Заметки к семинарам по статистике

https://github.com/bdemeshev/statistics_pro

11 декабря 2023 г.

Содержание

Это задачки по статистике. При везении подсказку, ответ или решение можно найти, кликнув по номеру задачи. Свежая версия доступна по ссылке https://github.com/bdemeshev/statistics_pro. Красивые и сложные олимпиадные задачи по вероятностям можно найти по ссылке https://github.com/bdemeshev/probability_dna, подборку прошлых экзаменов вшэ — https://github.com/bdemeshev/probability_hse_exams, а задачки к семинарам по вероятностями — https://github.com/bdemeshev/probability_pro.

Если github.com не доступен, то можно попробовать зеркало на gitlab.com.

1. Статистика ноль

> Среднее, $\bar{x} = (x_1 + \ldots + x_n)/n$. Медиану в случае непрерывной функции распределения находим из условия $\mathbb{P}(X \geqslant \operatorname{Med} X) = 0.5$, $\mathbb{P}(X \leqslant \operatorname{Med} X) = 0.5$. Медиану в общем случае находим из условия $\mathbb{P}(X \geqslant \operatorname{Med} X) \geqslant 0.5$, $\mathbb{P}(X \leqslant \operatorname{Med} X) \geqslant 0.5$.

1.1 Имеется пять действительных чисел: $x, 9, 5, 4, 7$. При каком значении x медиана будет равна среднему?

1.2 Измерен рост 25 человек. Средний рост оказался равным 160 см. Медиана оказалась равной 155 см. Машин рост в 163 см был ошибочно внесен как 173 см.

а) Как изменятся медиана и среднее после исправления ошибки?

б) А как могут измениться медиана и среднее, если истинный рост Маши равен 153?

1.3 Возможно ли чисто теоретически, что риск катастрофы в расчете на 1 час пути больше для самолета, чем для автомобиля, а в расчете на 1 километр пути — наоборот?

1.4 Деканат утверждает, что если Вовочку перевести из группы А в группу В, то средний рейтинг каждой группы возрастет.

Возможно ли это?

1.5 Есть три группы по 10 человек, две группы по 20 человек и одна группа из 40 человек. У каждой из групп свой преподаватель.

а) Каков средний размер группы, для которой читает лекции наугад выбранный профессор?

б) Каков средний размер группы, в которой учится наугад выбранный студент?

в) Творческий вопрос. Мы ловим студентов наугад и спрашиваем каждого размер группы, в которой он учится. Можно ли как-то восстановить средний размер группы с точки зрения преподавателя?

1.6 Приведите примеры случайных величин X и Y, для которых:

а) $\operatorname{Med}(X+Y) = \operatorname{Med}(X) + \operatorname{Med}(Y)$;

б) $\operatorname{Med}(X+Y) \neq \operatorname{Med}(X) + \operatorname{Med}(Y)$;

в) $\operatorname{Med}(X^k) = \operatorname{Med}(X)^k$ для всех k;

г) $\operatorname{Med}(X^2) \neq \operatorname{Med}(X)^2$.

1.7 Исследователь Вениамин измерил рост пяти случайно выбранных человек. Рост имеет непрерывное распределение.

Какова вероятность того, что истинная медиана роста лежит между минимумом и максимумом из этих пяти наблюдений?

1.8 Во время Второй Мировой войны американские военные собрали статистику попаданий пуль в фюзеляж самолёта. По самолётам, вернувшимся из полёта на базу, была составлена карта повреждений среднестатистического самолёта. С этими данными военные обратились к статистику Абрахаму Вальду с вопросом, в каких местах следует увеличить броню самолёта.

Что посоветовал Абрахам Вальд и почему?

1.9 Два лекарства испытывали на мужчинах и женщинах. Каждый человек принимал только одно лекарство. Общий процент людей, почувствовавших улучшение, больше среди принимавших лекарство А. Процент мужчин, почувствовавших улучшение, больше среди мужчин, принимавших лекарство В. Процент женщин, почувствовавших улучшение, больше среди женщин, принимавших лекарство В.

а) Возможно ли это?

б) Какое лекарству нужно порекомендовать больному, не зная его пола?

1.10 Из набора чисел $\{2, 4, 10, 14\}$ случайным образом равновероятно по очереди выбираются три числа с возможностью повторения.

а) Найдите закон распеределения (табличку с вероятностями) величины X_1. Найдите закон распределения величины X_2.

б) Найдите совместный закон распределения пары X_1, X_2. Найдите совместный закон распределения пары X_1, X_3.

в) Являются величины X_1, X_2, X_3 независимыми? Одинаково распределеннными?

г) Верно ли, что $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \mathbb{E}(X_3)$? Верно ли, что $\mathbb{V}\mathrm{ar}(X_1) = \mathbb{V}\mathrm{ar}(X_2) = \mathbb{V}\mathrm{ar}(X_3)$?

д) Верно ли, что $\mathbb{C}\mathrm{ov}(X_1, X_2) = \mathbb{C}\mathrm{ov}(X_1, X_3)$?

е) Как изменятся ответы на предыдущие вопросы, если числа выбираются без возможности повторения?

1.11 Из фиксированного множества N чисел случайным образом выбирают n чисел. Известно, что если бы выбирать наугад всего одно число, то тогда математическое ожидание и дисперсие этого одного случайного числа были бы равны $\mathbb{E}(X_1) = \mu$ и $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2$. Обозначим среднее арифметическое выбранных n чисел с помощью $\bar{X}_n$. Чему равны $\mathbb{E}(\bar{X}_n)$ и $\mathbb{V}\mathrm{ar}(\bar{X}_n)$, если:

а) Мы выбираем n чисел из N с возвращениями

б) Мы выбираем n чисел из N без возвращений

в) Во что превращаются полученные формулы при $n = 1$? при $n = N$? при $N \to \infty$?

1.12 Исследовательница Мишель подбрасывает кубик 100 раз. Обозначим с помощью X_i количество выпадений числа i на всех кубиках. Например, величина X_1 — это количество выпадений единицы, а X_6 — количество выпадений шестёрки.

а) Как распределена величина X_1? Величина X_6? Найдите $\mathbb{E}(X_1)$, $\mathbb{V}\mathrm{ar}(X_1)$.

б) Верно ли, что величины X_1 и X_6 независимы? Одинаково распередены?

в) Найдите $\mathbb{C}\mathrm{ov}(X_1, X_1 + X_2 + X_3 + X_4 + X_5 + X_6)$, $\mathbb{C}\mathrm{ov}(X_1, X_6)$.

г) Найдите $\mathrm{Corr}(X_1, X_6)$, проинтерпретируйте значение корреляции.

1.13 В множестве A всего два числ, $A = \{24, 42\}$. Случайным образом из множества A выбираются 3 числа с возможностью повторений. Явно найдите закон распределения выборочного среднего, выборочной медианы, выборочной моды, выборочного минимима и выборочного максимума.

1.14 Величины X и Y независимы и одинаково распределены на отрезке $[0; 1]$ с функцией плотности $f(x) = 2x$.

а) Найдите теоретическую медиану $\text{Med}(X)$.

б) Найдите теоретическую медиану $\text{Med}(X+Y)$.

1.15 Величины X_1, ..., X_n — независимы и равномерны на отрезке $[0; 1]$. Узнав значения этих величин, исследовательница Кассандра случайно равновероятно с возможностью повторений выбирает из них значения X_1^*, ..., X_n^*. То есть, X_1^* равновероятно равно X_1, X_2, ..., X_n. Аналогично, X_2^* равновероятно равно X_1, X_2, ..., X_n. И так далее.

а) Найдите $\mathbb{E}(\bar{X})$, $\mathbb{V}\text{ar}(\bar{X})$;

б) Найдите $\mathbb{E}(X_i^*)$, $\mathbb{V}\text{ar}(X_i^*)$, $\mathbb{C}\text{ov}(X_i, X_j^*)$;

в) Найдите $\mathbb{E}(\bar{X}^*)$, $\mathbb{V}\text{ar}(\bar{X}^*)$, $\mathbb{C}\text{ov}(\bar{X}, \bar{X}^*)$;

1.16 Аня, Боря и Вова поделили между собой натуральные числа от 1 до 27, включительно. У Ани среднее выборочное равно 15, у Бори — 3, у Вовы — 18.

Сколько чисел могло оказаться у Ани?

2. Случайная выборка

2.1 Создайте случайную выборку объемом $n = 1000$ из равномерного на отрезке $[0; 1]$ распределения.

а) Найдите выборочные характеристики: среднее, медиану, минимум и максимум, стандартную ошибку, 10%-й и 95%-й квантили.

б) Постройте гистограмму распределения, выборочную функцию распределения для первых 20 чисел из случайной выборки

в) Повторите данный опыт для нормального $\mathcal{N}(5, 1)$ распределения и для экспоненциального распределения с параметром $\lambda = 1$

г) Насколько сильно выборочные характеристики отличаются от истинных?

2.2 Придумайте способ, как сгенерировать 100 одинаково распределенных случайных величин, таких что $\sum_{i=1}^{100} X_i = 50$.

Будут ли эти величины X_i зависимы?

Модифицируйте способ, так чтобы он давал одинаково распределенные величины, такие что $\sum_{i=1}^{100} Y_i^2 = 50$. Будут ли эти новые величины Y_i зависимы?

2.3 Придумайте детерминистическую функцию, такую, которая бы превращала одну равномерную на $[0; 1]$ случайную величину X в

а) случайную величину Y, принимающую значения 1 и 0 с вероятностями 0.7 и 0.3 соответственно б) случайную величину Z с функцией плотности $f(z) = 2z$ на отрезке $z \in [0; 1]$

в) пару независимых одинаково распределенных случайных величин (Y_1, Y_2), принимающих значения 1 и 0 с вероятностями 0.7 и 0.3 соответственно

г) пару независимых равномерных на $[0; 1]$ случайных величин

2.4 Постройте случайную выборку в $n = 200$ наблюдений из двумерного нормального распределения с параметрами:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 25 \\ 125 \end{pmatrix}; \begin{pmatrix} 5 & 4 \\ 4 & 10 \end{pmatrix}\right) \tag{1}$$

а) Посчитайте выборочную ковариацию, выборочную корреляцию

б) Постройте диаграмму рассеяния, нанесите на диаграмму рассеяния линию $y(x) = \mathbb{E}(Y \mid X = x)$

в) Насколько сильно выборочные характеристики отличаются от истинных?

2.5 Создайте 500 выборок объемом $n = 20$ каждая из равномерного на отрезке $[0; 1]$ распределения и вычислите выборочное среднее для каждой из выборок.

а) Каково теоретическое математическое ожидание и дисперсия каждого из выборочных средних?

б) Постройте гистограмму выборочных средних

в) На фоне функции плотности стандартного нормального распределения изобразите в подходящем масштабе гистограмму стандартизированных выборочных средних

3. Проецируй!

Лемма Стейна. Если величина X стандартно нормально распределена $\mathcal{N}(0; 1)$ и $h(x)$ — многочлен, то $\mathbb{E}(Xh(X)) = \mathbb{E}(h'(X))$.

Если вектор Z имеет многомерное нормальное стандартное распределение, $Z \sim \mathcal{N}(0; I)$, $\hat{Z}$ — это проекция вектора Z на некоторое d-мерное подпространство V, Q — это квадрат длины проекции, $Q = \left\|\hat{Z}\right\|^2$, то говорят, что величина Q имеет *хи-квадрат распределение* с d степенями свободы, а пишут $Q \sim \chi^2_d$.

Если Z_1, Z_2, ... независимы и нормально распределены $\mathcal{N}(0; 1)$,

$$Q = Z_1^2 + Z_2^2 + \ldots + Z_d^2,$$

то говорят, что величина Q имеет *хи-квадрат распределение* с d степенями свободы, $Q \sim \chi^2_d$.

Если Z имеет нормальное распределение $\mathcal{N}(0; 1)$ и Q имеет хи-квадрат распределение χ^2_d, Z и Q независимы,

$$R = \frac{Z}{\sqrt{Q/k}},$$

то говорят, что величина R имеет *распределение Стьюдента* с d степенями свободы, $R \sim t_d$.

3.1 Вектор Z имеет многомерное нормальное распределение, $Z_i \sim \mathcal{N}(0; 1)$, и все Z_i независимы. Для каждого случая найдите проекцию $\hat{Z}$ вектора Z на подпространство V; найдите квадрат длины проекции, Q; укажите закон распределения величины Q:

а) $V = \text{span}(e)$, где $e = (1, 1, 1, 1, \ldots, 1)$;

б) $V = \text{span}(e)$, где $e = (1, 2, 3, 4, \ldots, n)$;

в) $V = \text{span}(e_1, e_2)$, где $e_1 = (1, 0, 0, 0, \ldots, 0)$, $e_2 = (0, 1, 1, 1, \ldots, 1)$;

г) $V = (\text{span}(e))^{\perp}$, где $e = (1, 1, 1, 1, \ldots, 1)$;

д) $V = \text{span}(e_5, e_7, e_9)$, вектор e_i содержит 1 на i-ом месте и 0 на остальных;

3.2 Найдите минимум функции $f(a, b, c) = (6-a)^2 + (3-b)^2 + (7-b)^2 + (8-c)^2 + (9-c)^2 + (10-c)^2 + 11^2$;

Проекцией какого вектора на какое пространство является вектор $\hat{Z} = (a^*, b^*, b^*, c^*, c^*, c^*, 0)$?

3.3 Вектор Z из n случайных величин имеет многомерное нормальное распределение, $Z_i \sim \mathcal{N}(0; 1)$, и все Z_i независимы. Определите, какое распределение имеет величина Q, на какое подпространство проецировали вектор Z, и найдите вероятность:

а) $Q = Z_1^2$, $\mathbb{P}(Q > 6.6)$;

б) $Q = n\bar{Z}^2$, $\mathbb{P}(Q < 3.8)$;

в) $Q = \sum (Z_i - \bar{Z})^2$;

г) $Q = Z_5^2 + Z_6^2 + Z_{32}^2$, $\mathbb{P}(Q < 0.58)$;

д) $Q = Z_2^2 + (Z_7 + Z_{11})^2/2$, $\mathbb{P}(Q > 6)$;

е) $Q = (Z_7 + Z_{11})^2/2 + (Z_3 + Z_9 + Z_{12})^2/3$, $\mathbb{P}(Q < 0.21)$;

Для каждого случая укажите подпространство, для которого величина Q будет квадратом длины проекции исходного вектора Z;

3.4 Вектор Z из n случайных величин имеет многомерное нормальное распределение, $Z_i \sim \mathcal{N}(0; 1)$, и все Z_i независимы.

а) Какое распределение имеет величина $Q = Z_1^2 + Z_2^2 + \ldots + Z_d^2$?

б) Чему равно $\mathbb{E}(Q)$?

в) Чему равна дисперсия $\mathbb{V}\text{ar}(Q)$?

г) Величина Q_a имеет хи-квадрат распределение с a степенями свободы, а величина Q_b — с b степенями свободы. Величины Q_a и Q_b независимы. Какое распределение имеет величина $S = Q_a + Q_b$?

3.5 Найдите функцию плотности χ-квадрат распределения с одной степенью свободы;

3.6 Вектор Z из n случайных величин имеет многомерное нормальное распределение, $Z_i \sim \mathcal{N}(0; 1)$, и все Z_i независимы. Вектор v имеет единичную длину.

а) Найдите вектор $\hat{Z}$, проекцию вектора Z на подпространство $\text{span}(v)$; Чему равно $\hat{Z}_i$?

б) Найдите дисперсию $\mathbb{V}\text{ar}(\langle Z, v \rangle)$;

в) Найдите ковариацию $\mathbb{C}\text{ov}(\hat{Z}_i, \hat{Z}_j)$;

г) Как выглядит ковариационная матрица вектора $\hat{Z}$?

3.7 Величины X и Y независимы и имеют хи-квадрат распределение с одной и двумя степенями свободы. Введём величины $R = X/(X + Y)$ и $S = X + Y$.

а) Выпишите совместную функцию плотности $f(x, y)$;

б) Найдите совместную функцию плотности $f(r, s)$;

в) Верно ли, что R и S независимы?

г) С точностью до сомножителя найдите функцию плотности S;

д) Какой закон распределения имеет величина S?

е) Предположите вид функции плотности хи-квадрат распределения с d степенями свободы и докажите догадку по индукции;

3.8 Величины X_1, X_2, … независимы и нормально распределены, $\mathcal{N}(0;1)$.

Укажите закон распределения следующих величин:

а) $X_1^2 + X_2^2 + X_3^2$;

б) $X_1^2 + X_{12}^2 + X_{100}^2$;

в) $(X_1 - X_2)^2/2$;

г) $(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2$, если $\bar{X} = (X_1 + X_2)/2$;

д) $\sum(X_i - \bar{X})^2$, если $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$;

е) X_{13}^2;

ж) $X_1/|X_2|$;

з) X_1/X_2;

и) $2X_1/\sqrt{X_2^2 + X_3^2 + X_4^2 + X_5^2}$;

к) $10X_1/\sqrt{X_2^2 + X_3^2 + \ldots + X_4^2 + X_{101}^2}$.

3.9 Рассмотрим пару величин $X \sim \mathcal{N}(0;1)$ и $Y \sim \mathcal{N}(1;4)$.

а) Найдите $\mathbb{E}(X^3)$, $\mathbb{E}(X^4)$, $\mathbb{E}(X^6)$, $\mathbb{E}(X^{2024})$.

б) Найдите $\mathbb{E}(Y^3)$, $\mathbb{E}(Y^4)$, $\mathbb{E}((Y-1)Y^{100})$.

3.10 Величина Q_1 имеет χ_1^2-распределение.

а) Найдите $\mathbb{E}(Q_1)$ и $\mathbb{V}\mathrm{ar}(Q_1)$.

Величина Q имеет χ^2-распределение с k степенями свободы.

б) Найдите $\mathbb{E}(Q)$, $\mathbb{V}\mathrm{ar}(Q)$.

в) Найдите $\mathbb{E}(Q^3)$.

3.11 Лемма Стейна утверждает, что для стандартно нормально распределенной величины $X \sim \mathcal{N}(0;1)$ и любого многочлена $h(x)$ выполнено равенство $\mathbb{E}(Xh(X)) = \mathbb{E}(h'(X))$.

а) Докажите лемму Стейна с помощью интегрирования по частям.

б) Лемма Стейна верна не только для многочленов. Какое более мягкое условие на функцию h является достаточным для леммы Стейна?

в) Обобщите лемму Стейна на случай $X \sim \mathcal{N}(\mu;\sigma^2)$.

3.12 а) Найдите $\mathbb{P}(R > 2.7)$, если $R \sim \chi_1^2$.

б) Найдите $\mathbb{P}(R \in [3.2;3.9])$, если $R \sim \chi_{10}^2$.

в) Найдите $\mathbb{P}(R < 6)$, если $R \sim \chi_2^2$.

г) Найдите $\mathbb{P}(R > 0.549)$, если $R \sim t_7$.

д) Найдите $\mathbb{P}(|R| > 0.727)$, если $R \sim t_5$.

е) Найдите $\mathbb{P}(R > 0.711)$, если $R \sim \chi^2_4$.

ж) Найдите b, если $\mathbb{P}(|R| < b) = 0.9$ и $R \sim t_8$.

з) Найдите b, если $\mathbb{P}(|R| > b) = 0.05$ и $R \sim t_5$.

и) Найдите b, если $\mathbb{P}(R > b) = 0.05$ и $R \sim \chi^2_5$.

к) Найдите b, если $\mathbb{P}(R < b) = 0.9$ и $R \sim t_{10}$.

3.13 Величина X имеет нормальное распределение $\mathcal{N}(0; 1)$, величина $Q_1 - \chi^2_9$, величина $Q_2 - \chi^2_{16}$. Величины X, Q_1 и Q_2 независимы.

Укажите закон распределения следующих величин:

а) $3X/\sqrt{Q_1}$;

б) $X/\sqrt{Q_2/16}$;

в) $Q_1 + Q_2$;

г) $5X/\sqrt{Q_1 + Q_2}$.

4. Максимально правдоподобно — 1!

4.1 Кот Матроскин каждый вечер ходит на рыбалку. Поймав одну «рыбку» кот Матроскин возвращается домой. В пруду встречаются караси, щуки и бегемоты. Кот Матроскин хочет оценить вероятность p поймать карася. От своей бабушки Кот Матроскин достоверно знает, что щуки встречаются в два раза чаще карасей. За ночь экосистема пруда успевает восстановиться от воздействия кота Матроскина.

а) Оцените $\hat{p}_{\text{КМ}}$ методом максимального правдоподобия, если Кот Матроскин ловил «рыбку» четыре дня и имеются наблюдения: $X_1 =$ щука, $X_2 =$ карась, $X_3 =$ карась, $X_4 =$ бегемот.

б) Постройте оценку $\hat{p}_{\text{КМ}}$ методом максимального правдоподобия в общем виде. Кот Матроскин ходил на пруд n дней, поймал $Y_{\text{к}}$ карасей, $Y_{\text{щ}}$ щук и $Y_{\text{б}}$ бегемотов.

в) Зависимы ли величины $Y_{\text{к}}$, $Y_{\text{щ}}$ и $Y_{\text{б}}$? Как распределена величина $Y_{\text{б}}$? Найдите $\mathbb{E}(Y_{\text{б}})$, $\mathbb{V}\text{ar}(Y_{\text{б}})$.

г) Найдите $\mathbb{E}(\hat{p}_{\text{КМ}})$, $\mathbb{V}\text{ar}(\hat{p}_{\text{КМ}})$

д) Является ли оценка $\hat{p}_{\text{КМ}}$ несмещенной, состоятельной?

е) Постройте аналогичную оценку $\hat{p}_{\text{ПШ}}$ для Пса Шарика. В отличие от Кота Матроскина Пёс Шарик не знает, что щуки встречаются в два раза чаще карасей. Является ли оценка Пса Шарика несмещенной и состоятельной?

ж) Какая из двух оценок является более эффективной? Почему?

4.2 «Про зайцев». В темно-синем лесу, где трепещут осины, живёт очень большое количество n зайцев. Мы не знаем n и случайным образом отловили 100 зайцев. Каждому из них на левое ухо мы завязали бантик из красной ленточки и потом всех отпустили.

Через неделю мы снова отловили 100 зайцев, из них $S = 80$ оказались с бантиками. Всех зайцев угостили морковкой и снова отпустили.

а) Постройте оценку для неизвестного параметра n методом моментов.

б) Постройте оценку для неизвестного параметра n методом максимального правдоподобия.

в) Постройте оценки для неизвестного параметра n методом моментов и методом максимального правдоподобия при произвольном S.

4.3 Вася и Петя независимо друг от друга прочитали всю Википедию. Вася всего нашёл 100 опечаток, Петя — 200 опечаток. При этом 80 опечаток оказались найдены и Петей, и Васей.

а) Оцените количество опечаток в Википедии методом моментов.

б) Оцените количество опечаток в Википедии методом максимального правдоподобия.

Назовём *внимательностью* вероятность, с которой человек находит опечатки.

в) Оцените внимательность Васи методом моментов.

г) Оцените внимательность Васи методом максимального правдоподобия.

4.4 У Васи есть два одинаковых золотых слитка неизвестной массы m каждый и весы, которые взвешивают с некоторой погрешностью. Сначала Вася положил на весы один слиток и получил результат $Y_1 = m + u_1$, где u_1 — случайная величина, ошибка первого взвешивания. Затем Вася положил на весы сразу оба слитка и получил результат $Y_2 = 2m + u_2$, где u_2 — случайная величина, ошибка второго взвешивания. Оказалось, что $y_1 = 0.9$, а $y_2 = 2.3$.

Используя метод максимального правдоподобия оцените вес слитка m и параметр погрешности весов b при разных предпосылках:

а) величины u_i — независимы и нормально распределены $\mathcal{N}(0; b)$;

б) величины u_i — независимы и равномерно распределены $\mathrm{U}[-b; b]$.

4.5 Задача о немецких танках. [1]

Предположим, что все выпущенные танки имеют порядковый номер. От самого первого выпущенного танка, имеющего номер 1, до самого последнего танка, имеющего номер n. В бою удалось подбить танки с номерами 15, 29 и 23.

а) Постройте оценку количества танков методом моментов.

б) Постройте оценку количества танков методом максимального правдоподобия.

в) Постройте несмещённую оценку количества танков с наименьшей дисперсией.

5. Максимально правдоподобно — 2!

Метод моментов (method of moments). Если величины X_1, X_2, ..., X_n независимы и одинаково распределены с $\mathbb{E}(X_1) = m(\theta)$, то оценка метода моментов находится из уравнения $\bar{X} = m(\hat{\theta})$.

Метод максимального правдоподобия (maximum likelihood). Задади функцию правдоподобия,

$$L(\theta) = \begin{cases} \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \theta) \text{ в дискретном случае,} \\ f(x_1, x_2, \ldots, x_n \mid \theta) \text{ в непрерывном случае.} \end{cases}$$

[1] Незадолго до высадки союзников в Нормандии, 6 июня 1944 года, в распоряжении союзников было всего два (!) немецких танка «Пантера V». По серийным номерам на шасси танков союзники оценили выпуск в феврале 1944 в 270 танков. Фактический выпуск «Пантер V» согласно немецким документам в феврале 1944 составил 276 танков, [RB47].

Лог-функция правдопдобия равна $\ell(\theta) = \ln L(\theta)$. Метод максимального правдоподобия максимизирует $\ell(\theta)$.

Для оценки информации Фишера используют один из двух подходов. Чаще всего используют *наблюдаемую информацию Фишера*, $\hat{I} = -\frac{\partial^2 \ell}{\partial \theta^2}$. Во-вторых, можно в формулу теоретической информации Фишера подставить оценку параметра, при этом получится $I(\hat{\theta})$.

$\widehat{\mathbb{V}\mathrm{ar}}(\hat{\theta}_{ML}) = \hat{I}^{-1}$.

При выполнений условий регулярности, $\hat{\theta}_{ML} \dot{\sim} \mathcal{N}(\theta, 1/I(\theta))$. Следовательно, при выполнении условий регулярности в качестве оценки дисперсии можно взять $1/\hat{I}$ или $1/I(\hat{\theta})$.

5.1 Величины X_i — независимы и имеют закон распределения, заданный табличкой:

x	0	1	2
$\mathbb{P}(X_i = x)$	$1.3 - 3\theta$	$2\theta - 0.2$	$\theta - 0.1$

а) Найдите оценку метода максимального правдоподобия $\hat{\theta}_{ML}$ для выборки $X_1 = 0$, $X_2 = 2$.

б) Найдите оценку метода максимального правдоподобия $\hat{\theta}_{ML}$ для произвольной выборки X_1, X_2, ...X_n, содержащей N_0 значений $X_i = 0$, N_1 значений $X_i = 1$ и N_2 значений $X_i = 2$.

в) Найдите оценку метода моментов $\hat{\theta}_{MM}$ для выборки $X_1 = 0$, $X_2 = 2$.

г) Найдите оценку метода моментов $\hat{\theta}_{MM}$ для произвольной выборки X_1, X_2, ...X_n.

д) Первоначально ничего о θ не было известно и поэтому исследователь предположил, что θ распределена равномерно на отрезке $[0.1; 1.3/3]$. Как выглядит условное распределение θ, если известно что $X_1 = 0$, $X_2 = 2$?

5.2 Величины Y_1 и Y_2 независимы и распределены по Пуассону с $\mathbb{E}(Y_1) = e^a$ и $\mathbb{E}(Y_2) = e^{a+b}$.

а) Найдите оценки максимального правдоподобия $\hat{a}$ и $\hat{b}$ для случая $y_1 = 7$ и $y_2 = 3$.

б) Найдите оценки максимального правдоподобия $\hat{a}$ и $\hat{b}$ в общем виде.

5.3 Величины X_1, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} \theta x^{\theta-1}, & \text{при } x \in [0; 1] \\ 0, & \text{иначе.} \end{cases}$$

а) Оцените значение θ с помощью метода максимального правдоподобия.

б) Оцените дисперсию оценки $\hat{\theta}_{ML}$ метода максимального правдоподобия.

в) Как примерно распределена $\hat{\theta}_{ML}$?

г) Оцените значение θ с помощью метода моментов.

д) Оцените дисперсию оценки $\hat{\theta}_{MM}$ метода моментов.

е) Как примерно распределена $\hat{\theta}_{MM}$?

5.4 Величины X_1, ..., X_n независимы и экспоненциально распределены с интенсивностью λ.

а) Постройте оценку $\hat{\lambda}_{MM}$ методом моментов.

б) Постройте оценку $\hat{\lambda}_{ML}$ методом максимального правдоподобия.

в) Оцените дисперсию оценки $\hat{\lambda}_{ML}$ максимального правдоподобия.

5.5 Величины X_1, ..., X_n независимы и нормально распределены $\mathcal{N}(\mu; 4)$.

а) Постройте оценку $\hat{\mu}_{MM}$ методом моментов.

б) Оцените дисперсию оценки $\hat{\mu}_{MM}$ метода моментов.

в) Постройте оценку $\hat{\mu}_{ML}$ методом максимального правдоподобия.

г) Оцените дисперсию оценки $\hat{\mu}_{ML}$ максимального правдоподобия.

5.6 Величины X_1, X_2, ..., X_n независимы и одинаково распределены $\mathcal{N}(a, 2a)$.

а) Постройте оценку $\hat{a}_{ML}$ методом максимального правдоподобия.

б) Оцените дисперсию оценки $\hat{a}_{ML}$ максимального правдоподобия.

в) Постройте оценку $\hat{a}_{MM}$ методом моментов.

г) Оцените дисперсию оценки $\hat{a}_{MM}$ метода моментов.

5.7 У нас есть всего одно наблюдение $Y_1 \sim N(0; 1/(1-\theta^2))$.

а) Постройте оценку метода максимального правдоподобия для θ.

б) Оцените дисперсию оценки метода максимального правдоподобия.

5.8 Величины X_1, X_2, ..., X_n независимы и их функции плотности имеет вид

$$f(x) = \begin{cases} (k+1)x^k, & \text{при } x \in [0; 1]; \\ 0, & \text{иначе.} \end{cases}$$

а) Постройте оценку $\hat{k}_{ML}$ методом максимального правдоподобия.

б) Оцените дисперсию оценки $\hat{k}_{ML}$ максимального правдоподобия.

в) Постройте оценку $\hat{k}_{MM}$ методом моментов.

г) Оцените дисперсию оценки $\hat{k}_{MM}$ метода моментов.

5.9 Величины X_1, X_2, ..., X_n независимы и равномерно распределены на отрезке $[0; \theta]$.

а) Постройте оценку метода моментов для неизвестного θ.

б) Постройте оценку метода максимального правдоподобия для неизвестного θ.

в) Дополнительно предположим, что $\theta > 1$. Как изменятся ответы на пункты (а) и (б), если исследователь не знает значений самих X_i, а знает только количество X_i, оказавшихся больше единицы?

5.10 В озере водятся караси, окуни, щуки и налимы. Вероятности их поймать занесены в табличку:

x	Карась	Окунь	Щука	Налим
$\mathbb{P}(X_i = x)$	0.1	p	p	$0.9 - 2p$

Рыбак поймал 100 рыб и посчитал количества рыб каждого вида, $N_{\text{карась}}$, $N_{\text{окунь}}$, $N_{\text{щука}}$, $N_{\text{налим}}$.

а) Постройте оценку $\hat{p}_{ML}$ методом максимального правдоподобия.

б) Найдите ожидаемую и наблюдаемую информацию Фишера.

в) Некая несмещённая оценка $\hat{\theta}$ получена по 100 наблюдениям: X_1, ..., X_{100}. В каких пределах может лежать $\mathbb{V}\text{ar}(\hat{\theta})$?

5.11 Величины X_1, X_2, ..., X_n независимы и имеют закон распределения, заданный таблицей:

x	1	2
$\mathbb{P}(X_i = x)$	p	$1-p$

а) Постройте оценку максимального правдоподобия $\hat{p}_{ML}$ для выборки $X_1 = 1$, $X_2 = 1$, $X_3 = 2$.

б) Постройте оценку максимального правдоподобия $\hat{p}_{ML}$ для произвольной выборки X_1, ..., X_n.

в) Постройте оценку метода моментов $\hat{p}_{MM}$ для произвольной выборки X_1, ..., X_n.

г) Найдите ожидаемую информацию Фишера. Постройте график функции $I(p)$.

д) Для произвольной выборки найдите наблюдаемую информацию Фишера.

е) Пусть $\hat{\theta}$ — несмещенная оценка, полученная по 100 наблюдениям: X_1, ..., X_{100}. В каких пределах может лежать $\mathbb{V}\mathrm{ar}(\hat{\theta})$?

5.12 Величины X_1, X_2, ..., X_n независимы и имеют экспоненциальное распределение с интенсивностью λ.

а) Найдите информацию Фишера $I(\lambda)$.

Перепараметризуем задачу, рассмотрим теперь параметр $\theta = 1/\lambda$.

б) Найдите информацию Фишера $I(\theta)$.

5.13 Величины X_i независимы и одинаково распределены. Пусть $I_{X_i}(\theta)$ — информация Фишера о θ, получаемая при наблюдении X_i.

а) Как связаны между собой величины $I_{X_1}(\theta)$ и $I_{X_2}(\theta)$?

б) Как связаны между собой величины $I_{X_1}(\theta)$ и $I_{X_1, X_2, \ldots, X_n}(\theta)$?

5.14 Величины X_1, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} \frac{\theta(\ln x)^{\theta-1}}{x}, & \text{при } x \in [1; e], \\ 0, & \text{иначе.} \end{cases}$$

а) Найдите оценку параметра θ методом максимального правдоподобия, если по выборке из 100 наблюдений оказалось $\sum \ln(\ln(X_i)) = -30$.

б) Найдите оценку параметра θ методом максимального правдоподобия для произвольной выборки.

в) Найдите ожидаемую и наблюдаемую информацию Фишера.

г) Постройте 95% доверительный интервал для параметра θ.

5.15 Величины X_1, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} \frac{\theta \exp\left(-\theta^2/2x\right)}{\sqrt{2\pi x^3}} & \text{при } x \in [0; +\infty), \\ 0, & \text{иначе.} \end{cases}$$

а) Найдите оценку параметра θ методом максимального правдоподобия, если по выборке из 100 наблюдений оказалось $\sum 1/X_i = 12$.

б) Найдите оценку параметра θ методом максимального правдоподобия для произвольной выборки.

в) Найдите теоретическую информацию Фишера $I(\theta)$.

г) Пользуясь данными по выборке постройте оценку $\hat{I}$ для информации Фишера.

д) Постройте 90% доверительный интервал для θ. Подсказка: $\mathbb{E}(1/X_i) = 1/\theta^2$, интеграл берется, например, заменой $x = \theta^2 a^{-2}$.

5.16 Величины X_1, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} \frac{\exp(-|x|)}{2(1-\exp(-\theta))} \text{ при } x \in [-\theta; \theta], \\ 0, \text{ иначе.} \end{cases}$$

а) Найдите оценку $\hat{\theta}_{ML}$ максимального правдоподобия для выборки $X_1 = -5$, $X_2 = 4$, $X_3 = 2$.

б) Найдите оценку параметра θ методом максимального правдоподобия для произвольной выборки.

в) Выполнены ли условия регулярности правдоподобия в этой задаче?

5.17 Величины X_1, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} 2x/\theta^2, \text{ при } x \in [0; \theta], \\ 0, \text{иначе.} \end{cases}$$

а) Найдите оценку параметра θ методом максимального правдоподобия для выборки $X_1 = 2$, $X_2 = 7$.

б) Найдите оценку параметра θ методом максимального правдоподобия для произвольной выборки.

в) Выполнены ли условия регулярности правдоподобия в этой задаче?

г) Найдите оценку метода моментов для произвольной выборки.

5.18 Известно, что X_i независимы, $\mathbb{E}(X_i) = 5$, $\mathbb{V}\text{ar}(X_i) = 4$.

Рассмотрим величины

$$\bar{X}_n, \quad Y_n = (\bar{X}_n + 3)/(\bar{X}_n + 6), \quad Z_n = \bar{X}_n^2, \quad W_n = 1/\bar{X}_n.$$

Как эти величины примерно распределены при большом n?

5.19 Величины X_i независимы и равномерно распределены на отрезке $[0; 1]$.

а) Найдите $\mathbb{E}(\ln(X_i))$, $\mathbb{V}\text{ar}(\ln(X_i))$, $\mathbb{E}(X_i^2)$, $\mathbb{V}\text{ar}(X_i^2)$.

Рассмотрим величины

$$R_n = \frac{\sum \ln(X_i)}{n}, \quad Y_n = (X_1 \cdot X_2 \cdots X_n)^{1/n}, \quad Z_n = \left(\frac{\sum X_i^2}{n}\right)^3.$$

б) Как примерно распределены эти величины при большом n?

5.20 Величины X_i независимы и одинаково распределены

x	0	a
$\mathbb{P}(X_i = x)$	$0.1a - 0.1$	$1.1 - 0.1a$

Параметр $a > 0$ неизвестен.

а) По выборке $X_1 = 0$, $X_2 = 0$ найдите оценку $\hat{a}_{ML}$ методом максимального правдоподобия.

б) По выборке $X_1 = 0$, $X_2 = 0$, $X_3 = 5$ найдите оценку $\hat{a}_{ML}$ методом максимального правдоподобия.

в) Найдите оценку $\hat{a}_{ML}$ методом максимального правдоподобия для произвольной выборки.

г) Найдите дисперсию $\hat{a}_{ML}$.

д) Выполнены ли условия регулярности правдоподобия в этой модели?

е) Выполнено ли неравенство Рао — Крамера в этой задаче?

ж) Найдите оценку $\hat{a}_{MM}$ методом моментов.

5.21 Начинающий каратист Вася тренируется бить кирпичи ударом ладони. Каждый день он бьёт ладонью по кирпичу до пор, пока тот не расколется от одного удара. Предположим, что вероятность разбить кирпич с одного удара равна p и неизменна во времени. Величины X_1, X_2, ..., X_n — количества ударов которые потребовались Васе в соответствующий день.

а) Найдите оценку $\hat{p}_{ML}$ методом максимального правдоподобия.

б) Найдите оценку $\hat{p}_{MM}$ методом моментов.

в) Найдите достаточную статистику T.

г) Выразите оценку дисперсии оценки $\hat{p}_{ML}$ через достаточную статистику T.

д) Найдите $\mathbb{P}(X_1 = 1 \mid T = t)$.

5.22 Продавщица Глафира отдаёт псу Шарику в конце каждого дня нерасфасованные остатки мясного фарша. Фарш фасуется упаковками по a грамм, поэтому нерасфасованный остаток в i-й день, X_i, случаен и равномерно распределен на отрезке $[0; a]$. Пёс Шарик хорошо помнит все X_1, ..., X_n.

Помогите псу Шарику!

а) Найдите оценку $\hat{a}_{ML}$ методом максимального правдоподобия.

б) Найдите оценку $\hat{a}_{MM}$ методом моментов.

в) Найдите достаточную статистику T.

г) Выразите $\widehat{\mathbb{V}\mathrm{ar}}(\hat{a}_{ML})$ через достаточную статистику T.

д) Найдите $\mathbb{P}(X_1 < 10 \mid T = t)$.

5.23 Величины X_1, X_2, ..., X_n независимы и равномерно распределены на отрезке $[-b; b]$.

а) Найдите оценку $\hat{b}_{ML}$ максимального правдоподобия для выборки $X_1 = -5$, $X_2 = 4$, $X_3 = 2$.

б) Найдите оценку $\hat{b}_{ML}$ максимального правдоподобия для произвольной выборки.

в) Найдите оценку $\hat{b}_{MM}$ методом моментов.

5.24 Величины X_i равномерны на отрезке $[-a; 3a]$ и независимы. Есть несколько наблюдений, $X_1 = 0.5$, $X_2 = 0.7$, $X_3 = -0.1$.

а) Найдите $\mathbb{E}(X_i)$ и $\mathbb{E}(|X_i|)$.

б) Постройте оценку метода моментов, используя $\mathbb{E}(X_i)$.

в) Постройте оценку метода моментов, используя $\mathbb{E}(|X_i|)$.

г) Постройте оценку обобщёного метода моментов используя моменты $\mathbb{E}(X_i)$, $\mathbb{E}(|X_i|)$ и взвешивающую матрицу

$$W = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

д) Найдите оптимальную теоретическую взвешивающую матрицу для обобщённого метода моментов.

е) Постройте двухшаговую оценку обобщённого метода моментов, начав со взвешивающей матрицы W.

5.25 Величины X_i имеют Пуассоновское распределение с параметром λ и независимы. Есть несколько наблюдений, $X_1 = 5$, $X_2 = 7$, $X_3 = 1$.

а) Найдите $\mathbb{E}(X_i)$ и $\mathbb{E}((X_i - \bar{X})^2)$.

б) Постройте оценку метода моментов, используя $\mathbb{E}(X_i)$.

в) Постройте оценку метода моментов, используя $\mathbb{E}((X_i - \bar{X})^2)$.

г) Постройте оценку обобщёного метода моментов используя моменты $\mathbb{E}(X_i)$, $\mathbb{E}((X_i - \bar{X})^2)$ и взвешивающую матрицу

$$W = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

д) Найдите оптимальную теоретическую взвешивающую матрицу для обобщённого метода моментов.

е) Постройте двухшаговую оценку обобщённого метода моментов, начав со взвешивающей матрицы W.

5.26 Величины X_1, ..., X_n независимы и равномерны на отрезке $[-1; 1]$. Величины Y_1, ..., Y_n независимы и имеют на отрезке $[-1; 1]$ функцию плотности

$$h(y) = \frac{3\exp(\alpha^{-4})}{4}(1 - \exp\left(2\alpha^{-4}\right)(y + \alpha - 1)^2).$$

Исследовательница Зинаида создаёт величины Z_1, ..., Z_n по простому принципу. Каждая Z_i равна Y_i с вероятностью α и X_i с вероятностью $(1 - \alpha)$.

Пусть $n = 500$.

а) Постройте выборку из X_i и изобразите результат с помощью гистограммы.

б) Постройте выборку из Y_i при $\alpha = 0.5$ и изобразите результат на гистограмме.

в) Постройте выборку из Z_i при $\alpha = 0.5$ и изобразите результат на гистограмме.

У рассеянной Зинаиды сохранилась только выборка из Z_i. Она не помнит ни α, ни X_i, ни Y_i. И поэтому Зинаида решила восстановить α с помощью максимального правдоподобия.

г) Для полученной выборки Z_i постройте график функции правдоподобия. Будьте осторожны! Если по-быстрому строить график командой встроенной в R/python/julia/..., то получится неверный график! Обратите внимание на значения функции в точках Z_i.

д) Найдите оценку максимального правдоподобия для параметра α по полученной выборке.

е) Найдите $\operatorname{plim}_{n\to\infty} \hat{\alpha}_{ML}(n)$ для произвольного α. Является ли оценка $\hat{\alpha}_{ML}$ состоятельной?

5.27 Величины X_1, ..., X_n независимы и одинаково распределены с функцией распределения

$$F(x) = 1/(1 + \exp(a - x))$$

Рассмотрим три оценки: оценку максимального правдоподобия, $\hat{a}_{ML}$, простую медиану $\hat{a}_{Med} = \operatorname{Med} X_1, \ldots, X_n$ и скорректированную медиану $\hat{a}_{1step} = \hat{a}_{Med} - \frac{\ell'(\hat{a}_{Med})}{\ell''(\hat{a}_{Med})}$.

а) Существует ли в явном виде формула для $\hat{a}_{ML}$ в этой задаче?

б) Сгенерируйте 1000 выборок по $n = 20$ наблюдений в каждой для $a = 2$. Для каждой выборки найдите оценки трёх видов. Таким образом получится 1000 оценок каждого вида. Рассчитайте выборочную дисперсию и выборочное среднее каждой оценки.

в) Прокомментируйте полученные результаты.

5.28 Екатерина Вторая оценивает два параметра, a_1 и a_2. Пусть $\ell(a)$ — это функция правдоподобия, а $\hat{a}_1$ и $\hat{a}_1$ — некая несмещённая оценка первого параметра, не обязательно оценка максимального правдоподобия.

а) Найдите $\mathbb{E}(\hat{a}_1)$, $\mathbb{E}(\ell'_1)$.

б) Найдите $\mathbb{C}\mathrm{ov}(\hat{a}_1, \ell'_1)$, $\mathbb{C}\mathrm{ov}(\hat{a}_1, \ell'_2)$.

в) Как связаны между собой $\mathbb{E}(\ell'_1 \cdot \ell'_2)$ и $\mathbb{E}(\ell''_{12})$?

г) Как связаны между собой $\mathbb{E}(\ell'_1 \cdot \ell'_1)$ и $\mathbb{E}(\ell''_{11})$?

5.29 Екатерина Первая оценивает параметр a. Пусть $\ell(a)$ — это функция правдоподобия, а $\hat{a}$ — некая несмещённая оценка параметра, не обязательно оценка максимального правдоподобия.

Екатерина хочет найти наилучшую линейную аппроксимацию для $\hat{a}$ с помощью производной правдоподобия. Другими словами Екатерина хочет представить оценку в виде:

$$\hat{a} = \gamma + \beta\ell' + u,$$

где u — ошибка линейной аппроксимации.

Екатерина ищет такие коэффициенты γ и β, при которых величина $\mathbb{E}(u^2)$ будет минимальна.

а) Выпишите условия первого порядка для задачи оптимизации.

б) Чем равно $\mathbb{E}(u)$, $\mathbb{C}\mathrm{ov}(u, \ell')$?

в) Найдите γ и β.

г) Чему равны ожидание и дисперсия наилучшей линейной аппроксимации?

д) Докажите неравенство Крамера-Рао: $\mathbb{V}\mathrm{ar}(\hat{a}) \geqslant 1/\mathbb{V}\mathrm{ar}(\ell')$.

5.30 Для анализа возьмём данные по распределению количества мальчиков в семьях с 12-ю детьми в Саксонии в 19-м веке,

https://github.com/vincentarelbundock/Rdatasets/raw/master/csv/vcd/Saxony.csv.

а) Нарисуйте симпатичную гистограммку. Сколько всего семей в наборе данных?

б) Предположим, что количества мальчиков в каждой семье распределены биномиально и независимо с общим параметром p. Оцените параметр p методом максимального правдопдобия.

в) Постройте 99%-й доверительный интервал для p.

г) Постройте 99%-й доверительный интервал для вероятности рождения ровно 6 мальчиков в семье с 12-ю детьми.

д) Сравните табличкой фактические частоты и прогнозируемые вероятности для каждого количества детей.

5.31 Для анализа возьмём данные по распределению количества мальчиков в семьях с 12-ю детьми в Саксонии в 19-м веке,

https://github.com/vincentarelbundock/Rdatasets/raw/master/csv/vcd/Saxony.csv.

а) Нарисуйте симпатичную гистограммку. Сколько всего семей в наборе данных?

б) Предположим, что вероятности рождения мальчика в каждой семье имеют бета распределение с параметрами a и b и независимы. Оцените параметры a и b методом максимального правдопдобия.

в) Постройте 99%-е доверительные интервалы для оценённых параметров.

г) Постройте 99%-й доверительный интервал для вероятности рождения ровно 6 мальчиков в семье с 12-ю детьми.

д) Сравните табличкой фактические частоты и прогнозируемые вероятности для каждого количества детей.

Ух, сколько параметров!

5.32 Величины X_1, X_2, ..., X_n независимы и нормальны $\mathcal{N}(\mu; d)$.

а) Найдите оценки μ и d методом максимального правдоподобия.

б) Найдите оценки μ и d методом моментов, используя $\mathbb{E}(X_1)$ и $\mathbb{E}(X_1^2)$.

5.33 Величины X_1, X_2, ..., X_n независимы и равномерны $\mathrm{U}[a; ka]$.

Найдите оценки a и k методом моментов, используя $\mathbb{E}(X_1)$ и $\mathbb{E}(X_1^2)$.

5.34 Величины X_1, X_2, ..., X_n независимы и равномерны $\mathrm{U}[\mu - \theta; \mu + \theta]$.

а) Найдите оценки μ и θ методом максимального правдоподобия по выборке $X_1 = -5$, $X_2 = 3$, $X_3 = 7$.

б) Найдите оценки μ и θ методом максимального правдоподобия для произвольной выборки X_1, X_2, ..., X_n.

в) Найдите оценки μ и θ методом моментов, используя $\mathbb{E}(X_1)$ и $\mathbb{E}(X_1^2)$.

5.35 Величины $X_1, X_2, \ldots, X_n$ распределены равномерно $\mathrm{U}[-a; a]$. Величины $Y_1, Y_2, \ldots, Y_n$ распределены экспоненциально с интенсивностью λ.

Все величины (X_i) и (Y_j) независимы.

Мы не знаем, ни $X_1, \ldots, X_n$, ни $Y_1, \ldots, Y_n$. Мы наблюдаем только $S_i = X_i + Y_i$.

Найдите оценки a и λ методом моментов, используя $\mathbb{E}(S_1)$ и $\mathbb{E}(S_1^2)$.

5.36 Величины $X_1, X_2, \ldots, X_n$ распределены равномерно $\mathrm{U}[-a; 3a]$. Величины $Y_1, Y_2, \ldots, Y_n$ распределены нормально $\mathcal{N}(1; d)$.

Все величины (X_i) и (Y_j) независимы.

Мы не знаем, ни $X_1, \ldots, X_n$, ни $Y_1, \ldots, Y_n$. Мы наблюдаем только $S_i = X_i + Y_i$.

Найдите оценки a и d методом моментов, используя $\mathbb{E}(S_1)$ и $\mathbb{E}(S_1^2)$.

6. LR-тест

6.1 В ограниченной модели при верной H_0 предполагается, что наблюдаемые Y_i независимы и нормально распределены $\mathcal{N}(0; \sigma^2)$ с неизвестной дисперсией. В неограниченной модели предполагается, что математическое ожидание Y_i может равняться произвольной неизвестной константе μ.

По 200 наблюдениям известно, что $\bar{Y} = 5$, $\sum Y_i^2 = 1000$. Альтернативная гипотеза состоит в том, что H_0 неверна, но верна неограниченная модель.

а) В предположении верности H_0 найдите ограниченные оценки метода максимального правдоподобия $\hat{\mu}_R$ и $\hat{\sigma}^2_R$.

б) Найдите неограниченные оценки метода максимального правдоподобия $\hat{\mu}_{UR}$ и $\hat{\sigma}^2_{UR}$.

в) Протестируйте гипотезу H_0 на уровне значимости 0.05 с помощью LR-теста и t-теста.

г) При верной H_0 найдите $\operatorname{plim} \ln \frac{\hat{\sigma}^2_R}{\hat{\sigma}^2_{UR}}$.

д) Воспользовавшись тем, что $\ln(1+u) \approx u$ при малых u, найдите

$$\operatorname{plim} \frac{LR_n}{T_n^2},$$

где LR_n — LR-статистика по n наблюдениям, а T_n — t-статистика по n наблюдениям.

7. Настоящий ценитель

7.1 Случайные величины X_1, X_2, ...независимы и одинаково распределены с неизвестными $\mathbb{E}(X_i) = \mu$ и $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$. Исследовательница Борислава хочет использовать оценку вида $\hat{\mu} = c(X_1 + X_2 + \ldots + X_n)$ для неизвестного параметра μ.

а) При каком c оценка Бориславы будет несмещённой? Возможно ли использовать такое c в практической задаче?

б) При каком c будет минимальной величина $MSE = \mathbb{E}((\hat{\mu} - \mu)^2)$? Возможно ли использовать такое c в практической задаче?

в) Святозар минимизирует по $\hat{\mu}$ штрафную функцию

$$Q(\hat{\mu}) = \sum (X_i - \hat{\mu})^2 + \lambda \hat{\mu}^2.$$

При каком λ оценка Святозара совпадёт с несмещённой оценкой Бориславы? С оценкой минимизирующей MSE?

7.2 Исследовательница Радомира размышляет о том, как оценить неизвестное математическое ожидание по выборке из независимых одинаково распределённых случайных величин. Она мучительно выбирает из нескольких оценок. Для каждой оценки определите, является ли она несмещённой, состоятельной и линейной по наблюдениям. Для линейных несмещённых оценок определите, являются ли они эффективными среди линейных несмещённых оценок.

а) Удалить из выборки наблюдение номер 13 и посчитать среднее арифметическое.

б) Удалить из выборки все нечётные наблюдения и посчитать среднее арифметическое.

в) Удалить из выборки все наблюдения после 13-го и посчитать среднее арифметическое.

г) Домножить наблюдение номер 13 на 13 и посчитать среднее арифметическое.

д) Прибавить число 13 к наблюдению номер 13 и посчитать среднее арифметическое.

е) Продублировать 13-ое наблюдение 13 раз и посчитать среднее арифметическое.

ж) Продублировать каждое наблюдение 13 раз и посчитать среднее арифметическое.

з) Домножить первое наблюдения на 1, второе — на 2, третье — на 3, и так далее и посчитать среднее арифметическое.

и) Прибавить к первому наблюдению 1, ко второму — 2, к третьему — 3, и так далее и посчитать среднее арифметическое.

к) Продублировать первое наблюдения 1 раз, второе — 2 раза, третье — 3 раза, и так далее и посчитать среднее арифметическое.

7.3 Величины Y_i независимы и имеют функцию плотности

$$f(y) = \begin{cases} 5y^4/\theta^5, & \text{если } y \in [0; \theta]; \\ 0, & \text{иначе.} \end{cases}$$

а) Найдите ML оценку неизвестного параметра θ.

б) Устно, не производя вычислений, определите, является ли оценка $\hat{\theta} = \max\{Y_1, Y_2, \ldots, Y_n\}$ несмещённой.

в) Найдите функцию распределения Y_1, функцию распределения $\hat{\theta}$, функцию плотности $\hat{\theta}$.

г) Найдите $\mathbb{E}(\hat{\theta})$.

д) Если $\hat{\theta}$ смещённая, то скорректируйте оценку так, чтобы она стала несмещённой.

7.4 Величина Y имеет биномиальное распределение $\text{Bin}(n, p)$.

а) Является ли оценка $\hat{p} = Y/n$ для p несмещённой? Если является смещённой, то скорректируйте оценку так, чтобы она стала несмещённой.

б) Чему равна теоретическая дисперсия σ^2 величины Y?

в) Является ли оценка $\hat{\sigma}^2 = n\hat{p}(1-\hat{p})$ для σ^2 несмещённой? Если является смещённой, то скорректируйте оценку так, чтобы она стала несмещённой.

7.5 Величины X_i независимы и одинаково распределены с неизвестными $\mathbb{E}(X_i) = \mu$ и $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$.

Рассмотрим четыре оценки:

$$\hat{\mu}_A = (X_1+X_2)/2, \quad \hat{\mu}_B = (X_1+X_2+X_3)/3, \hat{\mu}_C = 2X_1 - X2, \quad \hat{\mu}_D = (X_1+X_2+\ldots+X_{20})/21.$$

а) Какая из приведенных оценок для μ является несмещённой?

б) У какой несмещённой оценки самая маленькая дисперсия?

в) Выберите наилучшую из четырёх оценку по критерию MSE, если $\sigma = 0.5\mu$.

7.6 Величина X равномерна на $[0; a]$. Придумайте несмещённую оценку параметра a вида $\hat{a} = \beta_1 + \beta_2 X$.

7.7 Величины X_i — независимы и одинаково распределены. При каком значении параметра β

а) оценка $\hat{\mu} = 2X_1 - 5X_2 + \beta X_3$ будет несмещённой для $\mu = \mathbb{E}(X_i)$?

б) оценка $\hat{\sigma}^2 = \beta(X_1 + X_2 - 2X_3)^2$ будет несмещённой для $\sigma^2 = \mathbb{V}\mathrm{ar}(X_i)$?

7.8 Величины X_1 и X_2 независимы и равномерны на отрезке $[0; a]$ и $Z = \min\{X_1, X_2\}$.

а) Для величины Z найдите функцию распределения, функцию плотности, $\mathbb{E}(Z)$, $\mathbb{V}\mathrm{ar}(Z)$.

б) При каком β оценка $\hat{a} = \beta Z$ для параметра a будет несмещённой?

в) При каком β оценка $\hat{a} = \beta Z$ для параметра a будет обладать наименьшей среднеквадратичной ошибкой?

7.9 Величины X_i независимы и одинаково распределены с неизвестными $\mathbb{E}(X_i) = \mu$ и $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$.

Какие из приведенных оценок для σ^2 являются несмещёнными?

а) $X_1^2 - X_1 X_2$;

б) $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$;

в) $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$;

г) $\frac{1}{2}(X_1 - X_2)^2$;

д) $X_1 - 2X_2$;

е) $X_1 X_2$.

7.10 Величина X равномерна на отрезке $[3a - 2; 3a + 7]$.

При каких β_1 и β_2 оценка $\hat{a} = \beta_1 + \beta_2 X$ неизвестного параметра a будет несмещённой?

7.11 Закон распределения величины X имеет вид

а)

x_i	0	1	a
$P(X = x_i)$	$1/4$	$1/4$	$2/4$

;

б)

x_i	0	1	2
$P(X = x_i)$	$1/4$	a	$(3/4 - a)$

Постройте несмещённую оценку вида $\hat{a} = \beta_1 + \beta_2 X$ для неизвестного параметра a.

7.12 Время горения лампочки распределено экспоненциально с математическим ожиданием μ. Вася включил одновременно 20 лампочек. Величина $X_{\min}$ обозначает время самого первого перегорания.

Придумайте несмещённую оценку для μ на основе величины $X_{\min}$.

7.13 Величины X_i независимы и одинаково распределены с $\mathbb{V}\mathrm{ar}(X_i) = 16$ и $\mathbb{E}(X_i) = \frac{\theta}{\theta+1}$, где $\theta > 0$ — неизвестный параметр.

С помощью $\bar{X}$ постройте состоятельную оценку для θ.

7.14 Величины $Y_i = \beta x_i + \varepsilon_i$, константа β и случайные величины ε_i являются ненаблюдаемыми. Известно, что $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{V}\mathrm{ar}(\epsilon_i) = 1$, ϵ_i являются независимыми. Константы x_i наблюдаемы, и известно, что $20 < x_i < 100$. У исследователя есть две оценки для β: $\hat{\beta}_1 = \frac{\bar{Y}}{\bar{x}}$ и $\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2}$.

а) Проверьте несмещенность, состоятельность.

б) Определите, какая оценка из двух является наиболее эффективной.

7.15 Исследователи Иван да Марья интересуются, какая доля населения берёт взятки. Они независимо друг от друга задают разным людям вопрос: «Берёте ли Вы взятки?»

Иван использует следующий механизм: предлагает респонденту тайно подбросить правильную монетку, если монетка выпадает орлом, то предлагается ответить «да», если решкой — то предлагается ответить правду. Предположим, что все респонденты действую точно так, как предлагает Иван.

У Марьи кристально чистые голубые глаза, она юна и невинна, и солгать ей просто невозможно.

Пусть p — доля берущих взятки, а $\hat{q}_I$ и $\hat{q}_M$ — доля ответивших «да» на вопросы Ивана да Марьи. Марья использует оценку $\hat{p}_M = \hat{q}_M$, а Иван — $\hat{p}_I = 2\hat{q}_I - 1$.

а) Являются ли оценки $\hat{p}_M$, $\hat{p}_I$ несмещёнными? состоятельными?

б) Иван планирует опросить 100 человек. Сколько человек в завимости от p нужно опросить Марье, чтобы $\mathbb{V}\mathrm{ar}(\hat{p}_I) = \mathbb{V}\mathrm{ar}(\hat{p}_M)$?

Идея: Кирилл Фурманов

7.16 Величины X_i независимы от друг друга равны единице с вероятностью p и нулю с вероятностью $1 - p$.

Рассмотрим две оценки $\hat{p}_1 = \frac{\sum_{i=1}^{10} X_i}{10}$ и $\hat{p}_2 = \frac{\sum_{i=1}^{10} X_i}{11}$.

При каких значениях параметра p вторая оценка будет лучше по критерию среднеквадратичной ошибки MSE?

7.17 Машенькины весы измеряют вес с ошибкой, ошибки измерения — независимые одинаково распределённые случайные величины. Маша взвесилась до того, как выпила 100 грамм смузи, и после того. Однако весы показали на 200 грамм больше.

Найдите несмещённую оценку дисперсии ошибки весов.

7.18 Величины X_1, X_2 независимы и имеют распределение Бернулли с параметром p.

а) Постройте две разных несмещённых оценки для p.

б) Постройте две разных несмещённых оценки для p^2.

в) Существует ли несмещённая оценка для $\sqrt{p}$?

г) Для каких параметров существуют несмещённые оценки?

7.19 Величина X имеет нормальное распределение $\mathcal{N}(\mu, \sigma^2)$, с неизвестными параметрами.

Существует ли несмещённая оценка для $\theta = |\mu|$?

7.20 Рассмотрим две независимых случайных выборки, $X_1, X_2, \ldots, X_{n_x}, Y_1, Y_2, \ldots, Y_{n_y}$ с $\mathbb{E}(X_i) = \mu_x$, $\mathbb{E}(Y_i) = \mu_y$ и $\mathbb{V}\mathrm{ar}(X_i) = \mathbb{V}\mathrm{ar}(Y_i) = \sigma^2$.

Великий комбинатор Остап Бендер хочет с помощью констант a_x и a_y *скомбинировать* несмещённые оценки дисперсий по отдельным выборкам, $\hat{\sigma}_x^2$ и $\hat{\sigma}_y^2$, в одну общую

$$\hat{\sigma}_T^2 = a_x\hat{\sigma}_x^2 + a_y\hat{\sigma}_y^2,$$

чтобы оценить неизвестную дисперсию σ^2.

а) При каком условии на a_x и a_y оценка $\hat{\sigma}_T^2$ будет несмещённой?

б) При каком условии на a_x и a_y оценка $\hat{\sigma}_T^2$ будет состоятельной?

Дополнительно предположим, что исходные две выборки имеют нормальное распределение.

в) При каком условии на a_x и a_y оценка $\hat{\sigma}_T^2$ будет несмещённой с наименьшей дисперсией?

г) При каком условии на a_x и a_y оценка $\hat{\sigma}_T^2$ будет иметь распределение Стьюдента?

7.21 Величины X_1, X_2, ..., X_n независимы и равномерно распределены $\mathrm{U}[0; a]$ с неизвестным параметром a.

Определим величину $R = \max\{X_1, \ldots, X_n\}$.

а) Найдите функцию распределения $F_R(t)$ и функцию плотности $f_R(t)$.

б) Найдите $\mathbb{E}(R)$ и $\mathbb{V}\mathrm{ar}(R)$.

Рассмотрим несколько оценок для параметра a.

$$\hat{a}_1 = 2\bar{X}, \quad \hat{a}_2 = \frac{2(n+1)}{n}\bar{X}, \hat{a}_3 = R, \quad \hat{a}_4 = \frac{n+1}{n}R.$$

в) Какие из предложенных оценок являются несмещёнными?

г) Какие из предложенных оценок являются состоятельными?

7.22 Величины $X_1, X_2, \ldots, X_n$ независимы и экспоненциально распределены $\mathrm{Expo}(\lambda)$ с неизвестной интенсивностью λ.

Рассмотрим три оценки для параметра λ.

$$\hat{\lambda}_1 = \bar{X}, \quad \hat{\lambda}_2 = \frac{1}{\bar{X}}, \quad \hat{\lambda}_3 = \frac{n+1}{X_1 + \ldots + X_n}$$

Какие из предложенных оценок являются состоятельными?

7.23 Величины $X_1, X_2, \ldots, X_n$ независимы и нормально распределены $\mathcal{N}(\mu; \theta - \mu^2)$ с неизвестными параметрами μ и θ.

Рассмотрим три оценки для параметра θ.

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n X_i^2}{n-1}, \quad \hat{\theta}_2 = \frac{\sum_{i=1}^n X_i^2}{n}$$

$$\hat{\theta}_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} + (\bar{X})^2$$

а) Какие из предложенных оценок являются несмещёнными?

б) Какие из предложенных оценок являются состоятельными?

7.24 Величины $X_1, X_2, \ldots, X_n$ независимы с функцией распределения

$$F(x) = \begin{cases} \exp(x-a), \text{ если } x \leqslant a, \\ 1, \text{ иначе.} \end{cases}$$

а) Нарисуйте функцию распределения $F(x)$ и функцию плотности $f(x)$.

б) Какие значения принимают случайные величины X_i?

Рассмотрим величину $X_{\max} = \max\{X_1, \ldots, X_n\}$.

в) Найдите функцию распределения $F_{\max}(x)$ и функцию плотности $f_{\max}(x)$.

г) Найдите $\mathbb{E}(X_{\max})$.

д) Найдите $\mathbb{V}\mathrm{ar}(X_{\max})$.

е) Является ли величина $X_{\max}$ несмещённой оценкой для a?

ж) Является ли величина $X_{\max}$ состоятельной оценкой для a?

з) Придумайте несмещённую оценку $\hat{a}$ на основе величины $X_{\max}$.

7.25 Величины $X_1, X_2, \ldots, X_n$ независимы с функцией распределения

$$F(x) = \begin{cases} (x/a)^3, \text{ если } x \in [0; a], \\ 0, \text{ если } x < 0, \\ 1, \text{ если } x > a. \end{cases}$$

а) Нарисуйте функцию распределения $F(x)$ и функцию плотности $f(x)$.

б) Какие значения принимают случайные величины X_i?

Рассмотрим величину $X_{\max} = \max\{X_1, \ldots, X_n\}$.

в) Найдите функцию распределения $F_{\max}(x)$ и функцию плотности $f_{\max}(x)$.

г) Найдите $\mathbb{E}(X_{\max})$.

д) Найдите $\mathbb{V}\mathrm{ar}(X_{\max})$.

е) Является ли величина $X_{\max}$ несмещённой оценкой для a?

ж) Является ли величина $X_{\max}$ состоятельной оценкой для a?

з) Придумайте несмещённую оценку $\hat{a}$ на основе величины $X_{\max}$.

8. Идеальных оценок не бывает

Свойства производной лог-правдоподобия $S_1 = \frac{\partial \ln f(X_1)}{\partial \theta}$:

$$\mathbb{E}(S_1) = 0, \quad \mathbb{V}\mathrm{ar}(S_1) = \mathbb{E}(S_1^2) = -\mathbb{E}\left(\frac{\partial S_1}{\partial \theta}\right), \quad \mathbb{C}\mathrm{ov}(S_1, \hat{\theta}) = \frac{\partial \mathbb{E}(S_1)}{\partial \theta}.$$

Информация Фишера о неизвестном параметре θ, содержащаяся в одном наблюдении:

$$I_1(\theta) = \mathbb{V}\mathrm{ar}\left(\frac{\partial \ln f(X_1)}{\partial \theta}\right) = \mathbb{E}\left(\left(\frac{\partial \ln f(X_1)}{\partial \theta}\right)^2\right) = -\mathbb{E}\left(\frac{\partial^2 \ln f(X_1)}{\partial \theta^2}\right)$$

Для независимых наблюдений $I_n(\theta) = nI_1(\theta)$.

Неравенство Рао — Крамера («слишком хорошей оценки не бывает»)

$$\mathbb{V}\mathrm{ar}(\hat{\theta}) \geqslant \frac{\left(\mathbb{C}\mathrm{ov}\left(\frac{\partial \ln f(X_1,\ldots,X_n)}{\partial \theta}, \hat{\theta}\right)\right)^2}{I_n(\theta)} = \frac{\left(\frac{\partial \mathbb{E}(\hat{\theta})}{\partial \theta}\right)^2}{I_n(\theta)}$$

8.1 Величины X_1, X_2, ..., X_n независимы и одинаково распределены.

x	1	2	3
$\mathbb{P}(X_i = x)$	θ	2θ	$1-3\theta$

Обозначим вероятность $\mathbb{P}(X_i = x)$ с помощью функции $f(x)$. Рассмотрим величины $S_i = \frac{\partial \ln f(X_i)}{\partial \theta}$.

а) Какие значения с какими вероятностями принимает величина S_1?

б) Найдите $\mathbb{E}(S_1)$, $\mathbb{E}(S_1^2)$, $\mathbb{V}\mathrm{ar}(S_1)$.

в) Докажите, что $\mathbb{E}(S_1) = 0$ при любых вероятностях $f(x)$.

г) Какие значения с какими вероятностями принимает величина $\frac{\partial S_1}{\partial \theta}$?

д) Найдите $\mathbb{E}\left(\frac{\partial S_1}{\partial \theta}\right)$.

е) Найдите информацию Фишера о неизвестном параметре, содержащуюся в одном наблюдении, $I_1(\theta)$.

ж) Найдите информацию Фишера о неизвестном параметре, содержащуюся в n наблюдениях, $I_n(\theta)$.

з) Найдите минимально возможную дисперсию оценки $\hat{\theta}_A$ по n наблюдениям с $\mathbb{E}(\hat{\theta}_A) = 0.2 + 0.9\theta$.

и) Найдите минимально возможную дисперсию несмещённой оценки $\hat{\theta}_B$ по n наблюдениям.

8.2 Величины X_1, X_2, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} ax^{a-1}, \text{ если } x \in [0;1], \\ \text{иначе.} \end{cases}$$

Рассмотрим величины $S_i = \frac{\partial \ln f(X_i)}{\partial a}$.

а) Какие значения принимает величина S_1?

б) Найдите $\mathbb{E}(S_1)$, $\mathbb{E}(S_1^2)$, $\mathbb{V}\text{ar}(S_1)$.

в) Какие значения принимает величина $\frac{\partial S_1}{\partial a}$?

г) Найдите $\mathbb{E}\left(\frac{\partial S_1}{\partial a}\right)$.

д) Найдите информацию Фишера о неизвестном параметре, содержащуюся в одном наблюдении, $I_1(a)$.

е) Найдите информацию Фишера о неизвестном параметре, содержащуюся в n наблюдениях, $I_n(a)$.

ж) Найдите минимально возможную дисперсию оценки $\hat{a}_1$ по n наблюдениям с $\mathbb{E}(\hat{a}_1) = 0.3 + 0.9a$.

з) Найдите минимально возможную дисперсию несмещённой оценки $\hat{a}_2$ по n наблюдениям.

8.3 Величины X_1, X_2, ..., X_n независимы и экспоненциально распределены с интенсивностью $\lambda = 1/a$.

Обозначим функцию плотности каждой величины X_i с помощью $f(x)$. Рассмотрим величины $S_i = \frac{\partial \ln f(X_i)}{\partial a}$.

а) Найдите $\mathbb{E}(X_1)$, $\mathbb{V}\text{ar}(X_1)$.

б) Какие значения принимает величина S_1?

в) Найдите $\mathbb{E}(S_1)$, $\mathbb{E}(S_1^2)$, $\mathbb{V}\text{ar}(S_1)$.

г) Какие значения принимает величина $\frac{\partial S_1}{\partial a}$?

д) Найдите $\mathbb{E}\left(\frac{\partial S_1}{\partial a}\right)$.

е) Найдите информацию Фишера о неизвестном параметре, содержащуюся в одном наблюдении, $I_1(a)$.

ж) Найдите информацию Фишера о неизвестном параметре, содержащуюся в n наблюдениях, $I_n(a)$.

з) Найдите минимально возможную дисперсию оценки $\hat{a}_1$ по n наблюдениям с $\mathbb{E}(\hat{a}_1) = 0.3 + 0.9a$.

и) Найдите минимально возможную дисперсию несмещённой оценки $\hat{a}_2$ по n наблюдениям.

8.4 Величины X_1, X_2, ..., X_n независимы и одинаково распределены.

x	1	2
$\mathbb{P}(X_i = x)$	θ	$1-\theta$

а) Найдите $\mathbb{E}(X_1)$, $\mathbb{V}\text{ar}(X_1)$, $\mathbb{E}(\bar{X})$, $\mathbb{V}\text{ar}(\bar{X})$.

б) Является ли оценка $\bar{X}$ несмещённой для θ?

в) Является ли оценка $\bar{X}$ состоятельной для θ?

г) Найдите информацию Фишера в одном и в n наблюдениях, $I_1(\theta)$, $I_n(\theta)$.

д) Достигает ли оценка $\bar{X}$ границы Рао — Крамера?

е) Можно ли придумать другую оценку $\hat{\theta}$ с $\mathbb{E}(\hat{\theta}) = \mathbb{E}(\bar{X})$ и $\mathbb{V}\text{ar}(\hat{\theta}) < \mathbb{V}\text{ar}(\bar{X})$?

8.5 Величины X_1, X_2, ..., X_n независимы и нормально распределены $\mathcal{N}(0.9a, 1)$.

а) Найдите $\mathbb{E}(X_1)$, $\mathbb{V}\text{ar}(X_1)$, $\mathbb{E}(\bar{X})$, $\mathbb{V}\text{ar}(\bar{X})$.

б) Является ли оценка $\bar{X}$ несмещённой для a?

в) Является ли оценка $\bar{X}$ состоятельной для a?

г) Найдите информацию Фишера в одном и в n наблюдениях, $I_1(a)$, $I_n(a)$.

д) Достигает ли оценка $\bar{X}$ границы Рао — Крамера?

е) Можно ли придумать другую оценку $\hat{\theta}$ с $\mathbb{E}(\hat{a}) = \mathbb{E}(\bar{X})$ и $\mathbb{V}\text{ar}(\hat{a}) < \mathbb{V}\text{ar}(\bar{X})$?

8.6 Величины X_1, X_2, ..., X_n независимы и имеют геометрическое распределение Geom$(p = 1/a)$. Напомним, что у геометрического распределения $\mathbb{P}(X_i = k) = p(1-p)^{k-1}$ при $k \in \{1, 2, 3, \ldots\}$.

а) Найдите $\mathbb{E}(X_1)$, $\mathbb{V}\text{ar}(X_1)$, $\mathbb{E}(\bar{X})$, $\mathbb{V}\text{ar}(\bar{X})$ как функции от a.

б) Является ли оценка $\bar{X}$ несмещённой для a?

в) Является ли оценка $\bar{X}$ состоятельной для a?

г) Найдите информацию Фишера в одном и в n наблюдениях, $I_1(a)$, $I_n(a)$.

д) Достигает ли оценка $\bar{X}$ границы Рао — Крамера?

е) Можно ли придумать другую оценку $\hat{\theta}$ с $\mathbb{E}(\hat{a}) = \mathbb{E}(\bar{X})$ и $\mathbb{V}\text{ar}(\hat{a}) < \mathbb{V}\text{ar}(\bar{X})$?

9. Классические интервальные оценки

Здесь интервальные оценки без гипотез :)

9.1 Цыганка Роза ничего не понимает в статистике, но у неё всегда с собой колода из 36 карт.

Как научить цыганку Розу строить точный 95%-й доверительный интервал для неизвестной вероятности p того, что клиента ждёт дальняя дорога и казённый дом?

9.2 Кржемелик и Вахмурка прекрасно разбираются в математической статистике. Каждый из них построил 95%-й доверительный интервал для одной и той же неизвестной вероятности p. У Кржемелика получился интервал от 0.1 до 0.4, а у Вахмурки — от 0.6 до 0.9.
Возможно ли, что оба они правы?

9.3 Величина X распределена равномерно на отрезке $[0; 2a]$.

а) С какой вероятностью интервал $[0.9X; 1.1X]$ накрывает неизвестное a?

б) Постройте 95%-й доверительный интервал для a вида $[0; kX]$.

9.4 Величина X распределена экспоненциально с интенсивностью λ и ожиданием $\mu = \mathbb{E}(X)$.

а) С какой вероятностью интервал $[0.9X; 1.1X]$ накрывает μ?

б) Постройте 90%-й доверительный интервал для μ вида $[0; kX]$.

9.5 Случайные величины X_i независимы и нормально распределены $\mathcal{N}(\mu, 1)$.

а) Какова вероятность того, что интервал $[\bar{X}_{10} - 1; \bar{X}_{10} + 1]$ накроет неизвестное μ?

б) Постройте 90%-й доверительный интервал для μ вида $[\bar{X}_{10} - k; \bar{X}_{10} + k]$.

9.6 Вася наугад поймал 400 покемонов-девочек и 100 покемонов-мальчиков. Среди девочек 250 оказались ядовитыми, среди мальчиков — 60.

а) Найдите точечную оценку $\hat{p}$ для доли ядовитых покемонов среди мальчиков.

б) Найдите точечную оценку $\hat{\sigma}^2$ для дисперсии X_i, где X_i — индикатор ядовитости i-го покемона-мальчика.

в) Постройте 95%-й доверительный интервал для доли ядовитых покемонов среди мальчиков.

г) Постройте 95%-й доверительный интервал для доли ядовитых покемонов среди девочек.

д) Постройте 95%-й доверительный интервал для разницы долей ядовитых покемонов среди девочек и мальчиков.

е) Нужно ли предположение о нормальности X_i и Y_i для решения предыдущих пунктов?

9.7 Вася наугад поймал 400 покемонов-девочек и 100 покемонов-мальчиков. Средний рост покемонов-девочек равен 0.9 метра, и сумма квадратов ростов равна $\sum Y_i^2 = 1000$. Для покемонов мальчиков средний рост равен 1 метру, а сумма квадратов ростов равна $\sum X_i^2 = 2000$.

а) Постройте точечную оценку для ожидания роста покемона-мальчика μ_X.

б) Постройте точечную оценку для дисперсии роста покемона-мальчика σ_X^2.

в) Постройте 95%-й доверительный интервал для ожидаемого роста покемона-мальчика.

г) Постройте 95%-й доверительный интервал для ожидаемого роста покемона-девочки.

д) Постройте 95%-й доверительный интервал для разницы ожидаемых ростов покемона-мальчика и покемона-девочки.

е) Нужно ли предположение о нормальности X_i и Y_i для решения предыдущих пунктов?

9.8 В одном тропическом лесу длина удавов имеет нормальное распределение $\mathcal{N}(\mu, \sigma^2)$. По случайной выборке из 10 удавов оказалось, что $\sum Y_i = 20$ метрам, а $\sum Y_i^2 = 1000$.

а) Постройте 95%-й доверительный интервал для μ.

б) Постройте 95%-й доверительный интервал для σ^2.

в) Важна ли предпосылка о нормальности при решении предыдущих пунктов?

г) (*) Будут ли построенные интервалы иметь асимптотическую вероятность накрытия 95%, если Y_i не будут нормальными, но будут независимы и одинаково распределены с конечными моментами?

9.9 В одном тропическом лесу водятся удавы и питоны. Длина удавов имеет нормальное распределение $\mathcal{N}(\mu_X, \sigma_X^2)$. По выборке из 10 удавов оказалось, что $\sum X_i = 20$ метрам, а $\sum X_i^2 = 1000$. Длина питонов имеет нормальное распределение $\mathcal{N}(\mu_Y, \sigma_Y^2)$. По выборке из 20 питонов оказалось, что $\sum Y_i = 60$ метрам, а $\sum Y_i^2 = 4000$.

а) Анаконда — это питон или удав?

б) Постройте точечные оценки для μ_X, σ_X^2, μ_Y, σ_Y^2.

в) Постройте 95%-й доверительный интервал для σ_X^2/σ_Y^2.

г) Постройте 95%-й доверительный интервал для разницы $\mu_X - \mu_Y$ предполагая равенство дисперсий $\sigma_X^2 = \sigma_Y^2$.

д) Важна ли предпосылка о нормальности при решении предыдущих пунктов?

9.10 В реальности ни дисперсия, ни ожидание никогда не известны. Однако в мире розовых пони происходят чудеса. Там живут 50 розовых пони. Их веса X_1, X_2, ..., X_{50} можно считать независимыми и нормально распределенными случайными величинами, $\mathcal{N}(\mu; \sigma^2)$.

а) Постройте 95%-й доверительный интервал для μ. Величины μ и σ неизвестны, $\bar{X} = 60$, а несмещённая оценка дисперсии равна 16.

б) Постройте 95%-й доверительный интервал для σ. Величины μ и σ неизвестны, $\bar{X} = 60$, а несмещённая оценка дисперсии равна 16.

в) Постройте 95%-й доверительный интервал для μ. Величина μ неизвестна, $\bar{X} = 60$ и $\sigma = 4$.

г) Постройте 95%-й доверительный интервал для σ. Величина σ неизвестна, $\mu = 60$, а несмещённая оценка дисперсии равна 16.

9.11 Величины X_i независимы и нормально распределены $\mathcal{N}(\mu, 1)$. Кальямпуди Радхакришна Рао и Карл Харальд Крамер строят 95%-й доверительный интервал для μ по одной и той же выборке.

Рао строит интервал, зная дисперсию, а Крамер — не зная.

а) Правда ли, что центры доверительных интервалов совпадают?

б) Известно, что $\hat{\mu}$ оказалось очень-очень далеко от μ. У кого больше условная вероятность накрыть интервалом истинное значение μ?

в) Известно, что $\hat{\mu}$ оказолось очень-очень близко к μ. У кого больше условная вероятность накрыть интервалом истинное значение μ?

г) Правда ли, что Рао всегда строит более короткие доверительные интервалы в силу того, что обладает большей информацией?

д) Какова вероятность того, что Рао получит более короткий доверительный интервал при $n = 10$? К чему стремится эта вероятность при $n \to \infty$?

е) Что больше, $\mathbb{E}(\hat{\sigma}^2)$ или σ^2?

ж) Что больше, $\mathbb{E}(\hat{\sigma})$ или σ?

з) С помощью Монте-Карло экспериментов постройте график ожидаемой длины интервалов Рао и Крамера как функции от n в диапазоне $10 \leqslant n \leqslant 10000$.

9.12 Клиента просят оценить уровень сервиса по целочисленной шкале от 0 до 10. Традиционно критиками называют тех, кто оценивает услугу не выше 6, нейтральными тех, кто оценивает услугу на 7 или 8, и сторонниками тех, кто оценивает услугу на 9 или 10.

Допустим, что истинные доли критиков, нейтральных и сторонников равны π_-, π_0 и π_+.

а) Предложите точечную оценку для $\Delta = \pi_+ - \pi_-$. Этот индекс потребительской лояльности называется NPS (Net Promoter Score).

б) Как построить 95%-й асимптотический интервал для Δ?

10. Бутстрэп

Классные источники про бутстрэп: Тим Хестерберг, Что стоит знать преподавателю про бутстрэп, [Hes15], практический гид от аналитиков VK, [SM20], гид для медиков, [CB00].

10.1 Исходная выборка y — вектор из n независимых случайных величин, равновероятно принимающих значения 0 и 1. Пусть y^* — одна из бутстэп-выборок.

а) Просто для удобства выпишите $\mathbb{E}(y_i)$, $\mathbb{V}\text{ar}(y_i)$, $\mathbb{E}(\bar{y})$, $\mathbb{V}\text{ar}(\bar{y})$.

б) Найдите $\mathbb{E}(y_i^*)$, $\mathbb{V}\text{ar}(y_i^*)$, $\mathbb{E}(\bar{y}^*)$, $\mathbb{V}\text{ar}(\bar{y}^*)$.

в) Найдите $\mathbb{C}\text{ov}(y_i, y_i^*)$, $\mathbb{C}\text{ov}(\bar{y}, \bar{y}^*)$.

10.2 Исходная выборка y — вектор из n независимых случайных величин, равноверомерно принимающих значения из отрезка $[0; 1]$. Пусть y^* — одна из бутстэп-выборок.

а) Просто для удобства выпишите $\mathbb{E}(y_i)$, $\mathbb{V}\text{ar}(y_i)$, $\mathbb{E}(\bar{y})$, $\mathbb{V}\text{ar}(\bar{y})$.

б) Найдите $\mathbb{E}(y_i^*)$, $\mathbb{V}\text{ar}(y_i^*)$, $\mathbb{E}(\bar{y}^*)$, $\mathbb{V}\text{ar}(\bar{y}^*)$.

в) Найдите $\mathbb{C}\text{ov}(y_i, y_i^*)$, $\mathbb{C}\text{ov}(\bar{y}, \bar{y}^*)$.

10.3 Подробно опишите, как применять перцентильный бутстрэп для построения 99%-го доверительного интервала для каждого из неизвестных параметров.

а) $\mu = \mathbb{E}(X_i)$ по выборке $X_1, \ldots, X_n$;

б) $\sigma^2 = \mathbb{V}\text{ar}(X_i)$ по выборке $X_1, \ldots, X_n$;

в) $v = \sigma/\mu$ по выборке из неотрицательных $X_1, \ldots, X_n$;

г) $m = \text{Med}(X_i)$ по выборке $X_1, \ldots, X_n$;

д) $p = \mathbb{P}(X_i = 1)$ по выборке из берннулиевских $X_1, \ldots, X_n$;

е) $\mu_X - \mu_Y = \mathbb{E}(X_i) - \mathbb{E}(Y_i)$ по выборкам $X_1, \ldots, X_{n_X}$ и $Y_1, \ldots, Y_{n_Y}$;

ж) $m_X - m_Y = \text{Med}(X_i) - \text{Med}(Y_i)$ по выборкам $X_1, \ldots, X_{n_X}$ и $Y_1, \ldots, Y_{n_Y}$;

з) $p_X - p_Y = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1)$ по выборкам из берннулиевских $X_1, \ldots, X_{n_X}$ и $Y_1, \ldots, Y_{n_Y}$;

и) $\rho = \text{Corr}(X_i, Y_i)$ по выборкам $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$;

к) $c = \mathbb{C}\text{ov}(X_i, Y_i)$ по выборкам $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$;

10.4 Подробно опишите, как применять бутстрэп t-статистики для построения 99%-го доверительного интервала для каждого из неизвестных параметров.

а) $\mu = \mathbb{E}(X_i)$ по выборке $X_1, \ldots, X_n$;

б) $p = \mathbb{P}(X_i = 1)$ по выборке из берннулиевских $X_1, \ldots, X_n$;

в) $\mu_X - \mu_Y = \mathbb{E}(X_i) - \mathbb{E}(Y_i)$ по выборкам $X_1, \ldots, X_{n_X}$ и $Y_1, \ldots, Y_{n_Y}$;

г) $p_X - p_Y = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1)$ по выборкам из берннулиевских $X_1, \ldots, X_{n_X}$ и $Y_1, \ldots, Y_{n_Y}$.

10.5 Винни-Пух продаёт мёд правильных пчёл и проверяет гипотезу о том, что ворчалки и пыхтелки одинаково влияют на сумму покупок посетителей его заведения. Для этого Винни-Пух проводит свитчбэк эксперимент, через каждые 51 минуту он подкидывает монетку, если монетка падает орлом, то он включает ворчалки, а если решкой — то пыхтелки. Собрав данные по сумме покупок клиентов за продолжительный период времени, Винни-Пух призадумался, как ему лучше генерировать бутстрэп-выборки при проведении перестановочного теста.

Первый вариант: случайно присваивать клиентам метки контрольной и экспериментальной группы. Второй вариант: случайно присваивать метки интервалам времени. В первом варианте в бутстрэп-выборке клиенты, де-факто пришедшие в один свитчбэк интервал, могут получить разные метки. Во втором варианте клиенты, пришедшие в один свитчбэк интервал, обязательно получат одну и ту же метку в бутстрэп-выборке.

Как лучше поступить Винни-Пуху и почему?

11. Снижение дисперсии

Про cuped полезно прочитать [Den+13].

11.1. Стратифицированая выборка

11.2. DnD

11.3. cuped

12. Проверка гипотез: общая теория

12.1 Как распределено P-значение при верной H_0? Вовочка использует следующий статистический критерий: «Если P-значение больше 0.95, то H_0 отвергается». Чему де-факто равна вероятность ошибки первого рода в этом случае? Разумно ли использовать данный критерий?

12.2 Величины X_1 и X_2 независимы и равномерны на отрезке $[0; a]$. Есть две гипотезы, H_0: $a = 1$ и H_a: $a = 2$. Мальвина отвергает H_0 в том случае, если $X_1 + X_2 > 1.5$. Найдите вероятность ошибок первого и второго рода.

12.3 Величины X_1 и X_2 независимы и нормальны $\mathcal{N}(a; 1)$. Есть две гипотезы, H_0: $a = 1$ и H_a: $a = 2$. Мальвина отвергает H_0 в том случае, если $X_1 + X_2 > 1.5$. Найдите вероятность ошибок первого и второго рода.

12.4 Величины X_1 и X_2 независимы и распределены экспоненциально с интенсивностью a. Есть две гипотезы, H_0: $a = 1$ и H_a: $a = 2$. Мальвина отвергает H_0 в том случае, если $\min\{X_1, X_2\} < 1$. Найдите вероятность ошибок первого и второго рода.

12.5 Величины X_1 и X_2 независимы и распределены по Пуассону с интенсивностью a. Есть две гипотезы, H_0: $a = 1$ и H_a: $a = 2$. Мальвина отвергает H_0 в том случае, если $X_1 + X_2 \geqslant 2$. Найдите вероятность ошибок первого и второго рода.

12.6 Бабушка Аксинья утверждает, что обладает сверхспособностями и умеет слышать Внутренний Голос. Стоит лишь Аксинье глянуть на стакан с водой, как Внутренний Голос подсказывает, что налито в стакан, Бон-Аква или Аква Минерале.

Исследователь Кирилл проводит с Аксиньей следующий опыт. Правила опыта Аксинье известны. Кирилл в тайне от Аксиньи наливает в 3 стакана Акву Минерале и 2 стакана Бон-Аквы. Затем предлагает их Аксинье в случайной порядке. Задача Аксиньи после осмотра всех стаканов определить, в каком порядке они предлагались.

Кирилл проверяет гипотезу H_0 против H_a. Гипотеза H_0 состоит в том, что Аксинья не обладает сверхспособностями и её Внутренний Голос верно определяет содержимое каждого стакана с вероятностью $p = 0.5$. Альтернативная гипотеза H_a состоит в том, что Аксинья обладает сверхспособностями и $p = 0.9$.

Кирилл использует следующий критерий: если Аксинья ошиблась хотя бы один раз, то H_0 не отвергается; если не ошиблась ни разу, то H_0 отвергается.

а) Предположим, Аксинья, не задумываясь, говорит ровно то, что подсказывает ей Внутренний Голос. Найдите вероятности ошибок первого и второго рода.

б) (*) Предположим, Аксинья осознаёт что, Внутренний Голос может ошибаться с вероятностью $1 - p$. Поэтому при очевидной ошибке Внутренного Голоса старается её исправить. Например, если Внутренний Голос говорит ААААБ, то, следуя ему, угадать все стаканы невозможно. Найдите вероятности ошибок первого и второго рода.

12.7 В Нюрнберге в 1835 году «Общество мужчин любителей истины» провело один из первых двойных слепых медицинских экспериментов.

В то время в Баварии среди высшей аристократии было очень популярно лечение гомеопатией. Противник гомеопатии глава госпиталей Нюрнберга Фридрих Вильгельм фон Ховен (Friedrich Wilhelm von Hoven) публично выступил против сторонника гомеопатии врача Иоганна Якоба Рейтера (Johann Jacob Reuter) . Они договорились проверить утверждение Якоба Рейера, что с шансами 10:1 человек, выпивший гомеопатический раствор соли, почувствует что-то необычное.

Заранее были заготовлены 100 занумерованных пузырьков. Пузырьки занумеровали и расставлены в случайном порядке, затем в 50 из них налили дистилированную талую воду, а в 50 оставшихся — соль в разведении C100 (песчинка соли разводится в 100 мл воды, далее полученная смесь разводится 29 раз в соотношении 1 к 100). Соответствие номера содержимому запротоколировали, и протокол опечатали.

В таверне Красный Петух собрались 120 уважаемых жителей города, 47 добровольцев согласились участвовать и получили занумерованные пузырьки в таверне. Ещё на 7 пузырьков добровольцы нашлись после собрания в таверне.

Три недели спустя, 12 марта 1835 года, добровольцев попросили сообщить, ощущали ли они что-то необычное. Ответы были получены от 50 человек. Из восьми человек, ощущавших что-то необычное, трое получили простую воду, пятеро — разведённую соль.

а) Сформулируйте H_0 и H_a простыми словами.

б) Максимально аккуратно сформулируйте H_0 и H_a математически.

в) Укажите распределение наблюдаемых величин при верной H_0.

г) Проверьте гипотезу H_0, найтие точное p-значение.

12.8 Бариста Борис заметил, что в последнее время посетители заказывают только капуччино и раф. Предположим, что посетители выбирают напиток независимо друг от друга, а вероятность выбора капуччино постоянна и равна неизвестному числу p.

У Бориса есть только две гипотезы, $H_0 : p = 1/3$ и $H_a : p = 2/3$, в которые он до получения данных верит с вероятностями 0.6 и 0.4, соответственно.

Из первых 100 утренних посетителей $S = 40$ выбрали капуччино. Борис хочет измерить разными способами, насколько этот наблюдаемый результат соотносится с гипотезами.

а) Найдите $\mathbb{P}(H_0 \mid S = 40)$ и $\mathbb{P}(H_a \mid S = 40)$.

Борис решил на следующий день повторить эксперимент и снова посчитать S_{new}, количество клиентов из первых ста, которые выберут капуччино.

б) Найдите $\mathbb{P}(S_{\text{new}} \geqslant S \mid S = 40, H_0)$ и $\mathbb{P}(S_{\text{new}} \geqslant S \mid S = 40, H_a)$.

в) Какие из вероятностей можно посчитать без мнения Бориса о $\mathbb{P}(H_0)$ и $\mathbb{P}(H_a)$?

г) Какая из вероятностей называется p-значением для гипотезы H_0 и статистики S?

Примечание: можно использовать нормальную аппроксимацию или приложение для вычисления вероятностей, связанных с биномиальным распределением.

12.9 Бариста Борис заметил, что в последнее время посетители заказывают только капуччино и раф. Предположим, что посетители выбирают напиток независимо друг от друга, а вероятность выбора капуччино постоянна и равна неизвестному числу p.

У Бориса есть только две гипотезы, $H_0 : p = 1/3$ и $H_a : p = 2/3$, в которые он до получения данных верит с вероятностями 0.6 и 0.4, соответственно.

Борис решил использовать простой критерей: если из первой сотни посетителей больше 40 выберут капуччино, то он отвергнет H_0 в пользу H_a, а иначе — не отвергнет H_0.

а) Найдите $\mathbb{P}(H_0$ будет отвергнута $\mid H_0$ верна$)$ и $\mathbb{P}(H_0$ будет отвергнута $\mid H_a$ верна$)$.

б) Найдите $\mathbb{P}(H_0$ верна $\mid H_0$ отвергнута$)$ и $\mathbb{P}(H_0$ верна $\mid H_0$ не отвергнута$)$.

в) Найдите $\mathbb{P}(H_0$ будет отвергнута $)$, $\mathbb{P}($Борис примет ошибочное решение$)$.

г) Какие из вероятностей можно посчитать без мнения Бориса о $\mathbb{P}(H_0)$ и $\mathbb{P}(H_a)$?

д) Какая из вероятностей называется уровнем значимости теста? Частотой ложного отвержения нулевой гипотезы (FDR)?

Примечание: можно использовать нормальную аппроксимацию или приложение для вычисления вероятностей, связанных с биномиальным распределением.

12.10 У Андрея Николаевича есть выборка $X_1, X_2, \ldots, X_n$. Каждая величина X_i имеет непрерывную функцию распределения F.

а) Найдите закон распределения $Y_i = F(X_i)$ и закон распределения $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$.

б) Рассмотрим простейшую выборку из одного наблюдения. Найдите закон распределения статистики Колмогорова $D_1 = \sup \left|\hat{F}_1(x) - F(x)\right|$ при верной нулевой гипотезе о том, что истинный закон распределения задан функцией распределения F. Найдите квантиль порядка 0.95 для D_1.

в) Рассмотрим выборку из двух наблюдений. Найдите квантиль порядка 0.95 для статистики Колмогорова $D_2 = \sup \left|\hat{F}_2(x) - F(x)\right|$ при верной нулевой гипотезе о том, что истинный закон распределения задан функцией распределения F.

Лемма Неймана-Пирсона

12.11 Величины $X_1, \ldots, X_n$ независимы и нормально распределены с неизвестным математическим ожиданием μ и известной дисперсией $\sigma^2 = 4$.

Объем выборки $n = 16$. Тестируются основная гипотеза $H_0 : \mu = 0$ против альтернативной гипотезы $H_a : \mu = 2$. С помощью леммы Неймана–Пирсона найдите наиболее мощный критерий, имеющий уровень значимости $\alpha = 0.05$.

12.12 Величины X_1 и X_2 независимы и одинаково распределены на отрезке $[0;1]$. Есть две гипотезы, H_0: $X_i \sim U[0;1]$ и H_a: $f(x_i) = \begin{cases} 2x_i, & \text{если } x_i \in [0;1]; \\ 0, & \text{иначе} \end{cases}$. С помощью леммы Неймана–Пирсона найдите наиболее мощный критерий, имеющий уровень значимости $\alpha = 0.05$.

12.13 Рассмотрим две величины, X_1 и X_2. Есть две гипотезы, H_0: величины X_i независимы и равномерно распределены $U[0;1]$ и H_a: $f(x_1, x_2) = \begin{cases} x_1 + x_2, & \text{если } x_1, x_2 \in [0;1]; \\ 0, & \text{иначе} \end{cases}$. С помощью леммы Неймана–Пирсона найдите наиболее мощный критерий, имеющий уровень значимости $\alpha = 0.05$.

12.14 Величины $X_1, \ldots, X_n$ — случайная выборка из нормального распределения с известным математическим ожиданием $\mu = 1$ и неизвестной дисперсией σ^2.

Объем выборки $n = 16$. Тестируются основная гипотеза $H_0 : \sigma^2 = 4$ против альтернативной гипотезы $H_a : \sigma^2 = 9$. С помощью леммы Неймана–Пирсона найдите наиболее мощный критерий, имеющий уровень значимости $\alpha = 0.05$.

12.15 Количество мёда, которое вырабатывают пчёлы имеет функцию плотности $f(x) = (a-1)/x^a$ при $x \geqslant 1$ и $a > 1$. Известно, что у правильных пчёл $a = 3$, а у неправильных $a = 5$.

Винни-Пух успешно слазил за мёдом всего один раз, поэтому у него есть только одно наблюдение, X_1.

Помогите Винни-Пуху построить наиболее мощный тест для H_0: «мой мёд сделан правильными пчёлами» против H_a: «мой мёд сделан неправильными пчёлами» с вероятностью ошибки первого рода $\alpha = 0.01$.

Какая при этом получится вероятность ошибки второго рода?

12.1. Поиск оптимального n

12.16 У исследователя Пантелеймона n независимых наблюдений. Пантелеймон хочет поделить эти n наблюдений на экспериментальную (t, treatment) и контрольную группу (c, control), чтобы измерить эффект воздействия, $\mu_t - \mu_c$.

Дисперсия каждого наблюдения в контрольной группе равна σ_c^2, а в экспериментальной — σ_t^2.

Для упрощения будем считать, что дисперсии известны.

В каких пропорциях нужно разделить имеющиеся наблюдения?

12.17 Исследователь Пантелеймон разбил наблюдения на две группы, экспериментальную (t, treatment) и контрольную (c, control), по n наблюдений в каждой.

Наблюдения независимы, а дисперсия наблюдений одинакова и равна σ^2.

Пантелеймон проверяет гипотезу H_0: $\mu_t - \mu_c = 0$ против H_a: $\mu_t - \mu_c \neq 0$.

Пантелеймон хочет, чтобы вероятность ошибки первого рода равнялась α, а вероятность ошибки второго рода была равна β при величине эффекта равной $\mu_t - \mu_c = MDE > 0$. Сокращение MDE расшифровывается как minimal detectable effect, минимальные детектируемый эффект.

а) Какое n надо выбрать Пантелеймону?

б) Как изменится ответ, если размер контрольной группы в $r > 1$ раз больше, чем размер экспериментальной? В данном случае за n обозначим размер наименьшей из групп.

Почитать: [Ser+21] — обзор способов нахождения размера выборки.

13. Таблицы сопряжённости

13.1 Каждый день Винни-Пух добывает мёд. Количества правильного мёда, добытого в i-й день в килограммах, Y_i, образуют последовательность независимых и одинаково распределённых величин.

Кролик собрал статистику за 300 дней:

	$Y_i \in [0;1]$	$Y_i \in (1;2]$	$Y_i > 2$
Количество дней	120	130	50

У любопытного Кролика есть несколько гипотез:

H_a: Количество мёда распределено экспоненциально с $\lambda = 1$.

H_b: Количество мёда распределено экспоненциально с некоторым λ.

H_c: Вероятности сбора от 0 до 1 и от 1 до 2 килограмм равны.

H_d: Количество мёда имеет произвольное распределение.

H_e: Вероятности для трёх категорий равны $2/5$, $2/5$ и $1/5$.

а) Оцените вероятности попадания времени ожидания в каждый интервал с помощью максимального правдоподобия.

б) Оцените интенсивность λ с помощью максимального правдоподобия, предполагая, что Y_i распределено экспоненциально.

в) Проверьте гипотезу H_a против H_b.

г) Проверьте гипотезу H_a против H_d.

д) Проверьте гипотезу H_c против H_d.

е) Проверьте гипотезу H_b против H_d.

ж) Проверьте гипотезу H_e против H_d.

з) Проверьте гипотезу H_e против H_c.

и) Постройте доверительный интервал для вероятности попадания в каждый интервал.

к) Постройте доверительный интервал для разницы $\Delta = \mathbb{P}(Y_i \in [0;1]) - \mathbb{P}(Y_i \in [1;2])$.

л) Постройте асимптотический доверительный интервал для $r = \mathbb{P}(Y_i \in [0;1])/\mathbb{P}(Y_i \in [1;2])$.

м) Постройте интервал для $r = \mathbb{P}(Y_i \in [0;1])/\mathbb{P}(Y_i \in [1;2])$ с помощью перцентильного бутстрэпа.

13.2 Помимо широко известной системы групп крови AB0, существуют также и другие, например система MN. В этой системе существует три группы крови: M, N и MN. В гене существует всего два аллеля, M и N, кодоминантные по отношению друг к другу. Обладатели группы крови M имеют гомозиготный генотип MM, обладатели группы N — гомозиготный NN, обладатели группы MN — гетерозиготный MN. Наследование систем групп крови AB0 и MN происходит независимо друг от друга.

В идеальных условиях частоты генотипов должны находится в равновесии Харди-Вайнберга. Частоты генотипов должны быть равны: $p_{NN} = p_N^2$, $p_{MN} = 2p_N(1-p_N)$, $p_{MM} = (1-p_N)^2$, где p_N — частота аллеля N в популяции.

У 50 людей определили их генотипы: оказалось 10 человек с генотипом NN, 20 — с генотипом MM, и 20 — с генотипом MN.

а) С помощью критерия хи-квадрат Пирсона проверьте гипотезу, что условия Харди-Вайнберга выполнены и $p_N = 0.3$.

б) В рамках условий Харди-Вайнберга оцените неизвестную вероятность p_N с помощью максимального правдоподобия.

в) С помощью критерия хи-квадрат Пирсона проверьте гипотезу, что условия Хадри-Вайнберга выполнены.

г) Объясните, откуда берутся условия Харди-Вайнберга.

13.3 Космоохотник поймал n пришельцев. Каждый пришелец независимо от других относится к одному из r видов. Вероятность того, что очередной пришелец принадлежит i-му виду, равна p_i. Обозначим ν_i — количество пришельцев i-го вида.

Космоохотник нормирует частоты ν_i *внеземным* образом: вычитает ожидание и делит на корень из *математического ожидания*. Обозначим нормированные частоты как ν_i^*.

а) Найдите $\mathbb{E}(\nu_i)$, $\mathbb{V}\mathrm{ar}(\nu_i)$, $\mathbb{C}\mathrm{ov}(\nu_i, \nu_j)$.

б) Найдите ковариационную матрицу $V = \mathbb{V}\mathrm{ar}(\nu^*)$.

в) Представьте V в виде $V = I - vv'$, где I — единичная матрица, а v — некий вектор. Как выглядит вектор v и какую длину он имеет?

г) Найдите V' и V^2.

13.4 Глубоко проникнув в тайную суть вещей, почувствуйте теорему Пифагора в равенстве

$$\sum \frac{(a_i - b_i)^2}{b_i} + n = \sum \frac{a_i^2}{b_i},$$

где $n = \sum a_i = \sum b_i$.

а) Докажите быструю формулу нахождения хи-квадрат статистики Пирсона для гипотезы о законе распределения

$$\sum \frac{(\nu_i - np_i)^2}{np_i} = \sum \frac{\nu_i^2}{np_i} - n$$

б) Докажите быструю формулу нахождения хи-квадрат статистики Пирсона для гипотезы о независимости двух признаков

$$\sum \frac{(\nu_{ij} - \nu_{\cdot j}\nu_{i\cdot}/n)^2}{\nu_{\cdot j}\nu_{i\cdot}/n} = \sum \frac{\nu_{ij}^2 n}{\nu_{\cdot j}\nu_{i\cdot}} - n$$

13.5 Сформулируем аксиому адекватности для методов поддержки или опровержения научных теорий:

«Если из жёсткой гипотезы H_S следует мягкая гипотеза H_M, то любой эксперимент, отвергающий мягкую гипотезу H_M, должен отвергать жёсткую гипотезу H_S».

В далеком созвездии Тау-Кита живут инопланетяне трёх полов: A, B и C. Рассмотрим две нулевых гипотезы: H_S: все три пола встречаются равновероятно, и H_M: пол A и пол B встречаются равновероятно.

а) Придумайте данные по количеству инопланетян, а также уровень значимости так, чтобы нулевая гипотеза H_M отвергалась, а нулевая гипотеза H_S не отвергалась. Тем самым докажите, что при тестировании гипотез может нарушаться аксиома адекватности. В качестве альтернативной гипотезы возьмите гипотезу о произвольных вероятностях.

б) Будет ли аксиома адекватности нарушаться, если вне зависимости от H_0, использовать одинаковое число степеней свободы для хи-квадрат распределения?

По мотивам статьи о споре Фишера и Пирсона о числе степеней свободы хи-квадрат теста [Bai83].

13.6 В 1946 году актуарий Р. Д. Кларк опубликовал одну страничку, [Cla46], о результатах своей работы на английскую военную разведку в годы Второй Мировой войны.

С июня 1944 года немцы использовали ракеты фау-1 для бомбардировок Лондона. На Лондон было сброшено более двух тысяч ракет, погибло более пяти тысяч человек. При бомбардировках англичам было важно понять, является ли места падения ракет случайными, или имеет место более точное наведение, http://tiny.cc/london-V1.

Кларк разбил карту Южного Лондона на

14. К нам приполз питон...

14.1 Скачайте оценки по курсу теории вероятностей, https://github.com/bdemeshev/probability_hse_2020_21/raw/main/hse_probability_2020_2021.csv. Возьмём оценки первой и второй групп за первую контрольную работу.

Изобразите результаты графически!

Подсказка: основная идея визуализации при малом количестве наблюдений: показывайте каждую точку! Например, можно воспользоваться питоновским пакетом dabest, https://acclab.github.io/DABEST-python-docs/.

Проблемы при установке: если Windows ругается на запреты, то попробуйте установить dabest только для текущего пользователя, !pip install dabest --user. Если после установки пакет не импортируется, попробуйте перезапустить анаконду.

14.2 При построении каждого интервала выписывайте предпосылки полностью!! Номинальную вероятностью накрытия возьмите равной 95%.

а) Оцените μ_x, μ_y, $\Delta = \mu_x - \mu_y$, σ_x^2, σ_y^2.

б) Оцените вероятность получить отлично, p_x, p_y. Отлично получают те, кто пишет контрольную на 80 баллов и выше.

в) Постройте доверительный интервал для μ_x используя асимптотику.

г) Постройте доверительный интервал для p_x используя асимптотику.

д) Постройте доверительный интервал для μ_x используя t-распределение.

е) Постройте доверительный интервал для $\Delta\mu$ используя асимптотику.

ж) Постройте доверительный интервал для Δp используя асимптотику.

з) Постройте доверительный интервал для $\Delta\mu$ используя t-статистику и предполагая дисперсии равными.

и) Постройте доверительный интервал для $\Delta\mu$ используя тест Уэлча.

к) При построении интервала для Δ обратите внимание, лежит ли 0 в доверительном интервале. Какую гипотезу мы при этом формально проверяем?

л) Почему не надо строить интервал для p_x используя t-распределение?

Подсказка: scipy.stats.t.interval и scipy.stats.norm.interval.

14.3 Бутстрэп!

Номинальную вероятностью накрытия возьмите равной 95%.

а) Постройте доверительный интервал для μ_x используя наивный бутстрэп и t-бутстрэп.

б) Постройте доверительный интервал для σ_x^2 используя наивный бутстрэп и t-бутстрэп.

в) Постройте доверительный интервал для $\Delta\mu$ используя наивный бутстрэп и t-бутстрэп.

г) Постройте доверительный интервал для p_x используя наивный бутстрэп и t-бутстрэп.

д) Постройте доверительный интервал для Δp используя наивный бутстрэп и t-бутстрэп.

В питоне есть несколько пакетов для бутстрэпа. Одним из самых солидных будет arch, https://arch.readthedocs.io/. Он может пригодиться и для моделей волатильности финансовых активов когда-нибудь потом. Но всегда полезно и руками написать :)

Не скупитесь на число бутстрэповских выборок, хорошо — это $10000+$.

14.4 Бутстрэп-симуляции!

Номинальную вероятностью накрытия возьмите равной 95%. Число бутстрэповских выборок $n_{boot} = 10000$, число симуляций для оценки фактической вероятности накрытия — $n_{sim} = 1000$.

а) Сгенерируйте наборы данных по 30 наблюдений. Набор данных А: $X_i \sim \mathcal{N}(\mu = 20; \sigma = 1)$, набор данных B: $X_i \sim 20 + Cauchy()$, набор данных C: $X_i \sim 20 + Exp(\lambda = 1)$.
Не забудьте фиксировать зерно генератора случайных чисел!

б) Какое настоящее математическое ожидание, медиана и дисперсия у X_i в каждой выборке?

в) Где возможно, оцените фактическую вероятность накрытия μ асимптотическим интервалом.

г) Где возможно, оцените фактическую вероятность накрытия μ t-интервалом.

д) Где возможно, оцените фактическую вероятность накрытия μ бутстрэп t-интервалом.

е) Где возможно, оцените фактическую вероятность накрытия μ наивным бутстрэп интервалом.

ж) Оцените фактическую вероятность накрытия $\Delta\mu = \mu_a - \mu_c$ асимптотическим интервалом.

з) Оцените фактическую вероятность накрытия $\Delta\mu = \mu_a - \mu_c$ t-интервалом при равных дисперсиях.

и) Оцените фактическую вероятность накрытия $\Delta\mu = \mu_a - \mu_c$ интервалом Уэлча.

к) Оцените фактическую вероятность накрытия $\Delta\mu = \mu_a - \mu_c$ бутстрэп t-интервалом.

л) Оцените фактическую вероятность накрытия $\Delta\mu = \mu_a - \mu_c$ наивным бутстрэп интервалом.

15. Многомерщина

15.1 Известна ковариационная матрица и математическое ожидание вектора y:

$$\mathbb{E}(y) = \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix}, \mathbb{V}\mathrm{ar}(y) = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 4 & 2 \\ -1 & 2 & 9 \end{pmatrix}$$

Найдите $\mathbb{E}(Ay)$ и $\mathbb{V}\mathrm{ar}(Ay)$, где $A = \begin{pmatrix} 2 & 0 & 2 \\ -1 & 2 & 3 \end{pmatrix}$

15.2 Матрица X имеет размер $n \times k$, при этом матрица $X^T X$ обратима. Рассмотрим матрицу $A = X(X^T X)^{-1} X^T$.

а) Найдите A^2, A^{2017}, A^T.

б) В каком случае матрица A обратима?

15.3 Найдите матрицу A для каждой из ситуаций:

а) Матрица A проецирует n-мерные вектора на вектор $\mathbb{1} = (1, 1, 1, \ldots, 1)^T$.

б) Матрица A проецирует 3-мерные вектора на вектор $(1, 2, 9)^T$.

в) Матрица A проецирует 3-мерные вектора на плоскость, порождённую векторами $(1, 1, 1)^T$ и $(1, 2, 3)^T$.

15.4 Про каждую из матриц проверьте, является ли она проектором. … Для матриц-проекторов определите, на какие вектора они проецируют.

15.5 Известно, что … Найдите закон распределения $y^T B^{-1} y$, $y^T P^y$….

15.6 Определим информацию Фишера как ковариационную матрицу вектора $\frac{\partial \ell}{\partial \theta}$, $I = \mathbb{V}\mathrm{ar}(\frac{\partial \ell}{\partial \theta})$.

а) Найдите $\mathbb{E}(\frac{\partial \ell}{\partial \theta})$.

б) Докажите, что $I = -\mathbb{E}(\frac{\partial^2 \ell}{\partial \theta \partial \theta^T})$.

в) Докажите, что $I = \mathbb{E}(\frac{\partial \ell}{\partial \theta} \frac{\partial \ell}{\partial \theta}^T)$

г) Пусть $I_{n \times n}$ — единичная матрица. Докажите, что матрица $\mathbb{V}\mathrm{ar}(\hat{\theta})\, \mathbb{V}\mathrm{ar}(\frac{\partial \ell}{\partial \theta}) - I_{n \times n}$ является положительно определённой.

16. Самая главная компонента

16.1 Найдите максимум функции $2x^2 + 5y^2 + 3z^2 + 7w^2$ при ограничении $x^2 + y^2 + z^2 + w^2 = 1$.

16.2 Рассмотрим матрицу $A = \begin{pmatrix} 5 & 1 & 0 \\ 1 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$.

а) Найдите собственные векторы и собственные числа матрицы A.

б) Найдите максимум $v^T A v$ при ограничении $v^T v = 1$.

в) Найдите максимум $v^T A v$ при ограничениях $v^T v = 1$ и $v \perp a$, где $a^T = (1, 1, 0)$.

16.3 Вектора z_1 и z_2 имеют выборочную ковариационную матрицу

$$M = \begin{pmatrix} 6 & 2 \\ 2 & 10 \end{pmatrix}.$$

а) Как изменится коварционная матрица, если центрировать вектора z_1 и z_2?

б) Как изменится коварционная матрица, если центрировать и нормировать вектора z_1 и z_2?

16.4 Вектора-столбцы z_1 и z_2 содержат по пять наблюдений. Матрица X состоит из столбцов x_1 и x_2. Выборочная ковариационная матрица векторов z_1 и z_2 равна $\begin{pmatrix} 6 & 2 \\ 2 & 10 \end{pmatrix}$.

а) Найдите матрицу $X^T X$, если вектора x_i получены центрированием векторов z_i.

б) Найдите матрицу $X^T X$, если вектора x_i получены центрированием и нормирование векторов z_i.

16.5 Вениамин находит главные компоненты набора данных из трёх переменных. По каждой из переменных есть 100 наблюдений. Вениамин центрирует и не нормирует переменные, так как они изначально измеряются в одних и тех же единицах.

Собственные числа выборочной ковариационной матрицы исходных переменных равны 5, 4 и 1.

а) Найдите сумму выборочных дисперсий исходных переменных.

б) Найдите длины и выборочные дисперсии всех трёх главных компонент.

в) Какую долю от суммы выборочных дисперсий объясняют первые две главные компоненты?

16.6 Рассмотрим результаты пяти студентов за две контрольные работы:

ФИО	Контрольная 1	Контрольная 2
Маша	4	5
Вася	5	5
Лена	3	4
Коля	5	4
Рита	4	3

а) Найдите выборочную ковариационную матрицу.

б) Найдите собственные числа и собственные векторы единичной длины для ковариационной матрицы.

в) Выпишите первую и вторую главные компоненты.

г) Найдите сумму выборочных дисперсий исходных переменных.

д) Найдите длины и выборочные дисперсии всех главных компонент.

е) Какую долю от суммы выборочных дисперсий объясняет первая главная компоненты?

17. Я не врач!

Здесь задуман раздел по статистике в медицине, но он пока в зародыше :)

17.1 Какие 10 лекарств являются лидерами по объёму продаж в денежном выражении в России за прошлый год? Сколько фаз клинических испытаний каждое из них прошло согласно https://www.drugbank.ca?

17.2

что-то про шансы

17.3

здесь LR тест в качестве аналога MH теста для агрегирования отдельных таблиц сопряженности

17.4

множественное тестирование в ответе в Lancet на испытания Спутника

18. Сэр Томас Байес

Байесовский подход (bayesian approach):

а) Сделать изначальное априорное предположение о распределении θ, $p(\theta)$;

б) Сформулировать модель для данных, $p(data \mid \theta)$;

в) Получить апостериорный закон распределения θ по формуле условной вероятности, $p(\theta \mid data) \propto p(\theta) \cdot p(data \mid \theta)$.

18.1 В своём труде 1763 года «An Essay towards solving a Problem in the Doctrine of Chances» сэр Томас Байес решает следующую задачу: «Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named». А вам слабо?

Имеется монетка, возможно неправильная. Мы не знаем вероятность выпадения орла α, поэтому считаем, что α равномерно распределена на отрезке $[0; 1]$.

а) Какова безусловная вероятность того, что α лежит в диапазоне $[0; 0.5]$?

б) Какова условная вероятность того, что α лежит в диапазоне $[0; 0.5]$, если монетка выпала 5 раз орлом и 7 раз решкой?

в) В байесовском подходе α — это константа или случайная величина?

г) В каком году умер Томас Байес?

http://rstl.royalsocietypublishing.org/content/53/370.full.pdf

18.2 Время, которое Вася тратит на задачу — равномерно распределенная случайная величина: на простую — от 1 до 15 минут, на сложную — от 10 до 20 минут. Известно, что на некую задачу Вася потратил 13 минут.

а) С помощью метода максимального правдоподобия определите, простая она или трудная.

б) С помощью байесовского подхода посчитайте вероятности того, что задача была простая, если на экзамене было 7 легких и 3 трудных задачи.

18.3

19. MCMC

Изначально Курочки Ку, Ро и Чка стоят в точке $x = 1$. Каждая курочка перемещается по следующему алгоритму независимо от других.

Для начала курочка подбрасывает монетку. Если выпадает решка, то это означает *намерение* пойти на один шаг направо, если орёл — *намерение* пойти на один шаг налево. Обозначим текущую точку x_{now}, а точку *намерения* — x_{prop}.

Затем курочка считает отношение $a = p(x_{prop})/p(x_{now})$.

19.1

$$p(x) = \begin{cases} 0.1x, & \text{при } x \in \{1, 2, 3, 4\} \\ 0, & \text{иначе.} \end{cases}$$

Если $a \geqslant 1$, то курочка гарантированно отправляется туда, куда намеревалась. Если $a < 1$, то курочка отправляется туда, куда намеревалась с вероятностью a, а с вероятностью $(1 - a)$ отказывается от намерения. Затем курочка снова подбрасывает монетку и так далее до бесконечности.

После осуществления намерения или отказа от него, Ку откладывает одно яйцо в точку, где находится. Курочка Ро откладывает яйца только после того, как совершит успешное перемещение. Курочка Чка откладывает яйца только после того, как откажется от намерения перемещаться.

а) Нарисуйте функцию $p(x)$.

б) Как в долгосрочном периоде распределятся яйца, отложенные Ку?

в) Как в долгосрочном периоде распределятся яйца, отложенные Ро?

г) Как в долгосрочном периоде распределятся яйца, отложенные Чка?

20. Шпаргалка

20.1. Определения

а) Оценка $\hat{a}$ неизвестного параметра a называется *несмещённой*, если $\mathbb{E}(\hat{a}) = a$.

б) Последовательность случайных величин R_1, R_2, ...сходится к величине R по вероятности, если для любого положительного числа $\varepsilon > 0$ вероятность отклонения R_i от R больше, чем на ε, стремится к нулю:

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0, \ \forall \varepsilon > 0$$

Сходимость по вероятности обозначается с помощью оператора plim, $\operatorname{plim}_{n\to\infty} R_n = R$:

в) Последовательность оценок $\hat{a}_1$, $\hat{a}_2$, ...неизвестного параметра a называется *состоятельной*, если $\operatorname{plim} \hat{a}_n = a$.

г) Среднеквадратическое отклонение оценки $\hat{a}$ от истинного значение параметра a, $MSE(\hat{a}) = \mathbb{E}((\hat{a} - a)^2)$. По теореме Пифагора величина MSE представима в виде $MSE(\hat{a}) = \mathbb{V}\mathrm{ar}(\hat{a}) + (\mathbb{E}(\hat{a}) - a)^2$. Если оценка несмещённая, то $MSE = \mathbb{V}\mathrm{ar}(\hat{a})$.

д) Оценка $\hat{a}$ неизвестного параметра a называется эффективной среди некоторого набора оценок K, если оценка $\hat{a}$ обладает наименьшей среднеквадратичной ошибкой MSE среди рассматриваемого набора оценок K. Если рассматриваемые оценки являются несмещёнными, $K = \{\hat{a} \mid \mathbb{E}(\hat{a}) = a\}$, то эффективной оценкой среди них является та, которая обладает наименьшей дисперсией.

20.2. Гипотезы

20.2.1. Про единственную выборку

а) Гипотеза о математическом ожидании при большом количестве наблюдений

(a) Наблюдаем: X_1, X_2, ..., X_n;

(b) Предполагаем: X_i независимы и одинаково распределены (не обязательно нормально), количество наблюдений n велико.

(c) Проверяемая гипотеза: H_0: $\mu = \mu_0$ против H_a: $\mu \neq \mu_0$;

(d) Статистика:

$$Z = \frac{\bar{X} - \mu_0}{se(\bar{X})} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

(e) При верной H_0 оказывается, что $Z \to \mathcal{N}(0; 1)$;

б) Гипотеза о математическом ожидании при нормальных наблюдениях

(a) Наблюдаем: X_1, X_2, ..., X_n;

(b) Предполагаем: X_i независимы и одинаково нормально распределены $\mathcal{N}(\mu; \sigma^2)$, количество наблюдений n может быть мало.

(c) Проверяемая гипотеза: H_0: $\mu = \mu_0$ против H_a: $\mu \neq \mu_0$;

(d) Статистика:

$$t = \frac{\bar{X} - \mu_0}{se(\bar{X})} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

(e) При верной H_0 оказывается, что $t \sim t_{n-1}$;

в) Гипотеза о математическом ожидании при нормальных наблюдениях и известной дисперсии

(a) Наблюдаем: X_1, X_2, ..., X_n, знаем величину σ^2;

(b) Предполагаем: X_i независимы и одинаково нормально распределены $\mathcal{N}(\mu; \sigma^2)$, количество наблюдений n может быть мало.

(c) Проверяемая гипотеза: H_0: $\mu = \mu_0$ против H_a: $\mu \neq \mu_0$;

(d) Статистика:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

(e) При верной H_0 оказывается, что $Z \sim \mathcal{N}(0; 1)$;

г) Гипотеза о вероятности при наблюдениях с распределением Бернулли (0 или 1)

(a) Наблюдаем: X_1, X_2, ..., X_n

(b) Предполагаем: X_i независимы и имеют распределение Бернулли: равны 1 с вероятностью p и 0 с вероятностью $1 - p$. Количество наблюдений n велико.

(c) Проверяемая гипотеза: H_0: $p = p_0$ против H_a: $p \neq p_0$;

(d) Статистика:

$$Z = \frac{\hat{p} - p_0}{se(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Возможен вариант этой статистики:

$$Z = \frac{\hat{p} - p_0}{se(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

(e) При верной H_0 оказывается, что $Z \to \mathcal{N}(0; 1)$;

(f) Гипотеза о вероятностях является частным случаем гипотезы о математическом ожидании при большом количестве наблюдений. Можно заметить, что $\hat{p} = \bar{X}$ и $\hat{\sigma}^2 = \hat{p}(1 - \hat{p}) \cdot \frac{n}{n-1}$. И потому также корректен вариант статистики

$$Z = \frac{\bar{X} - \mu_0}{se(\bar{X})} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

д) Гипотеза о дисперсии при нормальных наблюдениях

(a) Наблюдаем: X_1, X_2, ..., X_n

(b) Предполагаем: X_i независимы и одинаково нормально распределены $\mathcal{N}(\mu; \sigma^2)$, количество наблюдений n может быть мало.

(c) Проверяемая гипотеза: H_0: $\sigma = \sigma_0$ против H_a: $\sigma \neq \sigma_0$;

(d) Статистика:

$$S = \frac{\sum(X_i - \bar{X})^2}{\sigma_0^2} = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2}$$

(e) При верной H_0 оказывается, что $S \sim \chi^2_{n-1}$;

20.2.2. Про пару выборок

е) Гипотеза о разнице ожиданий при большом количестве наблюдений

(a) Наблюдаем: X_1, X_2, ..., X_{n_x}, Y_1, Y_2, ..., Y_{n_y}. Возможно, что $n_x \neq n_y$. Дисперсии σ_x^2 и σ_y^2 не знаем и не уверены, что они равны.

(b) Предполагаем: X_i одинаково распределены между собой (не обязательно нормально), Y_i одинаково распределены между собой, но возможно совсем не так, как X_i (не обязательно нормально). Все величины независимы. Количества n_x и n_y велики.

(c) Проверяемая гипотеза: H_0: $\mu_x - \mu_y = \delta_0$ против H_a: $\mu_x - \mu_y \neq \delta_0$;

(d) Статистика:

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{se(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$$

(e) При верной H_0 оказывается, что $Z \to \mathcal{N}(0; 1)$;

ж) Гипотеза о разнице ожиданий при нормальности распределения обеих выборок и известных дисперсиях

(a) Наблюдаем: X_1, X_2, ..., X_{n_x}, Y_1, Y_2, ..., Y_{n_y}. Возможно, что $n_x \neq n_y$. Дисперсии σ_x^2 и σ_y^2 знаем. Возможно, что дисперсии не равны.

(b) Предполагаем: X_i одинаково распределены между собой $\mathcal{N}(\mu_x, \sigma_x^2)$, Y_i одинаково распределены между собой $\mathcal{N}(\mu_y, \sigma_y^2)$. Все величины независимы. Количества n_x и n_y любые.

(c) Проверяемая гипотеза: H_0: $\mu_x - \mu_y = \delta_0$ против H_a: $\mu_x - \mu_y \neq \delta_0$;

(d) Статистика:

$$Z = \frac{\bar{X} - \bar{Y} - \delta_0}{\sigma_{\bar{X}-\bar{Y}}} = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

(e) При верной H_0 оказывается, что $Z \sim \mathcal{N}(0; 1)$;

з) Гипотеза о разнице ожиданий при нормальности распределения обеих выборок и неизвестных но равных дисперсиях

(a) Наблюдаем: X_1, X_2, ..., X_{n_x}, Y_1, Y_2, ..., Y_{n_y}. Возможно, что $n_x \neq n_y$. Дисперсии σ_x^2 и σ_y^2 равны, но неизвестны.

(b) Предполагаем: X_i одинаково распределены между собой $\mathcal{N}(\mu_x, \sigma^2)$, Y_i одинаково распределены между собой $\mathcal{N}(\mu_y, \sigma^2)$. Все величины независимы. Количества n_x и n_y любые.

(c) Проверяемая гипотеза: H_0: $\mu_x - \mu_y = \delta_0$ против H_a: $\mu_x - \mu_y \neq \delta_0$;

(d) Статистика:

$$t = \frac{\bar{X} - \bar{Y} - \delta_0}{se(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y} - \delta_0}{\sqrt{\frac{\hat{\sigma}^2}{n_x} + \frac{\hat{\sigma}^2}{n_y}}},$$

где

$$\hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{n_x + n_y - 2}$$

(e) При верной H_0 оказывается, что $t \sim t_{n_x+n_y-2}$;

и) Гипотеза о разнице ожиданий в связанных парах

(a) Наблюдаем: X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_n. Количество X_i и Y_i одинаковое.

(b) Предполагаем: внутри пары X_i и Y_i зависимы, а наблюдения с разными номерами независимы. Рассматриваем разницу $D_i = X_i - Y_i$ и получаем одномерную выборку. Величины D_i независимы и одинаково распределены. Возможно три описанных ранее случая :) Здесь для примера рассмотрим случай, когда $D_i \sim \mathcal{N}(\mu_d, \sigma_d^2)$ с неизвестной дисперсией.

(c) Проверяемая гипотеза: H_0: $\mu_d = \mu_0$ против H_a: $\mu_d \neq \mu_0$;

(d) Статистика:

$$t = \frac{\bar{D} - \mu_d}{se(\bar{D})} = \frac{\bar{X} - \bar{Y} - \mu_d}{\sqrt{\frac{\hat{\sigma}_d^2}{n}}},$$

где

$$\hat{\sigma}_d^2 = \frac{\sum(D_i - \bar{D})^2}{n-1} = \frac{\sum(X_i - Y_i - (\bar{X} - \bar{Y}))^2}{n-1}$$

(e) При верной H_0 оказывается, что $t \sim t_{n-1}$;

к) Гипотеза о равенстве дисперсий при нормальности распределения обеих выборок

(a) Наблюдаем: X_1, X_2, ..., X_{n_x}, Y_1, Y_2, ..., Y_{n_y}. Возможно, что $n_x \neq n_y$. Дисперсии σ_x^2 и σ_y^2 не знаем. Возможно, что дисперсии не равны.

(b) Предполагаем: X_i одинаково распределены между собой $\mathcal{N}(\mu_x, \sigma^2)$, Y_i одинаково распределены между собой $\mathcal{N}(\mu_y, \sigma^2)$. Все величины независимы. Количества n_x и n_y любые.

(c) Проверяемая гипотеза: H_0: $\sigma_x = \sigma_y$ против H_a: $\sigma_x \neq \sigma_y$;

(d) Статистика:

$$F = \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2}$$

(e) При верной H_0 оказывается, что $F \sim F_{n_x-1, n_y-1}$;

21. Решения

1.1. Среднее равно $(x+25)/5$. Если $x<5$, получаем $x=0$. Если $x \in (5;7)$, получаем $x = 25/4$. Если $x > 7$, получаем $x = 10$.

1.2. Медиана не изменится, среднее упадёт на $10/25 = 0.4$. Для случая роста 153: среднее упадёт на 0.8, медиана упадёт произвольно на некое число из отрезка $[0;2]$.

1.3. да

1.4. да

1.5.

1.6.

а) Для $\mathrm{Med}(X+Y) = \mathrm{Med}(X) + \mathrm{Med}(Y)$ можно взять два независимых симметричных распределения.

б) Для $\mathrm{Med}(X+Y) \neq \mathrm{Med}(X) + \mathrm{Med}(Y)$ подойдёт практически любая сумма несимметричных распределений. Например, можно взять две независимых величины с плотностью $f(x) = 2 - 2x$ на отрезке $[0;1]$.

в) Для $\mathrm{Med}(X^k) = \mathrm{Med}(X)^k$ для всех k можно взять неотрицательную случайную величину.

г) Для $\mathrm{Med}(X^2) \neq \mathrm{Med}(X)^2$ подойдёт симметричная около нуля случайная величина.

1.7. Исключим те варианты, когда все пять наблюдений оказались или синхронно выше, или синхронно ниже медианы, получаем, $p = 1 - 2 \cdot 0.5^5 = 1 - 0.5^4$.

1.8. Укреплять те места, где не было следов пуль.

1.9.

1.10.

1.11.

а) $\mathbb{E}(\bar{X}_n) = \mu$, $\mathbb{V}\mathrm{ar}(\bar{X}_n) = \frac{\sigma^2}{n}$

б) $\mathbb{E}(\bar{X}_n) = \mu$, $\mathbb{V}\text{ar}(\bar{X}_n) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

в) При $N \to \infty$ получится формула для выборки с возвращениями.

1.12.

1.13.

1.14. $m^2 = 1/2$, $\text{Med}(X) = 1/\sqrt{2}$.

1.15. $\mathbb{E}(X_i^*) = \mathbb{E}(X_i)$, $\mathbb{V}\text{ar}(X_i^*) = \mathbb{V}\text{ar}(X_i)$

1.16.

2.1.

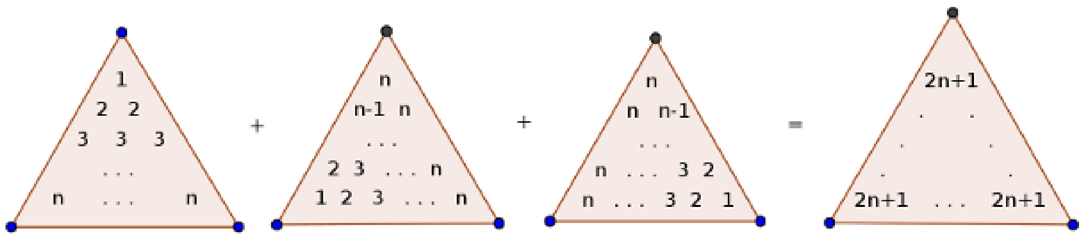
2.2.

2.3.

2.4.

2.5.

3.1. Здесь потребуется формула для $S = 1^2 + 2^2 + \ldots + n^2$.
На рисунке сложены три суммы:



На языке формул:

$$3S = (2n+1) \cdot (1 + 2 + \ldots + n) = (2n+1)n\frac{n+1}{2}.$$

3.2. Минимум, $a = 6$, $b = 5$, $c = 9$. Например, проекцией вектора $(6, 3, 7, 8, 9, 10, 11)$ на вектора вида $(a, b, b, c, c, c, 0)$.

3.3.

3.4.

а) χ^2_d

б) d

в) $2d$

г) χ^2_{a+b}

3.5. $Q = Z_1^2$, зная функцию плотности Z_1, $f(z_1) = \frac{1}{\sqrt{2\pi}} \exp(-z_1^2/2)$, находим функцию плотности Q;

3.6. $\hat{Z} = \langle Z, z\rangle \cdot v$; $\hat{Z}_i = \langle Z, z\rangle \cdot v_i$; $\mathbb{V}\mathrm{ar}(\langle Z, z\rangle) = 1$; $\mathbb{C}\mathrm{ov}(\hat{Z}_i, \hat{Z}_j) = v_i v_j \mathbb{V}\mathrm{ar}(\langle Z, z\rangle)_{ij} = v_i v_j$;

3.7.

3.8.

3.9.

а) $\mathbb{E}(X^3) = 0, \mathbb{E}(X^4) = 3, \mathbb{E}(X^6) = 5 \cdot 3, \mathbb{E}(X^{2024}) = 2023 \cdot 2021 \cdot \ldots \cdot 3 \cdot 1.$

б) $\mathbb{E}(Y^3), \mathbb{E}(Y^4), \mathbb{E}((Y-1)Y^{100})$

3.10. $\mathbb{E}(Q) = k$, $\mathbb{V}\mathrm{ar}(Q) = 2k$

3.11.

3.12.

3.13. $3X/\sqrt{Q_1} \sim t_9$, $X/\sqrt{Q_2/16} \sim t_{16}$, $Q_1 + Q_2 \sim \chi_{25}$, $5X/\sqrt{Q_1+Q_2} \sim t_{25}$

4.1. Метод максимального правдоподобия:

$$C_n^{Y_{\text{к}}} C_{n-Y_{\text{к}}}^{Y_{\text{щ}}} a^{Y_{\text{к}}} (2a)^{Y_{\text{щ}}} (1-3a)^{Y_{\text{б}}} \to \max_a$$

Решая задачу максимизации Кота Матроскина получаем

$$\hat{a}_{\text{КМ}} = \frac{Y_{\text{к}} + Y_{\text{щ}}}{3n}$$

Замечаем, что $Y_{\text{к}} + Y_{\text{щ}} \sim \mathrm{Bin}(n, 3a)$. Отсюда $\mathbb{E}(\hat{a}_{\text{КМ}}) = a$, $\mathbb{V}\mathrm{ar}(\hat{a}_{\text{КМ}}) = \frac{a(1-3a)}{n}$. Оценка несмещённая и состоятельная.

С точки зрения Пса Шарика, неизвестными являются две вероятности, a и b. Он решает задачу максимизации по двум переменным. В результате получается вполне себе интуитивная оценка $\hat{a}_{\text{ПШ}} = Y_{\text{к}}/n$.

4.2. Метод правдоподобия. Замечаем, что $L(n) = \mathbb{P}(S = 80 \mid \theta) = \frac{C_?^{80} C_?^{20}}{C_n^{100}}$. Ошибочно утверждать, что $S \sim \mathrm{Bin}\left(100, p = \frac{100}{n}\right)$, так как вероятность поймать очередного зайца с бантиком изменяется при поимке очередного зайца с бантиком.

Максимизируем вероятность, $\hat{n}_{ML} = 125$. Для максимизации полезно рассмотреть неравенство $L(n) > L(n-1)$ и понять, что это значит.

Максимизация ошибочной функции здесь даёт тот же ответ, но благородные доны и доньи так не поступают!

Метод моментов. Рассмотрим Y_1, Y_2, ..., Y_n. Величина Y_i равна 1 если при втором отлове i-й заяц оказался с бантиком и 0 иначе.

Метод моментов. Получаем теоретическое равенство:

$$\mathbb{E}(\bar{Y}) = \mathbb{E}(Y_i) = \frac{100}{n}.$$

Заменяем на выборочный аналог:

$$\frac{100}{\hat{n}_{MM}} = \bar{Y}$$

Отсюда

$$\hat{n}_{MM} = \frac{100}{\bar{Y}} = \frac{100^2}{S}$$

4.3.

4.4.

4.5.

5.1.

5.2. $\hat{a} = \ln(Y_1)$, $\hat{b} = \ln(Y_2) - \ln(Y_1)$

5.3.

5.4.

5.5.

5.6. $\hat{a}_{ml} = \sum X_i^2/2n$, $\hat{a}_{mm} = \bar{X}$.

5.7.

5.8.

$$\hat{k}_{MM} = \frac{2\bar{X} - 1}{1 - \bar{X}}$$

5.9.

5.10.

5.11.

5.12.

5.13. $I_{X_1}(\theta) = I_{X_2}(\theta)$, $I_{X_1,X_2,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$

5.14.

$$\ell(\theta) = \ln\theta + \sum(\theta - 1)\ln(\ln x_i) - \sum \ln x_i$$

$$\ell'(\theta) = \frac{n}{\theta} + \sum \ln(\ln x_i)$$

5.15.

5.16. Условия регулярности не выполнены, область значений величин X_i зависит от параметра θ.

5.17. Условия регулярности не выполнены, область значений величин X_i зависит от параметра θ.

5.18.

5.19.

5.20. Условия регулярности не выполнены, область значений величин X_i зависит от параметра a. Если среди X_i проскакивает значение $X_i \neq 0$, то оно и есть истинное a. Поэтому оценка выглядит довольно интересно,

$$\hat{a}_{ML} = \begin{cases} 11, \text{ если } X_1 = X_2 = \ldots = X_n = 0, \\ a, \text{ иначе.} \end{cases}$$

В большинстве реальных задач истинное значение параметра по выборке узнать невозможно.

5.21.

5.22.

5.23.

5.24.

а) $\mathbb{E}(X_i) = a$, $\mathbb{E}(|X_i|) = 5a/4$

б) $\hat{a} = 11/30$

в) $\hat{a} = 26/75$

г) $\hat{a}_{GMM} = 108/325$

д) $\begin{pmatrix} 37 & -44 \\ -44 & 64 \end{pmatrix}$

5.25.

5.26. $\operatorname{plim}_{n\to\infty} \hat{\alpha}_{ML}(n) = 0$, не является состоятельной

5.27.

5.28.

5.29. $\gamma = a$, $\beta = 1/\mathbb{V}\mathrm{ar}(\ell')$, $\mathbb{E}(u) = 0$, $\mathbb{C}\mathrm{ov}(u, \ell') = 0$, Ожидание наилучшей аппроксимации равно нулю, а дисперсия — $1/\mathbb{V}\mathrm{ar}(\ell')$.

5.30.

5.31.

5.32.

5.33.

5.34.

5.35.

5.36.

6.1.

$$\operatorname{plim} \ln \frac{\hat{\sigma}_R^2}{\hat{\sigma}_{UR}^2} = \ln 1 = 0.$$

$$LR_n = n \ln \frac{LR_n}{T_n^2} = n \ln \left(1 + \frac{\hat{\sigma}_R^2 - \hat{\sigma}_{UR}^2}{\hat{\sigma}_{UR}^2}\right) \approx n \frac{\hat{\sigma}_R^2 - \hat{\sigma}_{UR}^2}{\hat{\sigma}_{UR}^2}.$$

7.1.

а) $c = 1/n$, да;

б) $c = 1/(n + \sigma^2/\mu^2)$, нет, так как μ и σ неизвестны;

в) $\lambda = 0$ и $\lambda = \sigma^2/\mu^2$.

7.2.

а) несмещённая, состоятельная, линейная, неэффективная

б) несмещённая, состоятельная, линейная, неэффективная

в) несмещённая, несостоятельная, линейная, неэффективная

г) смещённая, состоятельная, линейная

д) смещённая, состоятельная, нелинейная

е) несмещённая, состоятельная, линейная, неэффективная ж) несмещённая, состоятельная, линейная, эффективная

з) смещённая, несостоятельная, линейная

и) смещённая, несостоятельная, нелинейная

к) несмещённая, состоятельная, линейная, неэффективная

7.3.

а) $\hat{\theta}_{ML} = \max\{Y_1, Y_2, \ldots, Y_n\}$.

б) Все Y_i меньше θ, значит и $\hat{\theta}$ всегда меньше θ, значит смещённая.

в) $F_{\hat{\theta}}(t) = \mathbb{P}(\hat{\theta} \leqslant t) = \mathbb{P}(Y_1 \leqslant t, Y_2 \leqslant t, \ldots) = (\mathbb{P}(Y_1 \leqslant t))^n$, $\mathbb{P}(Y_1 \leqslant t) = t^5/\theta^5$, $f_{\hat{\theta}}(t) = dF_{\hat{\theta}}(t)/dt = \frac{5nt^{5n-1}}{\theta^{5n}}$.

г) $\mathbb{E}(\hat{\theta}) = \frac{5n}{5n+1}\theta$.

д) $\hat{\theta}_{unbiased} = \frac{5n+1}{5n}\hat{\theta}$.

7.4.

а) $\hat{p}$ несмещённая

б) $\sigma^2 = np(1-p)$.

в) $\mathbb{E}(\hat{\sigma}^2) = (n-1)p(1-p)$, смещённая, $\hat{\sigma}^2_{unbiased} = \frac{n}{n-1}\hat{\sigma}^2$.

7.5. Несмещённые: $\hat{\mu}_A$, $\hat{\mu}_B$, $\hat{\mu}_C$. Среди несмещённых наименьшая дисперсия у оценки $\hat{\mu}_B$.

7.6.

7.7.

7.8.

а) $\mathbb{E}(Z) = a/3$, $\mathbb{V}\text{ar}(Z) = a^2/18$, $f_Z(t) = \frac{2a-2at}{a^2}$

б) $\beta = 3$

в) $\beta = 1/2$

7.9.

7.10.

7.11.

7.12. Закон распределения X_{min} также экспоненциальный, но с другой интенсивностью. Находим $\mathbb{E}(X_{\text{min}}) = \theta/20$, поэтому оценка $\hat{\theta} = 20X_{\text{min}}$ будет несмещённой.

7.13.
$$\operatorname{plim} \bar{X} = \frac{\theta}{\theta+1}$$
Следовательно, оценка $\hat{\theta} = \bar{X}/(1-\bar{X})$ будет состоятельной для θ.

7.14. Обе оценки несмещённые, состоятельные. Более эффективна оценка $\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2}$.

7.15.

7.16.

7.17. Разница показаний равна $u_2 + 100 - u_1$, где u_i — ошибки измерения. Замечаем, что $\mathbb{E}((u_2 - u_1)^2) = 2\sigma_u^2$. И, конечно, оценка $(u_2 - u_1)^2$ является несмещённой оценкой для параметра $2\sigma_u^2$. Отсюда получаем, что $\hat{\sigma}_u^2 = 5000$ грамм в квадрате.

7.18.

а) X_1, X_2

б) X_1X_2, $X_2 - X_1 + X_1X_2$

в) Нет, так как любая оценка определяется значением в четырёх возможных точках.

г) Линейные комбинации 1, p и p^2.

7.19. Нет

7.20.

а) $a_x + a_y = 1$;

б) $a_x + a_y = 1$;

в) $a_x = (n_x - 1)/(n_x - 1 + n_y - 1)$, $a_y = (n_y - 1)/(n_x - 1 + n_y - 1)$;

г) $a_x = 0$, $a_x = 1$, $a_x = (n_x - 1)/(n_x - 1 + n_y - 1)$ и $a_y = 1 - a_x$.

7.21.

$$F_R(t) = \begin{cases} 0, \text{ если } t < 0 \\ (t/a)^n, t \in [0; a] \\ 1, \text{ если } t > a \end{cases}$$

$$f_R(t) = \begin{cases} nt^{n-1}/a^n, t \in [0; a] \\ 0, \text{ иначе} \end{cases}$$

$ER = \frac{n}{n+1}a$, $E(R^2) = \frac{n}{n+2}a^2$.

7.22.

7.23.

7.24. Замечаем, что $\lim \mathbb{V}\mathrm{ar}(X_{\max}) = 0$. Поэтому, $\operatorname{plim} X_{\max} = \lim \mathbb{E}(X_{\max}) = a$. Оценка $X_{\max}$ является состоятельной.

7.25. Замечаем, что $\lim \mathbb{V}\mathrm{ar}(X_{\max}) = 0$. Поэтому, $\operatorname{plim} X_{\max} = \lim \mathbb{E}(X_{\max}) = a$. Оценка $X_{\max}$ является состоятельной.

8.1.

8.2.

8.3.

8.4. Оценка $\bar{X}$ достигает границу Рао — Крамера. Придумать оценку с таким же математическим ожиданием и меньшей дисперсией нельзя.

8.5. Оценка $\bar{X}$ достигает границу Рао — Крамера. Придумать оценку с таким же математическим ожиданием и меньшей дисперсией нельзя.

8.6. Оценка $\bar{X}$ достигает границу Рао — Крамера. Придумать оценку с таким же математическим ожиданием и меньшей дисперсией нельзя.

9.1. Удаляем из колоды вальтов, дам, королей и тузов. Если клиент позолотил ручку, то тасуем колоду и открываем первую карту. Если первая карта — семёрка червей, то объявляем доверительный интервал от 2 до 3, иначе объявляем интервал от 0 до 1.

9.2. Да, конечно! На первый взгляд, кажется, что если Кржемелик верно построил свой интервал, то вероятность накрытия неизвестной вероятности у Вахмурки не может превышать 5%. Однако это не так. Доверительная вероятность относится к процедуре построения доверительного интервала, а не к конкретным полученным числовым границам.

9.3. $\frac{1}{1.8} - \frac{1}{2.2} = \frac{10}{99}$, $[0; 10X]$.

9.4.

9.5.

9.6.

9.7.

9.8. Важна при обоих доверительных интервалах. Без предпосылки о нормальности интервал для дисперсии по данным формулам нельзя построить даже при больших n. При больших n можно отказаться от предпосылки о нормальности при построении интервала для μ и для σ^2 в силу центральной предельной теоремы.

9.9.

9.10.

9.11.

а) Да, центры интервалов равны $\bar{X}$.

б) У Машеньки, так как ширина интервала Вовочки постоянна и при большой разнице $\hat{\mu} - \mu$ окажется недостаточной, чтобы накрыть μ.

в) У Вовочки, так как ширина интервала Вовочки постоянна и при малой разнице $\hat{\mu} - \mu$ гарантированно достаточна, чтобы накрыть μ.

г) Нет, иначе интервалы не были бы оба 95%-е.

д) Стремится к $1/2$.

е) $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$

ж) $\mathbb{E}(\hat{\sigma}) < \sigma$

з) ...

9.12. При построении доверительного интервала важно понять, что $\mathbb{C}\mathrm{ov}(\hat{\pi}_+, \hat{\pi}_-) \neq 0$.

10.1.

10.2.

10.3.

10.4.

10.5. Второй вариант предпочтительнее. Мы хотим, чтобы дисперсия бутстрэп-выборки перестановочного теста равнялась дисперсии, наблюдаемой в исходных данных при верной нулевой гипотезе. А дисперсия первого варианта при довольно мягких предпосылках будет меньше, чем требуется. При отсутствии разницы ворчалок и пыхтелок первый вариант даст много ложных обнаружений эффекта.

12.1. равномерно, $\alpha = 0.05$; нет, он резко увеличивает ошибку второго рода

12.2. $\alpha = 1/8$, $\beta = 9/32$

12.3. $\alpha = \mathbb{P}(\mathcal{N}(0;1) > -0.35) \approx 0.64$, $\beta = \mathbb{P}(\mathcal{N}(0;1) \leqslant -1.76) \approx 0.04$.

12.4.

12.5.

12.6.

12.7.

12.8.

12.9.

12.10.

а) $Y_i \sim \mathrm{U}[0; 1]$;

б) $D_i \sim \mathrm{U}[0.5; 1]$, $q_{0.95} = 0.975$;

в) $q_{0.95} = 1 - \sqrt{0.05/2} \approx 0.842$.

12.11. Критерий Неймана-Пирсона сводится к сравнению $\bar{X}$ с порогом. При верной H_0 величина $\bar{X}$ распределена $\mathcal{N}(0; \frac{4}{n})$. Отсюда искомый критерий имеет вид: Если $\bar{X} \geqslant 0.825$, то гипотеза H_0 отвергается в пользу гипотезы H_a.

12.12. Упрощая неравенство из леммы Неймана-Пирсона, получаем критерий: если $X_1 \cdot X_2 > t$, то H_0 отвергается. Величину t находим из уравнения

$$\int_0^t (1 - t/x)\, dx = 0.05$$

12.13.

12.14.

12.15.

12.16.

12.17. Из условия на вероятность ошибки первого рода критическое значение для $(\hat{\mu}_t - \hat{\mu}_c - 0)/se(\hat{\mu}_t - \hat{\mu}_c)$ равно $Z_{1-\alpha/2}$.

В масштабе разницы средних критическое значение принимает вид:

$$Z_{1-\alpha/2} se(\hat{\mu}_t - \hat{\mu}_c).$$

Обеспечиваем заданную вероятность ошибки второго рода:

$$\mathbb{P}(\hat{\mu}_t - \hat{\mu}_c < Z_{1-\alpha/2} se(\hat{\mu}_t - \hat{\mu}_c) \mid H_a).$$

После стандартизации получаем равенство:

$$Z_\beta = Z_{1-\alpha/2} - MDE/se(\hat{\mu}_t - \hat{\mu}_c).$$

Выражаем n:
$$n = \frac{2\sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{MDE^2}.$$

В общем случае вместо 2 будет $(r+1)/r$.

13.1.

13.2. При p_N заданном в H_0 критерий Пирсона имеет хи-квадрат распределение с двумя степенями свободы. При оцениваемом p_N критерий Пирсона имеет хи-квадрат распределение с одной степенью свободы.

https://ru.wikipedia.org/wiki/Закон_Харди_—_Вайнберга

13.3.

а) $\mathbb{E}(\nu_i) = np_i$, $\mathbb{V}\text{ar}(\nu_i) = np_i(1-p_i)$, $\mathbb{C}\text{ov}(\nu_i, \nu_j) = -np_ip_j$.

б) $\mathbb{E}(\nu_i^*) = 0$, $\mathbb{V}\text{ar}(\nu_i^*) = 1 - p_i$, $\mathbb{C}\text{ov}(\nu_i^*, \nu_j^*) = -\sqrt{p_ip_j}$.

в) $v = (\sqrt{p_1}, \ldots, \sqrt{p_n})$.

г) $V' = V^2 = V$.

13.4. Одна положительная величина плюс вторая положительна величина дают в сумме третью положительную величину. Это торчат уши теоремы Пифагора!

Подберем вектора, чтобы переписать равенство $\sum \frac{(a_i-b_i)^2}{b_i} + n = \sum \frac{a_i^2}{b_i}$ в виде
$$\|x\|^2 + \|y\|^2 = \|x+y\|^2.$$
Нам подойдут вектор x с $x_i = (a_i - b_i)/sqrtb_i$ и ортогональный ему вектор y с $y_i = \sqrt{b_i}$.

13.5. Если использовать одинаковое число степеней свободы, то критическое значение для всех нулевых гипотез будет одинаковым. Хи-квадрат статистика для жёсткой гипотезы выше, чем для мягкой. Поэтому при таком способе аксиома адекватности нарушаться не будет.

13.6.

14.1.

14.2.

14.3.

14.4.

15.1. $\mathbb{E}(Ay) = A\,\mathbb{E}(y)$, $\mathbb{V}\text{ar}(Ay) = A\,\mathbb{V}\text{ar}(y)A^T$

15.2.

15.3.

15.4.

15.5.

15.6.

16.1. Вычтем $2x^2 + 2y^2 + 2z^2 + 2w^2$. Получим, что оптимальное $x = 0$. Далее, $z = 0$, $y = 0$. В итоге $w = 1$ или $w = -1$.

16.2.

16.3.

16.4.

16.5. $(5 + 4 + 1) = 10$; длины — $\sqrt{5} \cdot \sqrt{99}$, $\sqrt{4} \cdot \sqrt{99}$, $\sqrt{1} \cdot \sqrt{99}$; выборочные дисперсии — $5, 4, 1$; $(5 + 4)/10 = 0.9$.

16.6.

17.1.

17.2.

17.3.

17.4.

18.1.

18.2.

18.3.

19.1.

Модные хэштэги

22. Источники мудрости

[RB47] Richard Ruggles и Henry Brodie. «An empirical approach to economic intelligence in World War II». В: *Journal of the American Statistical Association* 42.237 (1947), с. 72—91. Подробности про то, как захватив всего два танка, можно оценить ежемесячный выпуск с точностью в пару процентов.

[Hes15] Tim C Hesterberg. «What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum». В: *The American Statistician* 69.4 (2015), с. 371—386. URL: http://arxiv.org/abs/1411.5279.

[SM20] Daniel Savenkov и Nikita Mar. *Practitioner's Guide to Statistical Tests*. 2020. URL: https://medium.com/@vktech/practitioners-guide-to-statistical-tests-ed2d580ef04f.

[CB00] James Carpenter и John Bithell. «Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians». В: *Statistics in medicine* 19.9 (2000), с. 1141—1164. URL: https://www.tau.ac.il/~saharon/Boot/10.1.1.133.8405.pdf.

[Den+13] Alex Deng и др. «Improving the sensitivity of online controlled experiments by utilizing pre-experiment data». В: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, с. 123—132. URL: www.exp-platform.com/Documents/2013-02-CUPED-ImprovingSensitivityOfControlledExperiments.pdf.

[Ser+21] Ceyhan Ceran Serdar и др. «Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies». В: *Biochemia Medica* 31.1 (2021), с. 27—53.

[Bai83] Davis Baird. «The Fisher/Pearson Chi-squared controversy: a turning point for inductive inference». В: *The British Journal for the Philosophy of Science* 34.2 (1983), с. 105—118.

[Cla46] RD Clarke. «An application of the Poisson distribution». В: *Journal of the Institute of Actuaries* 72.3 (1946), с. 481—481. URL: https://www.actuaries.org.uk/system/files/documents/pdf/0481.pdf.

[Buz+15] Nazar Buzun и др. «Stochastic Analysis in Problems, part 1 (in Russian)». В: *arXiv preprint arXiv:1508.03461* (2015). URL: https://arxiv.org/abs/1508.03461.