

O'REILLY®

2-е издание

# Практическая статистика для специалистов Data Science

50+ важнейших понятий с использованием R и Python



Питер Брюс  
Эндрю Брюс  
Питер Гедек

---

# Оглавление

<b>Об авторах.....</b>	<b>13</b>
<b>Предисловие .....</b>	<b>15</b>
Условные обозначения, принятые в книге .....	15
Использование примеров кода .....	16
Благодарности.....	16
Комментарии переводчика .....	17
<b>Глава 1. Разведывательный анализ данных.....</b>	<b>19</b>
Элементы структурированных данных .....	20
Дополнительные материалы для чтения .....	22
Прямоугольные данные .....	23
Кадры данных и индексы.....	24
Непрямоугольные структуры данных.....	25
Дополнительные материалы для чтения.....	26
Оценки центрального положения .....	26
Среднее .....	27
Медиана и робастные оценки .....	29
Выбросы .....	29
Пример: средние оценки численности населения и уровня убийств.....	30
Дополнительные материалы для чтения.....	32
Оценки вариабельности .....	32
Стандартное отклонение и связанные с ним оценки.....	33
Оценки на основе процентилей.....	35
Пример: оценки вариабельности населения штатов.....	36
Дополнительные материалы для чтения.....	37
Разведывание распределения данных .....	38
Процентили и коробчатые диаграммы .....	38
Частотные таблицы и гистограммы .....	40
Графики и оценки плотности.....	42
Дополнительные материалы для чтения.....	44
Разведывание двоичных и категориальных данных.....	44
Мода.....	46
Ожидаемое значение .....	47
Вероятность.....	47
Дополнительные материалы для чтения.....	48
Корреляция.....	48
Диаграммы рассеяния .....	52
Дополнительные материалы для чтения.....	53

Разведывание двух или более переменных .....	53
Сетка из шестиугольных корзин и контуры (сопоставление числовых данных с числовыми данными на графике) .....	54
Две категориальные переменные .....	57
Категориальные и числовые данные .....	58
Визуализация многочисленных переменных .....	60
Дополнительные материалы для чтения .....	62
Резюме .....	63
<b>Глава 2. Распределение данных и распределение выборок .....</b>	<b>65</b>
Случайный отбор и смещенная выборка .....	66
Смещение .....	68
Случайный отбор .....	69
Размер против качества: когда размер имеет значение? .....	70
Выборочное среднее против популяционного среднего .....	71
Дополнительные материалы для чтения .....	71
Систематическая ошибка отбора .....	72
Регрессия к среднему .....	73
Дополнительные материалы для чтения .....	75
Выборочное распределение статистической величины .....	75
Центральная предельная теорема .....	78
Стандартная ошибка .....	79
Дополнительные материалы для чтения .....	80
Бутстрап .....	80
Повторный отбор против бутстрапирования .....	84
Дополнительные материалы для чтения .....	84
Доверительные интервалы .....	84
Дополнительные материалы для чтения .....	87
Нормальное распределение .....	87
Стандартное нормальное распределение и квантиль-квантильные графики .....	89
Длиннохвостые распределения .....	91
Дополнительные материалы для чтения .....	93
<i>t</i> -Распределение Стьюдента .....	93
Дополнительные материалы для чтения .....	95
Биномиальное распределение .....	95
Дополнительные материалы для чтения .....	98
Распределение хи-квадрат .....	98
Дополнительные материалы для чтения .....	99
<i>F</i> -распределение .....	99
Дополнительные материалы для чтения .....	100
Распределение Пуассона и другие связанные с ним распределения .....	100
Пуассоновские распределения .....	101
Экспоненциальное распределение .....	101
Оценивание интенсивности отказов .....	102
Распределение Вейбулла .....	102
Дополнительные материалы для чтения .....	103
Резюме .....	104
<b>Глава 3. Статистические эксперименты и проверка значимости .....</b>	<b>105</b>
<i>A/B</i> -тестирование .....	105
Зачем нужна контрольная группа? .....	108

Почему только $A/B$ ? Почему не $C, D...?$ .....	109
Дополнительные материалы для чтения.....	110
Проверки гипотез .....	110
Нулевая гипотеза .....	112
Альтернативная гипотеза.....	112
Односторонняя проверка гипотезы против двухсторонней .....	113
Дополнительные материалы для чтения.....	114
Повторный отбор.....	114
Перестановочный тест.....	115
Пример: прилипчивость веб-страниц .....	115
Исчерпывающий и бутстраповский перестановочные тесты.....	119
Перестановочные тесты: сухой остаток для науки о данных .....	119
Дополнительные материалы для чтения.....	120
Статистическая значимость и $p$ -значения .....	120
$p$ -Значение .....	123
Альфа.....	124
Разногласия по поводу $p$ -значения.....	124
Практическая значимость .....	125
Ошибки 1-го и 2-го рода .....	125
Наука о данных и $p$ -значения .....	126
Дополнительные материалы для чтения.....	126
Проверки на основе $t$ -статистики.....	127
Дополнительные материалы для чтения.....	129
Множественное тестирование .....	129
Дополнительные материалы для чтения.....	132
Степени свободы .....	133
Дополнительные материалы для чтения.....	134
Дисперсионный анализ .....	134
$F$ -статистика.....	138
Двухсторонний дисперсионный анализ.....	139
Дополнительные материалы для чтения.....	140
Проверка на основе статистики хи-квадрат .....	140
Проверка хи-квадрат: подход на основе повторного отбора.....	141
Проверка хи-квадрат: статистическая теория .....	143
Точный тест Фишера.....	144
Релевантность для науки о данных .....	146
Дополнительные материалы для чтения.....	147
Алгоритм многоорукого бандита.....	147
Дополнительные материалы для чтения.....	150
Мощность и размер выборки.....	151
Размер выборки.....	152
Дополнительные материалы для чтения.....	155
Резюме .....	155
<b>Глава 4. Регрессия и предсказание .....</b>	<b>157</b>
Простая линейная регрессия.....	157
Уравнение регрессии.....	158
Подогнанные значения и остатки.....	161
Наименьшие квадраты .....	162
Предсказание против объяснения (профилирование).....	163
Дополнительные материалы для чтения.....	164

Множественная линейная регрессия .....	164
Пример: данные жилого фонда округа Кинг.....	165
Оценивание результативности модели .....	167
Перекрестный контроль .....	169
Отбор модели и пошаговая регрессия .....	170
Взвешенная регрессия .....	173
Дополнительные материалы для чтения.....	175
Предсказание с использованием регрессии .....	175
Опасности экстраполяции.....	175
Доверительный и предсказательный интервалы .....	176
Факторные переменные в регрессии.....	178
Представление фиктивных переменных.....	178
Факторные переменные с многочисленными уровнями.....	181
Упорядоченные факторные переменные .....	183
Интерпретирование уравнения регрессии.....	184
Коррелированные предсказатели .....	185
Мультиколлинеарность .....	186
Искажающие переменные .....	187
Взаимодействия и главные эффекты .....	188
Диагностика регрессии .....	190
Выбросы .....	191
Влиятельные значения .....	193
Гетероскедастичность, ненормальность и коррелированные ошибки .....	196
Графики частных остатков и нелинейность .....	199
Многочленная и сплайновая регрессия .....	201
Многочлены .....	202
Сплайны.....	203
Обобщенные аддитивные модели .....	206
Дополнительные материалы для чтения .....	207
Резюме .....	208
<b>Глава 5. Классификация .....</b>	<b>209</b>
Наивный Байес.....	210
Почему точная байесова классификация непрактична?.....	211
Наивное решение .....	211
Числовые предсказательные переменные .....	214
Дополнительные материалы для чтения.....	215
Дискриминантный анализ.....	215
Матрица ковариаций .....	216
Линейный дискриминант Фишера .....	217
Простой пример .....	217
Дополнительные материалы для чтения.....	221
Логистическая регрессия .....	221
Функция логистического отклика и логит .....	222
Логистическая регрессия и ОЛМ .....	223
Обобщенные линейные модели.....	225
Предсказанные значения из логистической регрессии .....	225
Интерпретирование коэффициентов и отношений перевесов.....	226
Линейная и логистическая регрессия: сходства и различия .....	228
Подгонка модели .....	228

Оценивание результативности модели .....	229
Анализ остатков .....	231
Дополнительные материалы для чтения .....	232
Оценивание классификационных моделей .....	233
Матрица путаницы .....	234
Проблема редкого класса .....	236
Прецизионность, полнота и специфичность .....	236
ROC-кривая .....	238
Площадь под ROC-кривой .....	240
Лифт .....	241
Дополнительные материалы для чтения .....	243
Стратегии для несбалансированных данных .....	243
Понижающий отбор .....	244
Повышающий отбор и повышающая/понижающая перевесовка .....	245
Генерация данных .....	246
Стоимостная классификация .....	247
Разведывание предсказаний .....	247
Дополнительные материалы для чтения .....	249
Резюме .....	249
<b>Глава 6. Статистическое машинное обучение .....</b>	<b>251</b>
<i>k</i> ближайших соседей .....	252
Небольшой пример: предсказание невыплаты ссуды .....	253
Метрики расстояния .....	255
Кодировщик с одним активным состоянием .....	256
Стандартизация (нормализация, <i>z</i> -оценки) .....	257
Выбор числа <i>k</i> .....	260
<i>k</i> ближайших соседей как механизм порождения признаков .....	261
Древесные модели .....	263
Простой пример .....	264
Алгоритм рекурсивного подразделения .....	267
Измерение однородности или загрязненности .....	268
Остановка выращивания дерева .....	270
Контроль за сложностью дерева в R .....	270
Контроль за сложностью дерева в Python .....	271
Предсказывание непрерывного значения .....	271
Каким образом используются деревья .....	272
Дополнительные материалы для чтения .....	273
Бэггинг и случайный лес .....	273
Бэггинг .....	274
Случайный лес .....	275
Важность переменных .....	279
Гиперпараметры .....	282
Бустинг .....	283
Алгоритм бустирования .....	285
XGBoost .....	286
Регуляризация: предотвращение перепогонки .....	288
Гиперпараметры и перекрестный контроль .....	292
Резюме .....	296

<b>Глава 7. Неконтролируемое самообучение.....</b>	<b>297</b>
Анализ главных компонент .....	298
Простой пример .....	299
Вычисление главных компонент.....	301
Интерпретирование главных компонент .....	302
Анализ соответствия .....	305
Дополнительные материалы для чтения.....	307
Кластеризация на основе $K$ средних .....	307
Простой пример .....	308
Алгоритм $K$ средних.....	310
Интерпретирование кластеров .....	311
Выбор числа кластеров .....	313
Иерархическая кластеризация.....	315
Простой пример .....	316
Дендограмма .....	317
Агломеративный алгоритм .....	318
Меры несхожести .....	319
Модельно-ориентированная кластеризация.....	321
Многомерное нормальное распределение.....	321
Смеси нормальных распределений .....	322
Выбор числа кластеров .....	325
Дополнительные материалы для чтения.....	327
Шкалирование и категориальные переменные.....	328
Шкалирование переменных.....	328
Доминантные переменные.....	330
Категориальные данные и расстояние Говера .....	332
Проблемы кластеризации смешанных данных .....	334
Резюме .....	336
<b>Библиография .....</b>	<b>337</b>
<b>Предметный указатель.....</b>	<b>339</b>