. Для решения второй задачи используются методы аппроксимации (без статистических оценок). Решение третьей и четвертой задачи производится с помощью алгоритмов прикладной статистики. При этом оптимальные оценки параметров модели уже получены при решении третьей задачи.

Стандартные отклонения вычисленных оценок параметров модели

$$SE(\hat{\theta}) = \sqrt{\text{diag} \left[ \left( P^T(X, \hat{\theta}) P(X, \hat{\theta}) \right)^{-1} MSE \right]},$$

где  $P(.,.)$  – матрица частных производных модели по параметрам,

$\hat{\theta}$  – вектор оценок параметров,

$X$  – заданный вектор аргументов,

$MSE$  – средняя квадратичная ошибка (дисперсия ошибки регрессии).



Матрица частных производных функции модели по параметрам (опуская номер итерации) имеет вид

$$P(X, \theta) = \begin{bmatrix} \frac{\partial f(x_1, \theta)}{\partial \theta_1} & \frac{\partial f(x_1, \theta)}{\partial \theta_2} & \dots & \frac{\partial f(x_1, \theta)}{\partial \theta_k} \\ \frac{\partial f(x_2, \theta)}{\partial \theta_1} & \frac{\partial f(x_2, \theta)}{\partial \theta_2} & \dots & \frac{\partial f(x_2, \theta)}{\partial \theta_k} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f(x_N, \theta)}{\partial \theta_1} & \frac{\partial f(x_N, \theta)}{\partial \theta_2} & \dots & \frac{\partial f(x_N, \theta)}{\partial \theta_k} \end{bmatrix},$$

где  $x_i, i = 1, 2, \dots, N$  – элементы вектора аргумента,

$f(., \theta)$  – выход модели, получающийся подстановкой в функцию модели заданного аргумента, при фиксированном значении вектора параметров.

Практически производные вычисляются либо методом конечных разностей (если вид модели заранее неизвестен), либо задаются в явном виде (если вид модели задан).

Дисперсия ошибки регрессии вычисляется по формуле

$$MSE = \frac{1}{N - k} \sum_{i=1}^N (y_i - f(x_i, \hat{\theta}))^2.$$

Также выводятся доверительные интервалы оценок параметров, которые вычисляются по формуле, опуская индексы,

$$I_{\theta} = [\hat{\theta} - \Psi((1 + \beta)/2) \cdot SE(\hat{\theta}); \hat{\theta} + \Psi((1 + \beta)/2) \cdot SE(\hat{\theta})],$$

где  $\Psi(.)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

Дополнительно выводятся значения  $t$ -статистики (при проверке гипотезы о равенстве нулю) и  $P$ -значения оценок параметров.

Доверительные интервалы оценок модели  $\hat{y}_i = f(x_i, \hat{\theta}), i = 1, 2, \dots, N$ , вычисляются как, опуская индексы,

$$I_{\hat{y}} = [\hat{y} - \Psi((1 + \beta)/2) \cdot SD(\hat{y}); \hat{y} + \Psi((1 + \beta)/2) \cdot SD(\hat{y})],$$

где  $SD(.)$  – стандартное отклонение ошибки регрессии – корень квадратный из дисперсии ошибки регрессии.

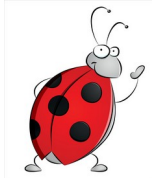
Дополнительно выводятся значения  $t$ -статистики (при проверке гипотезы о равенстве нулю) и  $P$ -значения стандартизованных остатков оценок модели.  $P$ -значения, не превышающие 0,05, отмечаются красным цветом, чтобы сигнализировать пользователю о возможном выбросе.

См. монографии Айвазяна с соавт., Полларда, Доугерти, Ферстера с соавт., Сергиенко с соавт., Райана (Ryan), Бородича.

### 9.2.3. Метод наименьших квадратов

Пусть записан функционал метода наименьших квадратов (method of least squares)





$$F(\theta) = \sum_{i=1}^N (y_i - f(x_i, \theta))^2,$$

где  $y_i, i = 1, 2, \dots, N$  – заданное экспериментальное значение, соответствующее значению абсциссы  $x_i, i = 1, 2, \dots, N$ ,

$f(.,.)$ ,  $i = 1, 2, \dots, N$ , – модельное значение, рассчитанное по теоретической формуле, заданной с точностью до параметров  $\theta$ ,

$N$  – число пар экспериментальных значений,

$\theta$  – вектор параметров, состоящий из подлежащих определению компонент  $\theta_i, i = 1, 2, \dots, r$ ,

$r$  – число параметров, зависящее от вида теоретической формулы.

Минимизация функционала  $F(\theta)$  по вектору параметров  $\theta_i, i = 1, 2, \dots, r$ ,

$$F(\theta) \rightarrow \min_{\theta}$$

приводит к системе  $r$  алгебраических уравнений:

$$\frac{\partial F(\theta)}{\partial \theta_j} = 0, j = 1, 2, \dots, r,$$

где в левой части уравнения находятся частные производные функционала  $F(\theta)$  по параметрам  $\theta_i, i = 1, 2, \dots, r$ .

Решив полученную линейную или нелинейную систему алгебраических уравнений, получим искомый оптимальный вектор параметров.

## 9.2.4. Полиномиальные модели

Модель представлена в виде полинома:

$$z(\theta, x) = \sum_{j=0}^r \theta_j x^j,$$

где  $\theta_i, i = 0, 1, \dots, r$  – коэффициенты полинома,

$x$  – заданная абсцисса,

$r$  – степень полинома.

Аналитически выразив частные производные функционала  $F(\theta)$  по параметрам, получаем  $\theta_i, i = 0, 1, \dots, r$ , систему  $r + 1$  алгебраических уравнений, линейных относительно параметров:

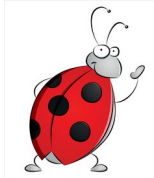
$$\sum_{j=0}^r (\theta_j \sum_{i=1}^N x_i^{j+k}) = \sum_{i=1}^N y_i x_i^k, k = 0, 1, \dots, r.$$

Может решаться более общая задача: наряду со степенью полинома можно указать минимальное значение степени члена полинома  $u$  (по умолчанию равно нулю):

$$z(\theta, x) = \sum_{j=u}^r \theta_j x^j,$$

Матрица частных производных функции модели по параметрам в общем случае будет





$$P(X, \theta) = \begin{bmatrix} x_1^u & x_1^{u+1} & \dots & x_1^r \\ x_2^u & x_2^{u+1} & \dots & x_2^r \\ \dots & \dots & \dots & \dots \\ x_N^u & x_N^{u+1} & \dots & x_N^r \end{bmatrix}.$$

## 9.2.5. Экспоненциально–степенная аппроксимация

Экспоненциальной функцией называется зависимость

$$z(x) = \theta_1 e^{\theta_2 x},$$

где  $\theta_i, i = 1, 2$  – параметры,

$x$  – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X, \theta) = \begin{bmatrix} e^{\theta_2 x_1} & \theta_1 x_1 e^{\theta_2 x_1} \\ e^{\theta_2 x_2} & \theta_1 x_2 e^{\theta_2 x_2} \\ \dots & \dots \\ e^{\theta_2 x_N} & \theta_1 x_N e^{\theta_2 x_N} \end{bmatrix}.$$

Степенной функцией называется зависимость

$$z(x) = \theta_1 x^{\theta_2}, x > 0,$$

где  $\theta_i, i = 1, 2$  – параметры,

$x$  – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X, \theta) = \begin{bmatrix} x_1^{\theta_2} & \theta_1 \theta_2 x_1^{\theta_2 - 1} \\ x_2^{\theta_2} & \theta_1 \theta_2 x_2^{\theta_2 - 1} \\ \dots & \dots \\ x_N^{\theta_2} & \theta_1 \theta_2 x_N^{\theta_2 - 1} \end{bmatrix}.$$

Гиперболой называется зависимость

$$z(x) = \theta_1 x^{-1} + \theta_2, x \neq 0,$$

где  $\theta_i, i = 1, 2$  – параметры,

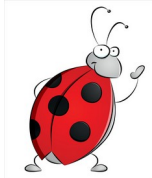
$x$  – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X, \theta) = \begin{bmatrix} x_1^{-1} & 1 \\ x_2^{-1} & 1 \\ \dots & \dots \\ x_N^{-1} & 1 \end{bmatrix}.$$

Экспоненциально–степенной называется зависимость





$$z(x) = e^{\theta_1 x} x^{\theta_2}, x > 0,$$

где  $\theta_i, i = 1, 2$  – параметры.

Матрица частных производных функции модели по параметрам будет

$$P(X, \theta) = \begin{bmatrix} e^{\theta_1 x_1} x_1^{\theta_2 + 1} & \theta_2 e^{\theta_1 x_1} x_1^{\theta_2 - 1} \\ e^{\theta_1 x_2} x_2^{\theta_2 + 1} & \theta_2 e^{\theta_1 x_2} x_2^{\theta_2 - 1} \\ \dots & \dots \\ e^{\theta_1 x_N} x_N^{\theta_2 + 1} & \theta_2 e^{\theta_1 x_N} x_N^{\theta_2 - 1} \end{bmatrix}.$$

При значениях аргумента, выходящих за указанные ограничения, соответствующие функции могут выдавать ошибку типа деления на нуль или выхода значений из допустимой области определения.

## 9.2.6. Логарифмическая функция

Логарифмической функцией называется зависимость

$$z(x) = \theta_1 + \theta_2 x + \theta_3 \ln(x), x > 0,$$

где  $\theta_i, i = 1, 2, 3$  – параметры,

$x$  – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X, \theta) = \begin{bmatrix} 1 & x_1 & \ln(x_1) \\ 1 & x_2 & \ln(x_2) \\ \dots & \dots & \dots \\ 1 & x_N & \ln(x_N) \end{bmatrix}.$$

При значениях аргумента, выходящих за указанные ограничения, представленная функция может выдавать ошибку выхода значений из допустимой области определения.

## 9.2.7. Логистический анализ

Пусть аргумент  $x$  означает время или величину растущего объекта, влияющего на размер  $y$  наблюдаемого явления. Тогда скорость роста может быть охарактеризована дифференциальным уравнением

$$\frac{dy}{dx} = f(x, y).$$

В частном случае введенная зависимость может иметь вид

$$\frac{dy}{dx} = f(y)g(x).$$

Аналитически можно получить различные примеры кривых роста, например

$$z(x) = \theta_1 \left[ 1 + e^{\theta_2 + \theta_3 x} \right]^{-1},$$

где  $\theta_i, i = 1, 2, 3$  – параметры, определяющие характер кривой,





$x$  – заданная абсцисса.

Матрица частных производных функции модели по параметрам будет

$$P(X, \theta) = \begin{bmatrix} \left[ \frac{1 + e^{\theta_2 + \theta_3 x_1}}{1 + e^{\theta_2 + \theta_3 x_2}} \right]^1 - \theta_1 e^{\theta_2 + \theta_3 x_1} \left[ \frac{1 + e^{\theta_2 + \theta_3 x_1}}{1 + e^{\theta_2 + \theta_3 x_2}} \right]^2 - \theta_1 x_1 e^{\theta_2 + \theta_3 x_1} \left[ \frac{1 + e^{\theta_2 + \theta_3 x_1}}{1 + e^{\theta_2 + \theta_3 x_2}} \right]^2 & \dots \\ \left[ \frac{1 + e^{\theta_2 + \theta_3 x_2}}{1 + e^{\theta_2 + \theta_3 x_2}} \right]^1 - \theta_1 e^{\theta_2 + \theta_3 x_2} \left[ \frac{1 + e^{\theta_2 + \theta_3 x_2}}{1 + e^{\theta_2 + \theta_3 x_2}} \right]^2 - \theta_1 x_2 e^{\theta_2 + \theta_3 x_2} \left[ \frac{1 + e^{\theta_2 + \theta_3 x_2}}{1 + e^{\theta_2 + \theta_3 x_2}} \right]^2 & \dots \\ \vdots & \ddots \\ \left[ \frac{1 + e^{\theta_2 + \theta_3 x_N}}{1 + e^{\theta_2 + \theta_3 x_N}} \right]^1 - \theta_1 e^{\theta_2 + \theta_3 x_N} \left[ \frac{1 + e^{\theta_2 + \theta_3 x_N}}{1 + e^{\theta_2 + \theta_3 x_N}} \right]^2 - \theta_1 x_N e^{\theta_2 + \theta_3 x_N} \left[ \frac{1 + e^{\theta_2 + \theta_3 x_N}}{1 + e^{\theta_2 + \theta_3 x_N}} \right]^2 & \dots \end{bmatrix}.$$

## 9.2.8. Пользовательская функция

Требования к пользовательской функции:

- Допустимость с точки зрения машины вычислений.
- Функция может содержать произвольное число параметров и стандартных элементарных функций.
- Аргумент в области допустимых значений. Некоторые варианты пользовательских функций могут содержать корректные формулы, но при некоторых значениях аргумента могут быть получены ошибки времени выполнения типа деления на нуль, переполнения.

Для аппроксимации пользовательской функцией применяются:

- Метод переменной метрики.
- Метод Гаусса– Ньютона.

Сравнивая данные методы, заметим, что иногда методы переменной метрики требуют меньшее число итераций, но время исполнения каждой итерации существенно выше (за счет вычисления оптимального параметра шага итераций). Достоинством метода переменной метрики является более широкая область сходимости (т. е. начальные значения параметров можно задать более удаленными от истинных их значений), если решение удастся получить вообще, причем метод Гаусса– Ньютона для тех же самых данных иногда дает решение. Поэтому для каждого набора данных и каждой модели может оказаться оптимальным свой метод. Может также встретиться случай, когда решение не удастся получить ни одним из представленных методов.

### 9.2.8.1. Метод Бroyдена–Флетчера–Голдфарба–Шанно

Применяется один из вариантов метода переменной метрики, а именно, метод Бройдена–Флетчера–Голдфарба–Шанно (метод BFGS). Согласно схеме метода, очередное приближение искомого вектора  $\theta$  решения нелинейной системы можно найти как

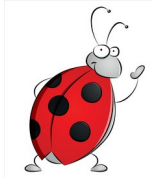
$$\theta^{(i+1)} = \theta^{(i)} + \rho_i d_i, i = 0, 1, 2, \dots,$$

где  $i, i = 0, 1, 2, \dots$  – номер итерации,

$\rho_i$  – параметр шага итераций,

$d_i$  – направление антиградиента (градиентом называют вектор, показывающий направление наибольшего роста скалярной функции  $n$  переменных), вычисляемое как





$$d_i = -A_i \nabla F(\theta^{(i)}),$$

где  $A_i$  – симметрическая положительно определенная матрица, аппроксимирующая матрицу, обратную к матрице Гессе системы  $[\nabla^2 F(\theta^{(i)})]^{-1}$ ,

$F(\theta)$  – квадратичный функционал невязок, построенный на основе выхода экспериментальных и модельных значений заданной пользователем функции.

Входящие в выражение градиента производные вычисляются методом конечных разностей.

Параметр  $\rho_i$  определяется из условия минимума

$$\varphi(\rho) = F(\theta^{(i)} + \rho_i d_i).$$

Для решения задачи минимизации

$$\varphi(\rho) \rightarrow \min_{\rho}$$

использован метод деления отрезка пополам.

Следующее приближение обращенной матрицы Гессе вычисляется по формуле

$$A_{i+1} = A_i + \frac{r_i r_i^T}{r_i^T g_i} - \frac{A_i g_i g_i^T A_i}{g_i^T A_i g_i},$$

где  $r_i = \theta^{(i+1)} - \theta^{(i)}$  – разность приближений,

$g_i = \nabla F(\theta^{(i+1)}) - \nabla F(\theta^{(i)})$  – разность градиентов.

Начальные значения входящих в формулы переменных берутся как  $g_0 = \nabla F(\theta^{(0)})$ ,  $A_0 = I$ , а  $\theta^{(0)}$  задано пользователем.

Вычисления прекращаются, если евклидова норма приращения очередного приближения вектора параметров меньше некоторого заранее заданного малого положительного числа  $\varepsilon$ .  
Методика и численные примеры представлены в монографии Носача.

## 9.2.8.2. Метод Гаусса– Ньютона

Метод Гаусса– Ньютона является одним из популярных квазиньютоновских методов.

Согласно схеме метода, очередное приближение искомого вектора  $\theta$  решения нелинейной системы можно найти как

$$\theta^{(i+1)} = \theta^{(i)} + [P_i^T(X, \theta) P_i(X, \theta)]^{-1} P_i^T(X, \theta) (Y - f(X, \theta^{(i)})), i = 0, 1, 2, \dots,$$

где  $i, i = 0, 1, 2, \dots$  – номер итерации,

$P_i(\dots)$ ,  $i = 0, 1, 2, \dots$  – матрица частных производных модели по параметрам,

$X$  – заданный вектор независимой переменной (аргумента),

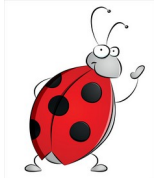
$Y$  – заданный вектор функции выхода эксперимента,

$f(\dots)$  – вектор выхода модели, получающийся подстановкой в функцию модели заданного вектора аргумента, при фиксированном значении вектора параметров.

Начальное значение  $\theta^{(0)}$  задано пользователем.

Вычисления прекращаются, если евклидова норма приращения очередного приближения вектора параметров меньше некоторого заранее заданного малого положительного числа  $\varepsilon$ .





Практически матрица частных производных вычисляется методом конечных разностей, т. к. вид модели заранее неизвестен.

Методика и численные примеры представлены в монографии Носача. Эффективные идеи даны в книгах Дэнниса (Dennis) с соавт., Дрейпера (Draper) с соавт.

## 9.2.9. Кусочно–линейная аппроксимация

Модель представлена (интерполирована) в виде кусочно–линейной функции

$$z(x) = \begin{cases} a_1 + b_1x, x_1 \leq x \leq x_2, \\ a_2 + b_2x, x_2 \leq x \leq x_3, \\ \dots \\ a_{N-1} + b_{N-1}x, x_{N-1} \leq x \leq x_N, \end{cases}$$

где  $a_i, i = 1, 2, \dots, N-1$  – массив вычисленных свободных членов,

$b_i, i = 1, 2, \dots, N-1$  – массив вычисленных коэффициентов.

Для  $x < x_1$  и для  $x > x_N$  модель не определена.

Вычисления коэффициентов на основе представленных опытных данных производятся по формулам:

$$a_i = \frac{y_i x_{i+1} - y_{i+1} x_i}{x_{i+1} - x_i}, i = 1, 2, \dots, N-1,$$

$$b_i = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, i = 1, 2, \dots, N-1.$$

Очевидно, что вычисления производятся «точно», поэтому вывод статистических характеристик для данной модели подобно тому, как это сделано для других моделей, не имеет смысла.

В приложениях иногда возникает необходимость вычисления значения ординаты по обращенной модели. Для расчета требуемого значения следует воспользоваться кусочной формулой

$$x(y) = \begin{cases} \frac{y - a_1}{b_1}, y_1 \leq y \leq y_2, \\ \frac{y - a_2}{b_2}, y_2 \leq y \leq y_3, \\ \dots \\ \frac{y - a_{N-1}}{b_{N-1}}, y_{N-1} \leq y \leq y_N. \end{cases}$$



## Глава 10. Корреляционный анализ

### 10.1. Введение

Исследуется корреляция и связи типа корреляции:

- количественных признаков,
- порядковых признаков,
- номинальных признаков,
- смешанных признаков,
- разнородных признаков.

Также выполняется канонический корреляционный анализ.

### 10.2. Теоретическое обоснование

В практических наблюдениях часто бывают случаи, когда зависимости не имеют функционального характера – равномерному изменению одного признака соответствует изменение величины другого признака в среднем. Такой вид соотношений называется корреляционной зависимостью, или корреляцией. Корреляционным анализом называется совокупность методов обнаружения корреляционной зависимости между случайными величинами или признаками. Считается, что исследование взаимной зависимости приводит к теории корреляции, тогда как изучение зависимости ведет к теории регрессии.

Выделяется также случай функциональной зависимости между величинами, измерения которых, возможно, подвержены ошибкам наблюдений или измерений. Под функциональной связью понимается такой род соотношения между двумя признаками, когда любому значению одного признака всегда соответствует определенное одно и то же значение другого. Функциональная зависимость отражает физические взаимосвязи изучаемого явления и может изучаться методами математического моделирования. Подробнее данные вопросы рассматриваются в главе «Регрессионный анализ».

Предполагается, что объекты располагаются по столбцам электронной таблицы. По строкам располагаются параметры, описывающие объекты. Это существенно для многомерных методов, оперирующих матрицами исходных данных. Если требуется повести исследование транспонированной матрицы исходных данных, для быстрого выполнения данной операции можно воспользоваться методом главы «Матричная и линейная алгебра».

Отметим, что возможность вычисления корреляционной матрицы, в том числе для признаков различных и смешанных типов, позволяет использовать корреляционную матрицу для факторного анализа указанных типов признаков, реализованного в главе «Факторный анализ».

#### 10.2.1. Корреляция количественных признаков

В данном разделе представлены методы исследования корреляции количественных признаков:





- коэффициент корреляционного отношения Пирсона, применяемый для измерения тесноты связи при прямолинейной корреляции,
- коэффициент корреляции Фехнера.

Дополнительно предоставлена возможность расчета ковариации и ковариационной матрицы. Данный показатель может быть необходим для применения в других методах, например, для «ручного» расчета критерия Уилкса, описанного в главе «Дисперсионный анализ».

Коэффициенты ранговой корреляции, которые исследуют корреляцию порядковых признаков (рангов), пусть и полученных из признаков количественных (путем применения операции присвоения рангов), помещены в раздел «Корреляция порядковых признаков».

Полученные в результате применения линейных методов корреляционного анализа выводы могут подтвердить или опровергнуть гипотезу о существовании линейной зависимости между рядами, но не связи другого типа. Вывод в этом случае такой: чем ближе вычисленная величина корреляционного отношения к 0, тем слабее сила линейной связи между рядами, чем ближе вычисленная величина к значению +1 (полная положительная корреляция) или к значению -1 (полная отрицательная корреляция), тем сильнее сила линейной связи.

## 10.2.1. Коэффициент корреляционного отношения Пирсона

Коэффициент корреляционного отношения Пирсона (коэффициент корреляции, выборочный коэффициент корреляции, коэффициент корреляции Бравайса – Пирсона) измеряет силу линейной корреляционной связи количественных признаков. Выборочная оценка коэффициента корреляции вычисляется по формуле

$$\hat{r} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

где  $x_i, i = 1, 2, \dots, n$  – варианты первой компоненты 2-мерной выборки,

$\bar{x}$  – соответствующее среднее значение,

$y_i, i = 1, 2, \dots, n$  – варианты второй компоненты 2-мерной выборки,

$\bar{y}$  – соответствующее среднее значение,

$n$  – численность 2-мерной выборки.

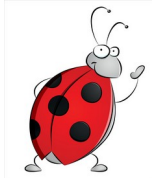
Иначе коэффициент корреляции может оказаться удобным вычислить как

$$\hat{r} = \frac{Cov(X, X)Cov(Y, Y)}{\sqrt{Cov(X, Y)}},$$

где  $X$  – первая компонента,

$Y$  – вторая компонента,

$Cov(.,.)$  – выборочная ковариация.



Использование коэффициента корреляции оправдано лишь тогда, когда совместное распределение пары количественных признаков соответствует 2–мерному нормальному распределению. Частой грубой ошибкой в публикациях является игнорирование этой предпосылки применения рассматриваемого показателя, поэтому перед вычислением коэффициента Пирсона следует проверить нормальность 2–мерной выборки с помощью методов из главы «Проверка нормальности распределения».

При  $|\hat{r}| < 1$  вычисляется еще ряд параметров. Доверительный интервал оцениваемого коэффициента корреляции нормальной двумерной генеральной совокупности вычисляется как

$$r \in \left[ \tanh \left( z(\hat{r}) - \frac{N_{(1+p)/2}}{\sqrt{n-3}} \right); \tanh \left( z(\hat{r}) + \frac{N_{(1+p)/2}}{\sqrt{n-3}} \right) \right],$$

где  $r$  – истинное значение коэффициента корреляции,

$N_{(1+p)/2}$  – квантиль нормального распределения,

$p$  – стандартное значение доверительного уровня,

$z(\cdot)$  –  $z$ –преобразование выборочного коэффициента корреляции.

Нормализующее  $z$ –преобразование выборочного коэффициента корреляции вычисляется как гиперболический арктангенс действительной переменной по формуле

$$z(\hat{r}) = \text{Arth}(\hat{r}) = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}} \quad \text{при } |\hat{r}| < 1.$$

Коэффициент корреляции может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. В случае нормального распределения исходных данных величина выборочного коэффициента корреляции считается значимо отличной от нуля, если выполняется неравенство

$$r^2 > \left[ 1 + (n-2)/t_\alpha^2 \right]^{-1},$$

где  $t_\alpha$  – критическое значение  $t$ –распределения с  $n-2$  степенями свободы.

Иначе говоря, величина

$$t_r = r \sqrt{\frac{n-2}{1-r^2}}$$

имеет  $t$ –распределение с  $n-2$  степенями свободы.

Распределение величины  $z(r)$  уже при небольших значениях  $n$  приближается нормальным распределением с математическим ожиданием, равным

$$Mz = \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{r}{2(n-1)} + \dots$$







и дисперсией

$$Dz = \frac{1}{n-3} + \dots,$$

где опущены слагаемые, малые по сравнению с оставленными слагаемыми.

Случайная величина

$$\frac{z(r) - Mz}{\sqrt{Dz}}$$

распределена приближенно по стандартному нормальному закону  $N(0,1)$ .

При исследовании многомерной совокупности случайных величин из коэффициентов корреляции, вычисленных попарно между случайными величинами, составляется квадратная симметрическая корреляционная матрица с единицами на главной диагонали, которая служит основным элементом при построении многих алгоритмов многомерной статистики, например, в факторном анализе.

Вывод см. в учебном пособии Львовского. О вычислении коэффициента корреляции, включая доверительные интервалы, см. монографию Айвазяна с соавт., работы Альтмана (Altman) с соавт., таблицы Большева с соавт., Мюллера с соавт. О проверке значимости см. также монографии Ферстера с соавт. О сравнении коэффициентов корреляции двух независимых совокупностей см. также монографии Мюллера с соавт., Родионова, а также работы Уильямса (Williams), Вольфе (Wolfe), Лемешко с соавт.

## 10.2.1.2. Коэффициент корреляции Фехнера

Коэффициент корреляции Фехнера (фехнеровский коэффициент корреляции, индекс Фехнера) был предложен для изучения корреляции количественных признаков. При вычислении коэффициента происходит понижение количественной шкалы до номинальной шкалы. В расчетах участвуют только количественные признаки (по ним вычисляются средние значения), поэтому метод представлен в разделе, посвященном количественным признакам. Вычисления производятся по формуле

$$r_F = \frac{C - H}{C + H},$$

где  $C$  – число совпадений знаков отклонений вариант от соответствующих средних значений,  $H$  – число несовпадающих знаков.

Коэффициент корреляции Фехнера с успехом применяется также и для изучения корреляции «чисто» номинальных признаков. В этом случае, в соответствующих обозначениях, приведенная формула может быть записана как

$$r_F = \frac{a + d - b - c}{a + b + c + d},$$

где  $a, b, c, d$  – значения в клетках таблицы  $2 \times 2$ .



Коэффициент корреляции Фехнера может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи). В этом случае можно произвести вычисление по формуле

$$t_F = \frac{a + d - b - c - 1}{\sqrt{a + b + c + d}} = \frac{r_F n - 1}{\sqrt{n}},$$

где  $n = a + b + c + d$ .

Критические значения статистики  $t_F$  приближенно распределены по стандартному нормальному закону  $N(0,1)$ .

См. монографии Лакина, Ферстера с соавт.

### 10.2.1.3. Ковариация

Ковариация (covariance) – числовая характеристика совместного распределения двух случайных величин  $X$  и  $Y$ . Иногда говорят о 2–мерной случайной величине, причем под  $X$  понимают первую компоненту, а под  $Y$  – вторую компоненту указанной случайной величины.

Ковариация определяется формулой

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)],$$

где  $E$  – символ математического ожидания.

Значение  $\text{Cov}(X, X)$  по определению является дисперсией случайной величины  $X$ .

Выборочная ковариация вычисляется как

$$\text{Cov}(X, Y) = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

где  $x_i, i = 1, 2, \dots, n$  – варианты первой компоненты 2–мерной выборки,

$\bar{x}$  – соответствующее среднее значение,

$y_i, i = 1, 2, \dots, n$  – варианты второй компоненты 2–мерной выборки,

$\bar{y}$  – соответствующее среднее значение,

$n$  – численность 2–мерной выборки.

Ковариация не может служить в качестве показателя типа корреляции, т. к. не обладает свойствами данного показателя. В частности, ковариация не является безразмерной величиной. Ее максимальное значение не ограничивается единицей.

Ковариация предлагается здесь из–за технического удобства вычисления. Необходимость вычисления ковариации вызвана тем, что в ряде разновидностей множественного статистического анализа (дисперсионный анализ, факторный анализ) находит применение ковариационная матрица, элементами которой служат попарные ковариации компонент случайного вектора.

Теоретическое обоснование см. у Ван дер Вардена.



## 10.2.2. Корреляция порядковых признаков

В данном разделе рассмотрены методы исследования связи типа корреляции признаков, измеренных в порядковой шкале, либо признаков, приведенных к порядковой шкале, (ранговой корреляции):

- показатель ранговой корреляции Спирмэна,
- коэффициент ранговой корреляции Кендалла.

Обзор коэффициентов ранговой корреляции (включая проверку значимости) см. в работах Филлера (Fieller) с соавт.

### 10.2.2.1. Показатель ранговой корреляции Спирмэна

Показатель ранговой корреляции Спирмэна (показатель корреляции рангов Спирмэна, коэффициент корреляции рангов, коэффициент корреляции Спирмэна, коэффициент ранговой корреляции  $\rho$ , Spearman rank correlation) применяется в случае, если изучается линейная связь между рядами, представленными в количественной или порядковой шкале. Следует заметить, что при анализе количественных признаков применять показатель Спирмэна вместо коэффициента корреляционного отношения Пирсона не следует, если для этого не существует веских оснований, так как при его вычислении происходит понижение количественной шкалы до порядковой шкалы. Поэтому наиболее широкое применение показатель Спирмэна нашел при анализе корреляции порядковых признаков.

Расчет выборочной оценки показателя ранговой корреляции ведется по формуле

$$\hat{\rho}_s = 1 - \frac{6(S_\rho + B_x + B_y)}{n^3 - n}; S_\rho = \sum_{i=1}^n (r_i - s_i)^2,$$

где  $r_i, s_i, i = 1, 2, \dots, n$  – массивы рангов анализируемых рядов,

$n$  – число пар вариант исследуемых рядов,

$B_x, B_y$  – поправки на объединение рангов в соответствующих рядах, вычисляемые по формуле

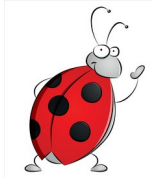
$$B_i = \frac{1}{12} \sum_{i=1}^m n_i (n_i^2 - 1),$$

где  $m$  – число групп объединенных рангов в ряду,

$n_i, i = 1, 2, \dots, m$  – число рангов в  $i$ -ой группе.

Доверительный интервал оцениваемого показателя Спирмэна вычисляется аналогично коэффициенту Пирсона.

Показатель ранговой корреляции Спирмэна может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. Можно также произвести вычисление по формуле



$$t_p = |\rho_s| \sqrt{\frac{n-2}{1-\rho_s^2}},$$

где критические значения статистики  $t_p$  имеют  $t$ -распределение с  $n-2$  степенями свободы. См. т. 2 Справочника под ред. Э. Ллойда и др., монографии Ферстера с соавт., Лакина, Малета с соавт., статью Артузи (Artusi) с соавт.

#### 10.2.2.2. Коэффициент ранговой корреляции Кендалла

Коэффициент ранговой корреляции Кендалла (коэффициент корреляции рангов, ранговый коэффициент корреляции, коэффициент корреляции Кендэла,  $\tau$  Кендалла, Kendall rank correlation) предназначен для вычисления силы корреляционной связи между двумя рядами при тех же условиях, что и рассмотренный выше показатель Спирмэна. Коэффициент Кендалла считается более строгой оценкой по сравнению с показателем ранговой корреляции Спирмэна.

Все основные замечания, данные при описании показателя Спирмэна, справедливы и в отношении коэффициента Кендалла.

Расчет выборочной оценки коэффициента ранговой корреляции ведется по формуле

$$\hat{\tau} = \frac{S_\tau}{\sqrt{\left(\frac{n(n-1)}{2} - B_x\right)\left(\frac{n(n-1)}{2} - B_y\right)}}, \quad S_\tau = \sum_{i=1}^n \sum_{j=i+1}^n \text{sign}(r_j - r_i),$$

где  $r_i, s_i, i=1, 2, \dots, n$  – массивы рангов анализируемых рядов,

$n$  – число пар вариант исследуемых рядов,

$B_x, B_y$  – поправки на объединение рангов в соответствующих рядах, вычисляемые по формуле

$$B_x = \frac{1}{2} \sum_{i=1}^m n_i(n_i - 1),$$

где  $m$  – число групп объединенных рангов в ряду,

$n_i, i=1, 2, \dots, m$  – число рангов в  $i$ -ой группе.

Доверительный интервал оцениваемого коэффициента Кендалла может вычисляться разными методами. Доверительный интервал вычисляется по формуле Нётера

$$\tau \in \left[ \hat{\tau} - \frac{2\sigma\Psi(1 - (1-p)/2)}{n(n-1)}; \hat{\tau} + \frac{2\sigma\Psi(1 - (1-p)/2)}{n(n-1)} \right],$$

где  $\Psi(\cdot)$  – функция, обратная функции стандартного нормального распределения,

$p$  – стандартное значение доверительного уровня,

$\sigma$  – величина, определяемая из формулы:



$$\sigma^2 = 4 \sum_{i=1}^n C_i^2 - 2 \sum_{i=1}^n C_i - \frac{2(2n-3)}{n(n-1)} \left( \sum_{i=1}^n C_i \right)^2,$$

где  $C_i, i=1,2,\dots,n$ , – вспомогательные величины, вычисляемые как

$$C_i = \sum_{\substack{j=1 \\ i \neq j}}^n \delta(r_i, r_j, s_i, s_j), i=1,2,\dots,n,$$

где  $\delta(\dots)$  – величины, вычисляемые по формуле

$$\delta(a,b,c,d) = \begin{cases} 1, (a-b)(c-d) > 0, \\ 0, (a-b)(c-d) < 0. \end{cases}$$

Проблема, однако, заключается в том, что для некоторых наборов данных величина  $\sigma^2$ , рассчитанная по показанной выше формуле, может оказаться отрицательной. Пример таких данных:

1,059	1,242
1,091	1,237
1,849	1,11
1,943	2,691
2,416	1,352
5,134	5,705
5,29	4,055
7,344	3,257
7,435	3,772

В этом случае метод Нётера оказывается несостоятельным и доверительный интервал вычисляется как

$$\tau \in \left( \hat{\tau} - t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}}; \hat{\tau} + t_{(1+\beta)/2} \frac{\sigma}{\sqrt{n}} \right),$$

где  $\sigma$  – стандартное отклонение,

$t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $n-1$  и  $(1+\beta)/2$ ,

$\beta$  – доверительный уровень, выраженный в долях.

При этом стандартное отклонение считается как

$$\sigma = \sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

Коэффициент ранговой корреляции Кендалла может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. В случае больших выборок можно произвести вычисление по формуле



$$t_r \approx \tau \sqrt{\frac{9n(n-1)}{2(2n+5)}},$$

где критические значения статистики  $t_r$  приближенно распределены по стандартному нормальному закону  $N(0,1)$ .

См. т. 2 Справочника под ред. Э. Ллойда и др., монографии Холлендера с соавт., статью Самара (Samara) с соавт.

### 10.2.3. Корреляция номинальных признаков

В данном разделе представлены методы исследования связи типа корреляции для признаков, измеренных в номинальной шкале либо приведенных к номинальной шкале. Особо отметим введенную выше поправку «типа корреляции», т. к. обычная корреляция для номинальных признаков не определена.

Рассмотрены коэффициенты (показатели подобия)

- Рассела–Рао,
- Бравайса.

Коэффициенты предназначены для оценки связи между дихотомическими (номинальными с числом градаций, равным двум, иначе качественными) признаками. Эти и другие показатели находят широкое применение в кластерном анализе, где они именуются также мерами сходства типа корреляции. Подробнее см. главу «Кластерный анализ».

Для исследования корреляции признаков, измеренных в номинальной шкале с числом градаций признаков больше двух (категоризированных данных), используются методы анализа двумерных таблиц сопряженности (кросстабуляции), выполняемого с помощью методов главы «Кросстабуляция». Представленные показатели принято именовать мерами связи.

Предполагается, что дихотомическая переменная может принимать только значения 1 (верхний уровень) и 0 (нижний уровень). Значение варианты выборки, равное 0, указывает на отсутствие переменной или признака, значение, равное 1 – на наличие. Например, в ячейку (клетку)  $a$  записано число пар элементов массивов 1 и 2, одновременно имеющих признак, равный 1. В ячейку  $c$  записано число пар элементов массивов 1 и 2, в которых значение элемента массива 1 равно 1, а значение элемента массива 2 равно 0 и т. д. Отметим, что при вычислении представленных показателей путаница между ячейками  $b$  и  $c$  не ведет к каким-либо неприятностям.

#### 10.2.3.1. Коэффициент Рассела–Рао

Коэффициент Рассела–Рао (показатель подобия Рассела–Рао) вычисляется по формуле

$$r = \frac{a}{a+b+c+d},$$

где  $a, b, c, d$  – значения в клетках таблицы  $2 \times 2$ .



Коэффициент Рассела–Рао может применяться для проверки гипотезы независимости признаков (значимости связи). В этом случае статистика

$$t_r = \frac{r \cdot n - 1}{\sqrt{n}},$$

где  $n = a + b + c + d$  – сумма таблицы,

приближенно распределена по стандартному нормальному закону  $N(0,1)$ .

### 10.2.3.2. Коэффициент сопряженности Бравайса

Специальная форма коэффициента корреляции – коэффициент сопряженности Бравайса ( $\phi$  – коэффициент ассоциации Пирсона, коэффициент контингенции Пирсона, тетрафорический показатель связи) – рассчитывается по формуле

$$\phi = \frac{ad - bc - 0,5 \cdot n}{\sqrt{(a+b)(a+c)(d+b)(d+c)}},$$

где  $a, b, c, d$  – значения в клетках таблицы  $2 \times 2$ .

$0,5 \cdot n$  – поправка на непрерывность Йэйтса,

$n = a + b + c + d$  – сумма таблицы,

В литературе встречаются и иные, эквивалентные, формулировки рассматриваемого коэффициента. Например, показанная формула фактически совпадает с формулой

$$\phi = \sqrt{\frac{\chi^2}{n}},$$

где  $\chi^2$  – статистика хи–квадрат Пирсона.

Коэффициент сопряженности Бравайса может применяться для проверки гипотезы независимости признаков (значимости связи). В этом случае статистика

$$t_\phi = |\phi| \sqrt{\frac{n-2}{1-\phi^2}}$$

приближенно имеет  $t$ –распределение с  $n - 2$  степенями свободы.

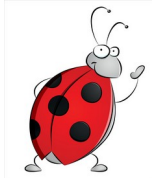
См. монографии Лакина, Ферстера с соавт., Малета с соавт.

### 10.2.4. Корреляция признаков, измеренных в различных шкалах

Настоящий раздел посвящен исследованию корреляции признаков, измеренных в различных (смешанных) шкалах. Рассмотрены:

- коэффициент Гауэра,
- точечно–бисериальная корреляция, позволяющая исследовать корреляцию в некоторых частных случаях.

Интересной возможностью является исследование корреляции разнородных признаков.



## 10.2.4.1. Коэффициент Гауэра

Коэффициент Гауэра допускает одновременное использование признаков, измеренных в шкалах: количественной, порядковой и дихотомической. Могут анализироваться выборки (например, описывающие параметры пациента), содержащие в себе признаки различных типов. Так, часть параметров может быть количественной (например, результаты инструментальных измерений), часть – порядковой (например, результаты исследований в баллах), часть – дихотомической (например, наличие или отсутствие некоторых симптомов). Вычисление элемента матрицы сходства, построенной на основе коэффициента Гауэра, производится по формуле:

$$s_{ij} = \frac{\sum_{k=1}^p S_{ijk}}{\sum_{k=1}^p W_{ijk}}, i = 1, 2, \dots, n, j = 1, 2, \dots, n,$$

где  $S_{ijk}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, n$ ;  $k = 1, 2, \dots, p$  – вклад признака в сходство объектов,  
 $W_{ijk}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, n$ ;  $k = 1, 2, \dots, p$  – весовая переменная признака,  
 $p$  – число признаков, характеризующих объект,  
 $n$  – число объектов.

## 10.2.4.1. Расчет вклада признаков

- Для дихотомических признаков алгоритм подсчета вклада признака и взятия весовых переменных совпадает с коэффициентом Жаккара

$$J = \frac{a}{a + b + c},$$

где  $a, b, c$  – значения в клетках таблицы  $2 \times 2$ .

- Для порядковых признаков алгоритм вычисления вклада признака совпадает с хемминговым расстоянием, если последнее мысленно обобщить на порядковые переменные, а весовые переменные берутся равными 1 для каждого участвующего в расчете порядкового признака.

$$H = a + d,$$

где  $a, b$  – значения в клетках таблицы  $2 \times 2$ .

- Для количественных признаков

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$

где  $x_{ik}$  и  $x_{jk}$  – значения  $k$ -й переменной для объектов  $i$  и  $j$ ,

$R_k$  – размах  $k$ -го признака, вычисленный по всем объектам,

а весовые переменные берутся аналогично случаю порядковых признаков.

Коэффициент Гауэра реализован только для метода «Матрица», поэтому в рассматриваемом случае следует выделить матрицу исходных данных в виде столбцов равной численности.







## 10.2.4.2. Точечно–бисериальная корреляция

Если одна переменная дихотомизирована, а другая измерена в количественной шкале, вычисляется точечно–бисериальный коэффициент корреляции (точечный двухсерийный коэффициент корреляции). Имеют место несколько эквивалентных формул вычисления выборочной оценки коэффициента, например,

$$\hat{r}_{pb} = \frac{(\bar{x}_1 - \bar{x}_0) \cdot \sqrt{n_1 n_0}}{s_n \cdot n},$$

где  $\bar{x}_1$  – среднее вариант количественной выборки, соответствующих событиям верхнего уровня дихотомической выборки,

$\bar{x}_0$  – среднее вариант количественной выборки, соответствующих событиям нижнего уровня дихотомической выборки,

$s_n$  – среднее квадратичное значение количественной выборки,

$n_1$  – число событий верхнего уровня,

$n_0$  – число событий нижнего уровня.

Средние значения вычисляются по формулам, соответственно,

$$\bar{x}_1 = \frac{1}{n_1} \sum_{\substack{i=1 \\ a_i=1}}^n x_i \quad \bar{x}_0 = \frac{1}{n_0} \sum_{\substack{i=1 \\ a_i=0}}^n x_i,$$

где  $x_i, i = 1, 2, \dots, n$  – количественная выборка,

$a_i, i = 1, 2, \dots, n$  – дихотомическая выборка,

$n = n_1 + n_0$  – численность пар анализируемых выборок.

Предполагается, что дихотомическая переменная может принимать только значения 1 (верхний уровень) и 0 (нижний уровень).

Выборочное среднее квадратичное отклонение вычисляется по формуле

$$s_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2},$$

где  $\bar{x}$  – выборочное среднее, которое вычисляется по формуле

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

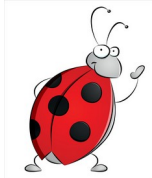
При  $|\hat{r}_{pb}| < 1$  вычисляется еще ряд параметров. Доверительный интервал оцениваемой точечно–бисериальной корреляции вычисляется как

$$r_{pb} \in \left[ \tanh \left( z(\hat{r}_{pb}) - \frac{N_{(1+p)/2}}{\sqrt{n-3}} \right); \tanh \left( z(\hat{r}_{pb}) + \frac{N_{(1+p)/2}}{\sqrt{n-3}} \right) \right],$$

где  $N_{(1+p)/2}$  – квантиль нормального распределения,

$p$  – стандартное значение доверительного уровня,





$z(\cdot)$  –  $z$ -преобразование выборочного коэффициента корреляции.

Нормализующее  $z$ -преобразование выборочного точечно-бисериального коэффициента корреляции вычисляется как гиперболический арктангенс действительной переменной по формуле

$$z(\hat{r}_{pb}) = \text{Arth}(\hat{r}_{pb}) = \frac{1}{2} \ln \frac{1 + \hat{r}_{pb}}{1 - \hat{r}_{pb}} \quad \text{при } |\hat{r}_{pb}| < 1.$$

Точечно-бисериальный коэффициент корреляции может применяться для проверки гипотезы независимости признаков (некоррелированности, значимости связи) следующим образом. Гипотеза о независимости признаков отвергается на выбранном уровне значимости, если вычисленное по опытным данным значение коэффициента корреляции превосходит (по модулю) критическое. В случае нормального распределения исходных данных величина выборочного коэффициента корреляции считается значимо отличной от нуля, если выполняется неравенство

$$r_{pb}^2 > \left[ 1 + (n - 2)/t_\alpha^2 \right]^{-1},$$

где  $t_\alpha$  – критическое значение  $t$ -распределения с  $n - 2$  степенями свободы.

Иначе говоря, величина

$$t_r = r_{pb} \sqrt{\frac{n - 2}{1 - r_{pb}^2}}$$

имеет  $t$ -распределение с  $n - 2$  степенями свободы.

Отметим, что результаты вычисления точечно-бисериального коэффициента корреляции и коэффициента корреляции Пирсона, хотя и при различных исходных предпосылках, в случае формальной подстановки в формулу последнего тех же числовых данных, совпадают. См. монографии Лакина, Мак-Немара (McNemar).

## 10.2.5. Корреляция разнородных признаков

Для исследования корреляции признаков, измеренных в смешанных шкалах (один объект описывается вектором данных, принадлежащих к различным шкалам), применяется коэффициент Гауэра. Метод применяется для исследования корреляции объектов. Не менее часто возникает задача исследования корреляции разнородных признаков, для решения которых предназначена описываемая ниже опция.

При построении корреляционной матрицы разнородных действуют следующие правила вычисления коэффициентов корреляции:

- При вычислении корреляции двух количественных параметров – коэффициент Пирсона.
- При вычислении корреляции порядковых/количественных и порядковых параметров – коэффициент ранговой корреляции Кендалла.
- При вычислении корреляции двух дихотомических признаков – коэффициент сопряженности Бравайса.



- При вычислении корреляции количественных/порядковых и дихотомических признаков – точно-бисериальная корреляция.

Из вычисленных корреляций формируется общая корреляционная матрица.

## 10.2.6. Канонический корреляционный анализ

Канонический корреляционный анализ выполняется между двумя совокупностями (группами) выборок и предназначен для определения линейной функции от первых  $p$  компонент и линейной функции от остальных  $q$  компонент так, чтобы коэффициент корреляции между этими линейными функциями принял наибольшее из возможных значений. Численности групп (количество выборок в первой и второй группах, обозначены как  $p$  и  $q$ ) могут различаться, однако необходимым требованием является равное количество вариант во всех выборках, составляющих обе группы. Матрица взаимной корреляции двух групп выборок имеет вид

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix},$$

где  $R_{11}$  – матрица взаимной корреляции  $p$  переменных 1-й группы, размер  $p \times p$ ,

$R_{22}$  – матрица взаимной корреляции  $q$  переменных 2-й группы, размер  $q \times q$ ,

$R_{12}$  – матрица взаимной корреляции переменных 1-й и 2-й группы, размер  $p \times q$ .

Решение задачи сводится к обобщенной проблеме собственных значений

$$R_{12}^T R_{11}^{-1} R_{12} \cdot v = \lambda \cdot R_{22} \cdot v,$$

где  $\lambda$  – вектор  $q$  собственных значений.

Так называемые канонические корреляции представляют собой квадратные корни из

собственных значений. Выводятся значения критерия  $\chi^2$  (массив длиной  $q$ ) и

соответствующие степени свободы (массив длиной  $q$ ), а также коэффициенты правой (массив размером  $q \times q$ ) и левой (массив размером  $q \times p$ ) стороны.

См. Сборник научных программ на Фортране, главу 10 монографии Итона (Eaton).

## Глава 11. Факторный анализ

### 11.1. Введение

Методы факторного анализа:

- метод главных факторов (если корреляционная матрица не редуцируется – метод главных компонент или компонентный анализ),
- метод максимального правдоподобия.

### 11.2. Теоретическое обоснование

Методами факторного анализа решаются три основных вида задач:





- отыскание скрытых, но предполагаемых закономерностей, которые определяются воздействием внутренних или внешних причин (факторов) на изучаемый процесс;
- выявление и изучение статистической связи признаков с факторами или главными компонентами;
- сжатие информации путем описания процесса при помощи общих факторов или главных компонент, число которых меньше количества первоначально взятых признаков (параметров), однако с той или иной степенью точности обеспечивающих воспроизводимость корреляционной матрицы.

Следует пояснить, что в факторном анализе понимается под сжатием информации. Дело в том, что корреляционная матрица получается путем обработки исходного массива данных. Корреляционная матрица образована из попарных коэффициентов корреляции компонент случайного вектора. Предполагается, что та же самая корреляционная матрица может быть получена с использованием тех же объектов, но описанных меньшим числом параметров. Якобы происходит уменьшение размерности задачи, хотя на самом деле это не так. Это не сжатие информации в общепринятом смысле – восстановить исходные данные по корреляционной матрице нельзя.

Основное требование к исходным данным для факторного анализа – это то, что они должны подчиняться многомерному нормальному распределению. По крайней мере, должно быть сделано допущение о многомерном нормальном распределении совокупности. Нормальность распределения может быть проверена с помощью методов, представленных в главе «Проверка нормальности распределения».

Основным объектом исследования методами факторного анализа является корреляционная матрица, построенная с использованием коэффициента корреляции (корреляционного отношения) Пирсона, разработанного для количественных признаков. Напомним, что коэффициентом корреляции называется безразмерная числовая характеристика совместного распределения двух случайных величин, выражающая их взаимосвязь. Чем ближе коэффициент корреляции к 1 или  $-1$ , тем сильнее эта взаимосвязь. Чем ближе к 0, тем взаимосвязь слабее. Подробнее см. главу «Корреляционный анализ».

Некоторые авторы предлагают использовать для факторного анализа дисперсионно-ковариационную матрицу, построенную из дисперсий-ковариаций. Дисперсионно-ковариационная (ковариационная) матрица образована из попарных ковариаций компонент случайного вектора. Ковариация случайной величины сама с собой, как известно, является дисперсией. В отличие от коэффициента корреляции, определяемого через ковариацию, последняя не является безразмерной величиной, поэтому менее удобна для применения в факторном анализе. В дальнейших рассуждениях, за исключением некоторых моментов (например, проблемы общности), говоря о корреляции, будем иметь в виду также и ковариацию.

В литературе предлагается также использование других коэффициентов типа корреляции, предназначенных для порядковых, качественных и смешанных признаков. Нужно только рассчитать заранее корреляционную матрицу с помощью методов, реализованных в главе «Корреляционный анализ». Дальнейший анализ – стандартный.



Основным требованием к построенной матрице является ее положительная полуопределенность. Свойства матриц подробно рассмотрены в соответствующих источниках. Из свойства положительной полуопределенности следует неотрицательность всех собственных значений.

Коэффициенты корреляции, составляющие корреляционную матрицу, по умолчанию вычисляются между параметрами (признаками, тестами), а не между объектами (индивидуумами, лицами), поэтому размерность корреляционной матрицы равна числу параметров. Это так называемая техника *R*. Однако может быть, например, изучена корреляция между объектами (точнее, их состояниями, описываемыми векторами параметров). Эта методика называется техникой *Q*. Проведение факторного анализа техникой *Q* обосновано тем, что состояния объектов могут иметь общую побудительную причину (причины), которая (которые) может быть выявлена с помощью факторного анализа. Существует также техника *P*, предполагающая факторный анализ результатов экспериментальных исследований, выполненных на одном и том же индивидууме в различные промежутки времени («объекты» – один и тот же индивидуум в различные промежутки времени), причем изучаются корреляции между состояниями индивидуума. Аналог техники *Q* для последнего случая составляет предмет исследования техники *O*. Применение техники *R* (*P*) или техники *Q* (*O*) или выбор техники *R* (*Q*) или *P* (*O*) осуществляются с помощью одних и тех же алгоритмов.

Получение матрицы факторного отображения в принципе является целью факторного анализа. Ее строки представляют собой координаты концов векторов, соответствующих  $m$  переменным в  $r$ -мерном факторном пространстве. Близость концов этих векторов дает представление о взаимной зависимости переменных. Каждый вектор в сжатой, концентрированной форме несет информацию о процессе. Близость этих векторов дает представление о взаимной зависимости переменных. Дополнительно, если число выделенных факторов больше единицы, производится вращение матрицы факторного отображения с целью получения так называемой простой структуры.

Для наглядности результаты можно изобразить графически, что, однако, проблематично для трех и более выделенных факторов. Поэтому дают изображение  $r$ -мерного факторного пространства в двумерных срезах.

При решении задачи факторного анализа возможна ситуация, когда вектора исходных данных коллинеарны (параметры линейно зависимы). Напомним, что два вектора называются коллинеарными, если они лежат на параллельных прямых или на одной прямой. В таком случае при решении возможно получение различных вычислительных проблем.

Корреляционная матрица для таких данных может оказаться вырожденной. Применяемый для определения собственных значений метод дает решение и в этом случае. При этом часть собственных значений, равная разности порядка матрицы и ее ранга, будет нулевой в вычислительном смысле, что делает метод главных факторов более устойчивым к таким «нехорошим» данным, чем метод максимума правдоподобия. Однако метод главных факторов уступает методу максимума правдоподобия в том, что он не позволяет получить точной оценки общности.



Для выявления мультиколлинеарности специально разработаны эффективные статистические методы, позволяющие выявить, при ее наличии, коллинеарность векторов исходных данных. После обнаружения таких параметров рекомендуется оставить в исходных данных только один из группы линейно зависимых параметров.

Лучшим руководством по факторному анализу является монография Хармана. Пример применения факторного анализа для исходных данных, измеренных не в количественной шкале, см. в работе Каплана.

## 11.2.1. Метод главных факторов

Рассмотрим подробнее метод главных компонент (компонентный анализ, principal components analysis), который по определению Лоули с соавт. представляет собой вариант метода главных факторов (когда корреляционная матрица не редуцируется), а затем сам метод главных факторов (principal factor analysis). В методе главных компонент в качестве исходного элемента анализа может быть использована как корреляционная, так и дисперсионно–ковариационная матрица, причем выводы по результатам анализа тождественны.

### 11.2.1. Компонентный анализ

Основная модель метода главных компонент Хотеллинга записывается в матричном виде следующим образом:

$$Z = AP,$$

где  $Z$  – матрица стандартизованных исходных данных, ее размер  $m \times n$ ,

$A$  – матрица факторного отображения, ее размер  $m \times r$ ,

$P$  – матрица значений факторов, ее размер  $r \times n$ ,

$m$  – количество переменных (векторов данных),

$n$  – количество индивидуумов (элементов одного вектора),

$r, r \leq m$  – количество выделенных факторов.

Как видно из приведенного выше выражения, модель компонентного анализа содержит только общие для имеющихся векторов факторы.

Матрица стандартизованных исходных данных определяется из матрицы исходных данных  $Y$  (ее размер  $m \times n$ ) по формуле

$$z_{ij} = \frac{y_{ij} - \bar{y}_i}{s_i}, i = 1, 2, \dots, m, j = 1, 2, \dots, n,$$

где  $y_{ij}$  – элемент матрицы исходных данных,

$\bar{y}_i$  – среднее значение,

$s_i$  – стандартное отклонение.

Для корреляционной матрицы имеет место соотношение:

$$\frac{1}{n-1} ZZ' = R,$$

где  $R$  – корреляционная матрица, ее размер  $m \times m$ ,





' – символ транспонирования.

На главной диагонали матрицы  $R$  стоят значения, равные 1. Эти значения называются общностями и обозначаются как  $h_i^2$ , являясь мерой полной дисперсии переменной. Для метода главных факторов общности отличны от 1 и вычисляются определенным образом. Неизвестными являются матрицы  $A$  и  $P$ . Матрица  $A$  может быть найдена из основной теоремы факторного анализа

$$R = ACA',$$

где  $C$  – корреляционная матрица, отражающая связь между факторами.

Если  $C = I$ , то говорят об ортогональных факторах, если матрица  $C$  не равна  $I$ , говорят о косоугольных факторах. Здесь  $I$  – единичная матрица.

Для матрицы  $C$  справедливо соотношение

$$\frac{1}{n-1} PP' = C.$$

Нами рассматривается только случай ортогональных факторов, для которых

$$R = AA'.$$

Модель классического факторного анализа содержит ряд общих факторов и по одному характерному фактору на каждую переменную. Число главных компонент всегда меньше либо равно числу переменных.

## 11.2.1.2. Факторный анализ методом главных факторов

По утверждению Хармана, «под методом главных факторов понимают приложение метода главных компонент к редуцированной корреляционной матрице (т. е. к матрице, у которой на главной диагонали вместо единиц стоят значения общностей)». Для метода главных факторов (факторного анализа методом главных факторов Томсона) основная модель записывается в виде

$$Z = FP',$$

где  $F$  – полная факторная матрица, ее размер  $m \times (r + m)$ ,

$P'$  – матрица значений факторов, включая значения характерных факторов, ее размер  $(r + m) \times n$ .

Матрица  $F$  может быть представлена в виде суммы двух матриц

$$F = A + U,$$

где  $A$  – матрица нагрузок общих факторов,

$U$  – матрица нагрузок характерных факторов.

Очевидно, что матрицы  $A$  и  $U$  имеют размер матрицы  $F$ . Матрица  $A$  (а именно ее часть размером  $m \times r$ , остальная же ее часть размером  $m \times m$  является нулевой) понимается как матрица факторного отображения.

Полная дисперсия переменной складывается из общности  $h_i^2$ , значение которой меньше либо равно 1 и означает часть полной дисперсии переменной, приходящейся на главные факторы,





и характерности, обозначаемой как  $u_i^2$ , приходящейся на характерные факторы.

Следовательно,

$$u_i^2 = 1 - h_i^2.$$

Часть размером  $m \times r$  матрицы  $U$  является нулевой, остальная ее часть (размером  $m \times m$ ) представляет собой диагональную матрицу с квадратными корнями из характерностей на главной диагонали, которые уже вычислены из общностей. Таким образом, может быть определена матрица  $U$ , а следовательно, и матрица  $F$ , если известна матрица  $A$ .

Используя введенную выше основную теорему для случая ортогональных факторов, можно записать

$$R = FF'.$$

Развернув выражение, получим:

$$R = R_h + U^2,$$

где  $R$  – корреляционная матрица с единицами на главной диагонали,

$R_h$  – корреляционная матрица с общностями на главной диагонали, определяемая выражением

$$R_h = AA',$$

$U^2$  означает  $UU'$ .

### 11.2.1.3. Проблема общности

Для метода главных факторов имеет место проблема общности, то есть на главной диагонали корреляционной матрицы, в отличие от метода главных компонент, необходимо проставить значения общностей, чтобы получить корреляционную матрицу  $R_h$ .

По определению, общность – сумма квадратов факторных нагрузок. Общность данной переменной – та часть ее дисперсии, которая обусловлена общими факторами. Это вытекает из предположения, что полная дисперсия складывается из общей дисперсии, обусловленной общими для всех переменных факторами, а также специфичной дисперсии, обусловленной факторами, специфичными только для данной переменной, и дисперсии, обусловленной ошибкой. Мы рассматриваем только методы, оперирующие общностями, не превышающими единицу.

Редукцией (редуцированием) корреляционной матрицы в методе главных факторов называется процесс замены единиц на главной диагонали корреляционной матрицы некоторыми величинами, называемыми общностями. Без редукции, то есть с единицами на главной диагонали корреляционной матрицы, мы получаем широко известный компонентный анализ (метод главных компонент).

Способы оценки общностей:

1. Способ наибольшей корреляции.
2. Коэффициент множественной корреляции, при этом общности вычисляются с помощью выражения

$$h_i^2 = 1 - \frac{1}{r^{ii}},$$







где в знаменателе стоит диагональный элемент матрицы, обратной к матрице  $R_h$ . Этот метод, однако, осложняется тем, что полученная в результате редукции корреляционная матрица может не являться матрицей Грама. Замена же диагональных членов оценками общностей считается допустимой, только если сохраняются свойства матрицы Грама. Напомним, что Эрмитова матрица называется положительно полуопределенной (матрицей Грама), если все ее главные миноры неотрицательны.

3. Средние по столбцу корреляционной матрицы коэффициенты корреляции.
4. Метод триад.

#### 11.2.1.4. Проблема факторов

Матрица факторного отображения определяется для компонентного анализа или для метода главных факторов методом множителей Лагранжа (максимизация функции, связанная с дополнительными условиями) из решения проблемы собственных значений матрицы  $R$  или  $R_h$ . Для простоты записи значок  $h$  далее опускаем, имея в виду  $R$  или  $R_h$ , в зависимости от применяемой разновидности метода факторного анализа.

Факторы пропорциональны собственным векторам матрицы  $R$ . Стандартная проблема собственных значений матрицы  $R$  записывается в виде:

$$(R - \lambda_l I) = 0,$$

где  $\lambda_l$ ,  $l = 1, 2, \dots, m$  –  $l$ -е собственное значение матрицы  $R$ ,

$l$  – номер собственного значения.

Результатом расчета будет матрица факторного отображения  $A$  размером  $m \times r$ ,  $m$  – количество переменных (векторов данных),  $r$ ,  $r \leq m$  – количество выделенных факторов.

Данная матрица состоит из элементов (векторов длиной  $m$ )  $a_l$ ,  $l = 1, 2, \dots, r$ ;  $r \leq m$  – соответствующих  $l$ -му собственному значению собственных векторов матрицы  $R$ .

Задача упрощается тем, что матрица  $R$  является действительной и симметрической, поэтому для решения проблемы собственных значений применимы хорошо разработанные эффективные устойчивые алгоритмы.

#### 11.2.1.5. Измерение факторов

Оценка значений факторов (так называемое измерение факторов) не является необходимой для интерпретации результатов процедурой. Остановимся на ней для полноты изложения.

Способ измерения главных компонент основан на применении основной модели факторного анализа:

$$Z = AP.$$

Умножив обе части равенства на  $A'$ , а затем на  $(A'A)^{-1}$ , получим

$$P = A^+ Z,$$

где  $A^+$  – матрица, определяемая по формуле

$$A^+ = (A'A)^{-1}A'.$$

Способ измерения главных факторов основан множественном регрессионном анализе (см. главу «Распознавание образов с обучением»).





## 11.2.2. Метод максимума правдоподобия

В факторном анализе может применяться метод максимума правдоподобия (метод максимального правдоподобия Лоули, maximum-likelihood solution). В методе максимума правдоподобия в качестве исходного элемента анализа может быть использована корреляционная, но не дисперсионно-ковариационная матрица, хотя мы предоставили пользователям возможность поэкспериментировать.

Оценка общностей до применения метода не производится – если исследователь отметит данную опцию, этап редуцирования корреляционной матрицы будет проигнорирован. Общности находятся в результате вычислений из условия полной воспроизводимости, с точностью до ошибки вычислений, редуцированной корреляционной матрицы (не путать с воспроизводимостью матрицы исходных данных!), причем процесс редукции и составляет есть итерационного процесса метода. В этом заключается основное преимущество метода максимума правдоподобия перед методом главных факторов.

Все основные выкладки рассматриваемого метода выполнены Лоули, однако мы дадим основные шаги алгоритма так, как они представлены Харманом:

1. Методом главных компонент вычисляется матрица факторного отображения (удерживаются заданное пользователем количество главных компонент), которая в схеме алгоритма обозначена  $A_{1/2}$ , причем индекс имеет смысл только удобного обозначения матрицы в итерационном процессе.
2. Вычисляется диагональная матрица характеристик  $D_i^2 = \text{diag}(I - A_{i-1/2}A_{i-1/2}')$ , где  $i = 1, 2, \dots$  номер итерации,  $I$  – единичная матрица.
3. Вычисляется матрица  $J_{i-1/2} = A_{i-1/2}'D_i^{-2}A_{i-1/2}$ , предварительно матрица характеристик обращается.
4. Вычисляется диагональная матрица  $J_i = Q_i'J_{i-1/2}Q_i$ , применяя метод вращения Якоби, где  $Q_i$  – матрица вращения.
5. Вычисляется факторная матрица  $A_i = A_{i-1/2}Q_i$ .
6. Вычисляется следующее приближение матрицы факторного отображения  $A_{i+1/2} = (RD_i^{-2} - \Pi)A_iJ_i^{-1}$ , предварительно матрица  $J_i$  обращается.
7. Итерации повторяются, начиная с шага 2, пока не будет выполнено условие  $|A_{i+1/2} - A_{i-1/2}| < \epsilon$ , где  $\epsilon$  – заранее заданное малое положительное число, например, 0,001.

Если задать число факторов равным числу параметров, то оценки общности будут совпадать с общностями нередуцированной корреляционной матрицы, то есть будут равны единице. За счет итерационного подбора общностей любое заданное пользователем число факторов обеспечит полное выделение дисперсий. Максимальное число удерживаемых факторов можно приблизительно установить из анализа процента дисперсии для каждого фактора. Основной недостаток рассматриваемого метода – неустойчивость к данным, содержащим совпадающие или линейно зависимые выборки (коллинеарные вектора исходных данных). С точки зрения многомерной статистики (в широком смысле) проблема заключается в том, что в данном случае, даже если исходные данные показывают многомерное нормальное распределение (см. главу «Проверка нормальности распределения»), оно будет вырожденным. С точки зрения факторного анализа (в узком смысле), будет вырожденной



матрица характеристик. Вряд ли можно избежать численной неустойчивости в данном случае до того, как будет устранена мультиколлинеарность. Об исследовании мультиколлинеарности рассказано в одноименном разделе главы «Матричная и линейная алгебра».

См. источники: Донг (Dong) и де Лью (de Leeuw).

### 11.2.3. Проблема вращения

Оси координат, соответствующие выделенным факторам, ортогональны, и их направления устанавливаются последовательно, по максимуму оставшейся дисперсии. Но полученные координатные оси большей частью содержательно не интерпретируются. Поэтому получают более предпочтительное положение системы координат путем вращения этой системы вокруг ее начала. Пространственная конфигурация векторов в результате применения этой процедуры остается неизменной. Целью вращения является нахождение одной из возможных систем координат для получения так называемой простой факторной структуры. Применяют метод вращения VARIMAX.

Результатом расчета является матрица факторного отображения  $A$  размером  $m \times r$ ,  $m$  – количество переменных (векторов данных),  $r$ ,  $r \leq m$  – количество выделенных факторов.

Данная матрица состоит из элементов (векторов длиной  $m$ )  $a_l$ ,  $l = 1, 2, \dots, r$ ;  $r \leq m$  – соответствующих  $l$ -му собственному значению собственных векторов матрицы  $R$ . В дальнейших рассуждениях более удобным обозначением элементов матрицы факторного отображения будет поэлементная запись  $a_{ij}$ ,  $i = 1, 2, \dots, m$ ;  $j = 1, 2, \dots, r$ .

Метод VARIMAX выполняет ортогональное вращение матрицы факторного отображения, чтобы удовлетворить выражение (нормальный критерий Кайзера, варимакс-критерий)

$$V = \sum_{j=1}^r \left\{ m \sum_{i=1}^m \left( \frac{a_{ij}^2}{h_i^2} \right)^2 - \left[ \sum_{i=1}^m \left( \frac{a_{ij}^2}{h_i^2} \right) \right]^2 \right\} \rightarrow \max,$$

где  $a_{ij}$ ,  $i = 1, 2, \dots, m$ ;  $j = 1, 2, \dots, r$  – факторные нагрузки,

$h_i^2$ ,  $i = 1, 2, \dots, m$ , – суммарная факторная нагрузка, вычисляемая по формуле

$$h_i^2 = \sum_{j=1}^r a_{ij}^2, i = 1, 2, \dots, m.$$

Результатом работы метода VARIMAX будет так называемая повернутая матрица факторного отображения, отличающаяся от исходной пространственной конфигурации переменных в пространстве выделенных факторов тем, что «гроздь» точек, описываемых матрицей, будут располагаться ближе к осям факторного пространства, сохраняя свое взаимное расположение. Считается, что такое вращение помогает улучшить интерпретируемость решения.

Подробное описание метода VARIMAX приводится в первом выпуске «Сборника научных программ на Фортране» и в монографии Хармана. Метод VARIMAX применяется также в



многомерном шкалировании, представленном в одноименной главе, в том смысле, что речь там идет не о пространстве факторов, а о пространстве шкал.

## 11.2.4. Критерии максимального числа факторов

Существует несколько критериев оценки максимального числа удерживаемых (значимых) факторов. Эффективные критерии, основанные на величине собственных значений корреляционной матрицы, в конечном счете, приводят к анализу процента дисперсии, выделенной факторами. Все общие факторы, число которых равно числу параметров, выделяют 100% дисперсии. Данное утверждение справедливо для всех методов факторного анализа. Если сумма процентов дисперсии превышает величину 100%, то это означает: при вычислении собственных значений корреляционной матрицы были получены отрицательные собственные значения и, как следствие, комплексные собственные вектора, что может означать некорректную редукцию исходной корреляционной матрицы.

### 11.2.4.1. Адекватность метода главных факторов

Для методов семейства главных факторов максимальное число удерживаемых факторов можно приблизительно установить из анализа процента дисперсии, выдаваемой для каждого фактора. Резюмируя сказанное, рекомендуем такой порядок действий:

- сначала пользователь проводит «разведочный» факторный анализ без указания максимального числа факторов,
- затем по величине дисперсий приблизительно оценивает необходимое число факторов,
- задавая число факторов, проводит повторный анализ, используя его результаты как окончательные.

### 11.2.4.2. Значимость числа факторов метода максимума правдоподобия

Для метода максимума правдоподобия с целью оценки значимости числа выделенных факторов предложен критерий Уилкса, статистика которого вычисляется по формуле:

$$U_m = n \ln \frac{|R|}{|R_h|},$$

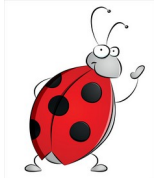
где  $|R|$  – определитель корреляционной матрицы с единицами на главной диагонали,  
 $|R_h|$  – определитель корреляционной матрицы с общностями на главной диагонали,  
 $n$  – количество индивидуумов (элементов одного вектора).

Из-за необходимости знания количества индивидуумов статистика Уилкса вычисляется только в том случае, если исходные данные представляют собой первичные выборки, а не заранее вычисленную корреляционную матрицу.

Распределение статистики Уилкса при больших значениях  $n$  стремится к распределению  $\chi^2$  с числом степеней свободы, равном

$$v = [(m - r)^2 + m - r] / 2,$$





где  $m$  – количество переменных (векторов данных),

$r, r \leq m$  – количество выделенных (или назначенных пользователем) факторов.

Критерий Уилкса применяется также в методе минимальных остатков, подробно описанном Харманом.

## Глава 12. Кластерный анализ

### 12.1. Введение

Рассматриваются методы кластерного анализа, относящиеся к категории методов обучения без учителя (автоматической классификации).

### 12.2. Теоретическое обоснование

Методами кластерного анализа решается задача разбиения (классификации, кластеризации) множества объектов, чтобы все объекты, принадлежащие одному кластеру (классу, группе) были более похожи друг на друга, чем на объекты других кластеров. В отечественной литературе синонимом термина «кластерный анализ» является термин «таксономия». В иностранной литературе под таксономией традиционно понимается классификация видов животных и растений.

Нами рассматриваются следующие методы кластерного анализа:

- Иерархические методы:
  - метод средней связи Кинга,
  - метод Уорда.
- Итеративные методы группировки:
  - метод  $k$ –средних Мак–Куина.

Классифицируемыми могут быть как параметры, так и объекты, поэтому по ходу изложения там, где идет речь о классификации объектов, вполне можно говорить о классификации параметров, и наоборот.

Меры различия накладывают жесткие ограничения на применяемые методы кластерного анализа:

- метод средней связи Кинга можно применять для признаков любых типов: количественных, порядковых, номинальных (как частный случай – экспертных ранжировок) и смешанных признаков
- метод Уорда можно применять только для количественных признаков, т. к. в его схеме применяется только евклидово расстояние
- метод  $k$ –средних Мак–Куина можно применять только для количественных признаков. Для использования метода с целью классификации данных в шкалах, отличной от количественной, требуется модификация метода.

Применяя формально метод, не соответствующий типу данных, пользователь рискует получить результаты, лишенные смысла.





## 12.2.1. Меры различия

Виды используемых в кластерном анализе мер сходства и различия перекликаются с философской дилеммой: «ищите сходство» или «ищите различие». Меры для кластерного анализа могут быть следующих видов:

- Мера сходства типа расстояния (функции расстояния), называемая также мерой различия. В этом случае объекты считаются тем более похожими, чем меньше расстояние между ними, поэтому некоторые авторы называют меры сходства типа расстояния мерами различия. При определенных условиях данная мера будет метрикой.
- Мера сходства типа корреляции, называемая связью, является мерой, определяющей похожесть объектов. В этом случае объекты считаются тем более похожими, чем больше связь между ними. Данные меры с помощью элементарных преобразований могут быть сведены к мерам сходства типа расстояния с целью единообразия.

Применяются меры различия, в зависимости от принадлежности параметров, описывающих объекты в различных шкалах измерения:

1. для количественных признаков:
  - евклидово расстояние
  - манхеттенское расстояние,
  - супремум–норма,
  - расстояние Махаланобиса,
  - расстояние Пирсона,
2. для порядковых признаков:
  - расстояние Спирмэна,
  - расстояние Кендалла,
3. для номинальных признаков:
  - расстояние Жаккара,
  - расстояние Рассела–Рао,
  - расстояние Бравайса,
  - расстояние Юла,
4. для смешанных и произвольных данных:
  - расстояние отношений.

Рассмотренными мерами могут оперировать различные методы кластерного анализа. Следует, однако, понимать, что меры различия накладывают жесткие ограничения на применяемые методы кластерного анализа:

- Для признаков любых типов: количественных, порядковых, номинальных (как частный случай – экспертных ранжировок) и смешанных можно применять метод средней связи Кинга.



- Только для количественных признаков можно применять метод k-средних Мак-Куина. Для использования метода с целью классификации данных в шкалах, отличной от количественной, требуется модификация метода.
- Только для количественных признаков можно применять метод Уорда, т. к. в его схеме применяется только евклидово расстояние.

Применяя формально метод, не соответствующий типу данных, пользователь рискует получить результаты, лишенные смысла.

Мера сходства типа расстояния называется метрикой, если она удовлетворяет определенным условиям:

- симметрии,
- неравенству треугольника,
- различимости нетождественных объектов,
- неразличимости тождественных объектов.

Термин «метрика» следует использовать с учетом данных условий.

## 12.2.1. Евклидово расстояние

Наиболее общей мерой является метрика Минковского

$$d_{ij} = \sqrt[r]{\sum_{k=1}^n |x_{ik} - x_{jk}|^r},$$

где  $x_{ij}$ ,  $x_{jk}$ ,  $k = 1, 2, \dots, n$  – выборочные совокупности  $i$  и  $j$ , соответственно,  $n$  – численность каждой выборки.

Если в метрике Минковского положить  $r = 2$ , мы получим стандартное евклидово расстояние (евклидову метрику)

$$d_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}.$$

### 12.2.1.2. Манхеттенское расстояние

При  $r = 1$  метрика Минковского дает манхеттенское расстояние (метрику города, city block, Manhattan distance)

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|,$$

где  $x_{ij}$ ,  $x_{jk}$ ,  $k = 1, 2, \dots, n$  – выборочные совокупности  $i$  и  $j$ , соответственно,  $n$  – численность каждой выборки.

### 12.2.1.3. Супремум–норма

При  $r \rightarrow \infty$  метрика Минковского дает метрику доминирования







$$d_{ij} = \max_k |x_{ik} - x_{jk}|, k = 1, 2, \dots, n,$$

где  $x_{ij}$ ,  $x_{jk}$ ,  $k = 1, 2, \dots, n$  – выборочные совокупности  $i$  и  $j$ , соответственно,  $n$  – численность каждой выборки.

что совпадает с супремум-нормой ( $\infty$ -нормой)

$$d_{ij} = \sup_k |x_{ik} - x_{jk}|, k = 1, 2, \dots, n.$$

#### 12.2.1.4. Расстояние Махаланобиса

Для различных ковариационных матриц в случае произвольного распределения дивергенция, оценивающая расхождение между статистическими распределениями  $i$  и  $j$ , выражается формулой

$$J_{ij} = \int_x [p_i(x) - p_j(x)] \ln \frac{p_i(x)}{p_j(x)} dx.$$

Иначе дивергенция называется полной средней информационной мерой различия двух классов или, более коротко, средней различающей информацией.

Практически дивергенция может быть вычислена по формуле

$$J_{ij} = \frac{1}{2} \text{tr}[(C_i - C_j)(C_j^{-1} - C_i^{-1})] + \frac{1}{2} \text{tr}[(C_i^{-1} + C_j^{-1})(m_i - m_j)(m_i - m_j)'],$$

где  $C_i, C_j$  – дисперсионно-ковариационные матрицы совокупностей  $i$  и  $j$ ,

$m_i, m_j$  – вектора средних совокупностей  $i$  и  $j$ .

Мера Махаланобиса (расстояние Махаланобиса, обобщенное Евклидово расстояние, обобщенное расстояние) является дивергенцией в предположении, что ковариационные матрицы классов равны

$$C_i = C_j = \Sigma,$$

а многомерная совокупность подчиняется многомерному нормальному распределению.

Соответствие совокупности многомерному нормальному распределению может быть протестировано с помощью методов главы «Проверка нормальности распределения», а равенство ковариационных матриц – с помощью методов главы «Дисперсионный анализ».

Указанные условия накладывают ограничения на применение рассматриваемой меры различия, что часто ошибочно не принимается во внимание исследователями, но несомненно, должно учитываться в практических расчетах. После должных преобразований получим, что мера Махаланобиса вычисляется как

$$d_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j),$$

где  $\Sigma$  – общая внутригрупповая дисперсионно-ковариационная матрица.





## 12.2.1.5. Расстояние Пирсона

Коэффициент корреляционного отношения Пирсона (коэффициент корреляции, выборочный коэффициент корреляции, коэффициент корреляции Бравайса–Пирсона) измеряет силу линейной корреляционной связи количественных признаков. Коэффициент корреляции вычисляется по формуле

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

где где  $x_i, y_i, i = 1, 2, \dots, n$  – выборочные совокупности  $i$  и  $j$ , соответственно,  
 $\bar{x}, \bar{y}$  – соответствующие выборочные средние значения,  
 $n$  – численность каждой выборки.

Использование коэффициента корреляции в качестве меры связи оправдано лишь тогда, когда совместное распределение пары признаков нормально или приближенно нормально.

Расстояние Пирсона как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

## 12.2.1.6. Расстояние Спирмэна

Показатель ранговой корреляции Спирмэна (показатель корреляции рангов Спирмэна, коэффициент корреляции рангов, коэффициент корреляции Спирмэна, коэффициент ранговой корреляции  $\rho$ , Spearman rank correlation) применяется в случае, если изучается линейная связь между рядами, представленными в количественной или порядковой шкале.

Практически при анализе количественных признаков применять показатель Спирмэна вместо коэффициента корреляционного отношения Пирсона не следует, так как при его вычислении происходит понижение количественной шкалы до порядковой. Поэтому наиболее широкое применение показатель Спирмэна нашел при анализе корреляции порядковых признаков.

Расчет ведется по формуле

$$\hat{\rho}_s = 1 - \frac{6(S_\rho + B_x + B_y)}{n^3 - n}; \quad S_\rho = \sum_{i=1}^n (r_i - s_i)^2,$$

где  $r_i, s_i, i = 1, 2, \dots, n$  – массивы рангов выборочных совокупностей,  
 $n$  – численность каждой выборки.

$B_x, B_y$  – поправки на объединение рангов в соответствующих совокупностях, вычисляемые по формуле

$$B_i = \frac{1}{12} \sum_{i=1}^m n_i(n_i^2 - 1),$$

где  $m$  – число групп объединенных рангов,  
 $n_i, i = 1, 2, \dots, m$  – число рангов в  $i$ -ой группе.





Расстояние Спирмэна как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

## 12.2.1.7. Расстояние Кендалла

Коэффициент ранговой корреляции Кендалла (коэффициент корреляции рангов, ранговый коэффициент корреляции, коэффициент корреляции Кендэла, Kendall rank correlation) предназначен для вычисления силы корреляционной связи между двумя рядами при тех же условиях, что и рассмотренный выше показатель Спирмэна. Коэффициент Кендалла считается более строгой оценкой по сравнению с показателем ранговой корреляции Спирмэна.

Все основные положения и замечания, данные при описании показателя Спирмэна, справедливы и в отношении коэффициента Кендалла. Расчет ведется по формуле:

$$\tau = \frac{S_\tau}{\sqrt{\left(\frac{n(n-1)}{2} - B_x\right)\left(\frac{n(n-1)}{2} - B_y\right)}}; S_\tau = \sum_{i=1}^n \sum_{j=i+1}^n \text{sign}(r_j - s_i),$$

где  $r_i, s_i, i = 1, 2, \dots, n$  – массивы рангов выборочных совокупностей,  
 $n$  – численность каждой выборки.

$B_x, B_y$  – поправки на объединение рангов в соответствующих совокупностях, вычисляемые по формуле

$$B_x = \frac{1}{2} \sum_{i=1}^m n_i(n_i - 1),$$

где  $m$  – число групп объединенных рангов,  
 $n_i, i = 1, 2, \dots, m$  – число рангов в  $i$ -ой группе.

Расстояние Кендалла как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

## 12.2.1.8. Расстояние Жаккара

Показатель подобия Жаккара (коэффициент Жаккара) вычисляется по формуле

$$J = a / (a + b + c),$$

где  $a, b, c$  – значения в клетках таблицы 2 x 2.

Расстояние Жаккара как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

## 12.2.1.9. Расстояние Рассела–Рао

Показатель подобия Рассела и Рао вычисляется по формуле

$$J = a / (a + b + c + d),$$

где  $a, b, c, d$  – значения в клетках таблицы 2 x 2.





Расстояние Рассела–Рао как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

## 12.2.1.10. Расстояние Бравайса

Специальная форма коэффициента корреляции – коэффициент сопряженности Бравайса (φ–коэффициент Пирсона) – рассчитывается по формуле

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}},$$

где  $a, b, c, d$  – значения в клетках таблицы  $2 \times 2$ .

Расстояние Бравайса как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

## 12.2.1.11. Расстояние Юла

Ориентировочную оценку корреляционной связи в случае исследования таблиц сопряженности  $2 \times 2$  может дать коэффициент ассоциации Юла. Вычисления производятся по формуле

$$Q = \frac{ad - bc}{ad + bc},$$

где  $a, b, c, d$  – значения в клетках таблицы  $2 \times 2$ .

Расстояние Юла как мера сходства может быть получено из рассмотренного коэффициента путем отнятия от единицы.

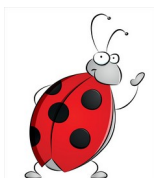
## 12.2.1.12. Расстояние отношений

Расстояние отношений (расстояние между матрицами отношений) построено на результатах исследования в теории множеств и может быть применено к объектам, измеренным в различных, в том числе смешанных, шкалах. Примерами такого рода объектов могут быть экспертные ранжировки. Смешанные данные часто возникают в медицинских исследованиях, когда вектор, описывающий объект (пациента), представляет собой совокупность количественных (результаты инструментальных исследований) и качественных (наличие–отсутствие симптомов) данных.

Расстояние отношений между объектами  $k$  и  $l$  определяется по формуле

$$d(P_k, P_l) = \sum_{i=1}^n \sum_{j=1}^n |p_{ij}^{(k)} - p_{ij}^{(l)}|,$$

где  $p_{ij}^{(k)}, i=1,2,\dots,n; j=1,2,\dots,n$  – элементы матрицы отношений частичного порядка, которые вычисляются на основе матрицы исходных данных (в экспертных оценках – матрицы опроса), как рассмотрено в главе «Обработка экспертных оценок».



При использовании представленной меры из методов кластерного анализа можно использовать только метод средней связи Кинга. Метод Уорда неприменим по понятной причине. Метод  $k$ -средних для использования предлагаемой меры нуждается в модификации. Подробно о данной мере рассказано в книге Литвака. О применении новых результатов теории множеств в кластерном анализе см. работы Петровского.

## 12.2.2. Метод средней связи Кинга

Метод средней связи Кинга (King) является одним из важнейших иерархических агломеративных методов кластерного анализа. Процесс классификации состоит из элементарных шагов:

- Поиск и объединение двух наиболее похожих объектов в матрице сходства.
- Основанием для помещения объекта в кластер является близость двух объектов, в зависимости от меры сходства.
- На каком-либо этапе ранее объединенные в один кластер объекты считаются одним объектом с усредненными по кластеру параметрами.
- На следующем этапе находятся два очередных наиболее похожих объекта, и процедура повторяется с шага 2 до полного исчерпания матрицы сходства.

При использовании представленного здесь метода не возникает проблемы возможного несоответствия применяемой меры и шкалы измерения, т. к. метод оперирует не исходными объектами, а построенной матрицей сходства, по определению являющейся количественной. Координаты центра тяжести кластера вычисляются не по исходным данным – они являются продуктом манипуляций с матрицей сходства.

В качестве меры различия для метода средней связи используется любая из представленных мер, чем и определяется универсальность метода для любых типов данных, в том числе для смешанных данных.

Обычно помимо общей информации (число объектов, число параметров, тип связи) выводится таблица номеров объединенных объектов и уровней соответствующих связей.

После объединения пары объектов второй объект каждой пары исключается из рассмотрения и делается перенумерация остальных объектов.

Выдается также таблица принадлежности объектов кластерам. Данная таблица строится на основе следующей идеи. Если взять построенную по результатам анализа дендрограмму (см. ниже раздел «Графическое представление результатов кластерного анализа»), то мысленно двигая воображаемую горизонтальную линию от самого верхнего значения ординаты, соответствующего максимальному уровню связи, вниз, мы последовательно пересекаем 1 (верхний уровень), 2, 3, ... вертикальных частей линий, соединяющих объекты. Как только мы достигаем заданного пользователем числа пересечений, начиная с данного уровня связи, можно «размотать» дендрограмму в обратном, нисходящем направлении (дендрограмма строилась снизу вверх) и установить, какие объекты принадлежат той или иной ветви. Гроздь данных объектов и будут составлять кластеры.



### 12.2.3. Метод Уорда

Метод Уорда (Ward) является одним из иерархических агломеративных методов кластерного анализа. Процесс классификации состоит из элементарных шагов:

- Поиск и объединение двух наиболее похожих объектов в матрице сходства.
- Основанием для помещения объекта в кластер является минимум дисперсии внутри кластера при помещении в него текущего классифицируемого объекта.
- На каком-либо этапе ранее объединенные в один кластер объекты считаются одним объектом с усредненными по кластеру параметрами.
- На следующем этапе находятся два очередных наиболее похожих объекта, и процедура повторяется с шага 2 до полного исчерпания матрицы сходства.

В качестве меры различия для метода Уорда используется только евклидово расстояние. Помимо общей информации (число объектов, число параметров) выдается таблица номеров объединенных объектов и уровней соответствующих связей. После объединения пары объектов второй объект каждой пары исключается из рассмотрения и делается перенумерация остальных объектов.

Также выдается таблица принадлежности объектов кластерам. Данная таблица строится на основе следующей идеи. Если взять построенную по результатам анализа дендрограмму (см. ниже раздел «Графическое представление результатов кластерного анализа»), то мысленно двигая воображаемую горизонтальную линию от самого верхнего значения ординаты, соответствующего максимальному уровню связи, вниз, мы последовательно пересекаем 1 (верхний уровень), 2, 3, ... вертикальных частей линий, соединяющих объекты. Как только мы достигаем заданного пользователем числа пересечений, начиная с данного уровня связи, можно «размотать» дендрограмму в обратном, нисходящем направлении (дендрограмма строилась снизу вверх) и установить, какие объекты принадлежат той или иной ветви. Гроздь данных объектов и будут составлять кластеры.

### 12.2.4. Метод $k$ -средних Мак-Куина

Теоретическое обоснование метода  $k$ -средних ( $k$  внутригрупповых средних) Мак-Куина (McQueen) сравнительно просто, логично и может быть найдено во многих источниках.

Принцип классификации сводится к следующим элементарным шагам:

- Некоторое, возможно, случайное, исходное разбиение множества объектов на заданное число кластеров (классов, групп, популяций). Расчет «центров тяжести» кластеров.
- Отнесение остальных объектов к ближайшим кластерам.
- Пересчет новых «центров тяжести» кластеров.
- Переход к шагу 2, пока новые «центры тяжести» кластеров не перестанут отличаться от старых.
- Получено оптимальное разбиение.

В качестве меры различия для метода средней связи используется любая из представленных мер, предназначенных для количественных данных.



Помимо общей информации (число объектов, число параметров, тип связи) выдают координаты «центров тяжести» кластеров и таблицу принадлежности объектов кластерам. Отметим, что в результате расчета может быть получено, что часть кластеров окажется пустой. Это – следствие того, что пользователем задано слишком много кластеров, причем это число превышает естественное количество кластеров, существующее в представленных исходных данных. Результаты анализа могут быть использованы, а пустые кластеры следует просто не принимать во внимание.

Результаты расчета могут быть использованы для графического построения пространственных эллипсоидов.

## 12.2.5. Модифицированный метод $k$ –средних

С целью использования метода  $k$ –средних для кластерного анализа данных, измеренных в шкалах, отличной от количественной, требуется модификация метода, заключающаяся в использовании в качестве центра тяжести кластера не среднего значения, вычисление которого корректно может быть выполнено только для количественных данных, а медианы. Универсальным решением может быть медиана множества, представленная в главе «Описательная статистика».

Принцип классификации сводится к следующим элементарным шагам:

1. Некоторое, возможно, случайное, исходное разбиение множества объектов на заданное число кластеров (классов, групп, популяций). Расчет «центров тяжести» кластеров, в качестве которых для смешанных и произвольных шкал может быть использована медиана Ойя (median Oja), для данных типа экспертных оценок может быть использовано среднее Кемени или медиана Кемени, представленные в главе «Обработка экспертных оценок».
2. Отнесение остальных объектов к ближайшим кластерам.
3. Пересчет новых «центров тяжести» кластеров.
4. Переход к шагу 2, пока новые «центры тяжести» кластеров не перестанут отличаться от старых.
5. Получено оптимальное разбиение.

В качестве меры различия для метода средней связи используется любая из представленных а также данные, измеренные в смешанных шкалах.

Помимо общей информации (число объектов, число параметров, тип связи) выдают координаты «центров тяжести» кластеров и таблицу принадлежности объектов кластерам. Отметим, что в результате расчета может быть получено, что часть кластеров окажется пустой. Это – следствие того, что пользователем задано слишком много кластеров, причем это число превышает естественное количество кластеров, существующее в представленных исходных данных. Результаты анализа могут быть использованы, а пустые кластеры следует просто не принимать во внимание.

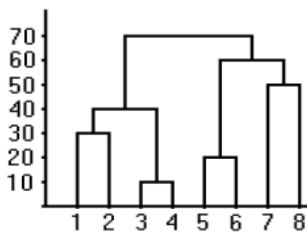


## 12.2.6. Графическое представление результатов кластерного анализа

Результаты кластерного анализа могут быть более наглядными, если их представить в виде графиков.

Результаты расчета методом  $k$ -средних могут быть очевидно использованы для графического построения пространственных эллипсоидов. Для этого достаточно изобразить координаты объектов в  $k$ -мерном пространстве (при числе «измерений», превышающем 2 или 3, можно изобразить 2 или 3-мерные срезы данного пространства).

По результатам расчета иерархическими методами можно построить специальный график, называемый дендрограммой (дендограммой). Предположим, после применения одного из иерархических методов получены результаты классификации в виде величин связи для пар объектов. Идея построения дендрограммы состоит в том, что пары объектов, отложенных по оси абсцисс, соединяются в соответствии с уровнем связи, отложенным по оси ординат. Ниже показан пример дендрограммы.



## Глава 13. Информационный анализ

### 13.1. Введение

Рассматривается вычисление основных показателей разведочного информационного анализа.

### 13.2. Теоретическое обоснование

Методы информационного анализа находят применение в различных научно-технических областях (примеры приводятся ниже). Основные показатели информационного анализа:

- Число классов.
- Число вариантов.
- Энтропия.
- Дисперсия энтропии.
- Максимальная энтропия.
- Относительная энтропия.
- Избыточность.
- Организация.







Исходными данными анализа являются дискретный или интервальный вариационные ряды. Они представляют собой таблицы распределения количеств вариантов по классам и различаются тем, что в первом случае количества относятся к определенным, возможно нечисловым, значениям признаков, а во втором случае – к интервалам изменения признака (классовым интервалам). Из исходной эмпирической выборки вариационный ряд удобно построить с помощью инструмента «Гистограмма» главы «Описательная статистика».

### 13.2.1. Число классов

Под группировкой (классификацией, разнесением вариантов по классам) понимается некоторое разбиение эмпирической выборки, содержащей все  $N$  наблюдавшихся вариантов  $x_1, x_2, \dots, x_N$ , на  $s$  интервалов (классов). Результатом группировки является вариационный ряд.

Исходными данными для расчета методами информационного анализа является не сама эмпирическая выборка, а построенные на ее основе, в зависимости от шкалы измерения исходных данных (см. главу «Введение»), дискретный или интервальный вариационные ряды. Они представляют собой таблицы распределения количеств вариантов по классам и различаются тем, что в первом случае количества относятся к определенным, возможно нечисловым, значениям признаков, а во втором случае – к интервалам изменения признака (классовым интервалам).

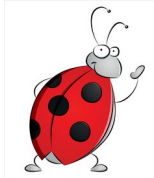
Количества вариантов, попавших в тот или иной класс в результате классификации, называют частотами. Частоты, отнесенные к численности  $N$ , называют частостями. Считается, что частоты могут служить оценками вероятностей. Это утверждение тем вернее, чем больше численность изучаемой выборочной совокупности.

Часто группировка является естественной (например, виды растений). В данном случае для дискретного вариационного ряда число классов  $s$  равно числу градаций переменной (в рассмотренном примере – числу видов). Отметим, что при анализе реальных данных может оказаться, что в конкретной совокупности некоторые классы окажутся с нулевыми частотами. Учитывать данные нулевые частоты (считать их нулями без уменьшения числа классов  $s$  на количество классов с нулевыми частотами) либо не учитывать (считать только классы с ненулевыми частотами, соответственно уменьшая  $s$ ), зависит от конкретной задачи. Для практического построения интервального вариационного ряда на основе численных данных, прежде всего, необходимо определить либо задать число классов (групп, интервалов)  $s$ . Субъективным критерием правильности выбора числа классов является верная передача типа распределения эмпирических частот данной совокупности. Если выбрано слишком мало классов, можно потерять характерную картину эмпирического распределения. При слишком подробном делении на классы можно затушевать реальную картину распределения частот случайными отклонениями.

### 13.2.2. Число вариант ряда

Совокупности состоят из отдельных элементов (объектов), которые объединены общностью некоторых свойств (признаков, переменных). В статистическом анализе данные объекты принято называть вариантами. Количество элементов (вариант) совокупности можно





называть по-разному. Так, если речь идет о выборке, количество ее элементов может называться численностью, величиной или размером.

Численность вариационного ряда определяется по формуле

$$N = \sum_{i=1}^s n_i,$$

где  $n_i, i = 1, 2, \dots, s$  – численности классов группировки,  
 $s$  – число классов группировки.

Численности классов группировки принято называть также частотами.

Знание числа вариантов может быть полезно для элементарного пересчета исходных данных, если исходные данные заданы в виде частот распределения.

### 13.2.3. Энтропия

В теории информации в качестве меры количества информации, возможности выбора (количества разнообразия) и неопределенности применяется величина, определяемая по формуле Шеннона

$$H = - \sum_{i=1}^s p_i \log_a p_i,$$

где  $p_i, i = 1, 2, \dots, s$  – вероятность появления дискретного события,  
 $s$  – число классов группировки,

$a$  – основание логарифма – единица, выбранная для оценки величины энтропии, обычно равная 2.

Величины  $p_i, i = 1, 2, \dots, s$ , образуют множество вероятностей, но для практических вычислений вероятности допустимо заменить частотами распределений:

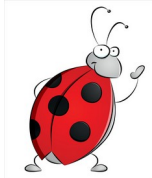
$$p_i \approx \frac{n_i}{N}, i = 1, 2, \dots, s,$$

$n_i, i = 1, 2, \dots, s$  – численности классов группировки (частоты),  
 $s$  – число классов группировки,  
 $N$  – число вариантов ряда.

В практических вычислениях некоторые частоты могут оказаться нулевыми, поэтому условились считать, что  $0 \cdot \log_a 0 = 0$ .

Величина  $H$  называется энтропией дискретного множества вероятностей (энтропией дискретной случайной величины, средней собственной информацией, энтропией Шеннона, энтропией Шеннона–Винера) и в источниках иногда обозначается как  $I$ , чтобы отличить ее от энтропии непрерывного распределения. Энтропия представляет собой количественную меру степени неопределенности исхода случайного опыта, зависящую не от индивидуальных свойств результатов опыта, а от соответствующих вероятностей. В дискретном случае энтропия равна нулю, когда одна из вероятностей равна 1, а остальные нулю.

Энтропия непрерывного распределения с функцией плотности распределения  $p(x)$  определяется как



$$H = - \int_{-\infty}^{\infty} p(x) \log_a p(x) dx$$

и называется также относительной или дифференциальной энтропией.

Если в качестве основания логарифма  $a$  выбрано число 2, энтропия вычисляется в битах, если число  $e$  – в нитах, если число 10 – в дитах. Если энтропия вычислена в нитах, для вычисления энтропии в битах нужно разделить значение в нитах на  $\ln 2$ . Утверждение вытекает из известной формулы замены основания логарифмов

$$\log_a c = \frac{\log_b c}{\log_b a}$$

при  $a > 0$  и  $a \neq 1$ .

Исходные данные для вычисления энтропии системы представляют собой дискретный или интервальный вариационный ряд.

Некоторыми авторами энтропия называется количеством информации или двоичной энтропией. Так называемая энтропия Колмогорова, характеризующая хаотическое движение в фазовом пространстве произвольной размерности, также определяется по формуле Шеннона. Введены обобщения энтропии, такие как энтропия Реньи порядка  $\alpha$ :

$$H_\alpha(p_1, \dots, p_n) = \frac{1}{1 - \alpha} \log_2 \sum_{i=1}^n p_i^\alpha,$$

при  $\alpha = 1$  называемая энтропией Шеннона,

при  $\alpha = 0$  – энтропией Хартли.

Двусторонний доверительный интервал оцениваемой энтропии вычисляется по формуле

$$I_H = \left[ H - \Psi((1 + \beta)/2) \frac{D_H}{\sqrt{s}}; H + \Psi((1 + \beta)/2) \frac{D_H}{\sqrt{s}} \right],$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$D_H$  – дисперсия энтропии.

Подробное теоретическое обоснование см. в книге Дмитриева.

## 13.2.4. Дисперсия энтропии

Дисперсия энтропии вычисляется по формуле

$$D_H = \frac{1}{N} \left[ \sum_{i=1}^s p_i \log_a^2 p_i - \left( \sum_{i=1}^s p_i \log_a p_i \right)^2 \right] + \frac{s-1}{2N^2},$$

где  $p_i, i = 1, 2, \dots, s$  – вероятность появления дискретного события,

$s$  – число классов группировки,

$N$  – число вариант ряда,

$a$  – основание логарифма, обычно равное числу 2.

Дисперсия энтропии находит применение при оценке значимости различий индексов

Шеннона, характеризующих видовое разнообразие, вычисленных для двух совокупностей, а



также для вычисления доверительных интервалов оцениваемой энтропии. Соответствующий критерий предложен Хатчесоном (Hutchenson).

См. статьи Хатчесона, Грениер (Grenier) с соавт., работу Шитикова с соавт., книгу Зара (Zar).

### 13.2.5. Максимальная энтропия

Максимальное разнообразие системы вычисляется по формуле Хартли

$$H_{\max} = \log_2 s,$$

где  $H_{\max}$  – максимальная энтропия,

$s$  – число классов группировки.

Справедлива формула

$$0 \leq H \leq H_{\max},$$

где  $H$  – энтропия.

Максимум энтропии соответствует наибольшей неопределенности или равенству вероятностей всех возможностей. Опыт имеет наибольшую энтропию при  $k$  равновероятных исходах с вероятностями (в практических вычислениях – частотами)  $1/k$ . Степень неопределенности опыта тем больше, чем больше число  $k$  его исходов.

Энтропия графического изображения зависит от количества уровней, а при одинаковом числе уровней – от закона распределения. Равномерный закон распределения соответствует полной хаотичности. В этом случае энтропия достигает максимума, который зависит только от количества уровней

$$H_{\max} = \log_a(h_{\max} - h_{\min} + 1),$$

где  $(.)$  – размах дискретной случайной величины,

$h_{\max}$  – максимальное значение уровня изображения,

$h_{\min}$  – минимальное значение уровня изображения.

Напомним, что размах выборки (размах вариации, амплитуда ряда) – это разность между максимумом и минимумом вариант выборки.

### 13.2.6. Относительная энтропия

Для сравнения систем, различающихся по количеству (!) элементов кода, простое сопоставление энтропий не будет корректным. Для решения задачи применяется относительная энтропия (коэффициент сжатия информации), определяемая как

$$h = H / H_{\max},$$

где  $H$  – энтропия,

$H_{\max}$  – максимальная энтропия.

Относительная энтропия определяет относительную степень информационной загруженности системы по отношению к возможной максимальной нагрузке. Кроме того, относительная энтропия, как и избыточность, может характеризовать степень близости закона распределения к равномерному.



## 13.2.7. Избыточность

Избыточность показывает, какая доля или процент передаваемой информации является избыточной. Она дает соотношение между полным количеством информации, шумом и сохранившейся упорядоченностью системы. Избыточность (в источниках обозначается также как  $R$ ) вычисляется по формуле

$$D = 1 - H / H_{\max},$$

где  $H$  – энтропия,

$H_{\max}$  – максимальная энтропия.

Избыточность характеризует степень близости закона распределения к равномерному распределению и может быть выражена в долях, как в показанной формуле, либо в процентах. Например, для полутонного изображения с 256 уровнями энтропия не может превышать 8, а избыточность при равномерном законе распределения равна нулю.

## 13.2.8. Организация системы

Под организацией системы понимают реализованную в ней неопределенность. Абсолютная организация системы вычисляется по формуле

$$O = H_{\max} - H,$$

где  $H_{\max}$  – максимальная энтропия,

$H$  – энтропия.

При организации, равной максимальной энтропии, система становится детерминированной, полностью стабильной.

## 13.2.9. Примеры информационного анализа

Ниже представлены несколько практических приложений информационного анализа:

- Разведочный информационный анализ.
- Исследование структурной перестройки объекта.
- Сравнение групп по индексу межвидового разнообразия.

### 13.2.9.1. Разведочный информационный анализ

В математической статистике и математическом моделировании важно точно обосновать возможность применения тех или иных методов анализа и адекватность модели.

Напомним, что математическим моделированием называют приближенное описание явлений, выраженное с помощью математической символики, а также процесс их изучения с помощью математических моделей. Математическое моделирование включает этапы:

- Формулировка законов, связывающих основные объекты модели, на основе изучения явлений и проникновения в их взаимосвязи.
- Составление уравнений модели. Исследование и математическое решение задачи, к которой приводит математическая модель.



- Сопоставление результатов расчета математической модели с данными изучаемого явления, полученными в результате наблюдения за этим явлением. При наличии в модели параметров, неизвестных на этапе составления или недоступных для прямого измерения, производится идентификация модели на основе экспериментальных данных.
- Исследование изучаемого явления с помощью математической модели. Уточнение модели на основе новых данных об изучаемом явлении.

Математическое моделирование строит модели в виде различных уравнений или систем уравнений. Уравнения математической модели могут быть алгебраическими, дифференциальными, интегральными и их совокупностями.

Математической же статистикой называют раздел математики, посвященный математическим методам сбора, систематизации, обработки и интерпретации статистических данных, а также использование их для научных и практических выводов. Напомним, что под статистическими данными понимают любую систему данных: числовую информацию, извлекаемую из результатов выборочных обследований; [эмпирические] выборки из любых генеральных совокупностей; результаты измерений и т.п. Математическая статистика строит вероятностно–статистическую, а не математическую в указанном выше смысле, модель. Статистической моделью называют описание выборочного пространства всех мыслимых исходов наблюдаемого случайного явления, выделение семейства распределений вероятностей этих исходов и определение другой априорной информации об этом семействе. Используя разведочный информационный анализ, в первом приближении проверку адекватности типа модели можно выполнить путем вычисления информационных показателей представленных статистических данных. Сделать выводы по вычисленным показателям необходимо в соответствии со следующей таблицей:

Величина избыточности, %	Характеристика системы	Адекватный тип модели
0–10	Вероятностная	Вероятностно–статистическая
10–30	Вероятностно–детерминированная	Дифференциальные уравнения
30–100	Детерминированная	Дифференциальные или интегральные уравнения

Это важные результаты. Например, при вычисленной избыточности, равной 25%, вероятностно–статистическая модель не будет адекватно описывать исследуемое явление. Исследователю придется заняться составлением математической модели в виде дифференциальных уравнений.

### 13.2.9.2. Исследование структурной перестройки объекта

Разность избыточности в норме и при патологии приводит к понятию ненадежности (эквивокации) передачи информации, что дает количественную характеристику структурной перестройки исследуемого объекта (системы). Вычисление эквивокации производится по формуле





$$D = R_{norm} - R_{pat} = \frac{H_{pat} - H_{norm}}{H_{max}},$$

где  $R_{norm}$  – избыточность в норме,

$R_{pat}$  – избыточность в патологии,

$H_{norm}$  – энтропия в норме,

$H_{pat}$  – энтропия в патологии.

### 13.2.9.2. Сравнение групп по индексам межвидового разнообразия

В качестве примера применения информационного анализа (в том числе его использования для проверки статистических гипотез) укажем так называемые [информационные] индексы видового разнообразия. В литературе часто используется индекс Шеннона (Shannon index), называемый авторами также индексом Шеннона–Винера (Shannon–Wiener diversity index) или индексом Шеннона–Уивера (Shannon–Weaver diversity index).

Индекс Шеннона представляет собой энтропию, вычисленную с использованием того или иного основания логарифма. Используется основание 2,  $e = 2,718281828...$  или 10.

Для сравнения групп по индексам Шеннона применяется специальная модификация критерия Стьюдента, предложенная Хатчесоном (Hutcheson). Статистика критерия вычисляется по формуле

$$t = \frac{|H_1 - H_2|}{\sqrt{D_{H_1}^2/n_1 + D_{H_2}^2/n_2}},$$

где  $H_1$  и  $H_2$  – индексы Шеннона (энтропии) совокупностей,

$D_{H_1}$  и  $D_{H_2}$  – соответствующие оценки дисперсий индексов Шеннона,

$n_1$  и  $n_2$  – соответствующие численности совокупностей.

Распределение статистики критерия Хатчесона близко к  $t$ -распределению Стьюдента при числе степеней свободы, равном

$$v = \frac{(D_1 + D_2)^2}{D_{H_1}^2/n_1 + D_{H_2}^2/n_2}.$$

В специальной литературе представлены и другие индексы (многие не имеющие отношения к теории информации).

См. работы Хатчесона, Грениер (Grenier) с соавт., Шитикова с соавт., Кейлок (Keylock), Магурран (Magurran), Розенцвейг (Rosenzweig), Пилу (Pielou), книгу Зара (Zar).



## Глава 14. Распознавание образов с обучением

### 14.1. Введение

Статистические методы, которые могут быть интерпретированы как методы распознавания образов с обучением:

- Линейный дискриминантный анализ Фишера.
- Канонический дискриминантный анализ.
- Линейный дискриминантный анализ.
- Линейный множественный регрессионный анализ.
- Логистическая регрессия.
- Пробит анализ.
- Регрессия Пуассона.

### 14.2. Теоретическое обоснование

Дискриминантный анализ представляет собой линейный метод распознавания данных с обучением. Линейным он является потому, что модель метода линейна относительно дискриминантных функций. Пользователь должен задать некоторое число объектов, указав их принадлежность к так называемым обучающим группам (классам, кластерам, популяциям). Поэтому применению методов распознавания обязательно должны предшествовать исследования методами классификации без обучения (кластерного анализа или эмпирической классификации, когда, например, врач на основании своего опыта выполняет отнесение диагноза того или иного пациента к определенному классу). В задачах, построенных на реальных экспериментальных данных, классы могут пересекаться, особенно если обучение производится на основании эмпирической классификации. Пересечение классов ухудшает качество классификации, а наилучшие для данного набора данных результаты распознавания получаются в случае непересекающихся классов.

Методы распознавания образов с обучением используют в качестве обучающих выборки объекты, заранее классифицированные тем или иным способом. Качество процедуры дискриминации определяется вероятностью правильной классификации. Очень хорошие результаты для распознавания образов с обучением дает предварительное применение метода (см. главу «Кластерный анализ») ближней связи, а применение метода  $k$ -средних для предварительного отнесения объектов классам дает практически 100% качество распознавания для любого метода дискриминантного анализа. Это обусловлено тем, что из рассмотренных нами методов распознавания без обучения только метод  $k$ -средних на основе выбранной метрики гарантированно строит непересекающиеся кластеры.

Методы распознавания образов с обучением вырабатывают некоторые решающие правила, позволяющие отнести предлагаемые объекты к заданным классам. Решающие правила могут быть получены:





- в виде простых классифицирующих функций, как это сделано в линейном дискриминантном анализе Фишера,
- в виде дискриминантных функций, как это сделано в каноническом дискриминантном анализе,
- в виде некоторых характеристик (групповая ковариационная матрица, групповой вектор средних и определитель ковариационной матрицы), как это сделано в линейном дискриминантном анализе,
- в виде набора коэффициентов регрессии (без свободного члена или со свободным членом), как это сделано в методе линейного множественного регрессионного анализа, логистической регрессии, пробит анализа или регрессии Пуассона.

## 14.2.1. Оценка качества моделей

Качество регрессионной модели зависит от ее типа и в литературе оценивается различными способами, в том числе особыми статистическими критериями, некоторые из которых представлены в данном разделе.

### 14.2.1. Количественные классификаторы

Наиболее простым и понятным способом, которым интуитивно пользуются, является оценка качества дискриминации относительным процентным содержанием, иначе числом верно классифицированных объектов обучающей выборки, отнесенным к объему обучающей выборки, выраженным в процентах. Хотя этот подход недостаточно правомерен, но он является единственно возможным для малого объема данных.

Корректно качество регрессионных моделей принято оценивать так. Обучающий массив данных (выборка) случайным образом (см. главу «Рандомизация и генерация случайных последовательностей») делится на две части. Одна часть используется для построения модели. Другая часть – для проверки ее качества. Подход может применяться, если представленный массив данных большую численность.

Качество построения множественной регрессионной модели удобно оценивать также средним квадратичным отклонением.

#### 14.2.1.2. Бинарные классификаторы

Для оценки качества модели бинарного классификатора предложен ряд параметров.

Дальнейшие обозначения проще всего пояснить с помощью таблицы 2 x 2, естественной для бинарных откликов.

Модель	Опыт	
	Положительный исход	Отрицательный исход
Положительный исход	$T_P$	$F_P$
Отрицательный исход	$F_N$	$T_N$

Суть обозначений ясна из первых букв английских терминов:

- True – истинно,







- False – ложно,
- Positive – положительный,
- Negative – отрицательный.

Рассчитываются такие параметры качества, как чувствительность и специфичность. Рассмотрим их подробнее.

Чувствительность показывает долю истинно положительных случаев, правильно идентифицированных моделью,

$$Se = \frac{T_P}{T_P + F_N} \cdot 100\%.$$

Специфичность показывает долю истинно отрицательных случаев, правильно идентифицированных моделью,

$$Sp = \frac{T_N}{T_N + F_P} \cdot 100\%.$$

Оптимальные величины порога отсечения вычисляются на основе критериев:

- Для метода 1 достигается минимум величины  $|Se - Sp|$ ,
- Для метода 2 достигается максимум величины  $Se + Sp$ . Данный критерий предложен Юденом (Youden).

Выбор того или иного оптимального порога отсечения (а также любого другого желаемого порога, в том числе предлагаемой некоторыми авторами величины 0,5) производится на основе требований, предъявляемых исследователем к прогностическим характеристикам модели.

По желанию пользователя выводятся таблица и график так называемой ROC кривой (Receiver Operating Characteristic Curve), отображающей величины  $Se$  и  $1 - Sp$  в зависимости от порога отсечения (параметрическая кривая) и стандартно применяемой для оценки бинарных классификаторов. Объективную оценку качества модели может показать также площадь под ROC кривой, в литературе называемая как AUC (Area Under Curve).

О сравнении ROC кривых см. статьи Вергара (Vergara) с соавт., Хэнли (Hanley) с соавт. О пороге отсечения см. статью Клотше (Klotsche) с соавт.

## 14.2.2. Оценка значимости модели

В данном разделе представлены методы оценки значимости бинарного классификатора.

### 14.2.2.1. Статистика Вальда

Статистическая значимость весовых коэффициентов бинарного классификатора может проверяться с помощью статистик Вальда, хотя этот способ и не рассматривается некоторыми авторами как абсолютно надежный. Статистики Вальда записываются как

$$W_i = \frac{|\hat{b}_i|}{\sqrt{\text{Var}(\hat{b}_i)}}, i = 1, 2, \dots, m,$$



где  $\hat{b}_i, i = 1, 2, \dots, m$ , – вычисленные оценки весовых коэффициентов,

$Var(\hat{b}_i), i = 1, 2, \dots, m$ , – дисперсии оценок весовых коэффициентов.

$m$  – количество измеряемых в эксперименте параметров объекта.

Дисперсии оценок находятся как диагональные члены матрицы  $I^{-1}(B)$ , обратной к информационной матрице Фишера  $I(B)$ . Информационная матрица Фишера в данном случае представляет собой матрицу, элементы которой являются взятыми с обратным знаком элементами матрицы Гессе функции максимального правдоподобия (далее – ФМП)

$$I(B) = -H(B),$$

где  $B = \{b_i\}, i = 1, 2, \dots, m$  – вектор–столбец весовых коэффициентов.

Статистики Вальда имеют стандартное нормальное распределение, что позволяет установить значимость вычисленных оценок весовых коэффициентов модели.

#### 14.2.2.2. Статистика G

Более надежный способ оценки значимости весовых коэффициентов бинарного классификатора основан на статистиках

$$G_i = -2 \ln \frac{L_i}{L}, i = 1, 2, \dots, m,$$

где  $L_i, i = 1, 2, \dots, m$  – ФМП системы с исключенным параметром  $i$ ,

$L$  – ФМП полной системы представленных данных.

Статистики  $G_i, i = 1, 2, \dots, m$ , имеют  $\chi^2$  распределение со степенью свободы 1.

См. Хосмер (Hosmer), Давнис с соавт.

#### 14.2.3. Линейный дискриминантный анализ Фишера

Метод линейного дискриминантного анализа (линейная дискриминация Фишера, дискриминаторный анализ) предложен Фишером, который предположил, что классификация должна проводиться с помощью линейной комбинации дискриминантных (различающих) переменных. Основанием отнесения объекта к кластеру (классу, популяции) является наибольшее значение так называемой простой классифицирующей функции  $h_k$  для  $k$ -го класса, являющейся линейной комбинацией дискриминантных переменных:

$$h_k = b_{k0} + \sum_{i=1}^p b_{ki} X_i,$$

где  $p$  – число дискриминантных переменных,

$b_{ki}$  – коэффициент для  $i$ -й переменной  $k$ -го класса, определяемый как

$$b_{ki} = (n - g) \sum_{j=1}^p a_{ij} X_{jk},$$

где  $n$  – общее число наблюдений по всем классам,

$a_{ij}$  – элементы матрицы, обратной к матрице  $W$  разброса внутри классов (внутригрупповая матрица сумм попарных произведений), вычисляемой по формуле





$$w_{ij} = \sum_{k=1}^g \sum_{m=1}^{n_k} (X_{ikm} - X_{ik.})(X_{jkm} - X_{jk.}),$$

где  $g$  – число классов,

$n_k$  – число наблюдений в  $k$ -м классе,

$X_{ikm}$  – значение  $i$ -ой дискриминантной переменной (величина  $i$ -й переменной  $m$ -го наблюдения  $k$ -го класса),

$X_{jk.}$  – среднее  $j$ -й переменной  $k$ -го класса.

В основе метода лежат два предположения:

1. Популяции, среди которых производится дискриминация, подчиняются многомерному нормальному распределению. Данное предположение проверяется с помощью методов главы «Проверка нормальности распределения».
2. Популяции, среди которых производится дискриминация, имеют статистически неразличимые ковариационные матрицы.

При искусственном объявлении ковариационных матриц статистически неразличимыми могут оказаться отброшенными наиболее важные индивидуальные черты, имеющие большое значение для хорошей дискриминации. Однако введенное предположение позволяет получить решение и в случае, когда количество обучающих выборок в кластере оказывается меньшим количества дискриминантных функций – то есть при тех условиях, когда более точный линейный дискриминантный анализ не работает.

Результаты линейного дискриминантного анализа Фишера совпадают в смысле качества классификации с результатами более сложного в реализации канонического дискриминантного анализа.

#### 14.2.4. Канонический дискриминантный анализ

Канонический дискриминантный анализ основан на определении так называемых дискриминантных функций, количество которых меньше либо равно числу параметров объектов:

$$f_{km} = u_0 + \sum_{i=1}^p u_i X_{ikm},$$

где  $f_{km}$  – значение канонической дискриминантной функции для  $m$ -го объекта  $k$ -го класса,

$u_i$  – коэффициенты, определяемые по формуле

$$u_i = v_i \sqrt{n_i - g}, u_0 = - \sum_{i=1}^p u_i X_{i..},$$

где  $X_{i..}$  – среднее  $i$ -й переменной по всем классам,

$v_i, i = 1, 2, \dots, p$  – коэффициенты, вычисляемые как компоненты собственных векторов решения обобщенной проблемы собственных значений:

$$Bv = \lambda Wv,$$

где  $B$  – межгрупповая сумма квадратов отклонений,

$v$  – собственный вектор,

остальные обозначения те же, что и в предыдущем разделе.





Матрица  $B$  определяется как

$$B = T - W,$$

где  $T$  – матрица сумм квадратов и попарных произведений, элементы которой вычисляются как

$$t_{ij} = \sum_{k=1}^g \sum_{m=1}^{n_k} (X_{ikm} - X_{i..})(X_{jkm} - X_{j..})$$

Отнесение новых, неклассифицированных объектов к заданным кластерам производится после вычисления дискриминантных функций на основе Евклидовой метрики.

Количество дискриминантных функций в каноническом дискриминантном анализе может быть равным или меньшим количества параметров, описывающих каждый объект.

Результаты распознавания методом канонического дискриминантного анализа совпадают с результатами линейного дискриминантного анализа Фишера.

#### 14.2.5. Линейный дискриминантный анализ

Недостатком описанного метода линейной дискриминации Фишера является предположение о равенстве ковариационных матриц рассматриваемых выборок, вследствие чего могут оказаться отброшенными важные индивидуальные черты, имеющие большое значение для хорошей дискриминации. В методе линейного дискриминантного анализа, напротив, ковариационные матрицы для различных классов считаются различными.

Отказ от предположения о статистической неразличимости ковариационных матриц, лежащего в основе линейной дискриминации Фишера, для обучающих кластеров не позволяет получить решение в случае, когда количество обучающих выборок в кластере оказывается меньше количества дискриминантных функций.

В рассматриваемом методе основанием отнесения объекта к классу является наибольшее значение для данного объекта функции плотности нормального распределения среди всех классов. Вектор средних значений, входящих в формулу функции плотности нормального распределения, а также дисперсионно-ковариационная матрица для каждого обучающего класса оцениваются по исходным данным на этапе обучения.

Отсутствие простых решающих правил (для получения результата нужно проделать довольно объемные вычисления) было некоторым препятствием для широкого применения этого мощного метода в период, предшествовавший распространению персональных компьютеров. Метод использовался фактически редко, но здесь введен нами как серьезная альтернатива линейному дискриминантному анализу Фишера. В наших расчетах метод линейного дискриминантного анализа показал более высокое качество распознавания по сравнению с линейной дискриминацией Фишера и каноническим дискриминантным анализом.

#### 14.2.6. Линейный множественный регрессионный анализ

В ходе нетривиального эксперимента абсолютно точные измерения параметров, как правило, невозможны. Чтобы уменьшить влияние ошибок, производится большое число измерений.





Каждое измерение дает нам уравнение с известной из теоретических соображений структурой с точностью до коэффициентов, подлежащих определению. При числе измерений большем, меньшем или равном числу параметров, мы приходим к необходимости решения системы, число уравнений которой больше, меньше либо равно числу неизвестных параметров, соответственно.

Для решения задачи в первом приближении положим зависимость результата эксперимента от параметров линейной модели (примем линейную модель) и сформулируем задачу математически следующим образом. Требуется решить систему линейных уравнений:

$$\sum_{j=1}^k a_{ij} x_j \approx b_i, i=1,2,\dots,n,$$

(в поэлементной записи) либо

$AX = B$  (в матричной записи),

где  $a_{ij}$ ,  $i = 1,2,\dots,n$ ;  $j = 1,2,\dots,k$  – элемент матрицы экспериментальных данных  $A$  (матрицы регрессоров) размером  $n \times k$ ,

$x_j$ ,  $j = 1,2,\dots,k$  – элемент подлежащего определению вектора весовых коэффициентов  $X$  длиной  $k$ ,

$b_i$ ,  $i = 1,2,\dots,n$  – элемент вектора результатов эксперимента  $B$  длиной  $k$ ,

$k$  – количество измеряемых в эксперименте параметров,

$n$  – количество опытов.

В формуле знак  $\approx$  применен вместо знака равенства, чтобы подчеркнуть неточность определения результата эксперимента.

Случай 1. При  $k = n$  система в случае, если матрица системы не вырождена, имеет одно решение, поиск которого не вызывает трудностей и может быть осуществлен методом Гаусса или одним из итерационных методов.

Случай 2. При  $k > n$  система имеет бесчисленное множество решений. Ранг  $n$  матрицы системы меньше порядка системы. Число линейно независимых уравнений меньше количества неизвестных, поэтому возникает неопределенная (недоопределенная) система линейных уравнений порядка  $k$ .

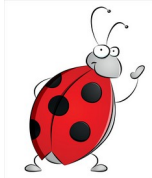
Если систему удастся решить, в общем случае полученный вектор решения системы уравнений не будет точно удовлетворять ни одному уравнению системы, однако можно получить решение, в каком-то смысле наилучшее. Существует бесчисленное множество решений рассматриваемой системы, однако из них можно выделить одно решение, наложив на систему дополнительные условия. Так, можно найти решение, обладающее минимальной Евклидовой нормой

$$\|B - AX\|_2 = \min.$$

Размерность матрицы  $A$  есть  $n \times k$ , причем  $n < k$ . Сначала нужно так преобразовать запись системы, чтобы  $k$  и  $n$  поменять местами, то есть создать предпосылки для поиска псевдообратной к  $A^T$  матрицы. После элементарных выкладок с применением свойств псевдообратной матрицы получаем формулу для нахождения решения, обладающего минимальной нормой, в виде

$$\hat{X}^T = B^T (A^T)^{\dagger},$$





где  $\hat{X}$  – вектор оценок весовых коэффициентов.

Случай 3. Данный случай  $k < n$  является наиболее важным практически. Система при этом является несовместной и может быть решена приближенно.

Ранг  $n$  матрицы системы больше порядка системы. Число линейно независимых уравнений больше количества неизвестных, поэтому возникает переопределенная система линейных уравнений порядка  $k$ .

Для решения системы достаточно домножить левую и правую части уравнения на матрицу  $A^T$  слева. Затем домножить левую и правую части полученного уравнения на матрицу  $(A^T A)^{-1}$  также слева. В результате получим готовую формулу решения

$$\hat{X} = (A^T A)^{-1} A^T B.$$

В поэлементной записи решение системы может быть представлено следующим образом.

Введем вектор ошибок  $\varepsilon_j, j = 1, 2, \dots, n$ . Тогда исходную систему можно переписать:

$$\sum_{j=1}^k a_{ij} x_j + \varepsilon_i = b_i, i = 1, 2, \dots, n.$$

Точное решение системы получить не удастся, поэтому обычно применяют метод наименьших квадратов, в котором минимизируют сумму квадратов ошибок  $\varepsilon_i, i = 1, 2, \dots, n$ .

Составим квадратичный функционал

$$I(x_1, x_2, \dots, x_k) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left[ b_i - \sum_{j=1}^k a_{ij} x_j \right]^2.$$

Потребуем минимума функционала  $I = I(x_1, x_2, \dots, x_k)$  по элементам вектора  $x_j, j = 1, 2, \dots, k$ :

$$I(x_1, x_2, \dots, x_k) \rightarrow \min_{x_j}, j = 1, 2, \dots, k.$$

Данное требование удовлетворяется в случае равенства нулю всех частных производных функционала  $I$  по элементам  $x_j, j = 1, 2, \dots, k$ , что приводит к системе  $k$  линейных

алгебраических уравнений, решив которые, мы найдем неизвестные величины  $x_j, j = 1, 2, \dots, k$ :

$$\frac{\partial I}{\partial x_j} = 0, j = 1, 2, \dots, k.$$

Подставив в последнюю формулу выражение для функционала, после элементарных преобразований получим:

$$\sum_{j=1}^k \left[ \sum_{i=1}^n a_{ij} a_{il} \right] x_j = \sum_{i=1}^n b_i a_{il}, l = 1, 2, \dots, k.$$

Это выражение представляет собой запись системы  $k$  линейных алгебраических уравнений для  $k$  неизвестных. Выражение в квадратных скобках – это запись элемента матрицы системы уравнений, а правая часть формулы – запись элемента столбца свободных членов.

Лучшие результаты в смысле более точного описания реального объекта исследований часто, хотя и не всегда, можно получить, если использовать уравнения со свободным членом:



$$x_0 + \sum_{j=1}^k a_{ij} x_j + \varepsilon_i = b_i, i=1,2,\dots,n.$$

Тогда остальные уравнения примут, соответственно, вид:

$$I(x_0, x_1, x_2, \dots, x_k) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left[ b_i - x_0 - \sum_{j=1}^k a_{ij} x_j \right]^2,$$

$$I(x_0, x_1, x_2, \dots, x_k) \rightarrow \min_{x_j}, j=0,1,\dots,k,$$

$$\frac{\partial I}{\partial x_j} = 0, j=0,1,\dots,k,$$

$$n x_0 + \sum_{j=1}^k \left[ \sum_{i=1}^n a_{ij} \right] x_j = \sum_{i=1}^n b_i,$$

$$\sum_{i=1}^n a_{il} x_0 + \sum_{j=1}^k \left[ \sum_{i=1}^n a_{ij} a_{il} \right] x_j = \sum_{i=1}^n b_i a_{il}, l=1,2,\dots,k.$$

Последнее выражение будет системой  $k + 1$  линейных алгебраических уравнений для  $k + 1$  неизвестных. Практически, однако, в применении последней формулы нет необходимости. Свободный член может быть введен, не меняя структуры уравнений регрессии (для всех случаев). При расчете достаточно ввести дополнительный  $(k + 1)$ -й параметр, которому будет соответствовать регрессор 1 в каждом векторе данных. Для определенности следует добавить единичный вектор в качестве первого столбца матрицы данных. Тогда первым коэффициентом в столбце коэффициентов регрессии, выводимом в рассматриваемом случае, как раз и будет вычисленный свободный член.

Дисперсии  $D\hat{x}_j, j=1,2,\dots,k$ , оценок коэффициентов  $\hat{x}_j, j=1,2,\dots,k$ , являющиеся диагональными элементами дисперсионно-ковариационной матрицы

$$C(\hat{X}) = MSE \cdot (A^T A)^{-1},$$

где  $MSE$  – средняя квадратичная ошибка (дисперсия ошибки регрессии), вычисляемая по формуле

$$MSE = \frac{1}{n - k} \sum_{i=1}^n e_i^2,$$

где  $e_i, i=1,2,\dots,n$  – остатки, вычисляемые как

$$e_i = b_i - \hat{b}_i, i=1,2,\dots,n,$$

где  $\hat{b}_i, i=1,2,\dots,n$ , – предсказанные результаты эксперимента.

Зная оценки коэффициентов и их дисперсии, можно вычислить статистики

$$t_j = \frac{\hat{x}_j}{\sqrt{D\hat{x}_j}}, j=1,2,\dots,k,$$





которые асимптотически имеют  $t$ -распределение с  $n - k$  степенями свободы, что позволяет сделать вывод о значимости оценок коэффициентов (значимом отличии от нуля).

Стандартное отклонение ошибки регрессии вычисляется как

$$SE = \sqrt{MSE}.$$

В дальнейших вычислениях понадобятся также стандартные ошибки остатков

$$m_i = SE \sqrt{1 - h_i}, i = 1, 2, \dots, n,$$

где  $h_i, i = 1, 2, \dots, n$  – показатели влияния (leverage, условные корреляции наблюдения и прогноза), вычисляемые как модули диагональных элементов матрицы

$$H = A(A^T A)^{-1} A^T.$$

Стандартизованные остатки (standardized residual) вычисляются как

$$E_i = \frac{e_i}{SE}, i = 1, 2, \dots, n.$$

Более точные оценки дают студентизированные остатки (studentized deleted residual), которые вычисляются как

$$E_i^* = \frac{e_i}{m_{(i)}}, i = 1, 2, \dots, n,$$

где индекс  $(i)$  здесь и далее означает, что вычисление показателя произведено при исключенном наблюдении с номером  $i$ .

В литературе выведена простая функциональная связь величин  $E_i, i = 1, 2, \dots, n$ , и

$E_i^*, i = 1, 2, \dots, n$ , поэтому последние можно вычислить через первые, теоретически намного уменьшив объем вычислений. Практически, однако, в этом нет необходимости, т. к. указанные вычисления все-таки необходимы для оценки других параметров решения.

## 14.2.6.1. Обработка выбросов

Анализ стандартизованных или студентизированных остатков может применяться для выявления выбросов наблюдений относительно статистической модели.

Студентизированные остатки  $E_i^*, i = 1, 2, \dots, n$ , асимптотически подчиняются  $t$ -распределению с  $n - k$  степенями свободы. Для удобства пользователей  $P$ -значения, не превышающие принятый для данного типа задачи стандартный порог 0,05 (что свидетельствует о значимости различий наблюдения и модельной оценки), выделяются красным цветом аналогично тому, как это сделано в главе «Обработка выбросов».

## 14.2.6.2. Выявление влияющих наблюдений

Для каждого наблюдения выводится мера Кука (Cook's distance, Cook's measure)

$$Q_i = \frac{E_i^2}{k} \frac{h_i}{1 - h_i}, i = 1, 2, \dots, n.$$





$$Q_i > \frac{4}{n-k}, i=1,2,\dots,n,$$

Значения т. е. превышающие порог тревожности, свидетельствуют о сильном влиянии данного наблюдения на смещение гиперплоскости регрессии.

Также для каждого наблюдения выводится значение меры Велча–Куха (Welsch–Kuh's distance, Welsch–Kuh's measure, *DFFITs*, *DFITS*)

$$DFFITs_i = E_i^* \sqrt{\frac{h_i}{1-h_i}}, i=1,2,\dots,n.$$

$$|DFFITs_i| > 2\sqrt{\frac{k}{n}}, i=1,2,\dots,n,$$

Значения т. е. превышающие порог тревожности, свидетельствуют о сильном влиянии данного наблюдения на смещение гиперплоскости регрессии.

Как уже упоминалось выше, в литературе выведена простая функциональная связь величин

$E_i, i=1,2,\dots,n$ , и  $E_i^*, i=1,2,\dots,n$ , поэтому меры Кука и Велча–Куха эквивалентны. Это означает, что результаты анализа данными методами в большинстве случаев должны давать одинаковые выводы.

Также для каждого наблюдения выводятся  $k$  значений меры *DFBETAS*

$$DFBETAS_{ij} = \frac{\hat{x}_j - \hat{x}_{j(i)}}{SE_{(i)}\sqrt{c_j}}, i=1,2,\dots,n; j=1,2,\dots,k,$$

где  $c_j, j=1,2,\dots,k$  – диагональные элементы матрицы  $C = (A^T A)^{-1}$ .

$$|DFBETAS_{ij}| > \frac{2}{\sqrt{n-k}}, i=1,2,\dots,n; j=1,2,\dots,k,$$

Значения т. е. превышающие порог тревожности, свидетельствуют о сильном влиянии данного наблюдения на оценки весовых коэффициентов.

#### 14.2.6.3. Автокорреляция остатков

Для оценки автокорреляции остатков множественной линейной регрессии, построенной на основе упорядоченных исходных данных (например, по латентной переменной – времени), разработан критерий Дарбина–Уотсона. Вычисление статистики критерия производится по формуле

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}.$$

Эквивалентная (матричная) форма статистики критерия

$$d = \frac{e^T A e}{e^T e},$$





где  $e$  – вектор остатков с элементами  $e_i, i = 1, 2, \dots, n$ ,

$A$  – матрица размером  $n \times n$  вида

$$A = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix}.$$

Модифицированная статистика критерия  $d/4$  асимптотически подчиняется бета-распределению с параметрами, определяемыми следующим образом.

Сначала вычисляются, соответственно, математическое ожидание и дисперсия статистики Дарбина–Уотсона по формулам

$$Ed = \frac{P}{n - k - 1},$$

$$Dd = \frac{2(Q - P \cdot Ed)}{(n - k - 2)(n - k + 1)},$$

где вспомогательные величины  $P$  и  $Q$  вычисляются, соответственно, как

$$P = \text{tr}A - \text{tr}[X^T A X (X^T X)^{-1}],$$

$$Q = \text{tr}A^2 - 2\text{tr}[X^T A^2 X (X^T X)^{-1}] + \text{tr}\{[X^T A X (X^T X)^{-1}]^2\}.$$

Идея состоит в том, что математическое ожидание и дисперсия статистики  $d/4$  должны быть равны, соответственно, математическому ожиданию и дисперсии функции бета-распределения. Функция бета-распределения с параметрами  $p$  и  $q$  имеет математическое ожидание и дисперсию, соответственно,

$$Eb = \frac{4p}{p + q},$$

$$Db = \frac{16pq}{(p + q)^2 (p + q + 1)}.$$

Приравнявая соответствующие выражения для  $Ed$  и  $Eb$ , а также  $Dd$  и  $Db$ , получим

$$(p + q) = \frac{Ed(4 - Ed)}{Dd} - 1,$$

$$p = \frac{1}{4}(p + q)Ed.$$

После вычисления из второго выражения величины  $p$  можно вычислить  $q$  из первого выражения как

$$q = \frac{Ed(4 - Ed)}{Dd} - 1 - p.$$





См. монографии Аллисона (Allison), Белсли (Belsley) с соавт., Дрейпера (Draper) соавт., Коэна (Cohen) с соавт., Кука (Cook) соавт., Райана (Ryan), Фон Ай (von Eye) с соавт., Усипайкка (Uusipaikka), Чаттерджи (Chatterjee) с соавт., доклад Галченковой. О выбросах и влияющих наблюдениях см. монографии Кука (Cook) с соавт., статьи Кука. О статистике Дарбина–Уотсона см. статью Дарбина (Durbin) с соавт., монографию Хеннана.

## 14.2.7. Логистическая регрессия

В практических приложениях возникает ситуация, когда отклик эксперимента представлен в бинарном виде (1 – наличие признака, 0 – отсутствие признака). Множественная линейная регрессия не учитывает данное ограничение на выход модели. Для решения задачи может использоваться логит анализ (множественная логистическая регрессия).

Множественная логистическая регрессия может быть представлена в виде следующей модельной формулы

$$P_j(B) = \text{Logit}(X_j B) = \frac{1}{1 + e^{-X_j B}}, j = 1, 2, \dots, n,$$

где  $P_j(B)$ ,  $i = 1, 2, \dots, n$  – выход модели,

$B = \{b_i\}$ ,  $i = 1, 2, \dots, m$  – вектор–столбец весовых коэффициентов,

$X_j = \{x_{ij}\}$ ,  $i = 1, 2, \dots, m$ ;  $j = 1, 2, \dots, n$  – вектор–строка параметров объекта  $j$ , измеренных в эксперименте,

$X_j B$ ,  $j = 1, 2, \dots, n$  – множественная линейная регрессия (замечание о модели со свободным членом см. ниже),

$m$  – количество измеряемых в эксперименте параметров объекта,

$n$  – численность обучающей выборки (число объектов).

Значение  $P_j(\cdot)$  может быть интерпретировано как вероятность получения логитом значения 1 при подстановке в уравнение определенного вектора  $X_j$ ,  $j = 1, 2, \dots, n$ , измеренного в эксперименте.

Оптимальные значения весовых коэффициентов могут быть найдены путем максимизации логарифмической функции максимального правдоподобия (далее – ФМП)

$$\ln L = \sum_{j=1}^n [Y_j \ln P_j(B) + (1 - Y_j) \ln(1 - P_j(B))],$$

где  $Y_j$ ,  $j = 1, 2, \dots, n$  – выход эксперимента, соответствующий измеренному в эксперименте вектору параметров  $X_j$ ,  $j = 1, 2, \dots, n$ .

Задача сводится к системе нелинейных алгебраических уравнений

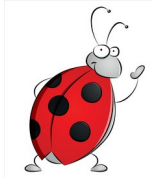
$$\frac{\partial \ln L}{\partial B} = 0,$$

для решения которой использован стандартный метод Ньютона–Рафсона. Итерационная схема метода записывается формулой

$$B_{k+1} = B_k - [H(B_k)]^{-1} g(B_k), k = 0, 1, \dots,$$

где  $k$  – номер итерации,

$H(\cdot)$  – матрица Гессе (матрица вторых производных) ФМП,



$g(.)$  – градиент (вектор производных) ФМП.

Вследствие аналитически доказанной сходимости итерационной схемы на всей области определения аргумента вектор начальных значений можно брать произвольным, для определенности и удобства вычислений – нулевым.

Решение задачи радикально упрощается благодаря известным выражениям вектора градиента и матрицы Гессе ФМП. Градиент ФМП имеет явное представление в виде (опуская номер итерации)

$$g(B) = \sum_{j=1}^n (Y_j - P_j(B)) X_j.$$

Матрица Гессе ФМП имеет явное представление в виде (опуская номер итерации)

$$H(B) = - \sum_{j=1}^n P_j(B)(1 - P_j(B)) X_j^T X_j.$$

Оценка качества регрессионной модели описана в одноименном разделе. Оценка значимости весовых коэффициентов также представлена.

Заметим, что если по условиям задачи требуется логит множественной линейной регрессии со свободным членом, в режиме обучения просто добавьте в массив исходных данных столбец из одних единиц, а при распознавании не забудьте в векторе каждого распознаваемого объекта также установить в данной позиции вектора единицу.

См. монографии Хосмер (Hosmer) с соавт., Шукри (Shoukri) с соавт., статьи Давнис с соавт., Дэвис (Davis) с соавт., Цвейг (Zweig) с соавт., Дэйвидсон (Davidson) с соавт., пособие Цыплакова. Введение в рассматриваемый предмет дано в обзорных статьях Паклина на сайте BaseGroup Labs. О решении систем нелинейных уравнений см. книгу Носача. О персептроне см. книги Осовского, Мандик (Mandic) с соавт., Минского с соавт., Круглова с соавт., Дюка с соавт.

## 14.2.8. Пробит анализ

В практических приложениях возникает ситуация, когда отклик эксперимента представлен в бинарном виде (1 – наличие признака, 0 – отсутствие признака). Множественная линейная регрессия не учитывает данное ограничение на выход модели. Для решения задачи может использоваться пробит анализ.

Пробит может быть представлен в виде следующей модельной формулы

$$P_j(B) = \text{probit}(X_j B) = \Phi(X_j B), j = 1, 2, \dots, n,$$

где  $P_j(B)$ ,  $i = 1, 2, \dots, n$  – выход модели,

$\Phi(.)$  – функция нормального распределения,

$B = \{b_i\}$ ,  $i = 1, 2, \dots, m$  – вектор–столбец весовых коэффициентов,

$X_j = \{x_i\}$ ,  $i = 1, 2, \dots, m$ ;  $j = 1, 2, \dots, n$  – вектор–строка параметров объекта  $j$ , измеренных в эксперименте,

$X_j B$ ,  $j = 1, 2, \dots, n$  – множественная линейная регрессия (замечание о модели со свободным членом см. ниже),

$m$  – количество измеряемых в эксперименте параметров объекта,





$n$  – численность обучающей выборки (число объектов).

Значение  $P_j(.)$  может быть интерпретировано как вероятность получения пробитом значения 1 при подстановке в уравнение определенного вектора  $X_j, j = 1, 2, \dots, n$ , измеренного в эксперименте.

Оптимальные значения весовых коэффициентов могут быть найдены путем максимизации логарифмической функции максимального правдоподобия (далее – ФМП)

$$\ln L = \sum_{j=1}^n [Y_j \ln P_j(B) + (1 - Y_j) \ln(1 - P_j(B))]$$

где  $Y_j, j = 1, 2, \dots, n$  – выход эксперимента, соответствующий измеренному в эксперименте вектору параметров  $X_j, j = 1, 2, \dots, n$ .

Задача сводится к системе нелинейных алгебраических уравнений

$$\frac{\partial \ln L}{\partial B} = 0,$$

для решения которой использован стандартный метод Ньютона–Рафсона. Итерационная схема метода записывается формулой

$$B_{k+1} = B_k - [H(B_k)]^{-1} g(B_k), k = 0, 1, \dots,$$

где  $k$  – номер итерации,

$H(.)$  – матрица Гессе (матрица вторых производных) ФМП,

$g(.)$  – градиент (вектор производных) ФМП.

Вследствие аналитически доказанной сходимости итерационной схемы на всей области определения аргумента вектор начальных значений можно брать произвольным, для определенности и удобства вычислений – нулевым.

Решение задачи радикально упрощается благодаря известным выражениям вектора градиента и матрицы Гессе ФМП. Градиент ФМП имеет явное представление в виде (опуская номер итерации)

$$g(B) = \sum_{j=1}^n V_j(B) X_j,$$

$$V_j(B) = f(X_j B) \frac{Y_j - P_j(B)}{P_j(B)(1 - P_j(B))}, j = 1, 2, \dots, n,$$

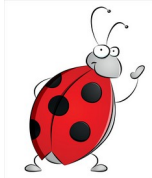
где  $f(X_j B), j = 1, 2, \dots, n$  – плотность вероятности нормальной случайной величины, – вспомогательные величины,

Матрица Гессе ФМП имеет явное представление в виде (опуская номер итерации)

$$H(B) = - \sum_{j=1}^n V_j(B) (V_j(B) + X_j B) X_j^T X_j.$$

Оценка качества регрессионной модели описана в одноименном разделе. Оценка значимости весовых коэффициентов также представлена.

Заметим, что если по условиям задачи требуется пробит со свободным членом, в режиме обучения просто добавьте в массив исходных данных столбец из одних единиц, а при



распознавании не забудьте в векторе каждого распознаваемого объекта также установить в данной позиции вектора единицу.

См. монографию и статью Кэмерон (Cameron) с соавт., работу Анселин (Anselin), монографии Дэвидсона (Davidson) с соавт., Лонга (Long), пособие Цыплакова. О решении систем нелинейных уравнений см. книгу Носача.

## 14.2.9. Регрессия Пуассона

Регрессия Пуассона является методом распознавания так называемых счетных данных, возникающих при подсчете количества каких-либо экспериментальных сущностей (например, числа бактерий в чашке Петри). Она может быть представлена в виде следующей модельной формулы

$$\mu_j(B) = \exp(X_j B), j = 1, 2, \dots, n,$$

где  $\mu_j(B)$ ,  $j = 1, 2, \dots, n$  – выход модели,

$B = \{b_i\}$ ,  $i = 1, 2, \dots, m$  – вектор–столбец весовых коэффициентов,

$X_j = \{x_i\}$ ,  $i = 1, 2, \dots, m$ ;  $j = 1, 2, \dots, n$  – вектор–строка параметров объекта  $j$ , измеренных в эксперименте,

$X_j B$ ,  $j = 1, 2, \dots, n$  – множественная линейная регрессия (замечание о модели со свободным членом см. ниже),

$m$  – количество измеряемых в эксперименте параметров объекта (регрессоров),

$n$  – численность обучающей выборки (число объектов).

В модели регрессии Пуассона параметр  $\mu$  интерпретируется как счетное количество, соответствующее вектору регрессоров  $X$ . При этом  $\mu$  является параметром распределения Пуассона, плотность которого имеет вид

$$P(Y = y) = \frac{e^{-\mu} \mu^y}{y!}, y = 0, 1, 2, \dots$$

Функция максимального правдоподобия (далее – ФМП) запишется как

$$L = \prod_{j=1}^n P(Y_j = y) = \prod_{j=1}^n \frac{\mu^{Y_j}}{Y_j!} e^{-\mu},$$

где  $Y_j$ ,  $j = 1, 2, \dots, n$  – выход эксперимента, соответствующий измеренному в эксперименте вектору параметров  $X_j$ ,  $j = 1, 2, \dots, n$ .

Тогда логарифмическая ФМП будет записана, с учетом модельной формулы регрессии, как

$$\ln L = \sum_{j=1}^n [-\exp(X_j B) + Y_j X_j B - \ln Y_j!]$$

Оптимальные значения весовых коэффициентов могут быть найдены путем максимизации логарифмической ФМП. Задача сводится к системе нелинейных алгебраических уравнений

$$\frac{\partial \ln L}{\partial B} = 0,$$

для решения которой использован стандартный метод Ньютона–Рафсона. Итерационная схема метода записывается формулой





$$B_{k+1} = B_k - [H(B_k)]^{-1}g(B_k), k = 0, 1, \dots,$$

где  $k$  – номер итерации,

$H(\cdot)$  – матрица Гессе (матрица вторых производных) ФМП,

$g(\cdot)$  – градиент (вектор производных) ФМП.

Вследствие аналитически доказанной сходимости итерационной схемы на всей области определения аргумента вектор начальных значений можно брать произвольным, для определенности и удобства вычислений – нулевым.

Решение задачи радикально упрощается благодаря известным выражениям вектора градиента и матрицы Гессе ФМП. Градиент ФМП имеет явное представление в виде (опуская номер итерации)

$$g(B) = \sum_{j=1}^n X_j(Y_j - X_j B).$$

Матрица Гессе ФМП имеет явное представление в виде (опуская номер итерации)

$$H(B) = - \sum_{j=1}^n \exp(X_j B) X_j^T X_j.$$

Оценка качества регрессионной модели описана в одноименном разделе. Оценка значимости весовых коэффициентов также представлена.

Если по условиям задачи требуется регрессия Пуассона со свободным членом, в режиме обучения просто добавьте в массив исходных данных столбец из одних единиц, а при распознавании не забудьте в векторе каждого распознаваемого объекта также установить в данной позиции вектора единицу.

См. монографию и статью Кэмерона (Cameron) с соавт., работу Анселин (Anselin), монографии Дэвидсона (Davidson) с соавт., Лонга (Long), пособие Цыплакова. О решении систем нелинейных уравнений см. книгу Носача. Большое число источников посвящено изучению явления сверхдисперсии (overdispersion), иногда возникающего при исследовании представленным методом и заключающегося в превышении дисперсии модельной оценки регрессии Пуассона над самой оценкой, т. е.  $D\lambda > E\lambda$ . О сверхдисперсии см. статьи Баррона (Barron), Бонинга (Bohning), Бреслоу (Breslow), Дина (Dean) и Дина с соавт., Дугласа (Douglas).

## 14.2.10. Оценка прогностической ценности параметров

Сравнительная оценка прогностической ценности параметров (применительно к логистической регрессии) представлена в работе Плавинской с соавт., причем в качестве альтернативы классической  $m$ -мерной множественной логистической регрессии использованы  $m$  логистических регрессий для каждого параметра объекта в отдельности.

Можно также упомянуть еще один эффективный способ оценки прогностической ценности, основанный на алгоритме Фаррара-Глаубера, представленном в главе «Матричная и линейная алгебра» (а именно, способ, который основан на вычислении коэффициентов детерминации, там же см. необходимые ссылки). Данный способ исследования мультиколлинеарности может оказаться практически полезным при решении проблемы оценки влияния того или



иного параметра, по предположению исследователя, характеризующего объект, в рассмотренных в настоящей главе методах распознавания образов с обучением. Параметры, имеющие значимые коэффициенты детерминации, рекомендуется исключить из рассмотрения как имеющие малое влияние на результаты распознавания. Исключение данных параметров помогает не только сократить объем вычислений и уменьшить вычислительную сложность решения (часто обеспечить саму возможность решения конкретным методом распознавания), но и, что самое важное, интерпретировать результат распознавания образов с привлечением существенно меньшего числа параметров (отбросив параметры, мало влияющие на качество распознавания), т. е. снизить размерность (а следовательно, и стоимость решения) задачи.

## Глава 15. Многомерное шкалирование

### 15.1. Введение

Рассматриваются классические методы многомерного шкалирования.

### 15.2. Теоретическое обоснование

Подобно методам факторного анализа, методы многомерного шкалирования используются для поиска структуры объектов, по терминологии многомерного шкалирования – стимулов – в многомерном пространстве (стимулом в многомерном шкалировании называют объект исследования, что соответствует понятию эмпирической выборки в прикладном статистическом анализе). Подобно методам кластерного анализа, методами многомерного шкалирования изучаются группировки объектов в многомерном пространстве. Методы многомерного шкалирования по цели расчета ближе к кластерному анализу – установлению пространственной конфигурации исследуемых стимулов.

Применяются следующие методы многомерного шкалирования:

- метрический метод Торгерсона,
- неметрический метод Краскела.

Методы отражают различные идеологические подходы к решению проблемы многомерного шкалирования.

#### 15.2.1. Метрики

Мера сходства  $d_{ij}$  между объектами  $i$  и  $j$  называется метрикой, если она удовлетворяет определенным условиям:

- симметрии  $d_{ij} = d_{ji}$ ,
- неравенству треугольника  $d_{ij} \leq d_{ik} + d_{kj}$ ,
- различимости нетождественных объектов и неразличимости тождественных объектов,





- иногда также ставят требование максимальной схожести объекта с «самим собой»

$$d_{ii} = \min_{i,j} d_{ij}, \text{ причем для рассматриваемых метрик всегда } d_{ii} = 0.$$

Метрика представляет собой меру сходства типа расстояния между стимулами, вычисленную по определенной формуле. В процессе попарного вычисления метрик между всеми объектами, составляющими матрицу исходных данных, получается так называемая матрица различий.

Основным элементом исследования в многомерном шкалировании является квадратная матрица различий  $D$  между стимулами, которая содержит  $p$  строк и столбцов, где  $p$  – количество стимулов, вычисленная на основе одной из метрик из матрицы первичных исходных данных. По диагонали матрицы различий обязательно располагаются нулевые значения (расстояния между стимулами «сами с собой»).

Допускается использовать готовые матрицы различий между стимулами, вычисленные заранее на основе любых других метрик или иных мер связи, поэтому возможности расчетов не ограничены используемыми метриками. Так, например, возможно получение элементов  $d_{ij}$ ,  $i = 1, 2, \dots, p$ ;  $j = 1, 2, \dots, p$ , матрицы различий  $D$  из корреляционной матрицы  $R$  по формуле

$$d_{ij} = \sqrt{1 - r_{ij}}, i = 1, 2, \dots, p; j = 1, 2, \dots, p,$$

где  $r_{ij}$ ,  $i = 1, 2, \dots, p$ ;  $j = 1, 2, \dots, p$  – коэффициент корреляции между стимулами  $i$ ,  $i = 1, 2, \dots, p$ , и  $j$ ,  $j = 1, 2, \dots, p$ .

Корреляционная матрица может быть построена с помощью методов главы «Корреляционный анализ», а все необходимые трансформации корреляционной матрицы – стандартными средствами.

В метрическом методе Торгерсона используется либо готовая матрица различий, либо матрица первичных исходных данных, по выбору. Напротив, неметрический метод Краскела работает только с матрицей первичных исходных данных и не работает с готовой матрицей различий.

## 15.2.1. Метрика Минковского

Наиболее общей классической мерой типа расстояния является метрика Минковского, которая определяет расстояние между стимулами  $i$ ,  $i = 1, 2, \dots, p$ , и  $j$ ,  $j = 1, 2, \dots, p$ .

$$d_{ij} = \sqrt[r]{\sum_{k=1}^n |x_{ik} - x_{jk}|^r},$$

где  $r$  – некоторая величина, причем  $r \geq 1$ ,

$n$  – размерность пространства,

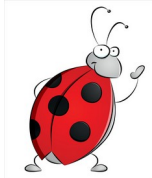
$x_{ik}$ ,  $i = 1, 2, \dots, p$  – проекция точки  $i$ ,  $i = 1, 2, \dots, p$ , на ось  $k$ ,  $k = 1, 2, \dots, p$ ,

$x_{jk}$ ,  $j = 1, 2, \dots, p$  – проекция точки  $j$ ,  $j = 1, 2, \dots, p$ , на ось  $k$ ,  $k = 1, 2, \dots, p$ .

Практически метрика вычисляется как мера между двумя выборками, численность каждой из которых равна  $n$ .

Используются некоторые частные случаи метрики Минковского:





- евклидова метрика,
- манхеттенское расстояние.

## 15.2.1.2. Евклидова метрика

Если в метрике Минковского положить  $r = 2$ , получим стандартное евклидово расстояние (евклидову метрику)

$$d_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2},$$

где  $n$  – размерность пространства,

$x_{ik}$ ,  $i = 1, 2, \dots, p$  – проекция точки  $i$ ,  $i = 1, 2, \dots, p$ , на ось  $k$ ,  $k = 1, 2, \dots, p$ ,

$x_{jk}$ ,  $i = 1, 2, \dots, p$  – проекция точки  $j$ ,  $j = 1, 2, \dots, p$ , на ось  $k$ ,  $k = 1, 2, \dots, p$ .

Практически метрика вычисляется как расстояние между двумя выборками, численность каждой из которых равна  $n$ . Геометрически евклидово расстояние представляет собой расстояние между двумя точками в  $n$ -мерном пространстве.

## 15.2.1.3. Манхеттенское расстояние

При  $r = 1$  метрика Минковского дает манхеттенское расстояние (метрику города, city block, Manhattan distance)

$$d_{ij} = \sum_{k=1}^n |x_{ik} - x_{jk}|,$$

где  $n$  – размерность пространства,

$x_{ik}$ ,  $i = 1, 2, \dots, p$  – проекция точки  $i$ ,  $i = 1, 2, \dots, p$ , на ось  $k$ ,  $k = 1, 2, \dots, p$ ,

$x_{jk}$ ,  $i = 1, 2, \dots, p$  – проекция точки  $j$ ,  $j = 1, 2, \dots, p$ , на ось  $k$ ,  $k = 1, 2, \dots, p$ .

Практически метрика вычисляется как мера между двумя выборками, численность каждой из которых равна  $n$ .

В многомерном шкалировании метрики, отличные от евклидова расстояния, применяются неохотно. Расчеты показывают, что матрицы различий, построенные на основе данных метрик, приводят к появлению отрицательных собственных значений при решении проблемы собственных значений матрицы скалярных произведений, что автоматически приводит к наличию комплексных собственных векторов, которые с трудом поддаются интерпретации в терминах многомерного шкалирования.

## 15.2.2. Метрический метод Торгерсона

Метрический метод многомерного шкалирования Торгерсона исходит в своих предпосылках из той идеи, что исходные данные об исследуемых объектах (параметры, посредством которых описаны объекты) являются результатами точных измерений, свободных от ошибок измерения. Задачей метода является представление конфигурации объектов в пространстве шкал меньшей размерности.





Метод основан на анализе так называемой матрицы скалярных произведений  $B$ . Данная матрица строится на основе матрицы различий  $D$  Метрики. Вычисления элементов матрицы скалярных произведений  $B$  производятся по формуле

$$b_{ij} = \frac{1}{2} \left( -d_{ij}^2 + \frac{1}{p} \sum_{i=1}^p d_{ij}^2 + \frac{1}{p} \sum_{j=1}^p d_{ij}^2 - \sum_{i=1}^p \sum_{j=1}^p d_{ij}^2 \right), i=1,2,\dots,p; j=1,2,\dots,p,$$

где  $d_{ij}$ ,  $i=1,2,\dots,p$ ;  $j=1,2,\dots,p$ , расстояния между стимулами  $i$ ,  $i=1,2,\dots,p$ , и  $j$ ,  $j=1,2,\dots,p$ , представляющие собой одну из разновидностей метрики Минковского, составляющие матрицу  $D$ ,

$p$  – количество стимулов.

Из решения стандартной проблемы собственных значений действительной симметрической матрицы скалярных произведений устанавливается представление стимулов в координатном пространстве осей шкальных значений. Подробнее о решении проблемы собственных значений и применяемых алгоритмах можно узнать из главы «Матричная и линейная алгебра».

Полученные координатные оси почти всегда могут быть содержательно интерпретированы. Кроме того, результаты анализа могут быть использованы для классификации стимулов. Основным результатом расчета является матрица шкальных значений  $Z$  (координат стимулов в пространстве шкал). Элементы матрицы координат стимулов вычисляются по формуле

$$z_{pq} = \sqrt{\lambda_q} y_{pq},$$

где  $q$ ,  $1 \leq q \leq n$  – количество удерживаемых максимальных собственных значений,

$n$  – размерность исходного пространства,

$\lambda_q$  –  $q$ -е собственное значение матрицы скалярных произведений,

$y_{pq}$  – собственный вектор матрицы скалярных произведений, соответствующий собственному значению  $\lambda_q$ .

Максимальное число координат стимулов  $q$  (количество осей шкальных значений, размерность пространства шкал) будет равно количеству  $p$  собственных значений матрицы скалярных произведений. Однако часть собственных значений  $\lambda_q$  на этапе вычислений может оказаться нулевой в вычислительном смысле. Кроме того, количество осей шкальных значений может быть скорректировано, исходя из анализа величин собственных значений (также их процентного содержания).

В реальных расчетах, как правило, происходит следующее: из величины собственных значений  $\lambda_q$  и их процентного содержания сразу ясно, сколько собственных значений нужно оставить, так как отброшенные значения являются очень малыми величинами, по модулю порядка  $10^{-10}$ .

Хотя в литературе описаны и другие методы выбора количества осей шкальных значений, общее правило состоит в том, что осей должно быть достаточно для содержательной интерпретации пространственной конфигурации стимулов. На практике часто ограничиваются двумя или тремя осями.

Если число осей шкальных значений больше или равно двум, дополнительно производится объективное вращение решения методом VARIMAX подобно тому, как это сделано в



факторном анализе, который подробно рассмотрен в одноименной главе. Процедура вращения не изменяет взаимную пространственную координацию стимулов, но часто улучшает интерпретируемость решения путем сдвига гроздей стимулов в координатном пространстве ближе к той или иной оси шкал.

Хотя это и не является необходимым этапом решения, результаты анализа рекомендуется изобразить графически, а для размерности пространства более двух или трех графики должны быть представлены двумерными срезами пространства.

### 15.2.3. Неметрический метод Краскела

Все методы неметрического многомерного шкалирования, в отличие от метрического многомерного шкалирования, исходят в своих предпосылках из той идеи, что данные об исследуемых объектах являются результатами измерений, искаженных ошибками. Поэтому причинами возможного неудовлетворительного представления конфигурации объектов в пространстве меньшей размерности в неметрическом шкалировании считаются ошибки в исходных данных. Параметры, описывающие объекты и заданные матрицей исходных данных, в неметрическом шкалировании в процессе решения изменяются, чтобы получить лучшее представление конфигурации стимулов в пространстве шкал меньшей размерности. Стрессом в неметрическом многомерном шкалировании называют функционал, подлежащий минимизации. Имеются различные формы стресса. Здесь мы используем стресс в одной из классических форм

$$S = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^p (d_{ij} - \hat{d}_{ij})^2}{\sum_{i=1}^p \sum_{j=1}^p d_{ij}^2}},$$

где  $d_{ij}$ ,  $i = 1, 2, \dots, p$ ;  $j = 1, 2, \dots, p$ , элементы матрицы различий, представляющие собой расстояния между стимулами  $i$ ,  $i = 1, 2, \dots, p$ , и  $j$ ,  $j = 1, 2, \dots, p$ , вычисленные по формуле одной из разновидностей метрики Минковского по полной матрице исходных данных, элементами которой являются величины  $x_{kl}$ ,  $k = 1, 2, \dots, p$ ;  $l = 1, 2, \dots, n$ ,

$n$  – размерность исходного пространства,

$\hat{d}_{ij}$ ,  $i = 1, 2, \dots, p$ ;  $j = 1, 2, \dots, p$ , элементы оценки матрицы различий, представляющие собой оценки расстояний  $d_{ij}$ ,  $i = 1, 2, \dots, p$ ;  $j = 1, 2, \dots, p$ , вычисленные в пространстве шкал меньшей размерности, элементами которой являются величины  $x_{kl}$ ,  $k = 1, 2, \dots, p$ ;  $l = 1, 2, \dots, q$ ,

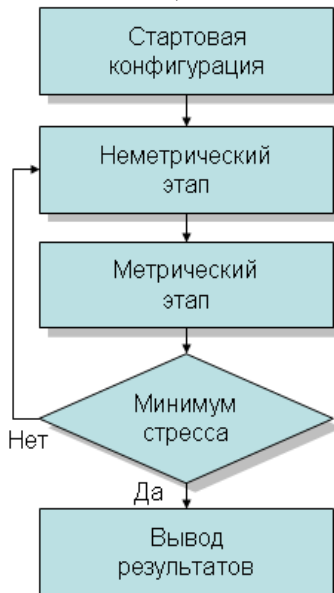
$q$  – размерность пространства шкал,

$p$  – количество стимулов.

Алгоритм представляет собой итерационный процесс. Каждая итерация включает в себя метрический этап, применяемый для получения оценки матрицы различий, и неметрический этап, заключающийся в подборе оценки матрицы исходных данных, которая будет минимизировать стресс. Целью вычислений является подбор такой конфигурации матрицы



исходных данных, чтобы обеспечить минимум стресса по величинам  $x_{kl}$ ,  $k = 1, 2, \dots, p$ ;  $l = 1, 2, \dots, n$ . Общая схема алгоритма показана на рисунке.



Метрический этап подробно описан в разделе, посвященном методу Торгерсона, поэтому подробнее остановимся на неметрическом этапе.

Ряд современных авторов предлагает для минимизации стресса использовать различные универсальные методы оптимизации, от градиентных методов и до нейронных сетей. Универсальные алгоритмы, использующие численное определение градиентов и других параметров схемы алгоритмов безусловной оптимизации, эффективно можно использовать, когда вид целевой функции неизвестен заранее. Если же вид целевой функции известен и, более того, весьма прост, то более плодотворным оказывается подход Краскела, который предполагает аналитическое вычисление градиента. Минимизация рассматриваемой формы стресса приводит к формуле, называемой градиентным алгоритмом Краскела. Алгоритм записывается в рекурсивном виде как

$$\hat{x}_{kl}^{(c+1)} = \hat{x}_{kl}^{(c)} - \frac{2\alpha}{B} \sum_{\substack{i=1 \\ i \neq k}}^p \left[ \frac{\hat{d}_{ik}^{(c)} - \hat{d}_{ik}^{(c+1)}}{\hat{d}_{ik}^{r-1}} \left| \hat{x}_{il}^{(c)} - \hat{x}_{kl}^{(c)} \right|^{r-1} \text{sign}(\hat{x}_{il}^{(c)} - \hat{x}_{kl}^{(c)}) \right], k=1, 2, \dots, p; l=1, 2, \dots, n,$$

где  $c$  – номер итерации, при  $c = 0$  стартовая конфигурация алгоритма задается метрическим методом Торгерсона, максимально допустимое число итераций задается, знак «крышечка» означает оценки величин, вычисленные на этапе итерации,  $\alpha$  – параметр, называемый шагом итерации, от которого может зависеть как скорость сходимости итерационного процесса, так и сама сходимость алгоритма; значение параметра «по умолчанию» может быть изменено пользователем,



$r$  – величина, применяемая при вычислении метрики Минковского в той или иной форме, причем  $r \geq 1$ , часто используются случаи евклидова расстояния ( $r = 2$ ) и манхеттенского расстояния ( $r = 1$ ),

$\text{sign}(\cdot)$  – знак выражения,

$$B = \sum_{i=1}^p \sum_{j=1}^p \hat{d}_{ij}^2.$$

Преобразования, выполненные над матрицей исходных данных в ходе итерационного процесса, представляют собой применение к исходным данным некоторой неизвестной монотонной функции. Итерационный процесс завершается при достижении минимума стресса, о чем свидетельствует отличие между стрессом на текущей и предыдущей итерациях на некоторую заранее заданную пользователем малую величину  $\varepsilon$ , например, 0,00001. Условие останова может быть записано в виде

$$|S^{(c+1)} - S^{(c)}| \leq \varepsilon,$$

где  $c$  – номер итерации.

Аварийным завершением процесса считают его завершение не по минимуму стресса, а по достижении определенного заранее заданного числа итераций.

Если число осей шкальных значений больше или равно двум, дополнительно производится объективное вращение решения методом VARIMAX подобно тому, как это сделано в факторном анализе. Процедура вращения не изменяет взаимную пространственную координацию стимулов, но часто улучшает интерпретируемость решения путем сдвига гроздей стимулов в координатном пространстве ближе к той или иной оси шкал.

Кроме возможности пользовательского ввода параметров, предусмотрен вывод динамики стресса по итерациям. Данная возможность может быть полезна при исследовании сходимости процесса.

Хотя это и не является необходимым этапом решения, результаты анализа рекомендуется изобразить графически, а для размерности пространства более двух или трех графики должны быть представлены двумерными срезами пространства. Динамика стресса по итерациям, при необходимости, также представляется графически.

#### 15.2.4. Проблема вращения

Оси координат пространства шкал ортогональны, и их направления устанавливаются последовательно, по максимуму оставшейся дисперсии. Более предпочтительное положение системы координат получают путем вращения этой системы вокруг ее начала.

Пространственная конфигурация стимулов в результате применения этой процедуры остается неизменной. Целью вращения является нахождение одной из возможных систем координат для получения так называемой простой структуры. Применяют популярный метод вращения VARIMAX. Подробное описание метода VARIMAX приводится во многих источниках по вычислительным методам статистики, например, у Магнуса, в первом выпуске Сборника научных программ на Фортране.



## Глава 16. Обработка экспертных оценок

### 16.1. Введение

Рассматриваются методы обработки экспертных оценок. Отметим, что представленные методы не исчерпывают всех возможностей обработки экспертных оценок. Примеры таких решений:

- Выявление однородных групп экспертов может быть выполнено с помощью соответствующих методов кластерного анализа, многомерного шкалирования, факторного анализа.
- Нахождение согласованного мнения группы экспертов может быть выполнено с помощью соответствующих методов дисперсионного анализа.
- Исследование корреляции экспертных оценок может быть выполнено с помощью соответствующих методов корреляционного анализа.

Остановимся более подробно на выявлении однородных групп экспертов. Для этого рекомендуется воспользоваться методами главы «Кластерный анализ». Из представленных методов классификации можно применять только метод средней связи Кинга в комбинации с мерой различия «Расстояние отношений», вычисляемое на основе матриц отношений частичного порядка.

См. книги Бешелева с соавт., Нейлора. Обзоры см. в книгах Литвака (1982), Тюрина, статье Шмерлинг с соавт. Об организации экспертной работы см. книгу Литвака (2004).

### 16.2. Теоретическое обоснование

Математико–статистические методы обработки экспертных оценок необходимо применять во всех случаях, когда исходные данные представляют собой результаты работы экспертов или экспертных комиссий, и требуется найти обоснованное согласованное мнение группы экспертов для представления результатов лицу, принимающему решение. Попутно могут быть решены и другие, частные, задачи типа получения весовых коэффициентов объектов и весовых коэффициентов компетентности экспертов.

Применяются различные методы обработки экспертных оценок, предназначенные для решения следующих основных задач:

Методы оценивания:

- Метод парных сравнений Терстоуна.
- Метод группового оценивания.

Методы исследования согласованности мнений экспертов:

- Коэффициент конкордации.
- Альфа Кронбаха.

Методы получения коллективного мнения:

- Метод средних рангов.







- Медиана Кемени.
- Среднее Кемени.

Представленные методы отражают различные подходы к решению однотипных задач и при определенных условиях могут давать одинаковые результаты.

Специфические для каждого метода алгоритмы требуют, чтобы исходные данные имели определенную структуру. Исходные данные, содержащие экспертные оценки, могут быть различных видов. Методы рассчитаны на обработку исходной матрицы определенной структуры, поэтому при обработке тех или иных исходных данных следует убедиться, что применяются адекватные методы. Конкретные требования к исходным данным представлены при описании методов расчета. Ниже рассмотрены основные матрицы исходных данных.

Матрица парных сравнений (матрица предпочтений)  $A$  должна иметь следующий вид:

Объекты	$A_1$	$A_2$	...	$A_n$
$A_1$	$a_{11}$	$a_{12}$	...	$a_{1n}$
$A_2$	$a_{21}$	$a_{22}$	...	$a_{2n}$
...	...	...	...	...
$A_n$	$a_{n1}$	$a_{n2}$	...	$a_{nn}$

Элементы матрицы парных сравнений получаются как:

$$a_{ij} = \begin{cases} 0, & A_i < A_j, \\ 1, & A_i \sim A_j, \\ 2, & A_i > A_j, \end{cases} \quad i, j = 1, 2, \dots, n,$$

где  $n$  – количество объектов,

$<, \sim, >$  – расширения на множества математических операций, соответственно,  $<, \approx, >$ .

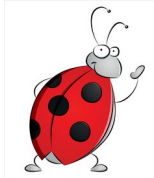
Операция  $A_i > A_j$  означает, что объект  $A_i$  превосходит, по мнению эксперта, объект  $A_j$ . Иначе говоря, запись  $A_i > A_j$  означает предпочтение объекта  $i$  над объектом  $j$ . Если эксперт затрудняется в выборе варианта предпочтений, имеет место так называемая связь, обозначаемая как  $\sim$ .

Матрица опроса  $P$  должна иметь следующий вид:

Объекты	Эксперты			
	$E_1$	$E_2$	...	$E_m$
$A_1$	$p_{11}$	$p_{12}$	...	$p_{1m}$
$A_2$	$p_{21}$	$p_{22}$	...	$p_{2m}$
...	...	...	...	...
$A_n$	$p_{n1}$	$p_{n2}$	...	$p_{nm}$

Элементы  $p_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, m$ , где  $n$  – количество объектов,  $m$  – количество экспертов, матрицы опроса могут быть как численными значениями (в том числе весами, приписанными экспертами данным объектам), так и ранжировками. Однако некоторые методы анализа могут иметь специфические требования к виду матрицы опроса, поэтому необходимо проявлять внимательность.





Ранжировка, данная одним экспертом, может быть представлена матрицей отношений частичного порядка  $P$  (существуют и другие типы матрицы отношений), элементы которой вычисляются по формуле, в терминологии и обозначениях множеств,

$$p_{ij} = \begin{cases} 1, \text{если } (a_i, a_j) \in P, (a_j, a_i) \notin P, \\ 0, \text{если } (a_i, a_j) \notin P, (a_j, a_i) \notin P, \\ -1, \text{если } (a_i, a_j) \notin P, (a_j, a_i) \in P. \end{cases}$$

Таким образом,  $m$  матриц отношений эквивалентны рассмотренной выше матрице опроса. Использование матриц отношений вызвано их удобством при изложении некоторых методов обработки экспертных оценок. Если объекты заданы числами, вычисление производится по формуле

$$p_{ij} = \begin{cases} 1, \text{если } a_i > a_j, \\ 0, \text{если } a_i = a_j, \\ -1, \text{если } a_i < a_j. \end{cases}$$

В литературе применяются также иные формы матрицы предпочтений. См. обзор Шмерлинга с соавт., книгу Литвака (1982). См. книгу Тюрина, учебные пособия Тиняковой, Эйтингона с соавт., статью Шмерлинга.

### 16.2.1. Парные сравнения

Метод парных сравнений Терстоуна в качестве исходных данных для анализа использует матрицу парных сравнений, применительно к результатам опроса одного эксперта.

Данная квадратная матрица обладает следующими основными свойствами:

- Матрица несимметрическая.
- Матрица действительная.
- Матрица неотрицательная.

Для таких матриц, согласно теореме Перрона и ее обобщению – теореме Фробениуса (Гантмахер, с. 334), всегда имеется действительное положительное собственное число. Этому положительному числу, превосходящему по модулю все остальные собственные числа, соответствует собственный вектор с положительными координатами. Данный вектор может быть интерпретирован в качестве весового вектора (вектора весовых коэффициентов), служащего решением задачи.

Для нахождения максимального по модулю собственного значения и соответствующего собственного вектора матрицы с указанными свойствами может быть применен достаточно просто реализуемый степенной метод (Деммель, с. 165), алгоритм которого записывается как

$$y_{i+1} = Ax_i,$$

$$x_{i+1} = y_{i+1} / \|y_{i+1}\|_2,$$

$$\lambda_{i+1} = x_{i+1}^T A x_{i+1},$$





где  $i$  – номер итерации,

$x$  – искомый собственный вектор весовых коэффициентов,

$\lambda$  – соответствующее собственное число (в решении задачи не используется),

$A$  – матрица парных сравнений.

Итерационный процесс повторяется циклически до достижения требуемой точности, заданной малой величиной типа 0,000001. Для оценки точности используется сравнение Евклидовых норм собственных векторов, вычисленных на текущей и на предыдущей итерации.

См. учебное пособие Тиняковой.

## 16.2.2. Групповое оценивание

Метод группового оценивания в качестве исходных данных для анализа использует матрицу опроса.

Пусть  $P$  – матрица опроса, имеющая размеры  $n$  строк на  $m$  столбцов, где  $n$  – количество объектов,  $m$  – количество экспертов.

Квадратные матрицы  $PP^T$  и  $P^TP$  обладают следующими основными свойствами:

- Матрицы симметрические.
- Матрицы действительные.
- Матрицы положительно полуопределенные.

Для таких матриц все собственные значения неотрицательны, в силу чего все собственные вектора действительные.

Собственный вектор  $p$  размером  $n$ , соответствующий максимальному собственному числу матрицы  $PP^T$ , может быть интерпретирован в качестве вектора групповой оценки (весовых коэффициентов объектов).

Собственный вектор  $v$  размером  $m$ , соответствующий максимальному собственному числу матрицы  $P^TP$ , может быть интерпретирован в качестве вектора компетентности экспертов (весовых коэффициентов компетентности). К данному параметру – коэффициентам компетентности, как и к прочим результатам анализа, следует относиться внимательно. Не следует понимать большое значение коэффициента как признак профессиональной компетентности эксперта. Данный весовой коэффициент всего лишь означает близость оценки этого эксперта к некоторой согласованной оценке всей группы экспертов. Не нужно пояснять, что большинство по объективным и субъективным причинам может ошибаться весьма часто.

Для нахождения всех собственных значений и соответствующих собственных векторов матрицы с указанными свойствами может быть применен один из многочисленных алгоритмов, например, достаточно эффективный для данного типа задач метод Якоби (Уилкинсон с соавт., с. 182).

Все изложенные в настоящем разделе методики анализа могут быть реализованы непосредственно с помощью методов главы «Матричная и линейная алгебра».

См. учебное пособие Тиняковой.



### 16.2.3. Коэффициент конкордации

Коэффициент конкордации (согласованности) Кендалла предназначен для исследования, хорошо ли согласуются друг с другом представленные экспертами ранжировки. Вычисление коэффициента конкордации производится по формуле

$$W = \frac{\sum_{i=1}^n \left( \sum_{j=1}^k x_{ij} - \frac{k(n+1)}{2} \right)^2}{\frac{1}{12} k^2 n(n^2 - 1) - k \sum_{j=1}^k B_j},$$

где  $x_{ij}$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, k$  – массив ранговых оценок,

$n$  – число объектов,

$k$  – число экспертов.

$B_j$ ,  $j = 1, 2, \dots, k$  – поправки на объединение рангов в оценках экспертов, вычисляемые по формуле

$$B_j = \frac{1}{12} \sum_{l=1}^m n_l(n_l^2 - 1),$$

где  $m$  – число групп объединенных рангов в данной экспертной оценке,

$n_l$ ,  $l = 1, 2, \dots, m$  – число рангов в  $l$ -ой группе.

Оценивается значимость вычисленного показателя на том основании, что величина  $k(n-1)W$  распределена как  $\chi^2$  с числом степеней свободы, равном  $n-1$ . Малая величина вычисленного  $P$ -значения означает, что представленные экспертами ранжировки хорошо согласованы. В противном случае можно предположить, что ранжировки неоднородны. В этом случае рекомендуется применить методы кластерного анализа (см. главу «Кластерный анализ») с использованием соответствующей случаю меры различия для выявления согласованных групп экспертов.

См. в книги Айвазяна с соавт., Большева с соавт., Джонсона с соавт., работы Шмерлинга, Лежандра (Legendre). Связь со статистикой Фридмана рассмотрена Тюриным.

### 16.2.4. Метод средних рангов

Метод средних рангов (средних арифметических рангов, упорядочение по сумме рангов) представляет собой разумный выбор согласованного мнения группы экспертов, матрица опроса которых представляет собой ранжировки.

Суть метода заключается в следующем:

- Мнения экспертов ранжируются (если это не было сделано заранее).
- Подсчитывается сумма рангов для каждого объекта.
- Массив сумм рангов объектов ранжируется, представляя решение задачи.

См. монографию Кендалла (Кендэла), статью Шмерлинга.



## 16.2.5. Медиана Кемени

Медиана Кемени (медиана Кемени–Снелла, Kemeny–Snell median) представляет собой выбор согласованного мнения группы экспертов, матрица опроса которых представляет собой ранжировки.

Пусть матрица опроса имеет размеры  $n$  строк на  $m$  столбцов, где  $n$  – количество объектов,  $m$  – количество экспертов. Запишем заданное множество ранжировок как  $\{P_1, P_2, \dots, P_m\}$ . Пусть  $d(P, P_i)$  – расстояние между произвольной ранжировкой  $P$  и ранжировкой  $P_i$ ,  $i = 1, 2, \dots, m$ . Тогда некоторая ранжировка  $P$ , принадлежащая множеству заданных ранжировок и удовлетворяющая выражению

$$M\{P_1, P_2, \dots, P_m\} = \arg \min_P \sum_{i=1}^m d(P, P_i),$$

называется медианой Кемени. Напомним, что обобщение понятия медианы (медианы множества) на произвольные шкалы введено нами в главе «Описательная статистика».

Расстояние между ранжировками  $k$  и  $l$  определяется по формуле

$$d(P_k, P_l) = \sum_{i=1}^n \sum_{j=1}^m |p_{ij}^{(k)} - p_{ij}^{(l)}|,$$

где  $p_{ij}^{(k)}$ ,  $i=1, 2, \dots, n$ ;  $j=1, 2, \dots, m$ , – элементы матриц отношений частичного порядка ранжировок  $k$  и  $l$ , соответственно, которые вычисляются на основе матрицы опроса, как это описано в разделе «Обработка экспертных оценок».

По определению медиана Кемени ищется только среди ранжировок, заданных анализируемой матрицей опроса. Решением будет ранжировка, представленная одним из экспертов.

Предыстория вопроса и основные методы рассмотрены в докладе Буры (Bury). О медиане Кемени см. монографии Литвака, книгу Тюрина, работу Сумкина. Примеры практического применения даны в статьях Богомолова, Глухова с соавт. Связь медианы Кемени с другими показателями (например, коэффициентами ранговой корреляции) рассмотрена в брошюре Тюрина, обзоре Шмерлинга с соавт.

## 16.2.6. Среднее Кемени

Среднее значение Кемени (среднее Кемени–Снелла, Kemeny–Snell mean) представляет собой выбор согласованного мнения группы экспертов, матрица опроса которых представляет собой ранжировки.

Пусть матрица опроса имеет размеры  $n$  строк на  $m$  столбцов, где  $n$  – количество объектов,  $m$  – количество экспертов. Запишем заданное множество ранжировок как  $\{P_1, P_2, \dots, P_m\}$ . Пусть  $d(P, P_i)$  – расстояние между произвольной ранжировкой  $P$  и ранжировкой  $P_i$ ,  $i = 1, 2, \dots, m$ . Тогда некоторую произвольную ранжировку (без связей, иначе, без совпадающих вариантов)  $P$ , удовлетворяющую выражению

$$M\{P_1, P_2, \dots, P_m\} = \arg \min_P \sum_{i=1}^m d(P, P_i),$$





назовем средним Кемени. Напомним, что обобщение понятия среднего значения на произвольные шкалы измерения введено нами в главе «Описательная статистика». Расстояние между ранжировками  $k$  и  $l$  определяется по формуле

$$d(P_k, P_l) = \sum_{i=1}^n \sum_{j=1}^n |p_{ij}^{(k)} - p_{ij}^{(l)}|,$$

где  $p_{ij}^{(k)}, i=1,2,\dots,n; j=1,2,\dots,n$ , – элементы матриц отношений частичного порядка ранжировок  $k$  и  $l$ , соответственно, которые вычисляются на основе матрицы опроса. По определению, среднее Кемени ищется среди всех  $n!$  возможных ранжировок весьма неэффективным методом полного перебора, поэтому ввиду трудоемкости вычислений число вариантов искусственно ограничено нами величиной 8. Данное ограничение – предельное для диалоговой системы с использованием метода генерации перестановок стандартным антилексикографическим методом и для современного уровня развития вычислительной техники. Данной величины достаточно для многих практических применений. О сложности задачи см. статью Вакабаяси (Wakabayashi). Отметим также разработки Литвака (1982), посвященные решению проблемы вычислительной сложности, а также Тюриня. Для ранжировок без связей среднее Кемени, как правило, совпадает с медианой Кемени. Принципиальное различие между данными показателями будет наблюдаться в случае обработки ранжировок со связями, поэтому применение среднего Кемени рекомендуется в задачах такого рода.

Предыстория вопроса и основные методы решения рассмотрены в докладе Буры (Bury). См. также Тюриня, Богомоллова, Глухова с соавт., обзор Шмерлинга с соавт. Методов методов генерации перестановок см. Липский, Скиена. О генерации перестановок см. также учебник Новикова.

## 16.2.7. Альфа Кронбаха

Статистика альфа Кронбаха (Cronbach's alpha) применяется для оценки надежности статистических тестов. Альфа рассчитывается по формуле

$$\alpha = \frac{SS_{row} - SS}{SS_{row}},$$

$$SS_{row} = \frac{1}{c} \sum_{i=1}^r T_i^2 - \frac{T_{..}^2}{rc}$$

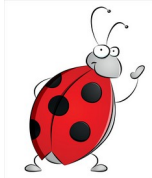
где – средний квадрат строк,

$$SS = \sum_{i=1}^r \sum_{j=1}^c x_{ij}^2 - \frac{T_{..}^2}{rc}$$

– средний квадрат погрешности,

$$T_i = \sum_{j=1}^c x_{ij}, i=1,2,\dots,r$$

– суммы строк,



$$T_{..} = \sum_{i=1}^r \sum_{j=1}^c x_{ij}$$

– общая сумма,

$c$  – число столбцов (выборок),

$r$  – число строк (параметров).

Рассчитываются также доверительные интервалы оцениваемой альфы. Нижняя граница доверительного интервала оцениваемой альфы считается как

$$L_{\alpha} = 1 - (1 - \alpha) F_{r-1, (r-1)(c-1)}^{-1}((1 - \beta)/2),$$

где  $F_{..}^{-1}(\cdot)$  – обратная функция  $F$ –распределения,

$\beta$  – доверительный уровень, выраженный в долях.

Верхняя граница доверительного интервала оцениваемой альфы считается как

$$H_{\alpha} = 1 - (1 - \alpha) F_{r-1, (r-1)(c-1)}^{-1}((1 - \beta)/2).$$

См. монографии Аванесова, Цыганова, Суен (Suen) и Суен с соавт., статьи Фельдт с соавт., Ван Зил (Van Zyl) с соавт., Якобуччи (Iacobucci) с соавт., Бланд (Bland) с соавт., Кистнер (Kistner) с соавт., Бонетт (Bonett), Панди (Pandey) с соавт., Суен с соавт., Аванесова.

## Глава 17. Анализ выживаемости

### 17.1. Введение

Анализ выживаемости (анализ данных типа времени жизни) исследует особые параметры, имеющие в различных отраслях знаний следующие наименования:

- в медико–биологических науках – время жизни,
- в общественных науках – длительность до момента прекращения,
- в технических науках – наработка до отказа.

### 17.2. Теоретическое обоснование

Методы анализа выживаемости могут применяться как в клинических исследованиях, так и при оценке надежности технических систем по цензурированным выборкам.

Параметрическая статистическая модель выживания описывается с помощью следующих функций:

- функция плотности распределения  $f(t)$ ,
- [кумулятивная или интегральная] функция распределения  $F(t) = P(T \leq t)$  – функция распределения длительностей до момента отказа  $t$ ,
- функция выживания  $S(t) = 1 - F(t)$  – вероятность безотказной работы до момента  $t$ ,

- функция интенсивности отказов (функция риска)  $h(t) = \frac{f(t)}{1 - F(t)}$ .





Под длительностью  $t$  может пониматься время жизни, количество циклов до отказа и т.п., в зависимости от конкретной задачи.

Предлагаются следующие методы расчета:

- Вычисление оценки функции выживания.
- Вычисление оценки функции риска.
- Подбор теоретического распределения.
- Критерий Кокса.
- Критерий Гехана.
- Модель пропорциональных рисков Кокса.

Все методы предполагают использование особых величин, называемых индикаторами цензурирования. Индикаторы могут принимать следующие стандартные значения:

- 1 – пациент умер по причине, связанной с исследуемой патологией,
- 0 – пациент цензурирован, т. е. выбыл из исследования, и его состояние либо неизвестно в момент исследования, либо он умер по причине, не связанной с исследуемой патологией.

Медицинская терминология, использованная в предыдущем абзаце, естественно обобщается на технические, социальные и любые другие системы.

См. монографию Хана с соавт.

## 17.2.1. Функция выживания

Оценка Каплана–Мейера функции выживания, в источниках называемая также множительной оценкой, вычисляется по формуле

$$\hat{S}(t \geq t_j) = \hat{S}(t_j) = \prod_{i=1}^j \left( 1 - \frac{d_i}{r_i} \right), j = 1, 2, \dots, K,$$

где  $d_i$  – количество наблюдений, моменты прекращения которых наблюдались с длительностью  $t_i$ ,  $i = 1, 2, \dots, K$ ,

$K$  – число моментов прекращения,

$r_i$  – количество наблюдений, незаконченных либо цензурированных к моменту  $t_i$ ,  $i = 1, 2, \dots, K$ , причем

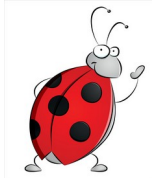
$$r_j = \sum_{i=j}^K (m_i + d_i), j = 1, 2, \dots, K,$$

где  $m_i$  – количество наблюдений, цензурированных между моментами  $t_i$  и  $t_{i+1}$ .

Дисперсия оценки Каплана–Мейера функции выживания вычисляется по формуле Гринвуда

$$D\hat{S}(t_j) = [\hat{S}(t_j)]^2 \sum_{i=1}^j \frac{d_i}{r_i(r_i - d_i)}, j = 1, 2, \dots, K.$$

На хвосте распределения оценка по формуле Гринвуда может не существовать, поэтому в данном случае используется формула Пето



$$D\hat{S}(t_j) = [\hat{S}(t_j)]^2 \frac{1 - \hat{S}(t_j)}{r_j}.$$

Доверительный интервал оцениваемой функции выживания вычисляется по асимптотической формуле

$$I_S(t_j) = \left( \hat{S}(t_j) - \Psi((1 + \beta)/2) \sqrt{D\hat{S}(t_j)}; \hat{S}(t_j) + \Psi((1 + \beta)/2) \sqrt{D\hat{S}(t_j)} \right), j = 1, 2, \dots, K,$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,  $\beta$  – доверительный уровень, выраженный в долях.

В случае малых выборок для вычисления доверительного интервала предлагается использовать уточненную формулу

$$I_S(t_j) = \left( \hat{S}(t_j)^{\exp\left[-\Psi((1+\beta)/2)\sqrt{D\hat{S}(t_j)}/(\hat{S}(t_j)\log\hat{S}(t_j))\right]}; \hat{S}(t_j)^{\exp\left[\Psi((1+\beta)/2)\sqrt{D\hat{S}(t_j)}/(\hat{S}(t_j)\log\hat{S}(t_j))\right]} \right), j = 1, 2, \dots, K.$$

Функция выживания изображается в виде ступенчатого графика.

См. монографии Кокса с соавт., Власова, Аален (Aalen) с соавт.

### 17.2.2. Функция риска

Оценка Каплана–Мейера функции риска вычисляется по формуле

$$\hat{H}(t \geq t_j) = \hat{H}(t_j) = \sum_{i=1}^j \frac{d_i}{r_i}, j = 1, 2, \dots, K,$$

где  $d_i$  – количество наблюдений, моменты прекращения которых наблюдались с длительностью  $t_i$ ,  $i = 1, 2, \dots, K$ ,

$K$  – число моментов прекращения,

$r_i$  – количество наблюдений, незаконченных либо цензурированных к моменту  $t_i$ ,  $i = 1, 2, \dots, K$ , причем

$$r_j = \sum_{i=j}^K (m_i + d_i), j = 1, 2, \dots, K,$$

где  $m_i$  – количество наблюдений, цензурированных между моментами  $t_i$  и  $t_{i+1}$ .

Дисперсия оценки Каплана–Мейера функции риска вычисляется по формуле

$$D\hat{H}(t_j) = \frac{D\hat{S}(t_j)}{\hat{S}(t_j)}, j = 1, 2, \dots, K,$$

где  $D\hat{S}(t_j)$ ,  $j = 1, 2, \dots, K$ , – дисперсия функции выживания,

$\hat{S}(t_j)$ ,  $j = 1, 2, \dots, K$ , – оценка Каплана–Мейера функции выживания.

Доверительный интервал оцениваемой функции риска вычисляется как

$$I_H(t_j) = \left( \hat{H}(t_j) - \Psi((1 + \beta)/2) \sqrt{D\hat{H}(t_j)}; \hat{H}(t_j) + \Psi((1 + \beta)/2) \sqrt{D\hat{H}(t_j)} \right), j = 1, 2, \dots, K,$$

где  $\Psi(\cdot)$  – обратная функция стандартного нормального распределения,







$\beta$  – доверительный уровень, выраженный в долях.

Функция риска изображается либо в виде ступенчатого графика либо в виде ломаной линии, соединяющей заданные точки, соответствующие моментам прекращения.

См. монографии Кокса с соавт., Власова, Аален (Aalen) с соавт., статью Аален.

### 17.2.3. Оценка параметра положения

Среднее значение (параметр положения) и дисперсия (параметр масштаба) длительности в задаче анализа данных типа времени жизни, как параметрические оценки параметра положения и параметра масштаба (если он имеет смысл), зависят от выбранного типа теоретического распределения представленных эмпирических данных. Для некоторых моделей оценок параметров в виде простых формул не существует – оценка возможна лишь вычислительными методами.

Непараметрические оценки параметра положения – точечная оценка медианы и ее интервальная оценка. Точечная оценка медианы имеет непосредственное отношение к оценке Каплана–Мейера:

$$\hat{m}_{0,5} = \inf\{t : \hat{S}(t) \leq 0,5\},$$

где  $\hat{S}(t)$  – оценка Каплана–Мейера функции выживания,  
 $t$  – длительность.

Последняя формула означает, что в качестве точечной оценки медианы берется такая минимальная длительность  $t$  (из представленных эмпирических длительностей), для которой выполняется неравенство  $\hat{S}(t) \leq 0,5$ .

Доверительный интервал оцениваемой медианы задается формулой

$$I_m = (y_c; y_{n+1-c}),$$

где  $y_i, i = 1, 2, \dots, n$ , т. е. упорядоченные по возрастанию длительности,

$c$  – параметр, вычисляемый по формуле

$$c = [n / 2 - \Psi((1 + \beta) / 2) n^{1/2} / 2],$$

где  $[.]$  – целая часть числа,

$\Psi(.)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

См. монографию Дезу (Desu) с соавт. О вычислении точечной и интервальной оценки медианы см. Холлендера с соавт.

### 17.2.4. Подбор распределения

Изучается проблема подбора теоретического распределения (подгонка, fitting distribution) для эмпирического распределения длительностей, в том числе цензурированных, следующими типами подходящих по теоретическим соображениям распределений, применяемыми для данной задачи:

- логнормальное распределение,
- логлогистическое распределение,





- гамма-распределение,
- распределение Вейбулла,
- экспоненциальное распределение,
- распределение Рэлея,
- распределение Гомпертца.

Функция распределения (с точностью до параметров) может быть известна из теоретических соображений. В таком случае задача вычисления параметров распределений может трактоваться как идентификация математической модели.

Теоретические значения частот [непрерывных] распределений элементарно вычисляются как произведение плотности теоретического распределения на численность выборки и на длину соответствующего классового интервала.

См. монографии Бьюри (Bury), Эванс (Evans) с соавт. Сводка параметров распределений и обзор алгоритмов подгонки теоретических распределений к экспериментальным данным представлена в монографии Кобзаря. См. также книги Кришнамурти (Krishnamoorthy), Кляйбер (Kleiber) с соавт., статьи Лу (Lu) с соавт., Айтчисон (Aitchison) с соавт.

#### 17.2.4.1. Общая методика

Аппроксимация эмпирического распределения опытных данных любым теоретическим стандартным распределением сводится к вычислению одним из математических методов (метод максимального правдоподобия, метод моментов, реже – метод наименьших квадратов) параметров теоретического распределения, ответственных за форму, масштаб и положение кривой распределения. Метод максимального правдоподобия является наиболее популярным методом решения рассматриваемого типа задач благодаря хорошей вычислительной устойчивости.

Реализация метода начинается с составления функции максимального правдоподобия (ФМП) в виде (возможны эквивалентные записи ФМП, с учетом введенных выше функций статистической модели, в зависимости от того, какие результаты интересуют автора исследования)

$$L(\theta) = \prod_{i=1}^n \left[ f(t_i) \right]^{\delta_i} \left[ S(t_i) \right]^{1-\delta_i},$$

где  $\theta$  – вектор неизвестных параметров статистической модели,

$t_i, i = 1, 2, \dots, n$  – массив длительностей,

$\delta_i, i = 1, 2, \dots, n$  – соответствующий массив индикаторов цензурирования (для конкретной длительности индикатор равен 1, если выборка нецензурирована, и 0, если цензурирована),  
 $n$  – численность массива длительностей.

Отметим, что в дальнейших выкладках число нецензурированных длительностей стандартно обозначено как

$$r = \sum_{i=1}^n \delta_i.$$





Вектор искомых параметров находится из условия максимума ФМП:

$$L(\theta) \rightarrow \max_{\theta}.$$

Максимум ФМП находится из условия равенства нулю частных производных ФМП по искомым параметрам, т. е. искомые параметры удовлетворяют уравнениям

$$\frac{\partial L(\theta)}{\partial \theta} = 0.$$

Для упрощения максимизируют не саму ФМП, а логарифм ФМП. Эта возможность основана на том факте, что ФМП и логарифм достигают максимума при одних и тех же значениях искомых параметров, однако работать с логарифмом ФМП значительно проще:

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0.$$

Задача сводится к аналитическому либо численному (одним из методов оптимизации) решению полученной линейной или нелинейной системы уравнений.

См. монографии Коллетт (Collett) и Лелесс (Lawless), статью Юзеф (Yousef), пособие Цыплакова.

#### 17.2.4.2. Логарифмические модели

Решение статистической модели может быть выполнено различными методами. Однако в любом случае стараются использовать наиболее эффективную модификацию общего метода, иногда позволяющую радикально упростить решение. Не конкретизируя тип модели, представим общий метод решения логарифмической двухпараметрической модели.

ФМП логарифмической двухпараметрической модели общего вида имеет вид

$$L(\theta) = \prod_{i=1}^n \left\{ \sigma^{-1} f\left(\frac{y_i - \mu}{\sigma}\right) \right\}^{\delta_i} \left\{ S\left(\frac{y_i - \mu}{\sigma}\right) \right\}^{1 - \delta_i},$$

где  $y_i = \ln t_i$ ,  $i = 1, 2, \dots, n$  – массив логарифмов длительностей,

$\theta = \{\mu, \sigma\}$  – вектор параметров,

$\mu$  – параметр положения,

$\sigma$  – параметр масштаба.

Логарифмическая ФМП может быть записана как

$$\ln L(\theta) = -r \ln \sigma + \sum_{i=1}^n [\delta_i \ln f(z_i) + (1 - \delta_i) \ln S(z_i)],$$

$$z_i = \frac{y_i - \mu}{\sigma}, i = 1, 2, \dots, n,$$

где – массив стандартизированных логарифмов длительностей.

В дальнейших выкладках понадобятся следующие очевидные выражения для производных

$$\frac{\partial z}{\partial \mu} = -\frac{1}{\sigma}, \quad \frac{\partial z}{\partial \sigma} = -\frac{z}{\sigma}.$$

Тогда компоненты вектора градиента  $G(\theta)$  логарифмической ФМП по параметрам





$$g_1 = \frac{\partial \ln L(\theta)}{\partial \mu} = -\frac{1}{\sigma} \sum_{i=1}^n \left[ \delta_i \frac{\partial \ln f(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \ln S(z_i)}{\partial z_i} \right],$$
$$g_2 = \frac{\partial \ln L(\theta)}{\partial \sigma} = -\frac{r}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^n \left[ \delta_i z_i \frac{\partial \ln f(z_i)}{\partial z_i} + (1 - \delta_i) z_i \frac{\partial \ln S(z_i)}{\partial z_i} \right].$$

Компоненты матрицы вторых производных  $H(\theta)$  логарифмической ФМП по параметрам (матрицы Гессе) вычисляются как

$$h_{11} = \frac{\partial^2 \ln L(\theta)}{\partial \mu^2} = \frac{1}{\sigma^2} \sum_{i=1}^n \left[ \delta_i \frac{\partial^2 \ln f(z_i)}{\partial z_i^2} + (1 - \delta_i) \frac{\partial^2 \ln S(z_i)}{\partial z_i^2} \right],$$
$$h_{12} = h_{21} = \frac{\partial^2 \ln L(\theta)}{\partial \mu \partial \sigma} = \frac{1}{\sigma^2} \sum_{i=1}^n \left[ \delta_i \frac{\partial \ln f(z_i)}{\partial z_i} + (1 - \delta_i) \frac{\partial \ln S(z_i)}{\partial z_i} \right] +$$
$$+ \frac{1}{\sigma^2} \sum_{i=1}^n \left[ \delta_i z_i \frac{\partial^2 \ln f(z_i)}{\partial z_i^2} + (1 - \delta_i) z_i \frac{\partial^2 \ln S(z_i)}{\partial z_i^2} \right],$$
$$h_{22} = \frac{\partial^2 \ln L(\theta)}{\partial \sigma^2} = \frac{r}{\sigma^2} + \frac{2}{\sigma^2} \sum_{i=1}^n \left[ \delta_i z_i \frac{\partial \ln f(z_i)}{\partial z_i} + (1 - \delta_i) z_i \frac{\partial \ln S(z_i)}{\partial z_i} \right] +$$
$$+ \frac{1}{\sigma^2} \sum_{i=1}^n \left[ \delta_i z_i^2 \frac{\partial^2 \ln f(z_i)}{\partial z_i^2} + (1 - \delta_i) z_i^2 \frac{\partial^2 \ln S(z_i)}{\partial z_i^2} \right].$$

С учетом введенных обозначений итерационная схема максимизации логарифмической ФМП алгоритма метода Ньютона–Рафсона может быть записана как

$$\theta^{j+1} = \theta^j - [H(\theta)]^{-1} G(\theta), j = 0, 1, 2, \dots,$$

где  $j, j = 0, 1, 2, \dots$  – номер итерации.

При численной реализации метода Ньютона–Рафсона должна быть учтена особенность данного метода при решении рассматриваемой задачи, заключающаяся в весьма узкой области сходимости. Поэтому начальные приближения параметров должны быть заданы достаточно близкими к оптимальному решению. В этом случае метод сходится очень быстро. Для грубой же локализации начальных приближений может применяться один из глобальных методов. В простейшем случае можно применить метод перебора с небольшим шагом по разумной области определения параметров либо, для упрощения численной реализации, один из вариантов метода спуска. Низкое быстродействие данных примитивных, но надежных методов компенсируется высоким быстродействием современных компьютеров.

#### 17.2.4.2.1. Логнормальное распределение

Плотность логнормального (логарифмически нормального) распределения с двумя параметрами имеет вид





$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln t - \mu)^2}{\sigma^2}\right], \sigma > 0.$$

Введем нормированную величину

$$z = \frac{\ln t - \mu}{\sigma}.$$

Тогда можно записать

$$f(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

где  $\varphi(\cdot)$  – функция плотности стандартного нормального распределения.

Соответствующая функция выживания

$$S(z) = 1 - \Phi(z),$$

где  $\Phi(\cdot)$  – функция стандартного нормального распределения.

С учетом нормировки функция максимального правдоподобия (ФМП) запишется как

$$L(\mu, \sigma) = \prod_{i=1}^n \left\{ \frac{1}{\sigma} \varphi(z_i) \right\}^{\delta_i} \{1 - \Phi(z_i)\}^{1-\delta_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\mu, \sigma) = -n \ln \sigma - \frac{1}{2} \sum_{i=1}^n \delta_i z_i^2 + \sum_{i=1}^n (1 - \delta_i) \ln[1 - \Phi(z_i)].$$

Производные, необходимые для итерационной схемы метода Ньютона–Рафсона, представленной в разделе «Общая методика для логарифмических моделей», запишутся как

$$\frac{\partial \ln f(z)}{\partial z} = -z, \quad \frac{\partial \ln S(z)}{\partial z} = -\frac{f(z)}{S(z)},$$

$$\frac{\partial^2 \ln f(z)}{\partial z^2} = -1, \quad \frac{\partial^2 \ln S(z)}{\partial z^2} = \frac{zf(z)}{S(z)} - \left[ \frac{f(z)}{S(z)} \right]^2.$$

## 17.2.4.2.2. Логлогистическое распределение

Плотность логлогистического (логарифмически логистического) распределения с двумя параметрами имеет вид

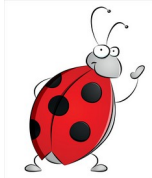
$$f(t) = \frac{1}{\sigma} \exp\left(\frac{\ln t - \mu}{\sigma}\right) \left[1 + \exp\left(\frac{\ln t - \mu}{\sigma}\right)\right]^{-2}, \sigma > 0.$$

Введем нормированную величину

$$z = \frac{\ln t - \mu}{\sigma}.$$

Тогда можно записать





$$f(z) = \frac{e^z}{(1 + e^z)^2}.$$

Соответствующая функция выживания

$$S(z) = \frac{1}{1 + e^z}.$$

С учетом нормировки функция максимального правдоподобия (ФМП) запишется как

$$L(\mu, \sigma) = \prod_{i=1}^n \left\{ \frac{1}{\sigma} \frac{e^z}{(1 + e^z)^2} \right\}^{\delta_i} \left\{ \frac{1}{1 + e^z} \right\}^{1 - \delta_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\mu, \sigma) = -r \ln \sigma + \sum_{i=1}^n \delta_i [z_i - 2 \ln(1 + e^{z_i})] - \sum_{i=1}^n (1 - \delta_i) \ln(1 + e^{z_i}).$$

Производные, необходимые для итерационной схемы метода Ньютона–Рафсона, представленной в разделе «Общая методика для логарифмических моделей», запишутся как

$$\frac{\partial \ln f(z)}{\partial z} = 1 - \frac{2e^z}{1 + e^z}, \quad \frac{\partial \ln S(z)}{\partial z} = -\frac{e^z}{1 + e^z},$$
$$\frac{\partial^2 \ln f(z)}{\partial z^2} = -\frac{2e^z}{(1 + e^z)^2}, \quad \frac{\partial^2 \ln S(z)}{\partial z^2} = -\frac{e^z}{(1 + e^z)^2}.$$

### 17.2.4.3. Гамма– распределение

Плотность гамма–распределения имеет вид

$$f(t) = \frac{1}{\alpha \Gamma(\kappa)} \left( \frac{t}{\alpha} \right)^{\kappa-1} e^{-t/\alpha}, t \geq 0, \alpha > 0, \kappa > 0,$$

Соответствующая функция выживания

$$S(t) = 1 - I(\kappa, t / \alpha),$$

где  $I(.,.)$  – неполная гамма–функция.

Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\alpha, \kappa) = \prod_{i=1}^n \left\{ \frac{1}{\alpha \Gamma(\kappa)} \left( \frac{t_i}{\alpha} \right)^{\kappa-1} e^{-t_i/\alpha} \right\}^{\delta_i} \left\{ 1 - I(\kappa, t_i / \alpha) \right\}^{1 - \delta_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\alpha, \kappa) = -r \kappa \ln \alpha - r \ln \Gamma(\kappa) + (\kappa - 1) \sum_{i=1}^n \delta_i \ln t_i - \frac{1}{\alpha} \sum_{i=1}^n \delta_i t_i + \sum_{i=1}^n (1 - \delta_i) \ln [1 - I(\kappa, t_i / \alpha)].$$

Аналитическое представление производных логарифмической ФМП выполнить сложно, поэтому задача решается численно одним из вариантов метода спуска, не использующим производных.



## 17.2.4.4. Распределение Вейбулла

Плотность распределения Вейбулла с двумя параметрами имеет вид

$$f(t) = \lambda \gamma t^{\gamma-1} \exp(-\lambda t^\gamma), \quad t \geq 0, \lambda > 0, \gamma > 0.$$

Соответствующая функция выживания

$$S(t) = \exp(-\lambda t^\gamma).$$

Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\lambda, \gamma) = \prod_{i=1}^n \left\{ \lambda \gamma t_i^{\gamma-1} \exp(-\lambda t_i^\gamma) \right\}^{\delta_i} \left\{ \exp(-\lambda t_i^\gamma) \right\}^{1-\delta_i}.$$

После преобразований окончательно получаем

$$L(\lambda, \gamma) = \prod_{i=1}^n \left\{ \lambda \gamma t_i^{\gamma-1} \right\}^{\delta_i} \exp(-\lambda t_i^\gamma).$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\lambda, \gamma) = \ln(\lambda \gamma) \sum_{i=1}^n \delta_i + (\gamma - 1) \sum_{i=1}^n \delta_i \ln t_i - \lambda \sum_{i=1}^n t_i^\gamma.$$

Логарифмическая ФМП примет окончательный вид

$$\ln L(\lambda, \gamma) = r \ln(\lambda \gamma) + (\gamma - 1) \sum_{i=1}^n \delta_i \ln t_i - \lambda \sum_{i=1}^n t_i^\gamma.$$

Для вычисления значений искомых параметров найдем частные производные логарифмической ФМП по искомым параметрам и приравняем их нулю. Сначала найдем производную по  $\lambda$ :

$$\frac{\partial \ln L(\lambda, \gamma)}{\partial \lambda} = \frac{r}{\lambda} - \sum_{i=1}^n t_i^\gamma = 0.$$

Отсюда уравнение для вычисления параметра  $\lambda$  получается как

$$\lambda = r \left[ \sum_{i=1}^n t_i^\gamma \right]^{-1}.$$

Вычислив производную по параметру  $\gamma$ ,

$$\frac{\partial \ln L(\lambda, \gamma)}{\partial \gamma} = \frac{r}{\gamma} + \sum_{i=1}^n \delta_i \ln t_i - \lambda \sum_{i=1}^n t_i^\gamma \ln t_i = 0,$$

с учетом выражения для параметра  $\lambda$ , получаем нелинейное уравнение для поиска параметра  $\gamma$  в виде:

$$\frac{r}{\gamma} + \sum_{i=1}^n \delta_i \ln t_i - r \left[ \sum_{i=1}^n t_i^\gamma \right]^{-1} \sum_{i=1}^n t_i^\gamma \ln t_i = 0.$$

Решение уравнения может быть произведено одним из методов оптимизации – в простейшем случае методом деления отрезка пополам.



## 17.2.4.5. Экспоненциальное распределение

Плотность экспоненциального распределения имеет вид

$$f(t) = \lambda e^{-\lambda t}, t \geq 0, \lambda > 0.$$

Соответствующая функция выживания

$$S(t) = e^{-\lambda t}.$$

Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\lambda) = \prod_{i=1}^n \left\{ \lambda e^{-\lambda t_i} \right\}^{\delta_i} \left\{ e^{-\lambda t_i} \right\}^{1-\delta_i}.$$

После преобразований окончательно получаем

$$L(\lambda) = \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda t_i}.$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\lambda) = r \ln \lambda - \lambda \sum_{i=1}^n t_i.$$

Для вычисления значения искомого параметра найдем производную логарифмической ФМП по данному параметру и приравняем ее нулю. Производная по параметру  $\lambda$  имеет вид:

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{r}{\lambda} - \sum_{i=1}^n t_i = 0.$$

Отсюда уравнение для вычисления параметра  $\lambda$  получается как

$$\lambda = r \left[ \sum_{i=1}^n t_i \right]^{-1}.$$

## 17.2.4.6. Распределение Рэлея

Плотность распределения Рэлея имеет вид

$$f(t) = \frac{t}{\beta^2} \exp\left(-\frac{t^2}{2\beta^2}\right), t \geq 0, \beta > 0.$$

Соответствующая функция выживания

$$S(t) = \exp\left(-\frac{t^2}{2\beta^2}\right).$$

Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{t_i}{\beta^2} \exp\left(-\frac{t_i^2}{2\beta^2}\right) \right\}^{\delta_i} \left\{ \exp\left(-\frac{t_i^2}{2\beta^2}\right) \right\}^{1-\delta_i}.$$

После преобразований окончательно получаем







$$L(\beta) = \prod_{i=1}^n \left( \frac{t_i}{\beta^2} \right)^{\delta_i} \exp \left( - \frac{t_i^2}{2\beta^2} \right).$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\beta) = \sum_{i=1}^n \delta_i \ln \frac{t_i}{\beta^2} - \frac{1}{2} \beta^{-2} \sum_{i=1}^n t_i^2.$$

Для вычисления значения искомого параметра найдем производную логарифмической ФМП по данному параметру и приравняем ее нулю. Производная по параметру  $\beta$  имеет вид:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = -2\beta^{-1} \sum_{i=1}^n \delta_i + \beta^{-3} \sum_{i=1}^n t_i^2 = 0.$$

Уравнение для вычисления параметра  $\beta$  получается как

$$\beta = \left[ \frac{1}{2r} \sum_{i=1}^n t_i^2 \right]^{-2}.$$

#### 17.2.4.7. Распределение Гомпертца

Плотность распределения Гомпертца имеет вид

$$f(t) = \beta e^{\alpha t} \exp \left[ \frac{\beta}{\alpha} (1 - e^{\alpha t}) \right], t \geq 0, \beta > 0, \alpha \in ]-\infty; 0[ \cup ]0; \infty[.$$

Соответствующая функция выживания

$$S(t) = \exp \left[ \frac{\beta}{\alpha} (1 - e^{\alpha t}) \right].$$

Поэтому функция максимального правдоподобия (ФМП) запишется как

$$L(\alpha, \beta) = \prod_{i=1}^n \left\{ \beta e^{\alpha t_i} \exp \left[ \frac{\beta}{\alpha} (1 - e^{\alpha t_i}) \right] \right\}^{\delta_i} \left\{ \exp \left[ \frac{\beta}{\alpha} (1 - e^{\alpha t_i}) \right] \right\}^{1 - \delta_i}.$$

После преобразований окончательно получаем

$$L(\alpha, \beta) = \prod_{i=1}^n \left\{ \beta e^{\alpha t_i} \right\}^{\delta_i} \exp \left[ \frac{\beta}{\alpha} (1 - e^{\alpha t_i}) \right].$$

Соответствующая логарифмическая ФМП имеет вид

$$\ln L(\alpha, \beta) = r \ln \beta + \alpha \sum_{i=1}^n \delta_i t_i + \frac{\beta}{\alpha} \left( n - \sum_{i=1}^n e^{\alpha t_i} \right).$$

Для вычисления значений искоемых параметров найдем частные производные логарифмической ФМП по искомым параметрам и приравняем их нулю. Сначала найдем производную по  $\beta$ :

$$\frac{\partial \ln L(\alpha, \beta)}{\partial \beta} = \frac{r}{\beta} + \frac{1}{\alpha} \left( n - \sum_{i=1}^n e^{\alpha t_i} \right) = 0.$$

Отсюда уравнение для вычисления параметра  $\beta$  получается как





$$\beta = \alpha r \left[ \sum_{i=1}^n e^{\alpha t_i} - n \right]^{-1}.$$

Вычислив производную по параметру  $\alpha$ ,

$$\frac{\partial \ln L(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^n \delta_i t_i - \frac{\beta}{\alpha^2} \left[ n + (1 - \alpha^2) \sum_{i=1}^n e^{\alpha t_i} \right] = 0,$$

с учетом выражения для параметра  $\beta$ , получаем нелинейное уравнение для поиска параметра  $\alpha$  в виде:

$$\alpha \sum_{i=1}^n \delta_i t_i - r \left[ \sum_{i=1}^n e^{\alpha t_i} - n \right]^{-1} \left[ n + (1 - \alpha^2) \sum_{i=1}^n e^{\alpha t_i} \right] = 0.$$

Решение уравнения может быть произведено одним из методов оптимизации – в простейшем случае методом деления отрезка пополам.

#### 17.2.4.8. Оценка качества подгонки модели

Адекватной является модель с  $P$ -значением, большим 0,05. В этом случае теоретическое и эмпирическое распределения значимо не различаются.

Качество статистической модели можно оценить (также сравнить между собой различные модели), используя информационный критерий Акаике (Akaike's information criterion, AIC)

$$AIC = -2 \ln L(\hat{\theta}) + 2k + \frac{2k(k+1)}{(n-k-1)},$$

где  $\ln L(\hat{\theta})$  – оценка логарифма функции максимального правдоподобия,

$\hat{\theta}$  – вектор оценок параметров статистической модели,

$k$  – число параметров модели.

Последний член в уравнении для AIC призван скорректировать значение статистики критерия для малых выборок и некоторыми авторами не используется.

Оценка логарифма функции максимального правдоподобия в рассматриваемом случае имеет теоретический вид

$$\ln L(\hat{\theta}) = \sum_{i=1}^n \delta_i \ln[f(t_i, \hat{\theta})] + \sum_{i=1}^n (1 - \delta_i) \ln[S(t_i, \hat{\theta})],$$

где  $t_i, i = 1, 2, \dots, n$  – эмпирический массив длительностей,

$\delta_i, i = 1, 2, \dots, n$  – соответствующий массив индикаторов цензурирования,

$n$  – численность массива длительностей,

$f(.,.)$  – оценка функции плотности теоретического распределения,

$S(.,.)$  – оценка соответствующей функции выживания.

При расчете здесь нет необходимости в явном выписывании упомянутых функций, т. к. формулы для логарифмов функций максимального правдоподобия всех изучаемых теоретических распределений известны из предыдущих выкладок (см. выше).



При сравнении нескольких статистических моделей лучшей считается модель с наименьшим значением AIC.

В литературе представлены и другие информационные критерии, имеющие интерпретацию, аналогичную AIC.

См. монографию и статью Бернхэм (Burnham) с соавт., статьи Акаике (Akaike), Аль-Фозан (Al-Fawzan), Анжиллетта (Angilletta), Боздоган (Bozdogan), Бидюк с соавт.

### 17.2.5. Критерий Кокса

Критерий Кокса (логарифмический ранговый критерий, обобщенный критерий Сэвиджа) является обобщением критерия Сэвиджа (см. главу «Непараметрическая статистика») на цензурированные выборки и вычисляется по формуле

$$S = \sum_{i=1}^N \left[ I(A_i) \sum_{j=1}^i \frac{1}{N+1-j} \right],$$

где  $N = n_1 + n_2$  – численность объединенной выборки,

$n_1$  – численность первой сравниваемой выборки (с наибольшей численностью),

$n_2$  – численность второй сравниваемой выборки (с наименьшей численностью),

$I(A_i)$ ,  $i = 1, 2, \dots, N$  – индикатор, что  $i$ -й член вариационного ряда, построенного по объединенной выборке, является нецензурированным (наработкой до отказа) и принадлежит первой из сравниваемых выборок; в этом случае значение индикатора равно 1, в противном случае – нулю.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{S - ES}{\sqrt{DS}},$$

где  $ES = r_1$  – математическое ожидание,

$$DS = \frac{r_1 r_2}{r_1 + r_2 - 1} \left( 1 - \frac{1}{N} \sum_{j=1}^N \frac{1}{j} \right) - \text{дисперсия,}$$

$r_1$  и  $r_2$  – количества нецензурированных элементов первой и второй сравниваемых выборок, соответственно,

распределена по стандартному нормальному закону.

См. монографию Скрипника с соавт.

### 17.2.6. Критерий Гехана

Критерий Гехана (обобщенный критерий Вилкоксона) является обобщением критерия Вилкоксона (см. главу «Непараметрическая статистика») на цензурированные выборки. В источниках могут быть даны различные эквивалентные формулы и схемы (часто оптимизированные для «ручного» счета) вычисления критерия. Статистика критерия вычисляется как



$$W = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} u_{ij},$$

где  $n_1$  – численность первой сравниваемой выборки,

$n_2$  – численность второй сравниваемой выборки.

Величины под знаками суммы вычисляются как

$$u_{ij} = \begin{cases} 1, x_i > y_j, \\ 1, x_i^* \geq y_j, \\ -1, x_i < y_j, \\ -1, x_i \leq y_j^*, \\ 0, \text{если } \_ \text{ иначе,} \end{cases} \quad i=1,2,\dots,n_1; j=1,2,\dots,n_2,$$

где  $x_i, i=1,2,\dots,n_1$  – элементы первой сравниваемой выборки,

$y_j, j=1,2,\dots,n_2$  – элементы второй сравниваемой выборки,

\* – знак, означающий, что элемент выборки цензурирован.

Практически значимость может вычисляться посредством нормальной аппроксимации критического значения критерия. При этом модифицированная статистика

$$\frac{|W|}{\sqrt{DW}},$$

где  $DW$  – дисперсия, вычисляемая по формуле

$$DW = \frac{n_1 n_2 \sum_{i=1}^N \sum_{j=1}^N U_{ij}^2}{N(N-1)},$$

$$U_{ij} = \begin{cases} 1, X_i > X_j, \\ 1, X_i^* \geq X_j, \\ -1, X_i < X_j, \\ -1, X_i \leq X_j^*, \\ 0, \text{если } \_ \text{ иначе,} \end{cases} \quad i, j=1,2,\dots,N,$$

$X_i, i=1,2,\dots,N$  – элементы объединенной выборки,

$N = n_1 + n_2$  – численность объединенной выборки,

распределена по стандартному нормальному закону.

См. монографию Ли (Lee) с соавт.

## 17.2.7. Модель пропорциональных рисков Кокса

Полупараметрическая (semi-parametric) модель пропорциональных рисков (пропорциональных интенсивностей) Кокса может быть записана в виде





$$\frac{h(t)}{h_0(t)} = \exp(\beta^T x),$$

где  $h(t)$  – функция риска,

$h_0(t)$  – базовая функция риска (функция риска при нулевых ковариатах),

$\beta$  – вектор (длиной  $m$ ) коэффициентов модели (коэффициентов при ковариатах),

$x$  – вектор (длиной  $m$ ) ковариат (предикторов независимых переменных) модели.

С помощью модели Кокса исследуется отношение  $h(t) / h_0(t)$ , поэтому базовая функция риска в модели Кокса не оценивается.

Для обучения модели пропорциональных рисков, помимо длительностей ( $n$ -мерный вектор  $t$ ) и индикаторов цензурирования ( $n$ -мерный вектор  $\delta$ ), должна быть представлена  $n \times m$  матрица  $X$  ковариат, представляющая собой  $n \times m$ -мерных векторов  $x_i$ ,  $i = 1, 2, \dots, n$ , ковариат для каждого индивидуума обучающей выборки, где  $n$  – численность выборки (количество пациентов). Целью обучения является определение оптимальных значений компонент  $m$ -мерного вектора коэффициентов модели  $\beta$  при ковариатах.

Предложенный Коксом метод частичного правдоподобия (partial likelihood) позволяет оценить значения компонент вектора коэффициентов модели, доставляющие максимум так называемой частичной функции максимального правдоподобия (ФМП). Частичная ФМП имеет вид

$$PL(\beta) = \prod_{i=1}^n \left[ \exp(\beta^T x_i) / \sum_{j \in R_i} \exp(\beta^T x_j) \right]^{\delta_i},$$

где  $R_i = \{j: t_j \geq t_i\}$  – множество таких длительностей  $j$ , для которых  $t_j \geq t_i$ .

Соответствующая логарифмическая ФМП имеет вид

$$\ln PL(\beta) = \sum_{i=1}^n \delta_i \left[ \beta^T x_i - \ln \left( \sum_{j=1}^n Y_j(t_i) \exp(\beta^T x_j) \right) \right],$$

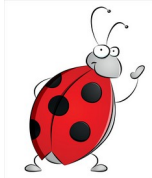
где  $Y_j(t_i)$ ,  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, n$  – индикатор, равный 1 для  $t_j \geq t_i$  (выборка упорядочена) или равный 0 в иных случаях.

Для дальнейших расчетов необходимы аналитические представления вектора первых производных (градиент) и матрицы вторых производных (матрица Гессе) логарифмической ФМП по искомым коэффициентам модели.

Градиент, т. е.  $m$ -мерный вектор первых производных логарифмической ФМП, вычисляется по формуле

$$G(\beta) = \frac{\partial \ln PL(\beta)}{\partial \beta} = \sum_{i=1}^n \delta_i [x_i - \bar{x}(t_i, \beta)],$$

где  $\bar{x}(\dots)$  – вектор взвешенных средних значений для длительности  $i$ ,  $i = 1, 2, \dots, n$ , вычисляется как



$$\bar{x}(t_i, \beta) = \frac{\sum_{j=1}^n Y_j(t_i) \exp(\beta^T x_j) x_j}{\sum_{j=1}^n Y_j(t_i) \exp(\beta^T x_j)}, i=1,2,\dots,n.$$

Фактически совокупность векторов взвешенных средних представляет собой  $n \times m$  матрицу  $\bar{X}$ , по структуре аналогичную заданной матрице ковариат.

Матрица Гессе, т. е.  $m \times m$  матрица вторых производных логарифмической ФМП, вычисляется по формуле

$$H(\beta) = \frac{\partial^2 \ln PL(\beta)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^n \delta_i \left\{ \frac{\sum_{j=1}^n Y_j(t_i) \exp(\beta^T x_j) [x_j - \bar{x}(t_i, \beta)] [x_j - \bar{x}(t_i, \beta)]^T}{\sum_{j=1}^n Y_j(t_i) \exp(\beta^T x_j)} \right\}.$$

Модель может быть решена с помощью метода Ньютона–Рафсона. Итерационная схема метода имеет вид

$$\beta^{l+1} = \beta^l - [H(\beta)]^{-1} G(\beta), l = 0, 1, 2, \dots,$$

где  $l, l = 0, 1, 2, \dots$  – номер итерации.

Начальные приближения для итерационного процесса можно взять нулевыми.

Дисперсии  $D\hat{\beta}_j, j = 1, 2, \dots, m$ , оценок коэффициентов при ковариатах, вычисленных, как показано выше, представляют собой соответствующие диагональные элементы матрицы Гессе (на последней итерации), взятые с обратным знаком. Доверительные интервалы данных оцениваемых коэффициентов считаются стандартно как

$$\beta_j = \left( \hat{\beta}_j - \Psi((1 + \beta)/2) \sqrt{D\hat{\beta}_j}; \hat{\beta}_j + \Psi((1 + \beta)/2) \sqrt{D\hat{\beta}_j} \right), j = 1, 2, \dots, m,$$

где  $\hat{\beta}_j, j = 1, 2, \dots, m$ , – оценки коэффициентов при ковариатах,

$\Psi(\cdot)$  – обратная функция стандартного нормального распределения,

$\beta$  – доверительный уровень, выраженный в долях.

Построенную модель пропорциональных рисков применяют также для вычисления регрессии Кокса, которая записывается в виде  $h(t) = h_0(t) \exp(\beta^T x)$ .

При необходимости явного построения функции риска с помощью данной регрессии для конкретного индивидуума может быть использована любая подходящая параметрическая модель из числа представленных в разделе «Подбор распределения». Получающаяся в результате регрессионная модель часто называется не регрессией Кокса, а по имени регрессии базовой функции риска (например, Вейбулла или Гомпертца).

Теоретическое обоснование см. в монографиях Кокса с соавт., Кляйн (Klein) с соавт., Дюпон (Dupont), Фортхофер (Forthofer) с соавт., Кемпбелл (Campbell). Техника вычислений представлена О'Квигли (O'Quigley), Лелесс (Lawless). См. также статьи Фан (Fan) с соавт.,



Гош (Ghosh D.), работы Биндер (Binder), Кларксон (Clarkson) с соавт., соответствующую статью энциклопедии под ред. Армитейдж (Armitage) с соавт. Примеры отбора ковариат и валидации модели см. в статьях Ли (Lee M.S.) с соавт., Ле (Le).

В модели пропорциональных рисков индикаторы цензурирования могут принимать только значения 0 (пациент цензурирован, т. е. выбыл из исследования, и его состояние неизвестно, или умер по причине, не связанной с исследуемой патологией) или 1 (пациент умер по причине, связанной с исследуемой патологией). Однако возможен такой случай, что причин смерти, связанной с исследуемой патологией, может быть выявлено более 1. Для такого случая разработана модель конкурирующих рисков (competing risks), рассмотренная в монографиях Краудера (Crowder), Хедекера (Hedeker) с соавт., Диггеля (Diggle) с соавт., Лачина (Lachin), Кальбфляйша (Kalbfleisch) с соавт., Пинтили (Pintilie), Штейерберга (Steyerberg), Фитцмауриса (Fitzmaurice) с соавт., Твиска (Twisk), Брауна (Brown), Чيانг (Chiang), Фиокко (Fiocco) с соавт.

## Глава 18. Анализ временных рядов и прогнозирование

### 18.1. Введение

Рассмотрены классические методы анализа временных рядов и прогнозирования.

### 18.2. Теоретическое обоснование

Анализ временных рядов оперирует зависимостью случайной величины  $y_i$ ,  $i = 1, 2, \dots, n$ , от контролируемой переменной  $t$ , в качестве которой может выступать время. В ряде моделей предполагается, что случайная величина состоит из истинного значения (тренда) с оценкой  $\eta_i$ ,  $i = 1, 2, \dots, n$ , и нормально распределенной случайной составляющей (ошибки измерений) с нулевым средним значением. Далее, предполагается, что случайные величины  $y_i$ ,  $i = 1, 2, \dots, n$ , наблюдаются через равные промежутки времени, а именно  $t_i - t_{i-1} = \text{const}$ ,  $i = 1, 2, \dots, n$ . Данное предположение значительно упрощает все выкладки, позволяя также избавиться от ввода временных отметок. Их роль играет номер отсчета. Введенные выше предположения являются общепринятыми.

Авторами выделяются основные задачи анализа временных рядов:

- исследование структуры временного ряда, в том числе описательные характеристики, выделение периодичностей, спектральный анализ,
- выделение сигнала на фоне шума,
- фильтрация и сглаживание.

Номенклатура задач не исчерпывается приведенным списком. Постоянно возникают новые прикладные задачи. Поставленные задачи решаются различными методами анализа, в том числе совокупностью представленных методов:





- метод скользящего среднего,
- сезонный разностный оператор,
- сингулярный спектральный анализ,
- гармонический анализ Фурье,
- автокорреляционная функция,
- периодограмма.

## 18.2.1. Метод скользящего среднего

Метод скользящего среднего (moving average) основан на следующих соображениях. С учетом сделанных ранее предположений, определим оценку тренда в виде полинома

$$\eta_j = \sum_{i=0}^l x_{i+1} j^i, j=1,2,\dots,n,$$

где  $n$  – численность временного ряда,

$x_i, i=1,2,\dots,l+1$ , – коэффициенты полинома, вычисленные в точке  $j$ ,

$l$  – степень полинома.

В некоторых источниках оценка тренда называется прогнозом (прогнозируемыми значениями), однако данное наименование конфликтует с понятием прогноза, под которым подразумевается продолжение линии тренда за пределы исходного временного ряда.

Обозначим вектор коэффициентов полинома как  $\bar{x}$ . Методом наименьших квадратов (МНК) найдено, что

$$\bar{x} = (A^T A)^{-1} A^T y,$$

где  $A$  – матрица размером  $(2k+1)(l+1)$ , элементы которой вычисляются по формуле

$$a_{ij} = (i - k - 1)^{j-1}, i=1,2,\dots,2k+1; j=1,2,\dots,l+1,$$

$k$  – полуширина окна (усредняющего интервала), особенность которого для данного метода показана ниже.

Коэффициенты полинома вычисляются на основе исходного временного ряда  $y_i, i=1,2,\dots,n$ , причем для вычислений использован интервал данного ряда с центром в точке  $j$ , имеющий протяженность на  $k$  значений  $y_i, i=1,2,\dots,n$ , влево и вправо от центра интервала.

Единственное требование к выбору полуширины окна определяется, согласно МНК, тем, что число точек усредняющего интервала должно быть  $l < 2k + 1$ .

После всех вспомогательных вычислений в качестве значения оценки тренда в точке  $j$  берется значение  $x_1$ . Остальные компоненты вектора коэффициентов полинома применяются при расчете крайних точек тренда и соответствующих средних квадратичных отклонений, как будет показано ниже.

Алгоритм в представленной выше форме не позволяет получить оценки тренда первых  $k$  и последних  $k$  точек временного ряда. Для вычислений крайних значений используются вектора коэффициентов полинома, вычисленные, соответственно, для точек с номерами  $k+1$  и  $n-k$  по формулам:





$$\eta_j = \sum_{i=0}^l x_{i+1}^{(k+1)} (j - k - 1)^i, j \leq k,$$

$$\eta_j = \sum_{i=0}^l x_{i+1}^{(n-k)} (j + k - n)^i, j > n - k.$$

Значения  $j$  в показанных формулах могут быть продолжены как влево (для первой формулы), так и вправо (для второй формулы), обеспечивая потребности задачи прогнозирования. Для вычисленного тренда определяются доверительные интервалы по формуле

$$\eta_j^{\pm} = \eta_j \pm c_{11} s_j t_{\beta}, j = 1, 2, \dots, n,$$

где  $c_{11}$  – элемент с индексами (1; 1) матрицы преобразования

$$C = (A^T A)^{-1},$$

$s_j, j = 1, 2, \dots, n$ , – оценка среднего квадратичного отклонения, определяемая как

$$s_j = \sqrt{\frac{1}{2k - l} \sum_{i=k}^j (y_i - \eta_i)^2}, j = 1, 2, \dots, n,$$

причем индекс суммирования в формуле является относительным,

$t_{\beta}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $2k - l$  и  $(1 + \beta) / 2$ ,  $\beta$  – доверительный уровень, выраженный в долях.

Доверительный уровень выбирается из стандартной линейки и по умолчанию равен 0,95.

Данный параметр может быть изменен пользователем.

Для первых  $k$  и последних  $k$  точек временного ряда оценка среднего квадратичного отклонения вычисляется на основе того же принципа использования полученных коэффициентов полинома, соответственно, для точек с номерами  $k + 1$  и  $n - k$ .

См. книги Брандта, Тюрина с соавт.

## 18.2.2. Сезонный разностный оператор

Сезонные разностные операторы предназначены для удаления сезонных компонент.

Процедура основана на формуле

$$y_i = x_i - x_{i-p}, i = p + 1, \dots, N,$$

где  $y_i, i = p + 1, \dots, N$ , – элементы преобразованного временного ряда,

$x_j, j = 1, 2, \dots, N$ , – элементы исходного временного ряда,

$N$  – численность ряда,

$p$  – период сезонности.

Процедура уменьшает численность ряда на величину  $p$ .

Описание см. в книге Тюрина с соавт.





### 18.2.3. Сингулярный спектральный анализ

Сингулярный спектральный анализ («Гусеница», singular spectrum analysis) предназначен для разделения исходного временного ряда на трендовые, сезонные и иные составляющие. Метод включает в себя ряд этапов: вложение, разложение по сингулярным числам, восстановление. Рассмотрим данные этапы подробно.

#### 18.2.3.1. Вложение

Рассмотрим временной ряд  $X$ , состоящий из элементов  $x_i, i = 1, 2, \dots, N$ . Выберем некоторое целое число  $L, 1 < L < N$ , которое назовем шириной окна. Затем будем двигать окно вдоль временного ряда. В результате применения данной процедуры вложения получится так называемая траекторная матрица  $A$  размером  $K \times L$ , где  $K = N - L + 1$ .

#### 18.2.3.2. Разложение по сингулярным числам

Каноническая формула разложения действительной прямоугольной матрицы  $A$  размером  $m \times n$  ( $m$  строк,  $n$  столбцов,  $m \geq n$ ) по сингулярным числам имеет вид

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

где  $U$  – матрица размером  $m \times m$ , сформированная из  $m$  ортонормированных собственных векторов, соответствующих собственным значениям матрицы  $AA^T$ ,  $U^T U = I_m$ ,

$\Sigma$  – диагональная матрица размером  $n \times n$ , диагональные элементы которой представляют собой так называемые сингулярные числа – квадратные корни из  $\lambda_i, i = 1, 2, \dots, n$  – неотрицательных собственных значений матрицы  $A^T A$ ,

$0$  – прямоугольная нулевая матрица размером  $(m - n)n$ ,

$V$  – матрица размером  $n \times n$ , состоящая из  $n$  ортонормированных собственных векторов матрицы  $A^T A$ ,  $V^T V = V V^T = I_n$ ,

$I$  – единичная матрица соответствующего порядка.

В другой записи разложение по сингулярным числам имеет более простой вид

$$A = U_n \Sigma V^T,$$

где  $U_n$  – матрица размером  $m \times n$ , сформированная из  $n$  ортонормированных собственных векторов, соответствующих  $n$  наибольшим из  $m$  собственным значениям матрицы  $AA^T$ ,

$$U_n^T U_n = I_n.$$

Раскрывая последнюю формулу, можно естественно получить, с учетом равенства нулю внедиагональных элементов матрицы  $\Sigma$ , что

$$A = \sum_{i=1}^m A_i,$$

где  $A_i, i = 1, 2, \dots, m$  – «элементарные» матрицы размером  $m \times n$ , элементы которых определяются согласно формуле





$$a_{ij} = \sqrt{\lambda_j} U_i V_j^T, i=1,2,\dots,m, j=1,2,\dots,n,$$

где  $U_i, i=1,2,\dots,m$  – столбец матрицы  $U_n$ ,  
 $V_j, j=1,2,\dots,n$  – столбец матрицы  $V$ .

См. статьи Голуба (Golub) с соавт., Стюарта (Stewart), книги Деммеля, Голуба с соавт.

### 18.2.3.3. Восстановление

Существует взаимно однозначное соответствие между матрицами  $A_i, i=1,2,\dots,m$ , размером  $K \times L$ , полученными на предыдущем этапе, и «элементарными» временными рядами каждый  $X_i, i=1,2,\dots,m$ , длиной  $N$ . Здесь величину  $m, m \leq L$ , можно интерпретировать как число выделяемых компонент временного ряда (гармоник). Если выбранное пользователем число гармоник превышает указанный предел, оно уменьшается до величины этого предела. Восстановление «элементарных» временных рядов производится методом диагонального усреднения матриц  $A_i, i=1,2,\dots,m$ , суть которого заключается в том, что каждый элемент ряда  $X_i, i=1,2,\dots,m$ , будет получен как среднее арифметическое величин, стоящих на «антидиагоналях» соответствующей матрицы  $A_i, i=1,2,\dots,m$ .

В отличие от гармонического анализа Фурье, описываемый метод не позволяет получить разложение исходного временного ряда на «чистые» гармоники. Пользователь может в этом убедиться, проведя сравнительные расчеты разными методами. Здесь полезно привести аналогию со ступенчатым регрессионным анализом, представленном во 2 книге монографии Дрейпера с соавт. и процитировать данное там положение: «Этот метод не дает правильного МНК-решения для переменных, включенных в итоговое уравнение».

См. работы Вотарда (Vautard) с соавт., Голяндиной с соавт., Александрова с соавт.

### 18.2.4. Гармонический анализ Фурье

Рассматриваемый метод называют гармоническим анализом Фурье (гармоническим регрессионным анализом). Конечный ряд Фурье представляет периодическую функцию  $y(t)$  в виде линейной комбинации  $r$  гармоник (гармонических векторов).

Пусть временной ряд задан в виде  $N$  отсчетов временного ряда  $y_n, n=1,2,\dots,N$ , в равноотстоящих точках  $t_n, n=1,2,\dots,N$ . Исходный временной ряд может быть представлен в виде конечного ряда Фурье, определяемого формулой

$$y_n = a_0 + \sum_{k=1}^r a_k \cos \frac{2\pi kn}{N} + \sum_{k=1}^r b_k \sin \frac{2\pi kn}{N}, n=1,2,\dots,N,$$

где коэффициенты вычисляются по формулам

$$a_0 = \frac{1}{N} \sum_{k=1}^N y_k,$$

$$b_0 = 0,$$

$$a_m = \frac{2}{N} \sum_{k=1}^N y_k \cos \frac{2\pi km}{N}, m=0,1,\dots,r,$$





$$b_m = \frac{2}{N} \sum_{k=1}^N y_k \sin \frac{2\pi km}{N}, m = 0, 1, \dots, r,$$

где  $y_k$ ,  $k = 1, 2, \dots, N$  – отсчеты временного ряда в точках  $t_k$ ,  $k = 1, 2, \dots, N$ .

$m$  – номер гармоники,

$N$  – количество наблюдений – число равных частей, на которые разделен период наблюдения,

$r$  – количество гармоник,  $r \leq N / 2$ .

Количество гармоник выбирается. Если выбранное число гармоник превышает указанный предел, оно уменьшается до величины этого предела.

Дополнительно для каждой гармоники:

$$A_m = \sqrt{a_m^2 + b_m^2}, m = 0, 1, \dots, r, \text{ – амплитуда,}$$

$$\theta_m = \arctg(-b_m / a_m) \cdot 180 / \pi, m = 0, 1, \dots, r \text{ – фаза.}$$

Рассмотренный метод описан в большом числе классических и современных источников.

См., например, главу 18 справочника по ред. Ллойда с соавт., с. 365 монографии

«Прикладной анализ случайных данных» Бендата с соавт., с. 85 книги Носача.

## 18.2.5. Автокорреляционная функция

Выборочная автокорреляционная функция (сериальная корреляция) строится по формуле

$$r_k = \frac{\sum_{i=1}^{N-1} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2}, k = 1, 2, \dots, N - 1,$$

где  $x_i$ ,  $i = 1, 2, \dots, N$  – элементы временного ряда,

$N$  – численность ряда,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \text{ – оценка среднего значения.}$$

График автокорреляционной функции – коррелограмма. На графике показывают также 95% доверительный интервал. Границы данного интервала называют корреляционной трубкой и вычисляют по формуле

$$r^{\pm} = \pm \frac{1}{n} \pm \frac{2}{\sqrt{n}}.$$

Подробное описание см. в книге Тюрина с соавт.

## 18.2.6. Периодограмма

Периодограмма временного ряда  $y_n$ ,  $n = 1, 2, \dots, N$ , состоит из  $r = N / 2$  значений, называемых интенсивностями и вычисляемых по формуле

$$I(f_i) = \frac{N}{2} (a_i^2 + b_i^2), i = 1, 2, \dots, r,$$

где  $a_i$ ,  $b_i$ ,  $i = 1, 2, \dots, r$  – коэффициенты ряда Фурье,





$f_i = i / N, i = 1, 2, \dots, r$  –  $i$ -я гармоника основной частоты.

См. монографию «Прикладной анализ случайных данных» Бендата с соавт., а также с. 52 первого выпуска книги Бокса и Дженкинса.

## Глава 19. Статистический контроль качества

### 19.1. Введение

Статистические методы контроля качества в лаборатории и на производстве предназначены для лабораторного контроля, для контроля качества выпускаемой продукции или оказываемых услуг с целью своевременного выявления нарушений в организации производства и в технологических процессах, приводящих к снижению качества продукции или услуг ниже норм, заданных техническими условиями. Контроль качества интересен как инструмент успешного эффективного решения задач, возникающих на этапе внедрения в производство передовых методов управления.

### 19.2. Теоретическое обоснование

Современное управление качеством основано на широком использовании статистических методов. Статистический контроль качества, по определению У. Деминга – это применение статистических принципов и приемов на всех стадиях производства, направленное на экономичное производство изделия, максимально полезного и имеющего сбыт. Статистическое управление качеством, по определению Дж. Мердока – это совокупность методов обнаружения неслучайных факторов, позволяющих диагностировать состояние процесса, провести его корректировку и, в конечном счете, способствующих улучшению качества продукции.

Рассмотренные методы статистического контроля качества:

- гистограмма качества,
- диаграмма Парето,
- контрольная карта –

принадлежат к совокупности семи элементарных методов контроля качества, введенной в классических источниках по контролю качества. К другим, не рассмотренным здесь статистическим методам контроля качества, относятся точечный график и диаграмма разброса. Последние два классических метода: диаграмма Исикавы (диаграмма «причины – результат») и таблица контроля – не относятся к статистическим методам.

Анализ Бланда–Альтмана предназначен для сравнения двух методов клинического или лабораторного контроля.

Данные методы просты, наглядны, удобны. Применение совокупности этих методов, по утверждению оригинальных источников, решает 95% всех производственных проблем. Не стоит, однако, понимать статистическое управление качеством в том смысле, что его



применение даст немедленный практический эффект. Статистическое управление качеством не предназначено для решения в принципе неразрешимых проблем.

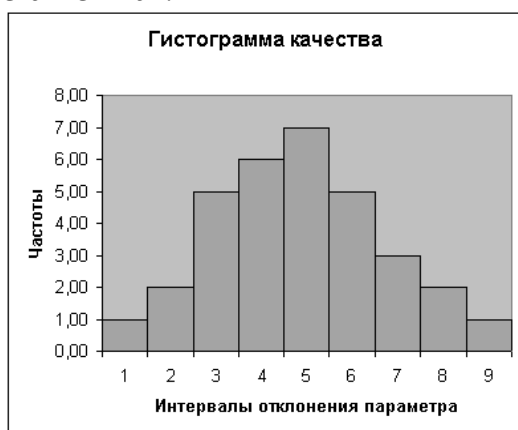
Отметим, что в статистическом управлении качеством используются разнообразные методы статистического анализа данных: описательная статистика; другие методы предварительной обработки данных; корреляционный анализ.

При анализе представленных теоретических материалов у пользователей, не являющихся специалистами в статистическом контроле качества и только начинающих знакомиться с теорией и практикой статистического контроля качества, может возникнуть впечатление, что в силу большого объема предлагаемых материалов их изучение может быть особенно трудным. На самом деле решение возникающих задач следует начать не с изучения всевозможных материалов, а с постановки задачи, вначале словесной («нужно улучшить качество»). Затем следует перевести постановку на некий промежуточный язык («от каких параметров зависит качество») и только потом переходить к математической формулировке задачи статистического контроля качества.

См. Шор, Коуден, Уилер с соавт., Миттаг с соав., Панде с соавт., Хэнсен, Шторм.

## 19.2.1. Гистограмма качества

Гистограмма качества позволяет в наглядной форме отобразить выявленный характер разброса значений контролируемого параметра. Дополнительно на гистограмме качества могут отображаться среднее значение, стандартное отклонение и заданные границы допуска. Исходными данными для построения гистограммы качества служат количества вариантов (частоты), относящиеся к каждому интервалу отклонения исследуемого параметра. На графике частоты откладываются по оси ординат. По оси абсцисс откладываются кодовые обозначения интервалов отклонения параметра (классов). Гистограмма может быть построена с помощью одноименного метода, представленного в главе «Описательная статистика».



Умение читать и анализировать гистограммы окажет неоценимую услугу специалисту не только в статистическом управлении качеством, но также и в других разделах статистического анализа данных.



Встречаются следующие основные типы гистограмм:

- Обычный тип. Гистограмма имеет симметричную колоколообразную форму. Среднее значение приходится примерно на середину размаха данных. Этот тип свидетельствует об однородности исходных данных, а в статистическом контроле качества – о нормальном протекании технологического процесса.
- Положительно (отрицательно) скошенное распределение. Форма асимметрична. Среднее значение локализуется справа (слева) от середины размаха. Такая форма встречается, когда нижняя (верхняя) граница регулируется либо теоретически, либо по значению допуска или когда левое (правое) значение недостижимо.
- Распределение с обрывом слева. Форма асимметрична. Среднее арифметическое локализуется слева (справа) от середины размаха. Эта форма встречается при 100% просеивании изделий из-за плохой воспроизводимости процесса.
- Плато (равномерное и прямоугольное распределения). Такая форма встречается в смеси нескольких распределений, имеющих различные средние.
- Бимодальное (двухпиковое) распределение. Такая форма встречается, когда смешиваются два распределения с далеко отстоящими друг от друга средними значениями.
- Распределение с изолированным пиком. Такая форма проявляется при наличии малых включений из другого распределения (из другого процесса), появления ошибки измерения или в случае нарушения нормальности процесса.

Рассмотрим влияние формы гистограммы качества на действия специалиста по контролю качества. В случае 1 процесс считается протекающим нормально. В случаях 2 и 3 требуется вмешательство специалистов для проверки и, если потребуется, наладки технологического процесса. Случаи 4, 5 и 6 свидетельствуют о неоднородности данных. Неоднородность может быть вызвана ошибками при сборе данных или свидетельствовать о нестабильности технологического процесса.

Как уже было отмечено, в некоторых источниках на гистограмму накладываются границы допуска (промежуток оси абсцисс между границами допуска называется полем допуска). Можно рассуждать о том, допустимо ли на одной диаграмме объединять график порядковой переменной (гистограмма) и количественной (границы допуска), а такие заблуждения распространены повсеместно. Однако так делают некоторые авторы, поэтому рассмотрим, как можно с некоторой пользой применить данную информацию. Процесс считается протекающим нормально, если почти вся гистограмма находится в границах допуска. В этом случае требуется лишь поддержание существующего состояния технологического процесса. Если гистограмма не удовлетворяет допуску, необходимо добиться смещения среднего значения ближе к центру поля допуска.

## 19.2.2. Диаграмма Парето

Диаграмма Парето (диаграмма распределения Парето) служит для наглядного выявления наиболее значимого фактора, влияющего на снижение качества продукции.



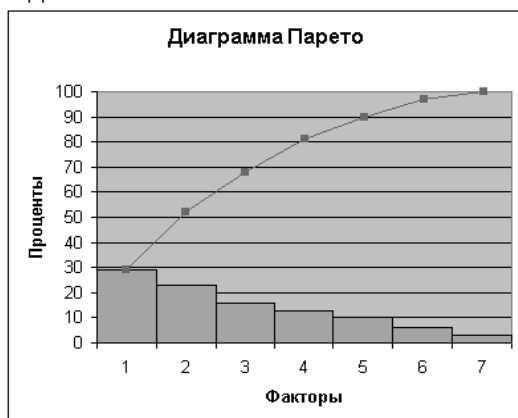




В разных источниках диаграмма Парето может изображаться по-разному, однако общим во всех источниках является комбинация на поле данной диаграммы двух графиков:

- графика типа гистограммы, изображающего процент брака по вине того или иного фактора,
- ломаной линии (полигоном, «кривой эффективности»), отражающей накопленные проценты.

Под браком здесь подразумевается несоответствие достигнутых показателей качества с показателями качества производственного процесса, определяемыми техническими заданиями.



Исходными данными для построения диаграммы Парето служат количества вариантов, относящиеся к каждому фактору, ответственному за снижение качества продукции, либо частоты. На графике по оси ординат откладываются частоты, выраженные в процентах. По оси абсцисс откладываются кодовые обозначения факторов. Частоты могут быть найдены так же, как указано в разделе, посвященном гистограмме качества.

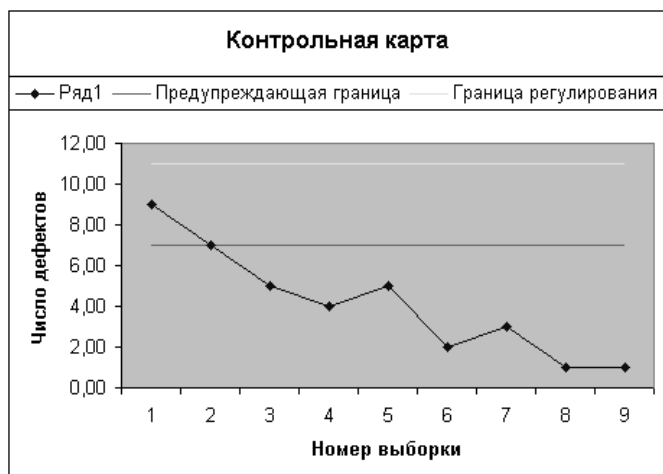
### 19.2.3. Контрольная карта

Контрольная карта предназначена для обнаружения отклонения характеристики качества выпускаемой продукции от заданных технологических норм и допусков.

Контрольная карта представляет собой график изменения исследуемой характеристики во времени. На график дополнительно наносятся предупреждающая граница и граница регулирования. Выделяют различные типы контрольных карт. Рассмотренная контрольная карта относится к типам:

- «рп-карта», т. е. когда показатель качества представлен числом дефектных изделий в последовательности выборок фиксированного объема,
- «с-карта», т. е. когда управление качеством контролируемого производственного процесса ведется по числу дефектов в изделиях одинакового размера.





Если характеристика качества производственного процесса находится ниже предупреждающей границы, то процесс протекает нормально. Если характеристика процесса находится между предупреждающей границей и границей регулирования, то технологический процесс функционирует, но не в соответствии с номиналом. Попадание характеристики в зону выше границы регулирования означает, что должна быть произведена коррекция технологического процесса.

За границу регулирования часто принимается величина, равная утроенному стандартному отклонению, что, как известно, означает попадание в данные границы 95% вариантов в том случае, если распределение нормальное. Подробнее о нормальном распределении и его проверке см. в главе «Проверка нормальности распределения».

Исходными данными для построения контрольной карты служат количества вариантов, относящиеся к каждому фактору, ответственному за снижение качества продукции. На графике по оси ординат откладываются частоты, выраженные в абсолютных величинах (в штуках). По оси абсцисс откладываются кодовые обозначения номеров выборок, отобранных в процессе производства с целью его контроля. Дополнительно в тех же единицах измерения, что и исходные данные, должны быть заданы предупреждающая граница и граница регулирования.

См. Андреев с соавт.

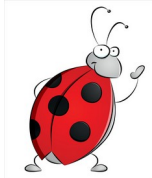
#### 19.2.4. Анализ Бланда–Альтмана

Метод Бланда–Альтмана предназначен для сравнения двух методов клинического или лабораторного контроля. Метод основан на анализе графика, который представляет собой зависимость разности измерений двух методов от среднего данных измерений с указанными средним разностей и 95% доверительными интервалами этого оцениваемого среднего. Средние значения двух измерений вычисляются по формуле

$$z_i = \frac{x_i + y_i}{2}, i = 1, 2, \dots, n,$$

где  $x_i$ ,  $i = 1, 2, \dots, n$  – измерения 1-го метода,





$y_i, i = 1, 2, \dots, n$  – измерения 2-го метода,

$n$  – число измерений каждого метода.

Разности значений двух измерений вычисляются по формуле

$$d_i = x_i - y_i, i = 1, 2, \dots, n,$$

если среднее значение 1-го метода больше среднего значения 2-го метода, либо

$$d_i = y_i - x_i, i = 1, 2, \dots, n,$$

если среднее значение 1-го метода меньше среднего значения 2-го метода.

При этом соответствующие средние значения вычисляются по формулам

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{и} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Среднее разности вычисляется по формуле

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i.$$

Вычисление двустороннего доверительного интервала оцениваемого среднего разности производится по формуле

$$I_{\bar{d}} = \left( \bar{d} - t_{(1+\beta)/2} \cdot \sqrt{\frac{DD}{n}}; \bar{d} + t_{(1+\beta)/2} \cdot \sqrt{\frac{DD}{n}} \right),$$

где  $DD$  – дисперсия разности,

$t_{(1+\beta)/2}$  – значение обратной функции  $t$ -распределения Стьюдента с параметрами  $n - 1$  и  $(1 + \beta) / 2$ ,

$\beta$  – доверительный уровень, выраженный в долях.

В обсуждаемом методе условились применять доверительный уровень 95%.

Дисперсия разности вычисляется как

$$DD = \frac{DX + DY}{2},$$

где  $DX$  – дисперсия 1-го метода,

$DY$  – дисперсия 2-го метода.

При этом соответствующие дисперсии вычисляются по формулам

$$DX = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{и} \quad DY = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2.$$

См. статьи Альтмана (Altman) и Бланда (Bland), Девиitte (Dewitte) с соавт., Манта (Mantha) с соавт., Стокла (Stockl) с соавт.



## Глава 20. Обработка пропущенных данных

### 20.1. Введение

Рассматривается обработка пропущенных значений различными методами. Перед применением метода необходимо убедиться, что он соответствует шкале измерения исходных данных (признаков).

### 20.2. Теоретическое обоснование

Под статистическими данными понимают любую систему данных: числовую и нечисловую информацию, извлекаемую из результатов выборочных обследований, выборки из любых генеральных совокупностей, результаты измерений и т. п. Однако в практической деятельности возникают ситуации, когда часть статистических данных по различным объективным или субъективным причинам оказывается утраченной. Существуют ряд методов, позволяющих специальным образом обработать пропущенные значения и вернуть утраченные данные в последующие процедуры анализа:

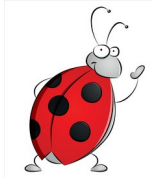
- Игнорирование пропусков.
- Заполнение средним значением.
- Заполнение регрессионными значениями.
- Заполнение случайными значениями.

Мы не рассматриваем причины возникновения пропущенных данных, однако хотелось бы думать, что они обусловлены обстоятельствами, не связанными с физикой явления. Перед применением методов обработки пропущенных данных исследователь должен точно знать, что данные на местах пропусков обязательно должны быть, но отсутствуют, скажем, из-за невнимательности лаборанта, отказа измерительного прибора, неявки пациента на очередное обследование или по другим форс-мажорным причинам. К таким причинам не относится досрочное выбытие из эксперимента объекта, если оно вызвано условиями эксперимента. Обработку таких данных производят с помощью методов анализа цензурированных выборок. Напомним, что цензурированными называются усеченные по условиям эксперимента выборки. Например, при испытании изделий часть их может отказать, а часть не отказать в течение периода испытаний.

Подход, предлагаемый некоторыми авторами и заключающийся в совместном анализе матрицы исходных данных и матрицы пропусков, вряд ли правомерен, т. к. предполагает взаимосвязь изучаемого процесса с причиной возникновения пропусков. Мы же постулируем, что эти явления совершенно независимы.

#### 20.2.1. Игнорирование пропусков

Самым простым и наиболее понятным способом обработки пропущенных значений является их игнорирование. На месте пропущенных значений, если не рассматривать физическую



картину изучаемого явления, а именно так и поступает наука статистика, и не делать никаких дополнительных предположений, могут стоять любые значения.

Метод рекомендуется применять для малых выборок с малым числом пропусков, причем выборки могут принадлежать любой шкале измерения.

## 20.2.2. Заполнение средним значением

Заполнение пропущенных значений некоторыми допустимыми значениями является распространенным способом их восстановления. В качестве допустимого значения выбирается выборочное среднее значение, вычисляемое по формуле

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

где  $x_i$ ,  $i = 1, 2, \dots, n$  – присутствующие варианты выборки,

$n$  – разность между численностью выборки и числом пропущенных значений.

Метод удобен тем, что в результате его применения важнейшая статистическая мера положения, а именно среднее значение, в выборке с заполненными пропусками не изменяется по сравнению со средним значением, вычисленным для выборки с пропущенными значениями (исходной выборки). Однако метод приводит к заниженной выборочной оценке дисперсии с заполненными значениями относительно дисперсии исходной выборки. Другим недостатком является искажение эмпирического распределения выборки независимо от типа эмпирического распределения исходной выборки.

Метод заполнения средним значением рекомендуется применять для больших выборок с малым числом пропусков. Выборка должна принадлежать количественной шкале. Данный метод представляет собой частный случай заполнения регрессионными значениями.

## 20.2.3. Заполнение регрессионными значениями

В практических наблюдениях бывают случаи, когда изменению одного признака соответствует изменение величины другого признака в среднем. Такой вид соотношений называется корреляционной зависимостью, или корреляцией. Считается, что исследование взаимной зависимости приводит к теории корреляции, тогда как изучение зависимости ведет к теории регрессии. Регрессионная модель позволяет выразить значения зависимой (регрессионной) переменной от независимой переменной без исследования функциональной (причинной) связи. Наличие статистической корреляционной зависимости не влечет зависимости причинной. Исследование причинной зависимости – предмет не статистики, а математического моделирования.

Заполнение регрессионными значениями базируется на идее заполнения пропусков, основываясь на информации о связи данной выборки с другой выборкой. При коэффициенте корреляции, по модулю близком к единице, можно предположить существующую тесную регрессионную зависимость между независимой  $x$  и регрессионной  $y$  переменными. Коэффициент корреляции Пирсона вычисляется по формуле



$$r = \frac{Cov(x, y)}{\sqrt{Cov(x, x)Cov(y, y)}},$$

где  $Cov(.,.)$  – выборочная ковариация, вычисляемая по формуле

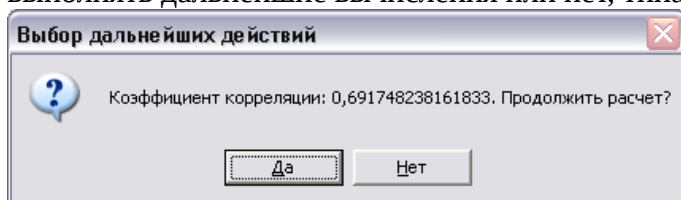
$$Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y}),$$

$x_i, i = 1, 2, \dots, n$  – присутствующие варианты выборки (независимая переменная),

$y_i, i = 1, 2, \dots, n$  – присутствующие варианты выборки (зависимая переменная),

$n$  – разность между численностью выборки и числом пропущенных значений.

После вычисления коэффициента корреляции пользователю предлагается принять решение, выполнять дальнейшие вычисления или нет, типа того, как показано на рисунке.



Если принято решение продолжить, производится вычисление линейной регрессии. В простейшем случае регрессионную зависимость можно представить полиномом 1-й степени  $y = ax + b$ ,

где  $x$  – независимая переменная,

$y$  – зависимая (регрессионная) переменная,

$a$  – коэффициент при первой степени  $x$ ,

$b$  – свободный член – коэффициент при нулевой степени  $x$ .

Иначе данная зависимость называется линейной, т. к. представляет собой уравнение прямой линии на плоскости. Коэффициенты полинома вычисляются по формулам

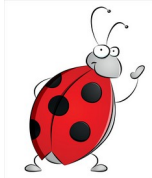
$$a = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i},$$

$$b = \bar{y} - a\bar{x}.$$

Если установлена линейная зависимость между переменными, можно попытаться восстановить отсутствующие значения независимой переменной по регрессионной, решив обратную задачу. В данном случае формула решения обратной задачи принимает вид:

$$x = \frac{y - b}{a}.$$

Данные вычисления возможны, конечно, если в парах значений регрессионные значения известны, а независимые утрачены. Если утраченными оказываются оба значения, заполнение пропусков производится средним значением. Выборка должна принадлежать количественной шкале в случае непрерывных признаков, однако можно предположить, что



представленный алгоритм может быть развит и для признаков других типов, в том числе для смешанных признаков.

## 20.2.4. Заполнение случайными значениями

Метод заполнения случайными значениями, вопреки бытовой трактовке наименования, на самом деле производит наиболее корректное заполнение пропусков из всех представленных методов в смысле сохранения несмещенности статистических параметров выборки. Данный метод теоретически обоснован гораздо лучше других методов. В основе метода, построенного на квазирандомизационном подходе, лежит предположение, что меры положения (средние значения) и меры разброса (средние квадратические отклонения) в присутствующей и в пропущенной частях выборки равны. Общий подход к его реализации заключается в определении типа теоретического распределения эмпирической выборки и последующей случайной генерации отсутствующей части с тем же теоретическим распределением.

Нами решается более частная задача: предполагается, что распределение является нормальным. Если распределение относится к другому стандартному типу, то приведенные ниже выводы должны быть с учетом этого скорректированы. Метод предлагает заполнять пропущенные значения случайными значениями, имеющими нормальное распределение с параметрами, вычисленными по исходной эмпирической выборке.

Генерация выборки, распределенной по нормальному закону  $N(\bar{x}, \sigma^2)$ , производится по выборке, распределенной по стандартному нормальному закону  $N(0,1)$ , с помощью формулы  $y_j = \bar{x} + \sigma x_j, j = 1, 2, \dots, m$ ,

где  $\bar{x}$  – выборочное среднее значение присутствующей части выборки,

$\sigma$  – выборочное среднее квадратичное отклонение, квадратный корень из дисперсии присутствующей части выборки,

$x_j, j = 1, 2, \dots, m$  – выборка, сгенерированная по стандартному нормальному закону  $N(0,1)$ ,

$m$  – число пропущенных значений.

Выборочное среднее определяется аналогично методу заполнения средним значением.

Выборочная дисперсия вычисляется по формуле

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где  $x_i, i = 1, 2, \dots, n$  – присутствующие варианты выборки,

$n$  – разность между численностью выборки и числом пропущенных значений.

В данной формуле в суммировании участвуют только присутствующие варианты выборки.

Выборка, сгенерированная по стандартному нормальному закону, получена из выборки с равномерным распределением в интервале (0,1), путем подстановки ее вариантов в нормальный интеграл (обратную функцию стандартного нормального распределения), представляющую собой решение обратной задачи



$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy,$$

т. е. по известному  $\Phi(x)$  определяется  $x$ .

Рассмотренный метод рекомендуется применять для больших выборок с большим числом пропусков, причем выборка должна принадлежать количественной шкале для случая непрерывных признаков.

## Глава 21. Обработка выбросов

### 21.1. Введение

Рассматривается обработка выбросов. Выбросом называют резко (экстремально) выделяющееся наблюдение. Предположительно данное наблюдение следует исключать из анализа. Хотя возможна ситуация, когда данное значение действительно наблюдалось в эксперименте и следует найти причину его появления.

Перед применением любого метода анализа данных необходимо убедиться, что он соответствует шкале измерения исходных данных (признаков). В случае применения представленных здесь методов анализа признаки должны принадлежать количественной шкале.

### 21.2. Теоретическое обоснование

По словам профессора Клиланд (С. Cleland) из университета Колорадо, «Сила науки – в экспериментальном подходе. История развития науки свидетельствует, что в конечном итоге именно изучение аномалий приводило к потрясающим открытиям и смене существующих научных парадигм». Однако в практической деятельности иногда возникают ситуации, когда экспериментальные статистические данные по объективным или субъективным причинам оказываются засоренными резко выделяющимися, аномальными наблюдениями (выбросами). Выбросы трактуются как грубые ошибки измерений, возникающие в результате просчета, неправильного чтения показаний прибора и т.п.

Существуют ряд методов, помимо непосредственной «ручной» проверки результатов наблюдений, позволяющих специальным образом обработать выбросы. Данные методы основаны на критериях исключения минимального (максимального) наблюдения и подробно описаны в литературе. Рассмотрены методы:

- Критерий Смирнова–Граббса.
- Критерий Титъена–Мура.
- Правило Томпсона.
- Критерий Диксона.
- Критерий Дина–Диксона.
- Критерий Шовене.





- Правило «ящик с усами».

Критерии Диксона и Дина–Диксона разработаны специально для обработки малых выборок, численностью от 3 до 30. Для локализации многочисленных выбросов в больших выборках может применяться критерий Уолша.

Критерий Кокрена, также применяющийся для обработки выбросов, представлен в главе «Дисперсионный анализ». Для обработки выбросов в многомерных данных могут применяться многомерные методы, например, «Факторный анализ» и «Кластерный анализ». Основные предположения при разработке представленных методов заключаются в следующем:

- Исходные данные имеют нормальное распределение.
- Рассматриваем случай, когда основные статистические параметры совокупности (мера положения – среднее значение – и мера разброса – дисперсия) неизвестны и вычисляются по выборке.
- В расчетах ограничиваемся, как это принято при обработке выбросов, стандартным уровнем значимости, равным 0,05.

О выборках, загрязненных выбросами, см. справочник Родионова с соавт. Об обработке выбросов в многомерных выборках см. Афифи с соавт.

### 21.2.1. Критерий Смирнова–Граббса

Критерий Смирнова–Граббса (критерий разногласий, Smirnov–Grubbs) предназначен для исключения одного выброса – резко выделяющегося максимального или минимального наблюдения из нормально распределенной выборки.

Статистика критерия Смирнова–Граббса основана, соответственно, на величине

$$T_N = \frac{\max_i x_i - \bar{x}}{s} \quad \text{или} \quad T_N = \frac{\bar{x} - \min_i x_i}{s},$$

где  $x_1, x_2, \dots, x_N$  – результаты  $N$  наблюдений,

$\bar{x}$  – выборочное среднее значение,

$s$  – выборочное стандартное отклонение, корень квадратный из выборочной дисперсии.

Выборочное среднее значение рассчитывается по формуле

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i,$$

а выборочная дисперсия

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2.$$

Величина статистики критерия сравнивается с критическим значением, точное распределение которого дается формулой (см. руководство NIST/SEMATECH)





$$G_N = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t_{\alpha/N, N-2}^2}{t_{\alpha/N, N-2}^2 + N-2}},$$

где  $t_{\alpha/N, N-2}$  – значение обратной функции  $t$ -распределения с параметрами  $\alpha / N$  и  $N-2$ ,  $\alpha$  – заданный уровень значимости, обычно 0,05.

При величине статистики, большей критического значения, наблюдение исключается. Находит применение еще один метод исключения максимального или минимального наблюдения – критерий Граббса (Груббса), связанный с представленным здесь критерием простой формулой

$$G_N = 1 - \frac{1}{N-1} T_N^2,$$

поэтому дающий точно такие же результаты своего применения.

Критерий Смирнова–Граббса, как и критерий Граббса, не годится для исключения нескольких ( $k > 1$ ) выбросов из-за т. н. маскирующего эффекта. Маскирующим эффектом выброса, при числе выбросов более 1, называют такое смещение параметров выборки, которое не позволяет методу обнаружить все выбросы. Поэтому выбросы уже не могут в рамках данного метода рассматриваться как нетипичные. Для исключения нескольких выбросов рекомендуется применять критерий Титъена–Мура.

См. Мюллера с соавт., Мотульски (Motulsky) с соавт.

## 21.2.2. Критерий Титъена–Мура

Критерий Титъена–Мура (Tietjen–Moore) является обобщением критерия Граббса на случай нескольких выбросов. Статистика критерия исключения  $k$  наибольших или наименьших аномальных значений основана, соответственно, на величине

$$L_k = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x}_k)^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad \text{или} \quad \tilde{L}_k = \frac{\sum_{i=k+1}^N (x_i - \hat{x}_k)^2}{\sum_{i=1}^N (x_i - \bar{x})^2},$$

где  $x_1, x_2, \dots, x_N$  – упорядоченные по возрастанию результаты  $N$  наблюдений,

$\bar{x}$  – выборочное среднее значение,

Выборочное среднее значение (для всей выборки) рассчитывается по формуле

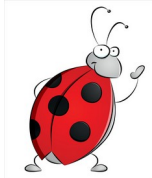
$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i,$$

а средние значения после отбрасывания  $k$  наибольших или наименьших значений рассчитываются, соответственно, по формулам

$$\bar{x}_k = \frac{1}{N-k} \sum_{i=1}^{N-k} x_i \quad \text{или} \quad \hat{x}_k = \frac{1}{N-k} \sum_{i=k+1}^N x_i.$$

Величина статистики критерия сравнивается с табличным значением. При величине статистики, меньшей табличного значения, наблюдения исключаются.





Критерий Титъена–Мура позволяет бороться с маскирующим эффектом. Маскирующим эффектом выброса, при числе выбросов более 1, называют такое смещение параметров выборки, которое не позволяет методу обнаружить все выбросы. Поэтому выбросы уже не могут в рамках данного метода рассматриваться как нетипичные.

Описание критерия см. у Айвазяна с соавт.

### 21.2.3. Правило Томпсона

В правиле Томпсона (критерии Рошера) для исключения выбросов используется статистика

$$t_i = \frac{|x_i - \bar{x}|}{s}, i = 1, 2, \dots, N,$$

где  $x_1, x_2, \dots, x_N$  – результаты  $N$  наблюдений,

$\bar{x}$  – выборочное среднее значение,

$s$  – выборочное стандартное отклонение, корень квадратный из выборочной дисперсии.

Выборочное среднее значение рассчитывается по формуле

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i,$$

а выборочная дисперсия – по формуле

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2.$$

По данным некоторых источников, при вычислении статистики может применяться несмещенная оценка дисперсии

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2.$$

Величина статистики критерия сравнивается с критическим значением, точное распределение которого дается формулой

$$T = \sqrt{\frac{(N-1)t_{1-\alpha/2, N-2}^2}{t_{1-\alpha/2, N-2}^2 + N-2}},$$

где  $t_{\dots}$  – значение обратной функции  $t$ -распределения с параметрами  $1 - \alpha / 2$  и  $N - 2$ ,  $\alpha$  – заданный уровень значимости, обычно 0,05.

При величине статистики, большей критического значения, наблюдение исключается.

Процедура повторяется для каждого наблюдения.

См. Мюллера с соавт.

### 21.2.4. Критерий Диксона

Критерий Диксона (критерий экстремальных значений) применяется для исключения одного максимального или минимального резко выделяющегося наблюдения в выборке численностью от 3 до 30. В критерии Диксона для исключения выбросов используются статистики





$$t_1 = \frac{z_N - z_{N-1}}{z_N - z_1},$$

$$t_2 = \frac{z_N - z_{N-1}}{z_N - z_2},$$

$$t_3 = \frac{z_N - z_{N-2}}{z_N - z_1},$$

где  $z_i, i = 1, 2, \dots, N$  – упорядоченные по возрастанию (при тестировании максимального значения) или по убыванию (при тестировании минимального значения)  $N$  наблюдений. Если хотя бы одна из статистик превышает соответствующее ей критическое значение на заданном уровне значимости (обычно 0,05), наблюдение, соответствующее  $z_N$ , исключается. Таблицы критических значений статистик критерия получены методом компьютерного моделирования. Для расчетов таблицы аппроксимированы гиперболами.

См. результаты Мак–Бейна (McBane).

## 21.2.5. Критерий Дина–Диксона

Критерий Дина–Диксона ( $Q$ –критерий Дина и Диксона) применяется для исключения одного максимального или минимального резко выделяющегося наблюдения в выборке численностью от 3 до 30. В критерии для исключения выбросов используется статистика

$$Q = \frac{|z_1 - z_2|}{|z_1 - z_N|},$$

где  $z_i, i = 1, 2, \dots, N$  – упорядоченные по убыванию (при тестировании максимального значения) или по возрастанию (при тестировании минимального значения)  $N$  наблюдений.

Если значение статистики превышает критическое значение, совпадающее с критическим значением статистики критерия Диксона  $t_1$  на заданном уровне значимости (обычно 0,05), anomальное наблюдение, соответствующее  $z_1$ , исключается.

## 21.2.6. Критерий Шовене

Критерий Шовене предназначен для исключения одного максимального или минимального anomального наблюдения. Статистика критерия основана на величине

$$S_N = N \left\{ 1 - I \left[ \frac{\max_i x_i - \bar{x}}{s} \right] \right\},$$

где  $x_1, x_2, \dots, x_N$  – результаты  $N$  наблюдений,

$\bar{x}$  – выборочное среднее значение,

$s$  – выборочное стандартное отклонение, корень квадратный из выборочной дисперсии,

$I[.]$  – нормальный интеграл.





Выборочное среднее значение рассчитывается по формуле

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i,$$

а выборочная дисперсия – по формуле

$$s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2.$$

Нормальный интеграл определяется формулой

$$I(x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^x e^{-t^2/2} dt$$

и с учетом симметрии функции плотности нормального распределения вычисляется через функцию стандартного нормального распределения как

$$I(x) = 2[\Phi(x) - 0,5].$$

Величина статистики критерия сравнивается со значением 0,5. При величине статистики, меньшей 0,5, наблюдение исключается.

Некоторыми авторами рекомендуется не применять критерий второй раз с использованием пересчитанных заново (после исключения одного аномального наблюдения) значений среднего и дисперсии.

Описание см. в монографии Тейлора.

### 21.2.7. Правило «ящик с усами»

Правило «ящик с усами» получило название от типа соответствующего графика, используемого для наглядного представления разброса эмпирических данных с нанесенными значениями медианы и квартилей.

Порядок вычисления следующий:

- Определяются выборочные значения межквартильного размаха  $f$  и медианы  $\mu$  (подробнее о данных показателях см. главу «Описательная статистика»).
- Выборочные значения, меньшие  $\mu - 1,5f$  и большие  $\mu + 1,5f$ , называются мягкими (подозрительными) выбросами.
- Выборочные значения, меньшие  $\mu - 3f$  и большие  $\mu + 3f$ , называются экстремальными выбросами и должны быть исключены.

Критерий удобен для автоматической идентификации любого числа экстремально малых и больших значений выборки. Критерий очень популярен, однако его рекомендуется применять только в случае, если численность выборки велика.

### 21.2.8. Критерий Кокрена

Критерий  $G$  Кокрена (статистика Кокрена) используется для проверки нулевой гипотезы о равенстве дисперсий нормальных генеральных совокупностей по независимым выборкам с одинаковыми численностями. Вычисление статистики критерия производится по формуле



$$G = \frac{\max_{1 \leq i \leq k} s_i^2}{\sum_{i=1}^k s_i^2},$$

где  $s_i^2, i = 1, 2, \dots, k$  – выборочные дисперсии совокупностей,  
 $k$  – число столбцов (выборок).

Установлено, что  $P$ –значение модифицированной статистики

$$G' = \frac{G(k-1)}{1-G}$$

может быть вычислено как

$$p = k \cdot F_{(n-1), (n-1)(k-1)}(G'),$$

где  $F_{..}(\cdot)$  – функция  $F$ –распределения,  
 $n$  – численность каждой совокупности.

Согласно ГОСТ<sup>10</sup>:

- Если значение тестовой статистики меньше (или равно) 5%–го критического значения, тестируемую позицию признают корректной.
- Если значение тестовой статистики больше 5%–го критического значения и меньше (или равно) 1%–го значения, тестируемую позицию называют квазивыбросом и отмечают одной звездочкой.
- Если значение тестовой статистики больше 1%–го критического значения, тестируемую позицию называют статистическим выбросом и отмечают двумя звездочками.

Метод реализован в «Дисперсионном анализе».

Описание критерия и примеры см. в монографиях Мюллера с соавт., Налимова, Siegel с соавт.

## Глава 22. Рандомизация и генерация случайных последовательностей

### 22.1. Введение

Рассмотрены методы рандомизации и генерирования массивов случайных чисел с равномерным распределением. Они могут применяться для компьютерного моделирования.

<sup>10</sup> ГОСТ Р ИСО 5725–2–2002. Точность (правильность и прецизионность) методов и результатов измерений. Часть 2. Основной метод определения повторяемости и воспроизводимости стандартного метода измерений.





## 22.2. Теоретическое обоснование

Под рандомизацией понимают случайное распределение объектов исследования по группам (например, контрольной и опытной). Рандомизация предполагает соблюдение двух условий: непредсказуемый (случайный) характер распределения пациентов по группам и «слепой» отбор. Сначала решается первая, математическая, часть задачи. Вторая часть задачи («слепой» отбор) носит в основном организационный характер.

Алгоритм рандомизации начинается с генерации псевдослучайной последовательности чисел. Затем, в соответствии с данной последовательностью, осуществляется «перетасовка» всего списка (например, списка пациентов, заданных номерами историй болезни, фамилиями или любым другим удобным для пользователя способом). Для этого меняются местами идентификаторы с номерами  $I$  и  $J$  списка пациентов в соответствии с формулой

$$J = \text{Int}(N \cdot R_i + 1),$$

где  $N$  – численность исходной выборки (под выборкой понимается совокупность объектов, идентифицирующих пациентов),

$R_i$  – число псевдослучайной последовательности, стоящее на месте с номером  $I$ .

После этого производится разделение «перетасованного» списка на заданное число групп. Это могут быть контрольная и опытная группы, группы с другими характеристиками, в зависимости от условий эксперимента.

### 22.2.1. Рандомизация в биомедицинских исследованиях

В настоящее время действуют «Единые требования к рукописям, представляемым в биомедицинские журналы» Всемирной ассоциации редакторов медицинских журналов (World Association of Medical Editors, WAME), объединяющей редакторов более 700 научных журналов из почти 80 стран. Оригинальная версия «Требований» под названием «Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication» доступна на сайте Международного комитета редакторов медицинских журналов (International Committee of Medical Journal Editors, ICMJE). Его повсеместное распространение в информационных целях и приведение требований для авторов всех медицинских журналов в соответствие с данным документом приветствуется ICMJE.

«Требования» гласят, что «сообщения о проведении рандомизированных контролируемых исследований должны содержать информацию обо всех основных элементах исследования, включая протокол (изучаемая популяция, способы лечения или воздействия, исходы и обоснование статистического анализа), назначение лечения (методы рандомизации, способы сокрытия формирования групп лечения) и методы маскировки (обеспечения «слепого» контроля). Авторы, представляющие обзоры литературы, должны включить в них раздел, в котором описываются методы, используемые для нахождения, отбора, получения информации и синтеза данных. Эти методы также должны быть приведены в резюме».

«Приведите детали процесса рандомизации. Опишите, какие методы были применены для обеспечения «слепого» контроля и насколько успешно».



Показано применение методов рандомизации в любых областях, где из некоторого списка необходимо получить тот же список, но в котором позиции будут расположены в случайном порядке, обеспечивая равноправие выбора.

О рандомизации в клинических исследованиях см. монографию Сергиенко с соавт.

## 22.2.2. Генерация случайных последовательностей

Различают случайные и псевдослучайные числа. Они различаются тем, что первые из них генерируются каким-либо стохастическим устройством, а вторые генерируются арифметическим численным алгоритмом. И в том, и в другом случае последовательности чисел обладают определенными статистическими свойствами. В нашем случае генерируются псевдослучайные числа, равномерно распределенные в интервале  $[0,1]$ . Используются 2 метода генерации псевдослучайной последовательности чисел, относящиеся к конгруэнтным генераторам:

- стандартный генератор, рекомендованный комитетом ANSI,
- мультипликативный линейный конгруэнтный датчик.

В теории чисел принята особая запись выражений, незнакомя многим специалистам, не имеющим специального математического образования. Напомним, что она означает. Пусть  $a$  и  $b$  – целые числа. Если их разность делится на число  $m$ , то этот факт выражается записью  $a = b(\text{mod } m)$ .

Запись читается « $a$  сравнимо с  $b$  по модулю  $m$ ». Делитель  $m$  положительный. Он называется модулем сравнения.

Показанная формула буквально означает

$$a - b = mk,$$

где  $k$  – целое число.

Проще говоря, функция  $\text{mod}$  в записи  $b(\text{mod } m)$  определяется как остаток от деления  $b$  на  $m$ .

Петрович с соавт. представили датчики случайных чисел с различными законами распределения. См. также книгу Эфроса. Из других широко применяемых современных алгоритмов назовем генераторы Фибоначчи, описанные в литературе.

### 22.2.2.1. Стандартный генератор ANSI

Стандартный генератор ANSI, он же линейный конгруэнтный датчик, дает высококачественные случайные числа, достаточные для некоторых практических применений.

Вычисление последовательности псевдослучайных чисел производится по формуле

$$R_{n+1} = (a \cdot R_n + c)(\text{mod } m),$$

где  $a$  – мультипликатор,

$c$  – инкремент,

$m$  – модуль.





Целые константы  $a$ ,  $c$ ,  $m$  выбираются определенным образом. Формула дает последовательность псевдослучайных чисел, пригодную для практических применений в некоторых областях, например, рандомизации, однако для целей компьютерного моделирования [распределений] последовательность, полученная с помощью рассматриваемого генератора, считается малоприменимой.

#### 22.2.2.2. Мультипликативный линейный конгруэнтный датчик

Если в соотношении для линейного конгруэнтного датчика положить значение инкремента  $c = 0$ , то оно упростится до

$$R_{n+1} = (a \cdot R_n) \pmod{m}.$$

Датчики, основанные на этой формуле, называются мультипликативными линейными конгруэнтными датчиками (МЛКД). МЛКД в источниках называется также генератором Парка–Миллера (Park, Miller) в честь авторов, исследовавших наборы констант, входящих в формулу.

Существуют различные версии реализаций генератора, в том числе объединения нескольких МЛКД. Одной из известных реализаций такого генератора составлена Лекуйе (L'Ecuyer). Начальное значение для рекуррентной формулы может выбираться различными способами.

## Глава 23. Преобразования данных

### 23.1. Введение

Представлены классические методы преобразования данных.

### 23.2. Теоретическое обоснование

Все преобразования данных можно разделить на 2 группы:

- универсальные,
- частные.

Универсальное преобразование – преобразование к нормальной линейной модели.

Представлены следующие методы универсального преобразования данных:

- преобразование Бокса–Кокса,
- преобразование Зеллера–Реванкара,
- преобразование гиперболического арксинуса,
- преобразование Йео–Джонсона,
- преобразование Джона–Дрейпера,
- преобразование Манли,
- многомерное преобразование Бокса–Кокса.

Частные преобразования подразделяются на преобразования:

- нормализующие ошибки,







- стабилизирующие дисперсии.

Частные методы преобразования рассматриваются при изучении критериев обнаружения гетероскедастичности и способов ее устранения, часто в курсе эконометрики. См. монографии Сокал (Sokal) с соавт., Зар (Zar).

Выбор представленной номенклатуры методов обусловлен характеристиками исходных данных, которые методы могут обрабатывать:

- классические преобразования Бокса–Кокса и Зеллнера–Реванкара – только неотрицательные варианты,
- остальные методы – любые варианты.

Вычисляется оптимальное значение параметра преобразования и соответствующее ему значение логарифмической функции максимального правдоподобия (ФМП), а также выводит преобразованные данные, структура которых соответствует исходным данным. ФМП характеризуют качество подгонки модели. ФМП для различных методов могут сравниваться между собой.

### 23.2.1. Одномерное преобразование

Классическое одномерное преобразование применяется в случаях, когда:

- Представлена одна выборка.
- Представлена совокупность выборок из одной и той же генеральной совокупности. Данная ситуация возникает, например, в однофакторном дисперсионном анализе (см. главу «Дисперсионный анализ»). Численности каждой выборки (каждого столбца таблицы) в данном случае могут совпадать или различаться. При этом с точки зрения оптимизации вся совокупность считается одной выборкой.

Параметр преобразования находится из условия максимума ФМП, с точностью до константы,

$$L(\lambda) = \sum_{i=1}^n \ln J_{\lambda}(x_i) - \frac{n}{2} \ln \hat{\sigma}^2(\lambda),$$

где  $\hat{\sigma}^2(\lambda)$  – оценка дисперсии преобразования,

$\lambda$  – скалярный параметр преобразования,

$n$  – численность выборки,

$x_i, i = 1, 2, \dots, n$  – исходная выборка,

$y_i(\lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$J_{\lambda}(\cdot)$  – якобиан преобразования (определитель матрицы производных  $y_i(\lambda)$  по  $x_i$ ),

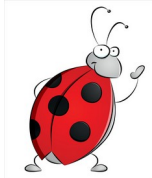
рассчитываемый по соответствующей формуле, зависящей от применяемого метода преобразования.

Оценка дисперсии может быть рассчитана по стандартной формуле

$$\hat{\sigma}^2(\lambda) = \frac{1}{n-1} \left[ \sum_{i=1}^n y_i^2(\lambda) - \frac{1}{n} \left( \sum_{i=1}^n y_i(\lambda) \right)^2 \right].$$

Поиск максимума ФМП осуществляется численно одним из методов локальной оптимизации. Применяется дихотомический поиск (метод деления отрезка пополам).





## 23.2.1. Преобразование Бокса–Кокса

Преобразование Бокса–Кокса (Box–Cox transformation) для каждой неотрицательной варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} \frac{x_i^\lambda}{\lambda}, \lambda \neq 0, \\ \ln x_i, \lambda = 0, \end{cases}$$

где  $\lambda$  – параметр преобразования,

$x_i, i = 1, 2, \dots, n$  – исходная выборка,

$y_i(\lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$n$  – численность выборки.

Якобиан преобразования рассчитывается по формуле

$$J_\lambda(x_i) = x_i^{\lambda-1}, i = 1, 2, \dots, n.$$

См. статьи Бокса (Box) с соавт., Спитцера (Spitzer), Йенг (Yang) с соавт., Линтона (Linton) с соавт., Сакиа (Sakia), Бикел (Bickel) с соавт., монографии Армитэйдж (Armitage) с соавт., Сокал (Sokal) с соавт., работы Зарембка (Zarembka), Джонсона (Johnson R.A.). Метод упоминается в учебниках по эконометрике, например, Доугерти.

## 23.2.1.2. Преобразование Zellner–Revankar

Преобразование Zellner–Revankar transformation) для каждой неотрицательной варианты исходной выборки запишется как

$$y_i(\lambda) = \ln(x_i) + \lambda x_i^2, i = 1, 2, \dots, n,$$

где  $\lambda$  – параметр преобразования,

$x_i, i = 1, 2, \dots, n$  – исходная выборка,

$y_i(\lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$n$  – численность выборки.

Якобиан преобразования рассчитывается по формуле

$$J_\lambda(x_i) = x_i^{-1} + 2\lambda x_i, i = 1, 2, \dots, n.$$

См. статьи Zellner (Zellner) с соавт., Linton (Linton) с соавт.

## 23.2.1.3. Преобразование гиперболического арксинуса

Преобразование гиперболического арксинуса (Arcsinh transformation) для каждой варианты исходной выборки запишется как

$$y_i(\lambda) = \text{Ash}(\lambda x_i) / \lambda, i = 1, 2, \dots, n,$$

где  $\lambda$  – параметр преобразования,  $\lambda \neq 0$ ,

$x_i, i = 1, 2, \dots, n$  – исходная выборка,

$y_i(\lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$n$  – численность выборки,

Ash(.) – функция гиперболического арксинуса.





Гиперболический арксинус может быть подсчитан по формуле

$$\text{Ash}(a) = \ln(a + \sqrt{1 + a^2})$$

Якобиан преобразования рассчитывается по формуле

$$J_{\lambda}(x_i) = (1 + \lambda^2 x_i^2)^{-1/2}, i = 1, 2, \dots, n.$$

См. статьи Линтона (Linton) с соавт., Джонсона (Johnson N.L.), Робинсона (Robinson).

#### 23.2.1.4. Преобразование Йео–Джонсона

Преобразование Йео–Джонсона (Yeo–Johnson transformation) для каждой варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} [(x_i + 1)^{\lambda} - 1] / \lambda, \lambda \neq 0, x_i \geq 0 \\ \ln(x_i + 1), \lambda = 0, x_i \geq 0 \\ - [(1 - x_i)^{2-\lambda} - 1] / (2 - \lambda), \lambda \neq 2, x_i < 0 \\ - \ln(1 - x_i), \lambda = 2, x_i < 0 \end{cases}, i = 1, 2, \dots, n,$$

где  $\lambda$  – параметр преобразования,

$x_i, i = 1, 2, \dots, n$  – исходная выборка,

$y_i(\lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$n$  – численность выборки.

Якобиан преобразования рассчитывается по формуле

$$J_{\lambda}(x_i) = \begin{cases} (x_i + 1)^{\lambda-1}, x \geq 0 \\ (1 - x_i)^{1-\lambda}, x < 0 \end{cases}, i = 1, 2, \dots, n.$$

См. главу 7 монографии Вайсберга (Weisberg).

#### 23.2.1.5. Преобразование Джона–Дрейпера

Преобразование Джона–Дрейпера (John–Draper transformation) для каждой варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} \text{sign}(x_i)[(|x_i| + 1)^{\lambda} - 1] / \lambda, \lambda \neq 0 \\ \text{sign}(x_i) \ln(|x_i| + 1), \lambda = 0 \end{cases}, i = 1, 2, \dots, n,$$

где  $\lambda$  – параметр преобразования,

$x_i, i = 1, 2, \dots, n$  – исходная выборка,

$y_i(\lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$n$  – численность выборки.

Якобиан преобразования рассчитывается по формуле

$$J_{\lambda}(x_i) = \text{sign}(x_i)(|x_i| + 1)^{\lambda-1}, i = 1, 2, \dots, n.$$

См. статьи Джона (John) с соавт., фон Хиппель (von Hippel), Чен (Chen) с соавт.



## 23.2.1.6. Преобразование Манли

Преобразование Манли (Manly transformation) для каждой варианты исходной выборки запишется как

$$y_i(\lambda) = \begin{cases} (e^{\lambda x_i} - 1) / \lambda, \lambda \neq 0 \\ x_i, \lambda = 0 \end{cases}, i = 1, 2, \dots, n,$$

где  $\lambda$  – параметр преобразования,

$x_i, i = 1, 2, \dots, n$  – исходная выборка,

$y_i(\lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$n$  – численность выборки.

Якобиан преобразования рассчитывается по формуле

$$J_\lambda(x_i) = e^{\lambda x_i}, i = 1, 2, \dots, n.$$

См. статью Манли (Manly).

## 23.2.2. Многомерное преобразование

Многомерное преобразование применяется в случае, если данные представлены в виде одной многомерной выборки. Предполагается, что по строкам таблицы при этом расположены варианты, по столбцам – параметры. Численности столбцов в данном случае должны совпадать.

Было бы неверным для каждого параметра многомерной выборки вместо многомерного преобразования применить одномерные маргинальные преобразования, т. к. целью многомерного преобразования является достижение требуемого эффекта (например, многомерной нормальности) всей многомерной выборки, а не только ее компонент. В многомерном случае удобно векторные параметры обозначать соответствующими прописными латинскими и греческими литерами. В этом случае векторный параметр преобразования находится из условия максимума ФМП, с точностью до константы,

$$L(\Lambda) = \sum_{i=1}^n \ln J_\Lambda(X_i) - \frac{n}{2} \ln |\Sigma(\Lambda)|,$$

где  $\Sigma(\Lambda)$  – оценка дисперсионно–ковариационной матрицы преобразования,

$|\cdot|$  – операция вычисления определителя,

$\Lambda$  – векторный параметр преобразования,

$n$  – численность многомерной выборки,

$X_i, i = 1, 2, \dots, n$  – исходная многомерная выборка,

$Y_i(\Lambda), i = 1, 2, \dots, n$  – преобразованные данные,

$J_\Lambda(\cdot)$  – якобиан преобразования (определитель матрицы производных  $Y_i(\Lambda)$  по  $X_i$ ),

рассчитываемый по соответствующей формуле, зависящей от применяемого метода преобразования.

Поиск максимума ФМП может осуществляться численно одним из методов локальной оптимизации. Применяется метод переменной метрики (метод Бройдена–Флетчера–



Голдфарба–Шанно). Определитель эффективно рассчитывается как побочный продукт разложения Грама дисперсионно–ковариационной матрицы.

Методы оптимизации см. в книгах Дэнниса с соавт., Носача, Ортега с соавт. О разложении матриц см. монографию Уилкинсона с соавт.

### 23.2.2.1. Многомерное преобразование Бокса–Кокса

Многомерное преобразование Бокса–Кокса (multivariate Box–Cox transformation) для каждой неотрицательной многомерной варианты исходной выборки в поэлементном виде запишется как

$$y_{ij}(\lambda_j) = \begin{cases} \frac{x_{ij}^{\lambda_j}}{\lambda_j}, \lambda_j \neq 0, \\ \ln x_{ij}, \lambda_j = 0, \end{cases}$$

где  $\lambda_j, j = 1, 2, \dots, p$  – компоненты векторного параметра преобразования,

$x_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, p$  – компоненты исходной многомерной выборки,

$y_{ij}(\lambda_j), i = 1, 2, \dots, n; j = 1, 2, \dots, p$  – компоненты преобразованных данных,

$n$  – численность многомерной выборки,

$p$  – размерность многомерной выборки.

Якобиан преобразования рассчитывается по формуле

$$J_{\lambda_j}(x_{ij}) = x_{ij}^{\lambda_j - 1}, i = 1, 2, \dots, n; j = 1, 2, \dots, p.$$

С учетом данной явной формы якобиана логарифмическая функция максимального правдоподобия для представленного метода преобразования запишется в виде

$$L(\Lambda) = \sum_{j=1}^p (\lambda_j - 1) \sum_{i=1}^n \ln x_{ij} - \frac{n}{2} \ln |\Sigma(\Lambda)|,$$

где  $\Sigma(\Lambda)$  – оценка дисперсионно–ковариационной матрицы преобразования,

$\Lambda$  – векторный параметр преобразования,

$|\cdot|$  – операция вычисления определителя.

При выборе представленного метода преобразования пользователю предоставляется напоминание о необходимости сохранить данные перед производством расчета. Это вызвано возможными вычислительными проблемами решения достаточно сложной оптимизационной задачи. Если такие проблемы возникли, аварийное снятие задачи с выполнения производится средствами операционной системы.

В отличие от одномерных методов, в результате работы данного многомерного метода выводятся компоненты вектора параметра преобразования. Расположение выведенных компонент соответствуют порядку параметров исходной многомерной выборки.

Метод представлен в статье Эндрюс (Andrews) с соавт., работе Уильямс (Williams) с соавт., препринте Рупперта (Ruppert).



## Глава 24. Матричная и линейная алгебра

### 24.1. Введение

Номенклатура матричных операций насчитывает лишь несколько наиболее употребительных методов. Однако более сложные операции могут быть реализованы последовательным применением представленных элементарных операций в требуемой комбинации. Представлены несколько основных алгоритмов, составляющих предмет линейной алгебры. Методы факторизации (разложения, декомпозиции) матриц помогут при конструировании алгоритмов, примеры которых приводятся.

### 24.2. Теоретическое обоснование

В раздел включены необходимые в повседневной работе исследователя алгоритмы матричных вычислений и линейной алгебры. От их качественного исполнения во многом зависит не только эффективность, но и сама работоспособность других алгоритмов. Напомним, что матрицей называется прямоугольная таблица чисел (скаляров) размером  $n$  строк на  $m$  столбцов. При этом матрица размером  $n \times 1$  называется столбцом (вектор–столбцом), а матрица размером  $1 \times m$  называется строкой (вектор–строкой). При  $m = n$  матрица называется квадратной.

Из предлагаемых алгоритмов матричных операций могут быть сконструированы почти все операции, встречающиеся на практике. Так, например, с использованием алгоритмов сложения матриц и умножения матрицы на число, в данном случае на  $-1$ , может быть получен алгоритм вычитания матриц.

#### 24.2.1. Транспонирование матрицы

Транспонированной называется матрица, полученная из данной прямоугольной матрицы путем замены ее строк соответствующими столбцами.

Транспонирование матрицы в матричной записи записывается как:

$$C = A^T,$$

где  $A$  – исходная матрица,

$C$  – транспонированная матрица.

Транспонирование матрицы ведется по формуле (в поэлементной записи):

$$c_{ji} = a_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, m,$$

где  $a_{ij}$  – элемент исходной матрицы,

$c_{ji}$  – элемент транспонированной матрицы,

$n$  – число строк в матрице  $A$  и столбцов в матрице  $C$ ,

$m$  – число столбцов в матрице  $A$  и строк в матрице  $C$ .

#### 24.2.2. Сложение матриц

Сложение двух матриц в матричной записи записывается как:





$$C = A + B,$$

где  $A$  – первое слагаемое,

$B$  – второе слагаемое,

$C$  – матрица – сумма двух матриц.

Сложение двух матриц ведется по формуле (в поэлементной записи):

$$c_{ij} = a_{ij} + b_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, m,$$

где  $a_{..}$  – элемент матрицы – первого слагаемого,

$b_{..}$  – элемент матрицы – второго слагаемого

$c_{..}$  – элемент матрицы – суммы,

$n$  – число строк в каждой из матриц,

$m$  – число столбцов в каждой из матриц.

### 24.2.3. Произведение матриц

Произведение двух матриц в матричной записи записывается как:

$$C = AB,$$

где  $A$  – матрица – первый сомножитель,

$B$  – матрица – второй сомножитель,

$C$  – матрица – произведение.

Произведение матриц ведется по формуле (в поэлементной записи):

$$c_{ij} = \sum_{k=1}^m a_{ik} b_{kj}, i = 1, 2, \dots, n; j = 1, 2, \dots, l,$$

где  $a_{..}$  – элемент матрицы – первого сомножителя,

$b_{..}$  – элемент матрицы – второго сомножителя,

$c_{..}$  – элемент матрицы – произведения,

$n$  – число строк в первом сомножителе,

$m$  – число столбцов в первом сомножителе и строк во втором,

$l$  – число столбцов во втором сомножителе.

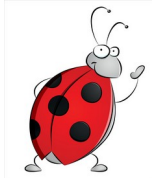
### 24.2.4. Обратная матрица

Пусть исходная квадратная матрица  $A$  имеет отличный от нуля определитель. Тогда существует матрица  $A^{-1}$  (или просто  $A^{-1}$ ), такая, что

$$A^{-1}A = A A^{-1} = I,$$

где  $I$  – единичная матрица, то есть матрица, имеющая единицы на главной диагонали и нули на месте остальных элементов.

Квадратная матрица  $A^{-1}$  называется матрицей, обратной данной матрице  $A$ . Критерий обратимости матрицы дает ее определитель: если он равен нулю, матрица вырождена, если не равен, то матрица обратима.



## 24.2.5. Определитель матрицы

Определителем (детерминантом) матрицы называется многочлен от элементов квадратной матрицы  $A$  порядка  $n$ , каждый член которого является произведением  $n$  элементов, взятых по одному из каждого столбца и каждой строки, и снабжен определенным знаком: плюсом, если перестановка четна, и минусом, если перестановка нечетна. Число  $n$  называется порядком определителя.

По определению, детерминант может быть разложен по элементам любого столбца (строки):

$$\det A = \sum_{j=1}^n a_{jk} A_{jk}, k = 1, 2, \dots, n,$$

где  $k$  – номер столбца (строки),

$A_{jk} = (-1)^{j+k} M_{jk}$  – алгебраическое дополнение элемента  $a_{jk}$ ,

$M_{jk}$  – минор элемента  $a_{jk}$ , то есть определитель порядка  $n - 1$ .

Минор получается при вычеркивании из исходной матрицы  $j$ -й строки и  $k$ -го столбца. При  $j = k$  миноры называются главными.

Определитель является одной из важнейших характеристик квадратной матрицы, определяющей ее поведение в различных алгоритмах. Определители находят и самостоятельное применение, например, при решении систем линейных уравнений методом Крамера.

Показанная выше формула называется рекурсивным определителем детерминанта. Однако на практике, особенно с развитием средств вычислительной техники, данная формула не применяется вследствие ее низкого быстродействия и тенденции накопления ошибки вычислений. Практически детерминант вычисляется через продукты некоторых видов факторизации (разложения) матриц, например, разложения Холецкого и разложения Краута.

## 24.2.6. Умножение матрицы на скаляр

С помощью данной операции производится умножение каждого элемента матрицы на заданную скалярную величину.

Умножение матрицы на скаляр в матричной записи записывается как:

$$C = kA,$$

где  $k$  – скаляр – первый сомножитель,

$A$  – исходная матрица – второй сомножитель,

$C$  – матрица – произведение.

Умножение матрицы на скаляр ведется по формуле (в поэлементной записи):

$$c_{ij} = ka_{ij}, i = 1, 2, \dots, n; j = 1, 2, \dots, l,$$

где  $k$  – значение скаляра – первого сомножителя,

$a_{ij}$  – элемент исходной матрицы – второго сомножителя,

$c_{ij}$  – элемент матрицы – произведения,

$n$  – число строк в каждой матрице,

$l$  – число столбцов в каждой матрице.







Операция находит применение в конструировании различных алгоритмов. Так, например, с использованием алгоритмов сложения матриц и умножения матрицы на скаляр, в данном случае на  $-1$ , может быть получен алгоритм вычитания матриц.

## 24.2.7. Псевдообратная матрица

Для действительной прямоугольной матрицы  $A$  размера  $m \times n$ , где  $m$  – число строк,  $n$  – число столбцов, вводится понятие псевдообратной матрицы  $A^+$ , то есть такой матрицы размера  $n \times m$ , что имеют место свойства

$$AA^+A = A,$$

$$A^+AA^+ = A^+,$$

$$(AA^+)^T = AA^+,$$

$$(A^+A)^T = A^+A,$$

Процедура вычисления псевдообратной матрицы (называемая также обращением Мура–Пенроуза) основана на функции разложения прямоугольной матрицы по сингулярным числам, алгоритм которой требует выполнения условия  $m \geq n$ .

Обсуждение см. в источниках: Голуб (Golub) с соавт., Уилкинсон, Дэннис с соавт., Магнус.

## 24.2.8. Решение системы линейных уравнений

Методы решения систем линейных алгебраических уравнений представлены здесь методом исключения Гаусса с выбором ведущего (главного) элемента. Система линейных алгебраических уравнений в матричной записи выглядит как

$$Ax = b,$$

где  $A$  – матрица системы,

$x$  – вектор подлежащих определению неизвестных,

$b$  – столбец свободных членов.

Решение возможно, если матрица системы не вырождена.

Предполагается, что система является определенной, т. е. число уравнений равно числу неизвестных (иначе матрица  $A$  – квадратная). О решении неопределенных (число уравнений меньше числа неизвестных) и переопределенных (число уравнений больше числа неизвестных) систем см. главу «Распознавание образов с обучением».

Часто возникает необходимость решить так называемую матричную систему линейных уравнений

$$AX = B,$$

где вектор неизвестных и столбец свободных членов обобщаются до матриц, соответственно,  $X$  и  $B$ .

Фактически последняя формула включает в себя несколько систем линейных уравнений, объединенных общей матрицей системы уравнений  $A$ , но имеющих различные столбцы свободных членов, составляющие матрицу  $B$ . В этом случае эффективнее решать систему с помощью одной универсальной функции. Манипуляции с матрицей системы, те же самые для каждого столбца свободных членов, выгоднее производить только один раз. Естественно,



и обычные системы, когда в правой части системы стоит столбец свободных членов, решаются с помощью данной функции.

## 24.2.9. Стандартная проблема собственных значений

Рассмотрим стандартную проблему собственных значений

$$Ax_i = \lambda_i x_i, i = 1, 2, \dots, n,$$

где  $A$  – квадратная матрица размером  $n \times n$ ,

$\lambda_i, i = 1, 2, \dots, n$  – собственные значения матрицы  $A$ ,

$x_i, i = 1, 2, \dots, n$  – соответствующие собственным значениям  $\lambda_i, i = 1, 2, \dots, n$ , вычисленные с точностью до множителя собственные вектора матрицы  $A$ .

Свойства собственных значений подробно представлены в обширной литературе и рассматриваются в вузовском курсе линейной алгебры (напомним, что линейную алгебру составляют два основных раздела: решение систем линейных уравнений и решение проблемы собственных значений).

Решается стандартная проблема собственных значений симметрической (симметричной относительно главной диагонали) действительной матрицы. Симметричность матрицы гарантирует, что вычисленные собственные значения будут действительными. В качестве исходных данных при вычислении используется только верхний треугольник матрицы. В связи с этим проверка симметричности не производится, а правильным может быть только верхний треугольник матрицы. Вычисляются все собственные значения и соответствующие им собственные вектора матрицы. Собственные значения выводятся в неупорядоченном виде. Расположение выводимых собственных векторов соответствует порядку собственных значений. Для решения используется метод Якоби.

См. Уилкинсон.

## 24.2.10. Обобщенная проблема собственных значений

Рассмотрим обобщенную проблему собственных значений

$$Ax_i = \lambda_i Bx_i, i = 1, 2, \dots, n,$$

где  $A$  – симметрическая матрица размером  $n \times n$ ,

$B$  – симметрическая положительно определенная матрица размером  $n \times n$ ,

$\lambda_i, i = 1, 2, \dots, n$  – собственные значения,

$x_i, i = 1, 2, \dots, n$  – соответствующие собственным значениям  $\lambda_i, i = 1, 2, \dots, n$ , вычисленные с точностью до множителя собственные вектора.

Обобщенная проблема собственных значений, путем замены переменных и используя разложение Холецкого, приводится к стандартной проблеме собственных значений.

См. Уилкинсон.

## 24.2.11. Разложение Холецкого

Разложение Холецкого (схема Холецкого, схема квадратного корня) основано на теореме Холецкого, а именно: если  $A$  – симметрическая положительно определенная матрица, то существует разложение





$$A = LL^T,$$

где  $L$  – действительная невырожденная нижняя треугольная матрица.

Рассматриваемое разложение получается из разложения

$$A = LDL^T,$$

где  $L$  – нижняя треугольная матрица с единичной диагональю,

$D$  – диагональная матрица с положительными диагональными элементами,  
как

$$A = LD^{1/2}D^{1/2}L^T = \bar{L}\bar{L}^T,$$

что с точностью до обозначений совпадает с показанной выше формулой.

Этот эффективный вид разложения может применяться в ряде процедур линейной алгебры, например, при решении задачи приведения обобщенной проблемы собственных значений к стандартной проблеме собственных значений и для решения систем линейных уравнений.

Различные аспекты схемы Холецкого и ее применения рассмотрены Мэйндоналдом.

Разложение может быть успешно использовано для вычисления определителя и в задаче генерации многомерного нормального распределения (Мюллер с соавт.).

См. также Уилкинсон с соавт., Гилл с соавт. О вычислении разложения см. Сборник научных программ на Фортране.

## 24.2.12. Разложение Краута

Если  $A$  – неособая матрица, то существует разложение:

$$A = LU,$$

где  $L$  – нижняя треугольная матрица,

$U$  – верхняя треугольная матрица с единичной диагональю.

Разложение не единственно. Если его записать как

$$LU = (LD)(D^{-1}U),$$

где  $D^{-1}U$  есть верхняя треугольная матрица с единичной диагональю, мы получим разложение Краута.

Приведем пример применения рассматриваемого разложения в задаче решения системы линейных уравнений:

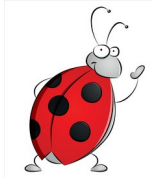
$$Ax = b.$$

Вычислив матрицы  $L$  и  $U$ , сводим задачу к двум элементарно решаемым системам с треугольными матрицами

$$Ly = b \text{ и } Ux = y.$$

Первая из двух последних формул называется прямым исключением, а вторая – обратной подстановкой, что полностью соответствует вычислительной схеме метода Гаусса.

Эквивалентность метода исключения Гаусса и рассматриваемого типа разложения подробно обсуждается Стренгом: рассмотренное разложение представляет собой просто удобную запись метода исключения Гаусса. См. также Уилкинсона с соавт., Форсайта с соавт. О вычислении разложения см. Сборник научных программ на Фортране.



## 24.2.13. Разложение QR

В некоторых приложениях применяется разложение произвольной невырожденной матрицы  $A$ , с помощью процесса ортогонализации Грама–Шмидта, следующего вида:

$$A = QR,$$

где  $Q$  – ортогональная матрица,

$R$  – верхняя треугольная матрица.

Данное  $QR$  разложение эквивалентно построению ортонормированной системы векторов в гильбертовом пространстве, порождающей то же самое линейное многообразие, что и заданная система. Процесс Грама–Шмидта может быть истолкован как разложение невырожденной матрицы в произведение ортогональной матрицы и верхней треугольной матрицы с положительными диагональными элементами. Решение ведется по рекуррентным формулам

$$u_i = \frac{v_i}{\|v_i\|}, v_1 = e_1, v_{i+1} = e_{i+1} - \sum_{k=1}^i (u_k, e_{i+1}) u_k, i = 1, 2, \dots,$$

где  $u$  – ортонормированная система векторов,

$e$  – заданная система векторов,

$n$  – число векторов.

Нетрудно видеть, что вычисленные векторы  $n$  представляют собой столбцы ортогональной матрицы  $Q$ .

См. Стренг.

## 24.2.14. Разложение по сингулярным числам

Рассмотрим разложение действительной прямоугольной матрицы  $A$  размером  $m$  строк на  $n$  столбцов, причем  $m \geq n$ , вида

$$A = U \begin{pmatrix} S \\ 0 \end{pmatrix} V^T,$$

где  $U$  – матрица размером  $m \times m$ , сформированная из  $m$  ортонормированных собственных векторов, соответствующих собственным значениям матрицы  $AA^T$ ,  $U^T U = I_m$ ,

$V$  – матрица размером  $n \times n$ , состоящая из  $n$  ортонормированных собственных векторов матрицы  $A^T A$  и обладающая свойствами  $V^T V = V V^T = I_n$ ,

$S$  – диагональная матрица, диагональные элементы которой представляют собой так называемые сингулярные числа – квадратные корни из неотрицательных собственных значений матрицы  $A^T A$ ,

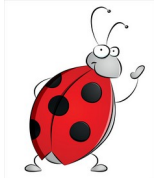
$0$  – прямоугольная нулевая матрица размером  $m - n$  строк на  $n$  столбцов,

$I$  – единичная матрица соответствующего порядка.

Рассмотренное разложение называется разложением по сингулярным числам (сингулярным разложением). В другой записи

$$A = U_n S V^T,$$





где  $U_n$  – матрица размером  $m \times n$ , далее для простоты обозначаемая, как  $U$ , и сформированная из  $n$  ортонормированных собственных векторов, соответствующих  $n$  наибольшим из  $m$  собственным значениям матрицы  $AA^T$ , и обладающая свойствами  $U^T U = V^T V = VV^T = I_n$ .

См. источники: Голуб (Golub) с соавт., Дэннис с соавт., Стренг, Уилкинсон с соавт.

## 24.2.15 Мультиколлинеарность

Вектора называются коллинеарными, если они лежат на параллельных прямых либо на одной прямой. Понятие коллинеарности, пришедшее из аналитической геометрии, тождественно линейной зависимости из линейной алгебры (линейная зависимость – также критерий мультиколлинеарности). Необходимость исследования коллинеарности возникает, например, перед применением ряда методов многомерного статистического анализа (множественная регрессия, факторный анализ) с целью исключения из рассмотрения линейно зависимых параметров. Это необходимо как для уменьшения размерности задачи, так и для снижения вычислительных сложностей.

Фаррар (Farrar) и Глаубер (Glauber) предложили совокупность статистических методов определения наличия мультиколлинеарности, известную под наименованием алгоритма Фаррара–Глаубера.

Пусть обозначено:

$n$  – число строк в матрице исходных данных (число наблюдений),

$m$  – число столбцов в матрице исходных данных (число векторов, в анализе данных – число параметров, в эконометрике – число объясняющих переменных),

$R$  – корреляционная матрица (о ее вычислении см. главу «Корреляционный анализ»),

$C$  – матрица, обратная корреляционной.

Алгоритм Фаррара–Глаубера статистически исследует проявления мультиколлинеарности, представленные далее.

### 24.2.15.1. Корреляция между параметрами

Вычисляется статистика критерия

$$S = -[n - 1 - (2m + 5) / 6] \ln|R|,$$

где  $|\cdot|$  – определитель.

Статистика имеет распределение хи-квадрат с  $m(m - 1) / 2$  степенями свободы. При значимой статистике хи-квадрат есть основание предположить наличие явления мультиколлинеарности в исследуемой системе векторов.

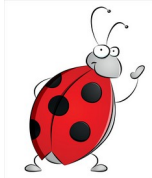
### 24.2.15.2. Коэффициенты детерминации векторов

Для каждого вектора вычисляется коэффициент детерминации

$$R_k^2 = 1 - 1 / c_{kk}, k = 1, 2, \dots, m,$$

где  $c_{kk}, k = 1, 2, \dots, m$ , – диагональный элемент матрицы  $C$ .





Также для каждого вектора вычисляется статистика

$$F_k = (c_{kk} - 1) \frac{n - m}{m - 1}, k = 1, 2, \dots, m.$$

Статистики  $F_k, k = 1, 2, \dots, m$ , подчиняются  $F$ -распределению со степенями свободы  $n - m$  и  $m - 1$ . При значимых  $F$ -статистиках есть основание предположить коллинеарность данного вектора с некоторыми или всеми остальными векторами. Такие векторы (в задаче распознавания образов соответствующие параметрам распознавания) следует исключить из матрицы исходных данных.

## 24.2.15.2. Частные коэффициенты корреляции

Для каждого внедиагонального элемента корреляционной матрицы (в силу симметрии исследуется только верхняя часть) вычисляется частный коэффициент корреляции

$$r_{kj} = - \frac{c_{kj}}{\sqrt{c_{kk} c_{jj}}}, k = 1, 2, \dots, m - 1; j = k + 1, \dots, m.$$

Также для каждого частного коэффициента корреляции вычисляется статистика

$$t_{kj} = \frac{r_{kj} \sqrt{n - m}}{\sqrt{1 - r_{kj}^2}}, k = 1, 2, \dots, m - 1; j = k + 1, \dots, m.$$

Статистики  $t_{kj}, k = 1, 2, \dots, m - 1; j = k + 1, \dots, m$ , подчиняются  $t$ -распределению с  $n - m$  степенями свободы. При значимых  $t$ -статистиках есть основание предположить коллинеарность в исследуемой паре векторов  $k$  и  $j$ .

Обзор методов исследования мультиколлинеарности см. в монографиях Ферстера с соавт., Айвазяна с соавт. См. также оригинальные статьи Фаррара с соавт., Карнеса (Carnes) с соавт., Уишера (Wichers). См. статью Рокуэлла (Rockwell), книгу Бородича (включая вывод основных формул и рекомендательные меры по устранению мультиколлинеарности).

## 24.2.16. Кронекеровское произведение

Кронекеровским произведением матриц  $A = (a_{ij})$  размером  $m \times n$  и  $B = (b_{st})$  размером  $p \times q$  называется матрица  $C = A \otimes B$  размером  $mp \times nq$  такая, что

$$\begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

См. монографию Магнуса.



## Глава 25. Обыкновенные дифференциальные уравнения

### 25.1. Введение

Представленные методы отражают многообразие подходов к решению проблемы и предназначены для решения уравнений соответствующих типов.

### 25.2. Теоретическое обоснование

Порядком дифференциального уравнения называют наибольший порядок входящих в него производных. Дифференциальное уравнение первого порядка называется также обыкновенным дифференциальным уравнением и может быть записано как

$$F(x, y, y') = 0,$$

где  $y$  – скалярная или  $s$ -компонентная векторная функция, подлежащая определению, в механике называемая вектором обобщенных координат,

$x$  – независимая переменная, при исследовании динамики протяженных во времени процессов в ее качестве берется время  $t$ ,

штрих означает полную производную по независимой переменной  $x$  и может быть заменен стандартным обозначением  $d / dx$ .

Если в явном виде независимая переменная в уравнения не входит, система уравнений называется автономной. Если система уравнений не автономна, то можно выбрать в качестве времени новую независимую переменную  $\tau$ , а затем, добавив к системе еще одно уравнение  $d\tau / d\tau = 1$ ,

превратить исходную систему уравнений в систему автономную, хотя явной выгоды от данного процесса можно и не ощутить. Выгода может быть получена при вычислениях, если составлять все алгоритмы только для автономных систем уравнений. Это несколько упростит интерфейс составленных функций. Можно также заранее предусмотреть автоматическое преобразование неавтономной системы уравнений к автономной системе уравнений по рассмотренному выше простому алгоритму.

Степенью дифференциального уравнения называют высший показатель степени, в которой старшая производная входит в уравнение после того, как уравнение приведено к рациональному виду. Если в уравнение искомая функция и все ее производные входят в первой степени, уравнение называется линейным, в противном случае уравнение будет нелинейным.

Если исходное уравнение разрешимо относительно производной, то получается уравнение  $y' = f(x, y)$ ,

где  $f(x, y)$  – функция правых частей (правая часть).

Общее решение дифференциального уравнения определяется с точностью до постоянных  $C$ , определяющих семейство кривых

$$y = y(x) + C.$$







При задании начального условия и в случае выполнения перечисленных требований решение существует и единственно. Задача Коши формулируется как аналитическое или численное решение системы  $s$  обыкновенных дифференциальных уравнений, в общем случае

нелинейных (с нелинейной функцией правых частей), на интервале  $x \in [a, b]$  с начальными условиями

$$y(x_0) = y_0,$$

где  $x_0 = a$ .

Иногда, чтобы отличить одно дифференциальное уравнение от системы дифференциальных уравнений, первое называют скалярным дифференциальным уравнением. Уравнение высшего порядка может быть сведено к системе уравнений первого порядка простым изменением обозначений и введением новых переменных. Размерность системы – это число уравнений, содержащих производные порядка не выше первого. Все выводы, сделанные относительно одного дифференциального уравнения, естественно обобщаются на систему, поэтому в теоретических выкладках не делается разницы, является ли скалярной или  $s$ -компонентной векторной функцией. Все показанные формулы могут быть использованы как для исследования одного уравнения, так и для исследования системы уравнений. В последнем случае можно просто считать, что величины, входящие в формулы, представляют собой не скаляры, а векторы.

Везде приводятся функции решения только одного уравнения, однако нетрудно обобщить их на решение системы уравнений. Для этого потребуется модернизация приведенных функций. Чтобы научить их решать системы уравнений, следует заменить векторами (массивами) функции и входящие в формулы скалярные параметры. Кроме того, там, где производится проверка точности, нужно в цикле производить проверку каждого элемента вектора (или их совокупность), ответственного за точность решения.

Единственная особенность системы уравнений относительно одного уравнения заключается в том, что принятие решения в случае недостаточной точности следует делать сразу же, как только один из элементов вектора даст для этого повод. Принятие же решения в случае чрезмерной точности делать только после того, как проверки всех элементов покажут, что точность чрезмерна (см. второй вариант метода Мерсона). Для систем, соответственно, должны быть изменены также краевые – начальные и граничные – условия задачи.

В заключение раздела отметим один научный факт, который будет использован в дальнейших рассуждениях. Система нелинейных дифференциальных уравнений может быть представлена в виде линеаризованной системы, которая получена с использованием только первых двух членов разложения решения в ряд Тейлора

$$y' = J(x)y + (**),$$

где  $J$  – [локально] постоянная матрица Якоби системы – матрица частных производных правых частей  $\partial f / \partial y$ ,

(\*\*) – нелинейные члены.





## 25.2.1. Математическое моделирование

Математическими формулами можно описать, но нельзя объяснить физическую картину явлений. Наличие формул, описывающих наблюдаемое явление, не компенсирует отсутствия точных знаний или хотя бы гипотез о причине явления.

Реальные объекты или явления представляют собой сущности с бесконечно богатым содержанием. Более того, они самостоятельно или в совокупности с другими объектами или явлениями способны порождать новые сущности. В математической модели переменные и константы моделируют реальные сущности, но не могут самостоятельно их порождать, представляя собой только символы на бумаге или на экране компьютера. При попытках более детального исследования или получении новых экспериментальных данных приходится создавать новые математические модели или совершенствовать старые. Поэтому главное в математическом моделировании – выделение свойств исследуемых объектов или явлений, существенных для решаемой в данный момент конкретной практической задачи. «Дать точное описание наблюдавшихся явлений природы, выхватить из многообразия деталей и мелочей главные, характерные черты, в резкой и краткой форме сформулировать все, что видел глаз и охватила мысль – это настолько сложная и важная задача, что перед ней бледнеют все трудности лабораторного исследования или теоретического анализа в кабинетах ученых» (А.Е. Ферсман).

Пытаясь описать всевозможные характеристики явления, до которых исследователь только смог добраться, он рискует запутаться в частностях и не описать их вовсе. Только опыт и здравый смысл помогут отличить влияние существенной части влияния на исследуемый процесс от совокупного влияния малозначительных черт. «Искусство быть мудрым состоит в умении знать, на что не следует обращать внимания» (У. Джеймс). Истина не может быть сложной для понимания – в противном случае не может быть даже установлено, что есть истина. Разумный уровень абстрактности позволит добиться успеха при исследовании даже очень сложных явлений.

Сказанное в предыдущем абзаце не означает, что исследование математических моделей позволяет только описать и наглядно отобразить явления, как это делают методы описательной статистики. Наоборот, исследование корректно составленных моделей позволяет выявить совершенно новые, часто неподдающиеся прямому исследованию в эксперименте, характеристики явлений, дополняя и даже иногда частично заменяя экспериментальные методы исследований, и породить новые неожиданные идеи.

В математическом моделировании, когда модель записана в виде системы обыкновенных дифференциальных уравнений, искомая функция  $y$  часто называется выходом модели в противоположность тому, что ее экспериментальный «аналог» называется выходом эксперимента. Выход модели в математическом моделировании может также представлять собой ту или иную, в том числе нелинейную, комбинацию всех или некоторых из  $s$  компонент функции  $y$  либо только наблюдаемую часть компонент векторной функции  $y$ , если по условиям эксперимента наблюдение всех компонент затруднительно.



Дадим несколько определений, которые могут пригодиться при математическом моделировании и исследовании математических моделей, в том числе на предельных (критических) режимах.

Детерминированная система – такая система, для которой существует правило в виде дифференциальных или разностных уравнений, определяющее ее будущее поведение, исходя из заданных начальных условий. Противоположностью детерминированной системы является система вероятностная (статистическая).

Выбор адекватной математической модели (детерминированной или вероятностной) может быть сделан на основе информационного анализа изучаемой физической системы.

Потенциальной называется сила, работа которой зависит только от начального и конечного положения точки ее приложения и не зависит от вида траектории и закона движения точки.

Гамильтонов подход к описанию динамики физических систем основан на системе обыкновенных дифференциальных уравнений

$$\dot{q}_i = \partial H / \partial p_i, \dot{p}_i = -\partial H / \partial q_i, i = 1, 2, \dots, n,$$

где точка означает полную производную по времени,

$q_i, p_i, i = 1, 2, \dots, n$ , – соответственно, обобщенные координаты и обобщенные импульсы, их совокупность называется каноническими переменными,

$n$  – число степеней свободы (число независимых обобщенных координат),

$H = H(q, p)$  – функция Гамильтона (гамильтониан), характеризующая физическое состояние системы.

Задача решается с начальными условиями при  $t = t_0$

$$p_i(t_0) = p_i^0, q_i(t_0) = q_i^0.$$

Если система автономна и действующие силы потенциальны, гамильтониан является полной энергией системы, выраженной через канонические переменные.

Решение представленной выше системы можно представить как движение точки в  $2n$  – мерном пространстве с координатами  $q$  и  $p$ . Такое пространство называется фазовым, его точки – фазовыми точками, а траектории движения фазовых точек – фазовыми траекториями (фазовыми кривыми). Совокупность фазовых кривых называют потоком.

Консервативная система – система, для которой имеет место закон сохранения энергии.

Другие законы сохранения (например, количества движения), могут не соблюдаться. Для консервативной системы элемент фазового пространства изменяет форму, но сохраняет объем. Примером консервативной системы является Солнечная система.

Для Гамильтоновых систем траектории в фазовом пространстве не пересекаются.

Гамильтоновы системы консервативны. Большинство изучаемых систем не являются Гамильтоновыми.

Диссипативная система – система, полная механическая энергия которой (кинетическая плюс потенциальная) при движении убывает (рассеивается), переходя в другие формы энергии, например, в энергию теплового хаотического движения молекул. К диссипативным системам относится большинство изучаемых систем. Примерами диссипативных систем являются



механические системы с трением, движение вязких жидкостей. Для диссипативной системы объем элемента фазового пространства сокращается (фазовый элемент сжимается) с течением времени. Сокращение фазового объема приводит к тому, что при  $t \rightarrow \infty$ , где  $t$  – параметр (время), все решения диссипативной системы будут стягиваться к некоторому подмножеству фазового пространства, называемому аттрактором.

Неконсервативная система может не быть диссипативной, если в ней рассеяние энергии компенсируется притоком энергии извне, хотя многие авторы не делают таких тонких отличий. Так, в одном литературном источнике утверждается, что в диссипативной системе переход к хаосу возможен якобы только при внешнем возбуждении (подводе энергии в открытую систему извне).

## 25.2.2. Основные предположения

Введем предположения относительно свойств исследуемой системы дифференциальных уравнений, позволяющие надеяться, что применение рассмотренных методов даст приемлемый результат.

Основное предположение относительно функции  $f(x, y)$  состоит в том, что она удовлетворяет условию Липшица

$$\|f(x, y_1) - f(x, y_2)\| \leq L \|y_1 - y_2\|,$$

где  $L$  – константа Липшица,

для всех значений  $x \in [a, b]$  и, в общем случае, всех соответствующих компонент векторов  $y_1$  и  $y_2$ .

Константа Липшица играет важную роль в теории численных методов, в частности, при исследовании их численной устойчивости, и может быть вычислена как

$$L = \|\partial f / \partial y\|,$$

в чем нетрудно заметить норму матрицы Якоби.

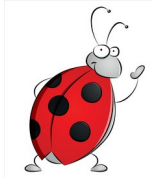
При выводе методов решения, основанных на разложении в ряд Тейлора, предполагается надлежащая дифференцируемость  $y$ . При этом говорят, что функция  $y$  должна обладать той степенью гладкости, которая требуется в конкретном частном случае. Проблема может возникнуть, если искомые функции имеют разрыв. Кроме того, разрыв могут иметь правые части дифференциальных уравнений. Такая ситуация может быть обусловлена конструктивными особенностями моделируемой физической системы.

Напомним, что точкой разрыва (особой точкой) функции называется такая точка, в которой функция не является непрерывной. Точка разрыва будет 1 рода, если существуют пределы

$$\lim_{x \rightarrow x_0 + 0} f(x) = f(x_0 + 0) \quad \text{и} \quad \lim_{x \rightarrow x_0 - 0} f(x) = f(x_0 - 0).$$

Величина  $f(x_0 + 0) - f(x_0 - 0)$  называется скачком функции.

Точка разрыва будет 2 рода, если функция определена в окрестности особой точки за исключением, быть может, самой точки. Также один из рассмотренных выше пределов не существует.



Разрыв 1 рода в правых частях дифференциальных уравнений преодолевается численным алгоритмом без какой-либо модернизации алгоритма.

### 25.2.3. Устойчивость

Применительно к дифференциальным уравнениям различают численную устойчивость методов решения и устойчивость системы дифференциальных уравнений. Различают устойчивость метода и устойчивость уравнений как математической модели некоторого физического явления, причем в последнем случае неустойчивость уравнений прямо означает неустойчивость моделируемого физического явления. При сходстве методов анализа устойчивости в том и в другом смысле совершенно различаются объекты, к которым применяется данный анализ.

Анализу устойчивости методов посвящен значительный объем источников по численным методам решения дифференциальных уравнений. Более того, описание вновь введенного метода принято сопровождать анализом его устойчивости. Анализ устойчивости метода сводится к изучению глобальной ошибки решения, складывающейся из ошибки усечения и ошибки распространения. На основе этого анализа для различных методов, если говорить о численных методах решения обыкновенных дифференциальных уравнений, выводятся рекомендации по выбору минимально допустимого шага интегрирования либо, в общем смысле, интервала устойчивости.

Другое определение устойчивого алгоритма проще: устойчивым называется алгоритм, в котором не накапливаются ошибки округления.

#### 25.2.3.1. Жесткие задачи

Устойчивое дифференциальное уравнение называется жестким, если оно имеет частное решение в виде убывающей экспоненты, постоянная времени которого очень мала по сравнению с длиной интервала, на котором разыскивается решение. Математически о высокой жесткости системы свидетельствует большое значение константы Липшица. Есть и другой показатель. Согласно Лэмберту, задача Коши для устойчивой системы называется жесткой, если локальный коэффициент жесткости задачи

$$S(x) = \max_{i=1,2,\dots,n} \operatorname{Re}(-\lambda_i) / \min_{i=1,2,\dots,n} \operatorname{Re}(-\lambda_i) \gg 1,$$

где  $\lambda_i, i=1,2,\dots,n$  – собственные значения матрицы Якоби системы.

Жесткой считается система уже при значении  $S(x) = 10$ , хотя на практике могут встречаться значения до  $S(x) = 10^6$ .

Жесткие системы трудны для решения, причем это проблема чисто вычислительная, не отражающая каких-то особых свойств моделируемой физической системы. Жесткие задачи решаются с помощью специально разработанных методов.



## 25.2.3.2. Устойчивость решения

Устойчивость решения системы дифференциальных уравнений соответствует устойчивости движения моделируемой физической системы. Описываемое системой дифференциальных уравнений движение (под которым можно понимать любой изменяющийся процесс) называется устойчивым асимптотически, если достаточно малым возмущениям будет соответствовать наперед заданная малость возмущенного движения. Для устойчивого движения имеет место также стремление амплитуды этого возмущенного движения к нулю при неограниченном росте времени.

Практически неустойчивое поведение физической системы ведет к ее разрушению, к выходу из интервала допустимых эксплуатационных параметров либо (в случае физической системы, описываемой системой нелинейных дифференциальных уравнений) к хаосу (беспорядку, нерегулярности) того или иного типа. «... маленькая ошибка в начале может стать большой в конце» (Фома Аквинский). Критерием перехода регулярной организованной структуры к хаосу может служить устойчивость структур по отношению к малым возмущениям. Если такая устойчивость отсутствует, детерминированное описание структур теряет смысл, и необходимо использовать статистические методы.

Часто для анализа поведения объекта или явления, описываемого системой дифференциальных уравнений, необязательно решать эту систему. Если исследователя, например, интересуют значения некоторых параметров системы или начальных условий задачи, определяющих, будут ли в системе затухать возмущения, или наоборот, малые возмущения будут приводить к нарастанию амплитуды возмущенного движения, значит, исследователю нужно исследовать устойчивость системы (интегрировать систему дифференциальных уравнений нет необходимости).

## 25.2.4. Численное решение дифференциальных уравнений

Если решается уравнение порядка выше первого, количество начальных условий будет равно порядку уравнения, т. к. начальные условия накладываются как на функцию  $y$ , так и на все ее производные. Сказанное естественно обобщается и на систему уравнений.

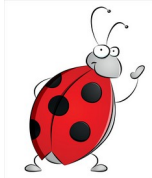
Среди численных методов наиболее употребительны разностные методы, требующие для своего функционирования знания значений функции в нескольких последовательных точках (многошаговые методы), и одношаговые методы (например, метод Рунге–Кутты), требующие знания значения функции только в одной предыдущей точке. И у одношаговых, и у многошаговых методов есть свои достоинства и недостатки, перечисленные при их описании.

Все рассмотренные численные методы являются дискретными, т. е. с их помощью ищется последовательность значений искомой функции  $y$  в  $N$  заданных точках

$$y_n \approx y(x_n),$$

где  $x_{n+1} = x_n + h_n, n = 0, 1, \dots, N - 1, x_0 = a, x_N = b$  – заданное множество точек интегрирования,  $h_n > 0$  – шаг сетки.





Чаще всего рассматривается случай  $h_n = h$ , где  $h$  – постоянная величина.

Следует отличать шаг сетки решения от шага вывода решения, т. е. шага, с которым решение должно выводиться. Хотя эти величины могут совпадать, в общем случае они различны, причем вторая намного (возможно, на порядки) может превосходить первую. Удобно, чтобы данные величины были кратными для гарантии от получения чрезмерно малого значения шага сетки в конце интервала. Это справедливо для всех методов, особенно для методов с автоматическим выбором длины шага, при вызове которых как начальный шаг решения, так и минимальный шаг должны быть кратны шагу вывода. Наиболее распространенные схемы решения обыкновенных дифференциальных уравнений без потери общности рассмотрим на примере решения одного дифференциального уравнения.

## 25.2.4.1. Одношаговые методы

Решение обыкновенного дифференциального уравнения в точке  $x_{n+1}$ , если оно известно в точке  $x_n$ , на основе разложения в ряд Тейлора ищется по формуле

$$y(x_{n+1}) = y(x_n) + h\Delta(x_n, y_n, h),$$

где

$$\Delta(x_n, y_n, h) = \sum_{k=1}^{\infty} \frac{h^{k-1}}{k!} y^{(k)}(x).$$

Если бесконечный ряд оборвать на некотором числе членов и точное значение  $y(x_n)$

заменить приближенным значением  $y_n$ , получится формула,

$$y_{n+1} = y_n + h\varphi(x_n, y_n, h),$$

где

$$\varphi(x, y, h) = \sum_{k=1}^p \frac{h^{k-1}}{k!} f^{(k-1)}(x_n, y_n),$$

где  $p$  – степень метода решения.

Считается (возможно, что это утверждение не является однозначно верным), что формулы более высоких степеней дают не только более высокую точность, но и уменьшают количество вычислительной работы, т. к. допускают большие величины шага. При  $p = 1$  имеем известный метод Эйлера.

С целью исключения производных из формулы для  $\varphi$  при построении методов со степенями  $p > 1$  Рунге, Хойн и Кутта предложили процесс «подгонки» рядов Тейлора выражениями, в общем случае имеющими вид

$$\varphi(x, y, h) = \sum_{r=1}^m c_r k_r,$$

где  $m$  – количество этапов решения.





В названии конкретного метода фигурирует как число этапов, означающее количество вычислений правой части, так и степень метода, означающее число удерживаемых членов разложения и характеризующего теоретическую точность метода. Для степени  $p > 4$  в источниках показано, что количество этапов  $m > p$ .

В источниках все одношаговые методы часто традиционно называют методами Рунге–Кутты.

### 25.2.4.1.1. Явные схемы

Пусть коэффициенты, входящие в предыдущую формулу, вычисляются как

$$k_1 = f(x, y), k_r = f(x + \alpha_r h, y + h \sum_{s=1}^{r-1} \beta_{r,s} k_s), r = 2, 3, \dots, m,$$

где  $\alpha, \beta$  – константы.

Данная формула определяет общий вид так называемых явных схем одношаговых методов. Подробный вывод громоздок и здесь не приводится.

Основное применение явные схемы находят при решении нежестких систем. Представлены метод Рунге–Кутты и методы Мерсона.

### 25.2.4.1.2. Неявные схемы

Для неявных схем отличие от явных схем в вычислении коэффициентов, входящих в общую формулу одношаговых методов, следующее

$$k_r = f(x + \alpha_r h, y + h \sum_{s=1}^m \beta_{r,s} k_s), r = 1, 3, \dots, m,$$

где  $\alpha, \beta$  – константы.

Неявные методы ориентированы на решение жестких систем. Они дают хорошие результаты и для нежестких систем, но по времени счета превосходят их.

Представлен метод Хаммера–Холлингсуорта.

### 25.2.4.1.3. Метод Рунге–Кутта

Расчетные формулы для схемы классического метода Рунге–Кутты имеют вид

$$y_{n+1} = y_n + (k_1 + 2k_2 + 2k_3 + k_4) / 6,$$

где

$$k_1 = hf(x_n, y_n),$$

$$k_2 = hf(x_n + h/2, y_n + k_1/2),$$

$$k_3 = hf(x_n + h/2, y_n + k_2/2),$$

$$k_4 = hf(x_n + h, y_n + k_3).$$

Наилучший прием проверки точности решения состоит в том, чтобы после окончания решения задачи методом Рунге–Кутта уменьшить шаг решения  $h$  в 2 раза и провести решение





с новым шагом. Если незначительность разницы между двумя решениями устраивает исследователя, можно удовлетвориться решением с первоначальным шагом. В противном случае алгоритм уменьшения шага решения следует повторить.

Классический метод Рунге–Кутты может рассматриваться, как обобщение на дифференциальные уравнения квадратурной формулы Симпсона, предназначенной для численного вычисления определенных интегралов.

#### 25.2.4.1.4. Методы Мерсона

Рассматриваемые далее формулы получены на основе идеи, что если схема решения дифференциального уравнения наряду с приращением функции будет содержать некоторое приближение более высокого порядка, последнее можно использовать для управления погрешностью решения и длиной шага интегрирования.

Созданная Мерсоном оригинальная модификация метода Рунге–Кутты предоставляет возможность автоматического выбора шага  $h$  для достижения заданной точности решения. Метод Мерсона (Кутты–Мерсона, Рунге–Кутты–Мерсона), являющийся пятиэтапным методом порядка 4, дает оценку погрешности решения в общем случае (т. е. для нелинейных уравнений) сверху, что позволяет управлять шагом сетки решения, добиваясь точности, не хуже заданной. Формулы первого варианта метода Мерсона могут быть представлены в виде:

$$y_{n+1} = y_n + y_1,$$

где

$$y_1 = k_1 / 6 + 2k_4 / 3 + k_5 / 6,$$

$$k_1 = hf(x_n, y_n),$$

$$k_2 = hf(x_n + h/3, y_n + k_1/3),$$

$$k_3 = hf(x_n + h/3, y_n + k_1/6 + k_2/6),$$

$$k_4 = hf(x_n + h/2, y_n + k_1/8 + 3k_3/8),$$

$$k_5 = hf(x_n + h, y_n + k_1/2 - 3k_3/2 + 2k_4).$$

Кроме того, подсчитывается

$$y_2 = k_1/2 - 3k_3/2 + 2k_4,$$

после чего производится вычисление величины

$$\varepsilon' = |y_1 - y_2|,$$

которая сравнивается с заданной абсолютной погрешностью  $\varepsilon$ . Если  $\varepsilon' > \varepsilon$ , шаг интегрирования уменьшается вдвое и вычисление повторяется с предыдущей точки.

Недостатком подхода является то, что не предусмотрено увеличение шага интегрирования при получении «слишком малой» погрешности. Это иногда ведет к более медленной работе алгоритма, чем могло бы быть, исходя из заданной погрешности.

Формулы второго варианта метода Мерсона, предусматривающие увеличение длины шага интегрирования в случае «слишком малой» вычисленной погрешности, имеют вид:







$$y_{n+1} = y_n + (k_1 + 4k_4 + k_5)/2,$$

где

$$k_1 = hf(x_n, y_n)/3,$$

$$k_2 = hf(x_n + h/3, y_n + k_1)/3,$$

$$k_3 = hf(x_n + h/3, y_n + k_1/2 + k_2/2)/3,$$

$$k_4 = hf(x_n + h/2, y_n + 3k_1/8 + 9k_3/8)/3,$$

$$k_5 = hf(x_n + h, y_n + 3k_1/2 - 9k_3/2 + 6k_4)/3.$$

Шаг решения выбирается автоматически в зависимости от оценки абсолютной погрешности решения, данной формулой

$$\varepsilon' = (k_1 - 9k_3/2 + 4k_4 - k_5/2)/5.$$

Возможны три случая:

- Если  $\varepsilon' > \varepsilon$ , точность неудовлетворительная, шаг уменьшается в 2 раза, и вычисление повторяется с новым, уменьшенным, шагом.
- Если  $\varepsilon' \leq \varepsilon/32$ , точность чрезмерно высока, шаг должен быть увеличен в 2 раза, и вычисление продолжается с новым, увеличенным, шагом.
- Если  $\varepsilon/32 < \varepsilon' \leq \varepsilon$ , точность в заданных пределах, шаг выбран (настроен) верно, и вычисление продолжается с текущим «правильным» шагом.

В настоящее время популярен другой метод решения с контролем погрешности – шестиэтапный метод 4 порядка Фельберга.

### 25.2.4.1.5. Метод Хаммера–Холлингсуорта

Рассмотрим двухэтапный неявный метод 4 порядка Хаммера–Холлингсуорта (Хаммера–Холлингсуорта):

$$y_{n+1} = y_n + (k_1 + k_2)/2,$$

где

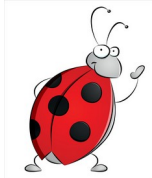
$$k_1 = hf(x_n + (1/2 - \sqrt{3}/6)h, y_n + k_1/4 + (1/4 - \sqrt{3}/6)k_2),$$

$$k_2 = hf(x_n + (1/2 + \sqrt{3}/6)h, y_n + (1/4 + \sqrt{3}/6)k_1 + k_2/4).$$

Две последних формулы определяют итерационный процесс, для нежестких уравнений сходящийся очень быстро.

Для нежестких систем представленный неявный алгоритм не эффективнее классического метода Рунге–Кутты. Однако в случае жестких систем, по данным литературы, явные одношаговые методы не работают вообще, а с помощью неявных одношаговых методов решение получить иногда удается.

Другие часто применяемые неявные методы: метод Кунцмана–Бутчера порядка 6 и 8, а также комбинированные алгоритмы Бутчера (Батчера) порядка 4 и 6.



## 25.2.4.2. Многошаговые методы

Многошаговые методы, в отличие от рассмотренных выше одношаговых методов, оперируют значениями функции в нескольких предыдущих точках. Общая формула линейного  $k$ -шагового метода имеет вид

$$y_{n+1} = \sum_{i=1}^k \alpha_i y_{n+1-i} + h \sum_{i=0}^k \beta_i f_{n+1-i},$$

где  $\alpha$  – коэффициенты перед значениями функции,  
 $\beta$  – коэффициенты перед значениями правых частей.

С помощью многошагового метода нельзя начать решения, т. к. для вычисления  $y_{n+1}$  должны быть известны все или некоторые (в зависимости от метода) величины

$$y_n, y_{n-1}, \dots, y_{n-k+1}, f_n, f_{n-1}, f_{n-k+1}.$$

Бывают явные и неявные многошаговые методы. Некоторые многошаговые методы включают в себя два шага: явный, называемый предиктором, и неявный (или несколько неявных), называемый корректором.

Многошаговые методы показали свою эффективность при решении жестких систем.

При  $\beta_0 = 0$  метод называется явным. Представлен метод Адамса.

Если  $\beta_0 \neq 0$ , метод будет называться неявным. Представлен метод Гира 4 порядка.

### 25.2.4.2.1. Метод Адамса

Формула метода Адамса 4 порядка имеет вид

$$y_{n+1} = y_n + h(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})/24.$$

В варианте предиктор–корректор данная формула дополняется выражением

$$y_{n+1} = y_n + h(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})/24.$$

### 25.2.4.2.2. Методы Гира

Для наиболее простых частных случаев рассматриваемый метод введен Кертисом и Хиршфельдером (Гиршфельдером). Общий случай рассмотрел Гир. Метод Гира 4 порядка относится к неявным многошаговым методам, использующим идею так называемого дифференцирования назад. Для рассматриваемого случая формула имеет вид

$$25/12 y_{n+1} - 4y_n + 3y_{n-1} - 4/3 y_{n-2} + 1/4 y_{n-3} = hf_{n+1}.$$

Методы Гира применяются для интегрирования жестких дифференциальных уравнений. Для нежестких систем получены также хорошие результаты.



## Глава 26. Многочлены

### 26.1. Введение

Полиномом (многочленом) называют выражение, состоящее из нескольких частей одного типа. Многочлены возникают в различных приложениях, например, при решении дифференциальных уравнений, интерполировании и т. д. Возможные приложения многочленов к статистическому анализу данных и математическому моделированию рассмотрены при описании конкретных примеров.

### 26.2. Теоретическое обоснование

Мы рассмотрим вычисление некоторых многочленов из класса многочленов Аппеля, содержащего такие важные системы многочленов, как многочлены Бернулли, Лагерра, Эрмита, а также вычисление многочленов Чебышева и Лежандра. Многочлены Чебышева первого и второго рода и многочлены Лежандра являются частными случаями многочленов Якоби, являющихся, в свою очередь, специальным случаем гипергеометрической функции. Связь многочленов Якоби и функции В-распределения обсуждается в литературе. Многие из рассмотренных многочленов обладают свойством ортогональности, которое заключается в том, что

$$\sum_{s=1}^n P_i(x_s)P_j(x_s) = 0, i \neq j.$$

Некоторые многочлены связаны с аналитическим решением специальных типов дифференциальных уравнений. Для асимптотических разложений статистических распределений применяются многочлены Эрмита и Лагерра.

#### 26.2.1. Многочлены Бернулли

Многочлены Бернулли определяются формулой

$$B_n(x) = \sum_{s=0}^n C_n^s B_s x^{n-s}, n = 0, 1, 2, \dots,$$

где  $B_s$  – числа Бернулли.

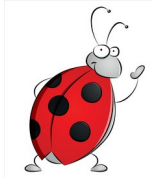
Так как все нечетные числа Бернулли, кроме числа с номером 1, равны нулю, для обозначения чисел и многочленов Бернулли применяют также обозначение  $B_{2m}(x), m = 0, 1, 2, \dots$ . Рекуррентные соотношения для определения чисел Бернулли имеют вид:

$$B_0(x) = 1,$$

$$B_n(x) = x^n - \frac{1}{n+1} \sum_{s=0}^{n-1} C_{n+1}^s B_s(x), n = 1, 2, \dots,$$

где можно было бы продолжить и составить рекурсию для числа сочетаний, чего мы пока делать не будем.





## 26.2.2. Многочлены Лагерра

Обобщенные (присоединенные) многочлены Лагерра являются решениями дифференциального уравнения

$$xy'' + (\lambda + 1 - x)y' + ny = 0$$

и определяются формулой

$$L_n^{(\lambda)}(x) = (-1)^n x^{-\lambda} e^x \frac{d^n}{dx^n} (x^{\lambda+n} e^{-x}), n = 0, 1, 2, \dots,$$

где  $\lambda > -1$ ,

$$0 \leq x < \infty.$$

Рекуррентные соотношения для вычисления обобщенных многочленов Лагерра выглядят как

$$L_0^{(\lambda)}(x) = 1,$$

$$L_1^{(\lambda)}(x) = x - \lambda - 1,$$

$$L_n^{(\lambda)}(x) = (x - \lambda - 2n + 1)L_{n-1}^{(\lambda)}(x) - (n-1)(\lambda + n - 1)L_{n-2}^{(\lambda)}(x), n = 2, 3, \dots$$

Связь многочленов Лагерра с интегралом вероятностей  $\chi^2$  и с  $\Gamma$ -распределением обсуждается в литературе.

## 26.2.3. Многочлены Эрмита

Многочлены Эрмита являются решениями дифференциального уравнения

$$y'' - 2xy' + 2ny = 0, n = 0, 1, 2, \dots,$$

и определяются формулой

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}.$$

Рекуррентные соотношения для вычисления многочленов Эрмита выглядят как

$$H_0(x) = 1,$$

$$H_1(x) = 2x,$$

$$H_n(x) = 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x), n = 2, 3, \dots$$

Рекуррентные соотношения для вычисления многочленов Эрмита можно записать по-другому. Тогда их называют многочленами Чебышева–Эрмита и вычисляют как

$$H_0(x) = 1,$$

$$H_1(x) = -x,$$

$$H_n(x) = -xH_{n-1}(x) - (n-1)H_{n-2}(x), n = 2, 3, \dots$$

В такой форме многочлены применяются для вычисления производных плотности нормального распределения по формуле

$$\varphi^{(n)}(x) = \frac{d^n \varphi(x)}{dx^n} = H_n(x) \varphi_n(x).$$

## 26.2.4. Многочлены Чебышева

Многочлены Чебышева первого рода являются решениями дифференциального уравнения





$$(1 - x^2)y'' - xy' + n^2y = 0$$

и определяются формулой

$$T_n(x) = \cos(n \arccos(x)) = \frac{2^n n!}{(2n)!} \sqrt{1-x^2} \frac{d^n}{dx^n} \left[ (1-x^2)^{n-1/2} \right], n = 0, 1, 2, \dots,$$

где  $-1 \leq x \leq 1$ .

Рекуррентные соотношения выглядят гораздо проще

$$T_0(x) = 1,$$

$$T_1(x) = x,$$

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), n = 2, 3, \dots$$

Многочлены Чебышева второго рода являются решениями дифференциального уравнения

$$(1 - x^2)y'' - 3xy' + n(n+2)y = 0$$

и определяются формулой

$$U_n(x) = \frac{\sin[(n+1) \arccos(x)]}{\sqrt{1-x^2}} = \frac{2^n (n+1)!}{(2n+1)!} \frac{1}{\sqrt{1-x^2}} \frac{d^n}{dx^n} \left[ (1-x^2)^{n+1/2} \right], n = 0, 1, 2, \dots,$$

где  $-1 \leq x \leq 1$ .

Рекуррентные соотношения выглядят как

$$U_0(x) = 1,$$

$$U_1(x) = 2x,$$

$$U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x), n = 2, 3, \dots$$

Многочлены Чебышева образуют систему, ортогональную на отрезке  $-1 \leq x \leq 1$ . В литературе дается формула связи многочленов Чебышева первого и второго рода.

## 26.2.5. Многочлены Лежандра

Рассмотрим дифференциальное уравнение

$$(1 - x^2)y'' - 2xy' + \left[ n(n+1) - \frac{\mu^2}{1-x^2} \right] y = 0,$$

где  $n$  и  $\mu$  – произвольные числа.

Если  $n = 0, 1, 2, \dots$ , а  $\mu = 0$ , то ограниченные на отрезке  $-1 \leq x \leq 1$  решения уравнения называются многочленами Лежандра (сферическими многочленами)

$$P_n(x) = \frac{1}{n! 2^n} \frac{d^n}{dx^n} (x^2 - 1)^n$$

и определяются по рекуррентным формулам

$$P_0(x) = 1,$$

$$P_1(x) = x,$$

$$P_n(x) = [(1-2n)xP_{n-1}(x) + (n-1)P_{n-2}(x)] / n, n = 2, 3, \dots$$

Благодаря тому, что многочлены Лежандра ортогональны на отрезке  $-1 \leq x \leq 1$ , они образуют полную систему функций и могут быть использованы для разложения в ряд произвольной функции, интегрируемой на отрезке  $-1 \leq x \leq 1$ .



## Приложение. Статистические распределения

В вычислительном аспекте наименование «функция распределения» употребляется для математического объекта, в который подставляются статистика, а также некоторый набор параметров (в том числе, возможно, так называемые степени свободы), а в результате получается значение, имеющее смысл вероятности и, следовательно, заключенное в интервале  $[0;1]$ . И наоборот, «обратная функция распределения» – это такая функция, в которую подставляется параметр, имеющий смысл вероятности, а также, возможно, некоторый набор дополнительных параметров, а в результате получается значение статистики.

Источники: Большев с соавт., Брандт, Де Гроот, Попов с соавт., Родионов, Родионов с соавт., Хан с соавт., Хастингс с соавт., Шор с соавт., Бьюри (Bury), Эванс (Evans) с соавт. Сводку распределений и аппроксимаций дал Кобзарь.

### П.1. Биномиальное распределение

Функция биномиального распределения вычисляется по формуле

$$P(k < K) = \sum_{k=0}^{K-1} W_k^n,$$

где  $n$  – число степеней свободы,  $n > 0$ ,

$W_k^n$  – вероятности биномиального распределения, вычисляемые по формуле

$$W_k^n = C_n^k p^k (1-p)^{n-k},$$

где  $C_n^k$  – число сочетаний из  $n$  по  $k$ .

Для обеспечения численной устойчивости алгоритма число сочетаний может вычисляться как (Брандт)

$$C_n^k = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)},$$

где  $\Gamma(\cdot)$  – гамма-функция.

### П.2. Гипергеометрическое распределение

Функция гипергеометрического распределения вычисляется по формуле

$$P(k < k') = \sum_{k=0}^{k'-1} W_k,$$

где  $W_k$  – вероятности гипергеометрического распределения, вычисляемые по формуле

$$W_k = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}, n \leq N, k \leq K,$$

где  $C_K^k$  – число сочетаний из  $K$  по  $k$ ,





$C_{N-K}^{n-k}$  – число сочетаний из  $N - K$  по  $n - k$ ,

$C_N^n$  – число сочетаний из  $N$  по  $n$ ,

$K, N, n$  – параметры распределения.

## П.3. Нормальное распределение

Нормальным называется одно из важнейших распределений вероятностей случайной величины. Теоретическое обоснование роли нормального распределения дается центральными предельными теоремами, рассматриваемыми в курсе «Теории вероятностей».

Функция плотности нормального распределения имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}},$$

$$-\infty < a < \infty, \sigma > 0, -\infty < x < \infty.$$

Путем введения нормированной величины

$$t = \frac{x - a}{\sigma},$$

где  $a$  – математическое ожидание (обычно его оценка – среднее значение, но могут применяться и другие параметры положения),

$\sigma^2$  – дисперсия (параметр разброса),

показанной выше формуле придан несколько иной вид. Этой формулой удобно пользоваться при расчете теоретических частот эмпирического распределения. К тому же таблицы обычно даются для функции, называемой также плотностью вероятности стандартизованной (стандартной) нормальной случайной величины,

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Функция стандартного нормального распределения равна

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy$$

и называется функцией Лапласа (вторым законом распределения Лапласа) либо интегралом вероятности Гаусса (законом Гаусса, гауссовым распределением) в честь применения данного закона распределения для изучения ошибок наблюдений.

Практически вычисление функции стандартного нормального распределения производится по формуле

$$\Phi(x) = \frac{1}{2} \left[ 1 + \operatorname{sign}(x) P_{x^2/2} \left( \frac{1}{2} \right) \right],$$

где  $P(.)$  – неполная гамма-функция.

Находит применение интеграл вероятностей





$$I(x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^x e^{-y^2/2} dy.$$

С использованием свойства симметрии подынтегральной функции стандартного нормального распределения  $\Phi(\cdot)$  расчетная формула интеграла вероятностей  $I(\cdot)$  сводится к простому выражению

$$I(x) = 2\Phi(|x|) - 1).$$

## П.4. Многомерное нормальное распределение

В случае многомерного нормального распределения плотность распределения совокупности определяется формулой

$$P(X) = \frac{1}{(2\pi)^{d/2} |S|^{1/2}} e^{-\frac{1}{2}(X - \bar{X})' S^{-1} (X - \bar{X})},$$

где  $S$  – дисперсионно–ковариационная матрица,

$\bar{X}$  – вектор математического ожидания,

$d$  – «число измерений» – порядок матрицы  $S$  и длина вектора  $\bar{X}$ ,

' – операция транспонирования.

Дисперсионно–ковариационная матрица в случае многомерного распределения является параметром, аналогичным дисперсии в одномерном случае. На диагонали данной матрицы располагаются дисперсии компонент случайного вектора. Внедиагональные члены матрицы являются ковариациями.

Иногда нормальное многомерное распределение ошибочно понимается в том смысле, что каждая переменная, составляющая многомерную совокупность (реализацию случайного многомерного вектора), имеет нормальное распределение. Это неверно: исследуя такое распределение «одномерных составляющих», анализируют только маргинальные распределения компонент случайного многомерного вектора, составляющих многомерное распределение, но не само многомерное распределение. Для исследования нормальности многомерного распределения разработаны специальные методы.

О многомерном нормальном распределении см. статью Мартынова.

## П.5. t–распределение

Функция  $t$ –распределения Стьюдента выражается формулой

$$F_n(x) = \frac{1}{\sqrt{n} B(1/2, n/2)} \int_{-\infty}^x \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2} dy,$$

где  $n$  – число степеней свободы,  $n > 0$ ,

$B(\cdot)$  – бета–функция.

Практически вычисление функции производится по формуле







$$F_n(x) = \frac{1}{2} \left[ 1 + \operatorname{sign}(x) \left[ 1 - I_{n/(n+x^2)}(n/2, 1/2) \right] \right],$$

где  $I(.,.)$  – регуляризованная неполная бета-функция.

## П.6. F-распределение

Функция F-распределения выражается формулой

$$F_x(n_1, n_2) = \frac{\Gamma((n_1 + n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left( \frac{n_1}{n_2} \right)^{n_1/2} \int_0^x y^{n_1-1} \left( 1 + \frac{n_1}{n_2} y \right)^{-(n_1+n_2)/2} dy,$$

где  $n_1$  – число степеней свободы,  $n_1 > 0$ ,

$n_2$  – число степеней свободы,  $n_2 > 0$ ,

$\Gamma(.)$  – гамма-функция.

Практически вычисление функции производится по формуле

$$F_x(n_1, n_2) = 1 - I_{n_2/(n_2+n_1x)}(n_2/2, n_1/2),$$

где  $I(.,.)$  – регуляризованная неполная бета-функция.

## П.7. Бета-распределение

Функция бета-распределения – эквивалентное наименование регуляризованной неполной бета-функции.

## П.8. Хи-квадрат распределение

Функция распределения  $\chi^2$  выражается формулой

$$F_n(x) = \frac{1}{2^{n/2} \Gamma(n/2)} \int_0^x y^{n/2-1} e^{-y/2} dy,$$

где  $n$  – число степеней свободы,  $n > 0$ ,

$\Gamma(.)$  – гамма-функция.

Практически вычисление функции производится по формуле

$$F_n(x) = 1 - P_{x/2}(n/2),$$

где  $P(.)$  – неполная гамма-функция.

## П.9. Нецентральное хи-квадрат распределение

Функция нецентрального распределения  $\chi^2$  выражается формулой

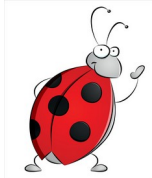
$$F'_n(x, \lambda) = e^{-\lambda/2} \sum_{k=0}^{\infty} \frac{\lambda^k}{k! 2^{n/2+2k} \Gamma(n/2+k)} \int_u^{\infty} y^{n/2+k-1} e^{-y/2} dy,$$

где  $n$  – число степеней свободы,  $n > 0$ ,

$\lambda$  – параметр нецентральности,  $\lambda \geq 0$ ,

$u$  – обратная функция распределения  $\chi^2$ ,





$\Gamma(\cdot)$  – гамма–функция.

При  $\lambda = 0$  нецентральное распределение  $\chi^2$  совпадает с распределением  $\chi^2$ .

Практически вычисление функции производится посредством аппроксимации, предложенной Пирсоном,

$$F'_n(x, a) = F_{n'}(x) \frac{n + 3\lambda}{n + 2\lambda} - \frac{\lambda^2}{n + 3\lambda},$$

где  $F_n(x)$  – функция распределения  $\chi^2$  с числом степеней свободы, равным

$$n' = \frac{(n + 2\lambda)^3}{(n + 3\lambda)^2}.$$

Свойства, аппроксимации и приложения распределения изучены Большевым с соавт., Оуэном, Кобзарем, Кульбаком. Один из частных случаев рассмотрен Фишером (Fisher).

## П.10. Обобщенное гамма–распределение

Функция гамма–распределения может иметь один, два или три параметра. Гамма–функция с тремя параметрами, называемая обобщенной гамма–функцией, вычисляется по формуле

$$F_x(a, b, c) = \frac{1}{b^a \Gamma(a)} \int_0^x (t - c)^{a-1} e^{-(t-c)/b} dt.$$

Практически вычисление функции производится по формуле

$$F_x(a, b, c) = P_{(x-c)/b}(a),$$

где  $P(\cdot)$  – неполная гамма–функция.

## П.11. Логнормальное распределение

Функция логнормального (логарифмически нормального) распределения с двумя параметрами вычисляется по формуле

$$P_x(a, b) = \frac{1}{b\sqrt{2\pi}} \int_0^x y^{-1} e^{-(\ln y - a)^2 / 2b^2} dy.$$

Заменой переменной  $\ln y = t$ ,  $y^{-1} dy = dt$  и, соответственно, меняя пределы интегрирования  $y \in [0; x]$  на  $t \in [-\infty; \ln x]$ , получаем формулу, которая пригодится в дальнейших выкладках,

$$P_x(a, b) = \frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{\ln x} e^{-(t-a)^2 / 2b^2} dt.$$

Функция нормального распределения от нестандартизованной случайной величины равна

$$F_x(a, b) = \Phi((x - a) / b),$$

где  $\Phi(\cdot)$  – функция стандартного нормального распределения.

Расчетная формула путем преобразований примет вид

$$P_x(a, b) = F_{\ln x}(a, b) = \Phi((\ln x - a) / b).$$

Рассмотренное распределение является частным случаем логнормального распределения с тремя параметрами, называемого также распределением  $S_L$  Джонсона.





Плотность логнормального распределения с двумя параметрами вычисляется по формуле

$$f(x, a, b) = \frac{1}{xb\sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{\ln(x) - a}{b} \right)^2}, x > 0, b > 0, -\infty < a < \infty.$$

## П.12. Распределение $S_U$ Джонсона

Функция распределения  $S_U$  Джонсона вычисляется по формуле

$$P_x(a, b, c, d) = \frac{b}{d\sqrt{2\pi}} \int_{-\infty}^x \frac{1}{\sqrt{((y-c)/d)^2 + 1}} e^{-\frac{1}{2} \left\{ a + b \ln \left[ (y-c)/d + \sqrt{((y-c)/d)^2 + 1} \right] \right\}^2} dy$$

Руководствуясь материалами монографии Хана с соавт. (с. 233) по распределениям Джонсона, устанавливаем, что расчетная формула будет иметь вид

$$P_x(a, b, c, d) = 1 - \Phi(a + b \cdot \text{Arsh}(x - c) / d),$$

где  $\Phi(\cdot)$  – функция стандартного нормального распределения,

$\text{Arsh}(\cdot)$  – функция гиперболического арксинуса.

## П.13. Распределение выборочного размаха

Функция распределения  $P_n(W \leq w)$  выборочного размаха (range)  $W$  для выборки численности  $n$ , иначе вероятность того, что он не превысит  $w$ , определяется формулой

$$P_n(W \leq w) = n \int_{-\infty}^{\infty} [F(x + w) - F(x)]^{n-1} dF(x).$$

Если совокупность распределена нормально, то выражение  $F(\cdot)$ , входящее в формулу, представляет собой функцию распределения нормально распределенной стандартизованной случайной величины

$$P(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$$

Для вычисления функции распределения размаха дополнительно необходимо выразить величину  $dF(x)$ , входящую в формулу ее вычисления, через  $dx$ . Можно записать

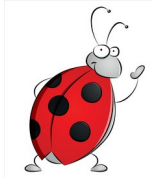
$$dF(x) = \frac{dF(x)}{dx} dx,$$

где  $\frac{dF(x)}{dx}$  – производная  $F(x)$  по  $x$  – плотность распределения вероятности – для стандартизованной нормальной случайной величины вычисляется по формуле

$$\frac{dP(x)}{dx} = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Сделав все необходимые подстановки, получаем пригодную для практических вычислений формулу





$$P_n(W \leq w) = \frac{n}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [P(x+w) - P(x)]^{n-1} e^{-x^2/2} dx.$$

Расчет производится численным интегрированием методом Симпсона. Метод см. у Лоренсель (Laurencelle) с соавт. См. Мюллер с соавт., Хальд, Оуэн, Барндорф–Нильсен с соавт.

## П.14. Распределение стьюдентизированного размаха

Пусть из нормальной совокупности извлекается выборка численностью  $n$  и по данной выборке вычисляется выборочный размах  $W$ . Затем из той же нормальной совокупности или из другой нормальной совокупности с тем же стандартным отклонением извлекается выборка численностью  $f$  и по данной выборке вычисляется выборочное стандартное отклонение  $s$ .

Тогда отношение  $W/s$  называется стьюдентизированным размахом (размахом Стьюдента, studentized range) и его распределение зависит только от величин  $n$  и  $f$ . Функция распределения  $P_{n,f}(W/s \leq q)$  выборочного стьюдентизированного размаха, иначе вероятность того, что он не превысит  $q$ , определяется формулой

$$P_{n,f}(W/s \leq q) = \frac{f^{f/2}}{\Gamma(f/2) \cdot 2^{f/2-1}} \int_0^{\infty} x^{f-1} e^{-fx^2/2} P_n(qx) dx,$$

где  $P_n(\cdot)$  – функция распределения выборочного размаха.

Вычисление рассматриваемого распределения производится численным интегрированием методом Симпсона. Метод см. у Лоренсель (Laurencelle) с соавт. См. также источники: Оуэн, Ликеш с соавт., Мюллер с соавт., Хальд, Дэвид, Дэвид (David) с соавт., Хартер (Harter) с соавт., Шеффе. Аппроксимации рассмотрены Копенгауэр (Copenhaver) с соавт., Глизон (Gleason), Рамсей (Ramsey) с соавт., Карри (Currie), Пирсон (Pearson), Типпетт (Tippett). Методику вычислений см. также в статье Копенховер (Copenhaver) с соавт., Баум (Baum) с соавт.

## П.15. Распределение стьюдентизированного максимума модулей

Функция распределения  $P(Q_{k,n} \leq q)$  стьюдентизированного максимума модулей (studentized maximum modulus)  $Q_{k,n}$  с параметром  $k$  и числом степеней свободы  $n$ , иначе вероятность того, что он не превысит  $q$ , определяется формулой

$$P(Q_{k,n} \leq q) = \int_0^{\infty} [2\Phi(qx) - 1]^k d\mu_n(x),$$

где  $\Phi(\cdot)$  – функция стандартного нормального распределения,

$\mu(\cdot)$  – деленная на  $\sqrt{n}$  плотность функции  $\chi$ -распределения.

Плотность  $\chi$ -распределения имеет вид





$$f(x, n) = \frac{x^{n-1} e^{-x^2/2}}{2^{n/2-1} \Gamma(n/2)}.$$

Выполнив необходимые преобразования, по смыслу аналогичные тем, что произведены при вычислении выборочного размаха, получим, что дифференциал  $d\mu_n(x)$  определяется формулой

$$d\mu_n(x) = \frac{n^{n/2} x^{n-1} e^{-nx^2/2}}{2^{n/2-1} \Gamma(n/2)} dx.$$

Тогда искомым вид функции распределения

$$P(Q_{k,n} \leq q) = \frac{n^{n/2}}{2^{n/2-1} \Gamma(n/2)} \int_0^q [2\Phi(qx) - 1]^k x^{n-1} e^{-nx^2/2} dx.$$

Вычисление рассматриваемого распределения производится численным интегрированием методом Симпсона. Метод см. у Лоренсель (Laurencelle) с соавт. Способ интегрирования методом разложения в ряд см. в статьях Пиллаи (Pillai) и Пиллаи с соавт. Таблицы и аппроксимацию см. в статье Юри (Ury) с соавт. Как указывают Сахаи (Sahai) с соавт., функция распределения студентизированного максимума модулей может быть получена также как корень квадратный из функции распределения студентизированного максимума хи-квадрат (studentized maximum chi-square) – см. Армитэдж (Armitage) с соавт. О студентизированном максимуме и минимуме хи-квадрат (studentized minimum chi-square) см. монографию Гупта (Gupta) с соавт., о студентизированном минимуме хи-квадрат см. статью Алан (Alam). См. также Бечхофер (Bechhofer) с соавт., Столайн (Stoline) с соавт.

## П.16. Распределение статистики критерия Колмогорова

Распределение статистики критерия Колмогорова ( $\lambda$ -распределение) вычисляется по точной формуле

$$K(x) = \begin{cases} \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 x^2}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

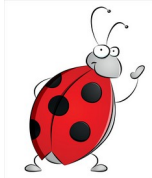
Бесконечная последовательность быстро сходится, и для получения приемлемой для практических вычислений точности критического значения достаточно небольшого числа ее членов.

## П.17. Распределение статистики критерия Койпера

Распределение статистики критерия Койпера вычисляется по точной формуле

$$Q(x) = \sum_{i=1}^{\infty} (4i^2 x^2 - 1) e^{-2i^2 x^2}.$$





Бесконечная последовательность быстро сходится, и для получения приемлемой для практических вычислений точности критического значения достаточно небольшого числа ее членов.

## П.18. Распределения статистик критериев Вилкоксона

Различают распределения статистик критерия Вилкоксона для независимых выборок и критерия Вилкоксона для связанных выборок.

Для независимых выборок рекуррентные формулы вычисления критических значений критерия Вилкоксона суть

$$f(n_1, n_2, W) = f(n_1, n_2 - 1, W - n_1) + f(n_1 - 1, n_2, W),$$

$$f(n_1, n_2, -x) = 0, x > 0,$$

$$f(n_1, n_2, 0) = 1,$$

$$f(n_1, 0, W) = 0,$$

где  $n_1$  – численность одной выборки,

$n_2$  – численность другой выборки.

$P$ –значение вычисляется как

$$P = \frac{n_2}{n_1 + n_2} f(n_1, n_2 - 1, W - n_1) + \frac{n_1}{n_1 + n_2} f(n_1 - 1, n_2, W).$$

Для связанных выборок рекуррентные формулы вычисления критических значений критерия Вилкоксона суть

$$f(N, W^+) = f(N - 1, W^+) + f(N - 1, W^+ - N),$$

$$f(N, 0) = 1,$$

$$f(N, -x) = 0, x > 0,$$

$$f(N, W^+) = f(N, N(N + 1) / 2), W^+ \geq N(N + 1) / 2,$$

где  $N$  – численность каждой выборки.

$P$ –значение вычисляется как

$$P = \frac{f(N, W^+)}{2^N}.$$

См. Оуэна.

## П.19. Распределение статистики критерия Манна–Уитни

Рекуррентные формулы вычисления критических значений статистики критерия Манна–Уитни суть

$$f(n_1, n_2, U) = f(n_1 - 1, n_2, U - n_1) + f(n_1, n_2 - 1, U),$$

$$f(n_1, n_2, -x) = 0, x > 0,$$

$$f(n_1, n_2, 0) = 1,$$

$$f(n_1, 0, U) = 1,$$

$$f(n_1, n_2, U) = f(n_2, n_1, U),$$

где  $n_1$  – численность одной выборки,





$n_2$  – численность другой выборки.

$P$ –значение вычисляется как

$$P = \frac{n_1!n_2!}{(n_1 + n_2)!} f(n_1, n_2, U).$$

См. работы Ван де Вилия (Van de Wiel) с соавт., Ди Буччианико (Di Bucchianico) с соавт.

## П.20. Распределение статистики критериев типа омега–квадрат

Предельная функция распределения  $a_1$  критериев типа омега–квадрат вычисляется как

$$a_1(x) = \frac{1}{\sqrt{2x}} \sum_{j=0}^{\infty} \frac{\Gamma(j+1/2)}{\Gamma(1/2)\Gamma(j+1)} \sqrt{4j+1} \cdot \exp\left(-\frac{(4j+1)^2}{16x}\right) \cdot \left\{ I_{-1/4}\left(\frac{(4j+1)^2}{16x}\right) - I_{1/4}\left(\frac{(4j+1)^2}{16x}\right) \right\},$$

где  $I(.)$  – модифицированная функция Бесселя.

См. Большева с соавт.

## П.21. Маргинальные распределения

Маргинальным (частным) распределением называют проекцию многомерного распределения на подпространство, порожденное некоторым набором координатных векторов. Пусть  $F(x_1, x_2, \dots, x_n)$  – функция распределения случайного  $n$ –мерного вектора  $(X_1, X_2, \dots, X_n)$ . Функция распределения  $(X_{i_1}, X_{i_2}, \dots, X_{i_m}), 1 \leq i_1 < i_2 < \dots < i_m \leq n, m < n$ , называется маргинальной функцией распределения по отношению к  $F(.)$ , а соответствующее распределение – маргинальным.

См. энциклопедию «Вероятность и математическая статистика» (с. 299).

## П.22. Специальные функции

Гамма–функция Эйлера определяется формулой

$$\Gamma(a) = \int_0^{\infty} y^{a-1} e^{-y} dy.$$

Неполная гамма–функция (с одним параметром) определяется формулой

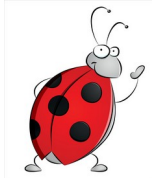
$$P_x(a) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt.$$

Бета–функция определяется формулой

$$B(a, b) = \int_0^1 y^{a-1} (1-y)^{b-1} dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Неполная бета–функция определяется формулой





$$B_x(a, b) = \int_0^x y^{a-1} (1-y)^{b-1} dy.$$

Регуляризованная неполная бета-функция (иногда для краткости именуемая просто регуляризованной бета-функцией) определяется формулой

$$I_x(a, b) = \frac{B_x(a, b)}{B(a, b)},$$

причем для целых значений аргументов имеет место простая формула

$$I_x(a, b) = \sum_{j=a}^{a+b-1} \frac{(a+b-1)!}{j!(a+b-1-j)!} x^j (1-x)^{a+b-1-j}.$$

Модифицированная функция Бесселя 1 рода вычисляется по формуле

$$I_\nu(z) = \sum_{k=0}^{\infty} \frac{(z/2)^{2k+\nu}}{k! \Gamma(k+\nu+1)}.$$

Вычисления с гарантированной точностью производят при помощи разложения в ряд и взятием конечного числа его членов, либо при помощи непрерывных усеченных, при достижении заданной точности, цепных дробей, либо при помощи аппроксимаций. Некоторые специальные случаи рекурсивного вычисления рассмотрены Де Гроотом. См. Абрамовиц с соавт., Брандт, Де Гроот.