Введение в математическую статистику II

Леонид Иосипой

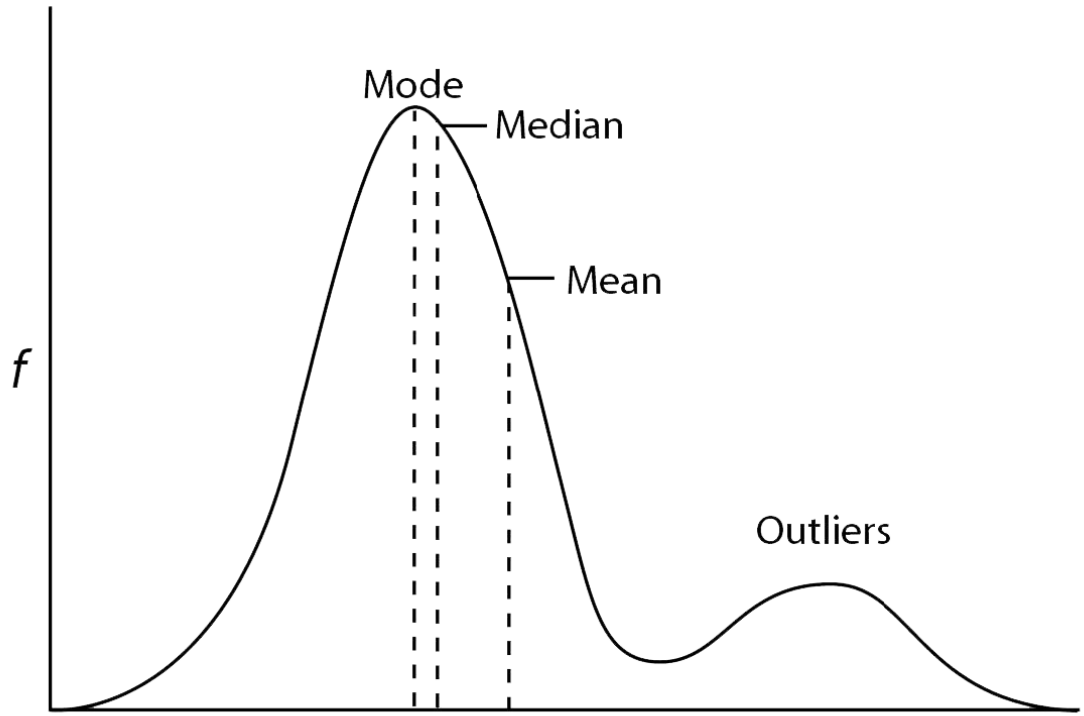
Программа «Математика для анализа данных»
Центр непрерывного образования, ВШЭ

31 Октября 2018

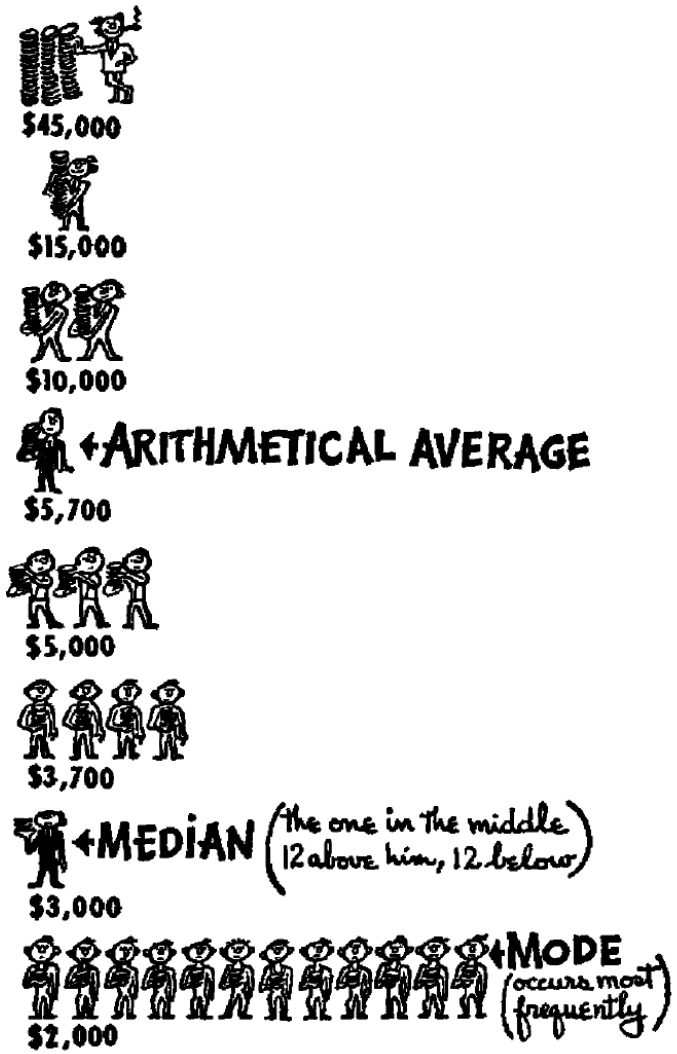
- Оценка математического ожидания распределения
- Проверка гипотез
- Линейная регрессия
- Спасибо за внимание

Оценка математического ожидания

Выборочное среднее — среднее арифметическое по выборке.
Выборочная мода — самое распространённое значение.
Выборочная медиана — центральный элемент вариационного ряда.



Оценка математического ожидания



Проверка гипотез

- Делается предположение о процессе, генерирующем данные, и задача состоит в том, чтобы определить, содержат ли данные достаточно информации, чтобы отвергнуть это предположение.

- Если информации не достаточно, то говорится, что опытные данные предположению (гипотезе) не противоречат.

Проверка гипотез

Пример

Пусть $X_1, \ldots, X_n \sim \text{Ber}(p)$.
$H_0: \ p = \frac{1}{2}$ (основная гипотеза).
$H_1: \ p \neq \frac{1}{2}$ (альтернативная гипотеза).

Рассмотрим статистику $T = |\bar{X} - 1/2|$.
Если она достаточно большая, то основная гипотеза отклоняется.

Проверка гипотез

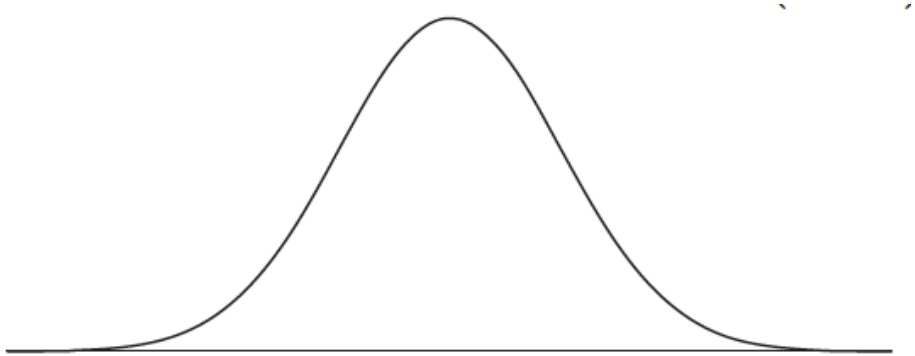
выборка: $\mathbf{X} = (X_1, \ldots, X_n)$

нулевая гипотеза: $H_0 : \ X_i \sim F_0$

альтернатива: $H_1 : \ X_i \sim F_1 \neq F_0$

статистика: $T(x_1, \ldots, x_n), \ T(\mathbf{X}) \sim G$ при $\mathbf{X} \sim F_0$

$T(\mathbf{X}) \nsim G$ при $\mathbf{X} \sim F_1$

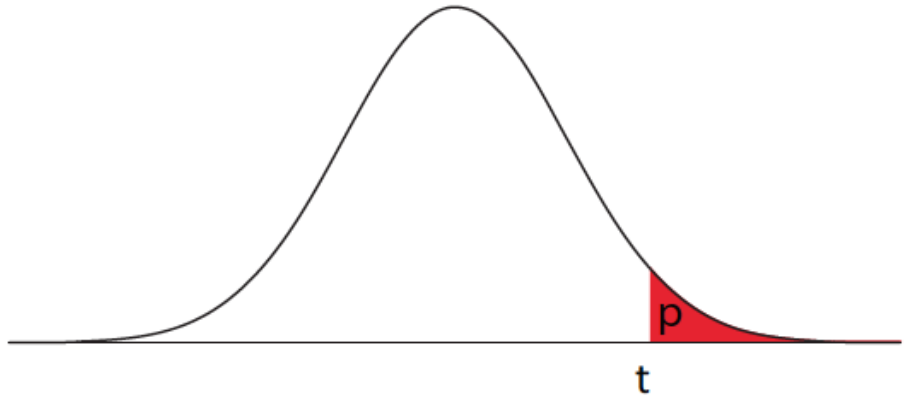


Проверка гипотез

реализация выборки: $\mathbf{x} = (x_1, \ldots, x_n)$

реализация статистики: $t = T(\mathbf{x})$

достигаемый уровень значимости: $p(\mathbf{x}) = \mathbb{P}(T(\mathbf{X}) \geq t \mid H_0)$



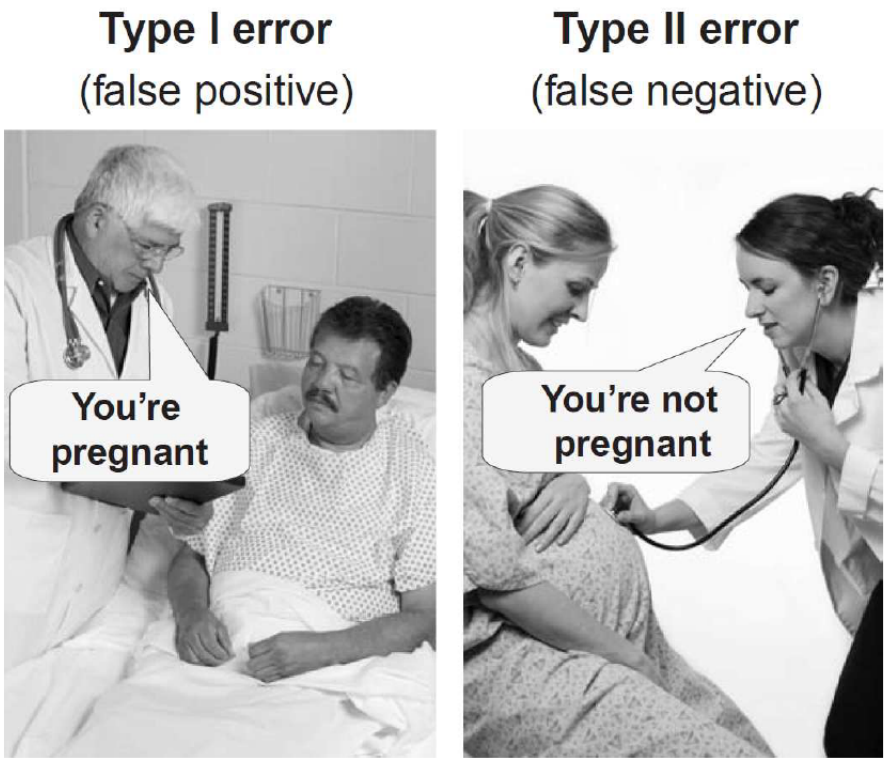
$p(\mathbf{x})$ — вероятность при H_0 получить $T(\mathbf{x}) = t$ или ещё более экстремальное значение. Гипотеза отвергается при $p(\mathbf{x}) \leq \alpha$, α — уровень значимости.

Проверка гипотез

Величина $p(\mathbf{x})$ называется p-value.

Проверка гипотез

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода (False negative)
H_0 отвергается	Ошибка первого рода (False positive)	H_0 верно отвергнута



Проверка гипотез

Если величина p-value достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Если величина p-value недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы!

Проверка гипотез

Вероятность отвергнуть нулевую гипотезу зависит не только от того, насколько она отличается от истины, но и от размера выборки.

По мере увеличения n нулевая гипотеза может сначала приниматься, но потом выявятся более тонкие несоответствия выборки гипотезе H_0, и она будет отвергнута.

Пример проверка гипотез

Джеймс Бонд говорит, что предпочитает мартини взболтанным, но не смешанным. Проведём слепой тест: n раз предложим ему пару напитков и выясним, какой из двух он предпочитает.

Выборка: бинарный вектор длины n:
1 — Джеймс Бонд предпочёт взболтанный, 0 — смешанный.

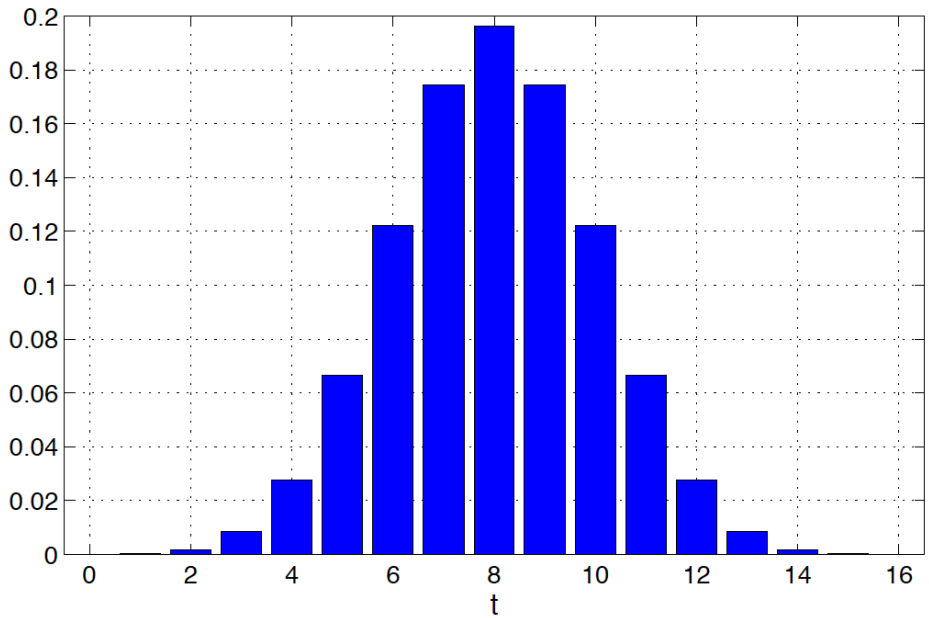
Нулевая гипотеза: Джеймс Бонд не различает два вида мартини, т.е. выбирает наугад.

Статистика: $t =$ число единиц в выборке.

Пример проверка гипотез

Если нулевая гипотеза справедлива и Джеймс Бонд не различает два вида мартини, то равновероятны все выборки длины n из нулей и единиц.

Пусть $n = 16$, тогда существует $2^{16} = 65536$ равновероятных варианта. Статистика t принимает значения от 0 до 16:



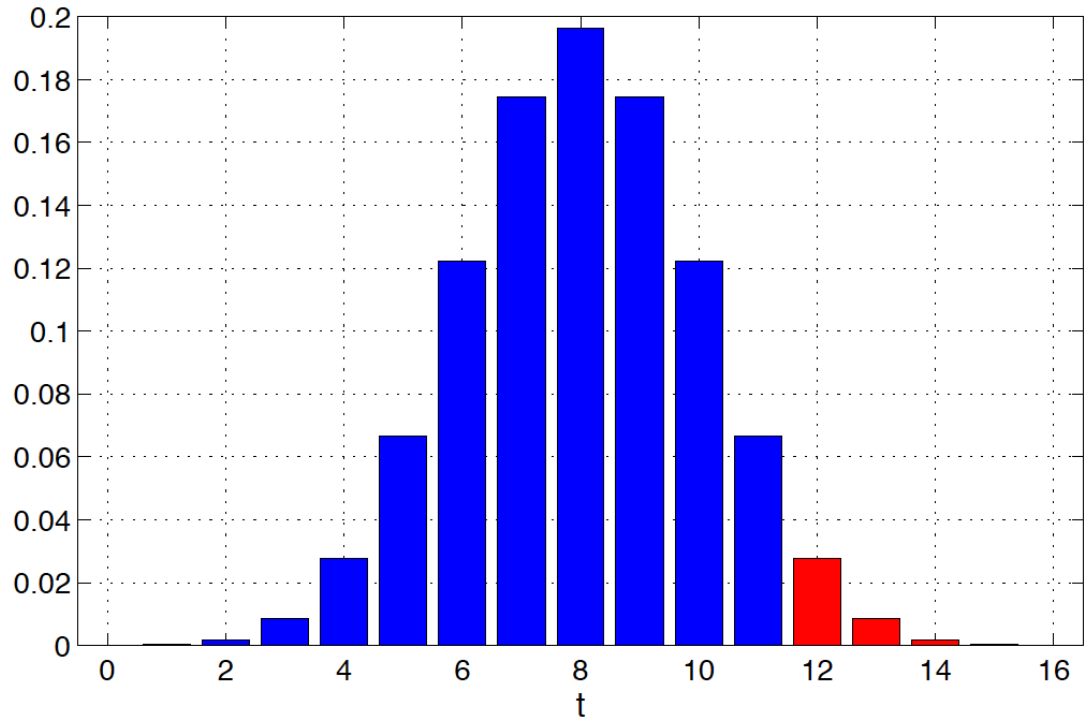
Пример проверка гипотез

H_1: Джеймс Бонд предпочитает взболтанный мартини. При справедливости такой альтернативы более вероятны большие значения t (т.е., большие t свидетельствуют против H_0 в пользу H_1).

Вероятность того, что Джеймс Бонд предпочтёт взболтанный мартини в 12 или более случаях из 16 при справедливости H_0, равна $\frac{2517}{65536} \approx 0.0384$.

Пример проверка гипотез

0.0384 — достигаемый уровень значимости при реализации $t = 12$.



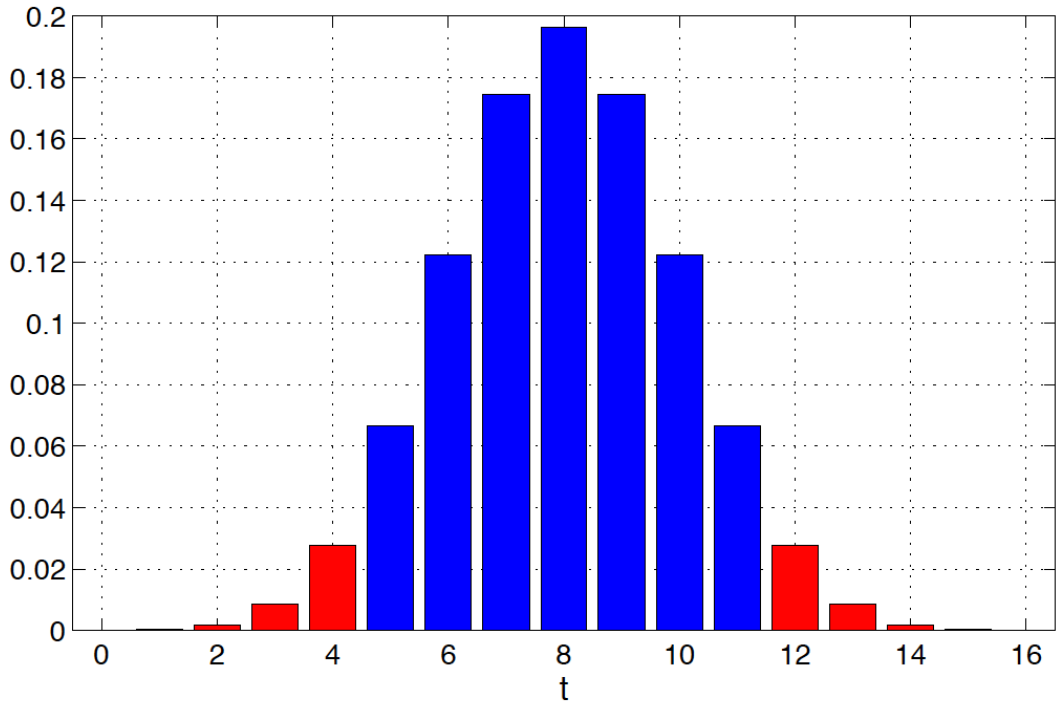
Пример проверка гипотез

H_1: Джеймс Бонд предпочитает какой-то определённый вид мартини. При справедливости такой альтернативы и очень большие, и очень маленькие значения t свидетельствуют против H_0 в пользу H_1).

Вероятность того, что Джеймс Бонд предпочтёт взболтанный мартини в 12 или более случаях из 16 при справедливости H_0, равна $\frac{5034}{65536} \approx 0.0768$.

Пример проверка гипотез

0.0768 — достигаемый уровень значимости при реализации $t = 12$.



Пример проверка гипотез

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Достигаемый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

Линейная регрессия

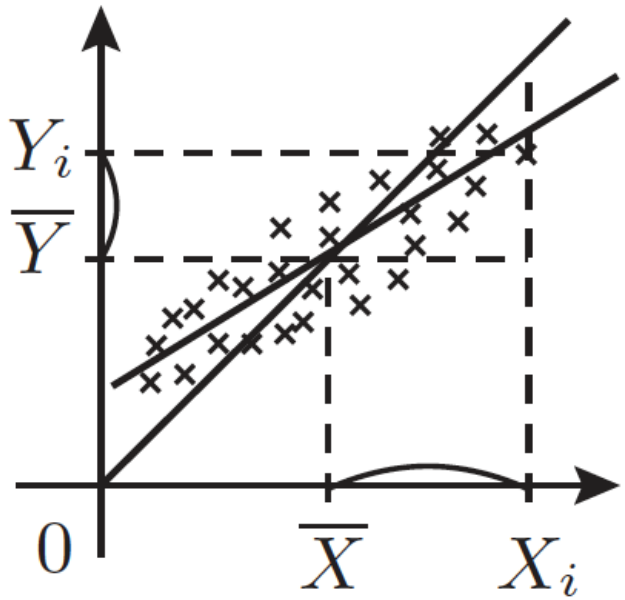
Термин «регрессия» ввел Ф. Гальтон в своей статье «Регрессия к середине в наследовании роста» (1885 г.), в которой он сравнивал средний рост детей Y со средним ростом их родителей X (на основе данных о 928 взрослых детях и 205 их родителях).

Гальтон заметил, что рост детей у высоких (низких) родителей обычно также выше (ниже) среднего роста популяции $\mu \approx \overline{X} \approx \overline{Y}$, но при этом отклонение от μ у детей меньше, чем у родителей. Другими словами, экстремумы в следующем поколении сглаживаются, происходит возвращение назад (регрессия) к середине.

Линейная регрессия

По существу, Гальтон показал, что зависимость Y от X хорошо выражается уравнением

$$Y - \overline{Y} = \frac{2}{3}(X - \overline{X}).$$



Линейная регрессия

Проиллюстрируем основные идеи регрессии на примере изучения зависимости между скоростью автомобиля V и расстоянием Y, пройденным им после сигнала об остановке.

Линейная регрессия

Для каждого отдельного случая результат определяется в основном тремя факторами:

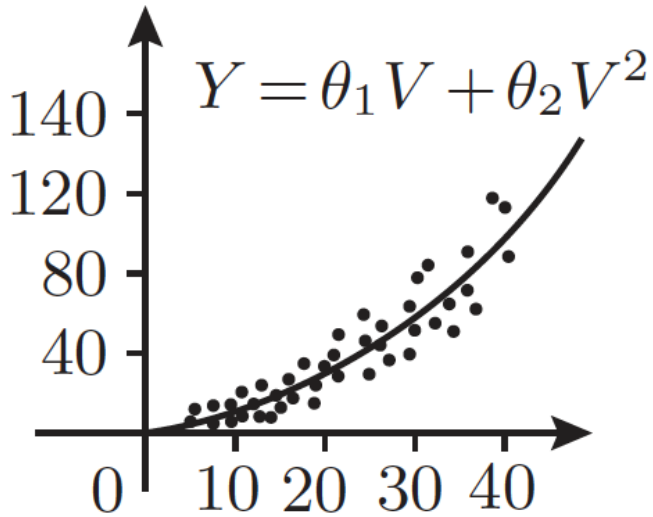
- скоростью автомобиля V в момент подачи сигнала;
- временем реакции θ_1 водителя на этот сигнал;
- тормозами автомобиля.

Автомобиль успеет проехать путь $\theta_1 V$ до момента включения водителем тормозов и еще $\theta_2 V^2$ после этого момента, поскольку согласно элементарным физическим законам теоретическое расстояние, пройденное до остановки с момента торможения, пропорционально квадрату скорости.

Линейная регрессия

Таким образом, в качестве модели годится $Y = \theta_1 V + \theta_2 V^2$.

Для экспериментальных данных **методом наименьших квадратов** были подсчитаны значения $\theta_1 = 0.76$ и $\theta_2 = 0.056$.



Парадоксы регрессии

Есть несколько **типичных ошибок** («тонких мест»), которые следует иметь в виду, применяя регрессионный анализ. Сами по себе они достаточно очевидны. Тем не менее, о них часто забывают при работе с реальными данными и в результате приходят к неверным выводам.

Существуют три вида лжи: ложь, наглая ложь и статистика.
(Марк Твен)

Парадоксы регрессии

Пример

При исследовании зависимости *веса* Z студентов двух групп от их *роста* X и *размера обуви* Y в первой группе было получено регрессионное уравнение

$$Z - \overline{Z} = 0.9(X - \overline{X}) + 0.1(Y - \overline{Y}),$$

а для второй группы:

$$Z - \overline{Z} = 0.2(X - \overline{X}) + 0.8(Y - \overline{Y}).$$

Как объяснить существенное различие коэффициентов этих двухмоделей?

Парадоксы регрессии

Ответ: дело здесь в том, что X и Y сильно зависимы, вследствие чего общий «весовой» коэффициент при $(X - \overline{X}) + (Y - \overline{Y})$ случайным образом распределился между слагаемыми.

Парадоксы регрессии

Пример

Во время второй мировой войны англичане исследовали зависимость *точности бомбометания* Z от ряда факторов, в число которых входили *высота бомбардировщика* H, *скорость ветра* V, *количество истребителей противника* X.

Как и ожидалось, Z увеличивалась при уменьшении H и V. Однако (что поначалу представлялось необъяснимым), точность бомбометания Z возрастала также и при увеличении X.

Парадоксы регрессии

Ответ: дальнейший анализ позволил понять причину этого парадокса. Дело оказалось в том, что первоначально в модель не был включен такой важный фактор, как Y — облачность. Он сильно влияет и на Z (уменьшая точность), и на X (бессмысленно высылать истребители, если ничего не видно).

Парадоксы регрессии

Пример

Если найти зависимость между *ежегодным количеством родившихся в Голландии детей* Z и количеством прилетевших аистов X, то она окажется довольно значительной. Можно ли на основе этого статистического результата заключить, что детей приносят аисты?

Парадоксы регрессии

Ответ: рассмотрим проблему на содержательном уровне. Аисты появляются там, где им удобно вить гнезда; излюбленным же местом их гнездовья являются высокие дымовые трубы, какие строят в голландских сельских домах. По традиции новая семья строит себе новый дом — появляются новые трубы и, естественно, рождаются дети. Таким образом, и увеличение числа гнезд аистов, и увеличение числа детей являются следствиями одной причины Y — образования новых семей.

Спасибо за внимание!