МдАД: Математическая статистика **Осень 2018**

Математическая статистика 1: 27 октября

Преподаватель: Антон Савостьянов *Ассистент: Даяна Мухаметшина*

Контакты: *Антон Савостьянов, почта:* a.s.savostyanov@gmail.com, *telegram:* @mryodo
Даяна Мухаметшина, почта: dayanamuha@gmail.com, *telegram:* @anniesss1

Правила игры: *Домашние задания следует присылать в читаемом виде на почту преподавателя не позднее указанного при выдаче задания крайнего срока (дедлайна).*
При выполнении домашнего задания приветствуется использование среды LaTeX; допустим набор в редакторах Word (Libreoffice, Google Docs) и отсканированные письменные материалы.
Выполненное домашнее задание должно содержать решение задачи, по которому возможно восстановить авторский ход решения, а не только ответ.

1.1 Введение в математическую статистику

В этой части мы поговорим про основные представления, связанные с разделом математическая статистика, а также введем основные понятия.

1.1.1 Основные понятия

Определение 1.1. Все множество объектов, для которых требуется получить некую информацию, будем называть популяцией (генеральной совокупностью).

Определение 1.2. Каждого участника популяции будем характеризовать некоторой случайной величиной X_i, в большинстве случаев полагая, что все X_i распределены одинаково.

Определение 1.3. Маленькая часть популяции, доступная для изучения, X_1, $X_2, \dots X_n$, называется выборкой. Следует понимать, что под выборкой можно понимать и последовательность случайных величин, и их конкретные реализации.

Определение 1.4. Статистикой будем называть любую функцию от выборки $f(X_1, X_2, \dots X_n)$.

Определение 1.5. Мощностью выборки будем называть число n, количество объектов в ней.

Следует понимать, что когда говорят о статистических методах изучения, то в нашей терминологии имеют ввиду методы для получения заключений (выводов) о популяции, основываясь на статистиках, полученных из выборок.

Меры центральности Договоримся, что выборка имеет вид $\{X_1, X_2, \ldots X_n\}$. Существует несколько способов оценить «среднее» значение по нашей выборке. Для всего этого применяется набор статистик (то есть функций от выборки), не всегда записанных элементарными функциями:

- выборочное среднее: $\overline{X} = \frac{X_1+X_2+\ldots+X_n}{n}$
- медиана: допустим для нашей выборки, что мы отсортировали реализацию в порядке неубывания:

  $$\{X_1, X_2, \ldots X_n\} \to \left\{X_{(1)}, X_{(2)}, \ldots X_{(n)}\right\}$$

  Таким образом $X_{(k)}$ обозначает $k-$ый по возрастанию элемент выборки, также называемый $k-$ой порядковой статистикой. Тогда медианой называется серединная (говорят «50%-ая») порядковая статистика:

  $$MED = X_{\left(\left[\frac{n+1}{2}\right]\right)}$$

  Вообще говоря, если число $\frac{n+1}{2}$ нецелое, то определение медианы является предметом договоренности
- мода: наиболее частое значение

Упражнение 1. Найдите выборочное среднее, моду и медианы выборки 1, 5, 4, 4, 3, 2, 4, 6, -2, 1.

Меры разброса

Определение 1.6. $p-$персентилем называется число, ниже которого лежит p процентов значений выборки.

Определение 1.7. Верхним квартилем называется 75%-персентиль:

$$UP = X_{([0.75(n+1)])}$$

Определение 1.8. Нижним квартилем называется 25%-персентиль:

$$LP = X_{([0.25(n+1)])}$$

Определение 1.9. Межквартильным расстоянием IQR называется величина:

$$IQR = UP - LP$$

Между верхним и нижним квартилем лежат 50% наиболее центральных значений выборки.

Определение 1.10. Разбросом выборки называется величина:

$$RANGE = X_{max} - X_{min} = X_{(n)} - X_{(1)}$$

Определение 1.11. Вариацией называется величина $VarX = s_X^2$, равная

$$s_X^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1}$$

Также как и в случае дисперсии, вариация имеет физический смысл момента инерции.

Величина $s_X \geq 0$ называется стандартным отклонением.

Заметим, что

$$\begin{aligned} s_X^2 &= \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n-1} = \\ &= \frac{X_1^2 - 2\overline{X}X_1 + \overline{X}^2 + X_2^2 - 2\overline{X}X_2 + \overline{X}^2 + \ldots + X_n^2 - 2\overline{X}X_n + \overline{X}^2}{n-1} = \\ &= \frac{\sum_{i=1}^n X_i^2 - 2\overline{X}\sum_{i=1}^n X_i + n\overline{X}^2}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) = \\ &= \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right) \end{aligned}$$

Эта формула часто гораздо удобнее для подсчета выборочной дисперсии, чем определение.

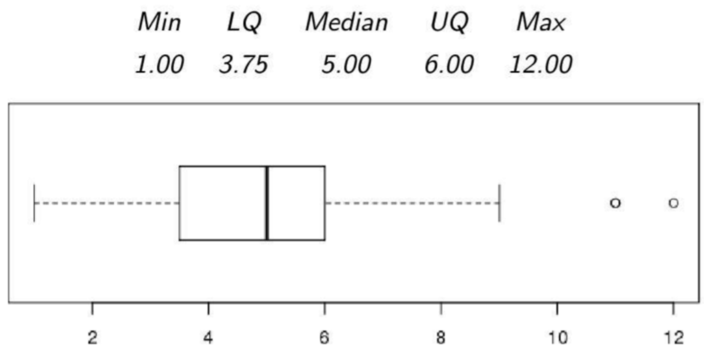
Определение 1.12. Выбросом называется наблюдаемое значение в выборке, которое очень далеко от других значений. Формально: если расстояние от точки до ближайшего квартиля (верхнего или нижнего) больше, чем $\frac{3}{2}IQR$.

Чувствительны к выбросам: среднее, вариация, стандартное отклонение, разброс

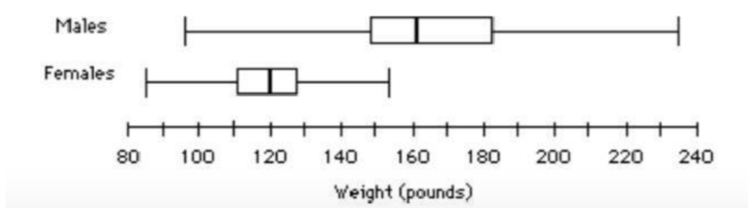
Нечувствительны к выбросам: медиана, квартили, межквартильное расстояние

Упражнение 2. Найдите межквартильное расстояние и стандартное отклонение в следующей выборке: -2, 1, 1, 2, 3, 4, 4, 4, 5, 6.

Частым методом отображения выборки являет так называемый боксплот (boxplot):



Упражнение 3. Веса мужчин и женщин (в фунтах) проиллюстрированы на следующем графике:



Какое из этих утверждение неверно:

1. Около половины всех мужчин весят между 150 и 185 фунтами
2. Около 25% всех женщин весят более 130 фунтов
3. Медианный вес мужчин — 162 фунта
4. Средний вес женщин — примерно 120 фунтов из-за симметрии
5. Веса женщин менее дисперсны чем веса мужчин

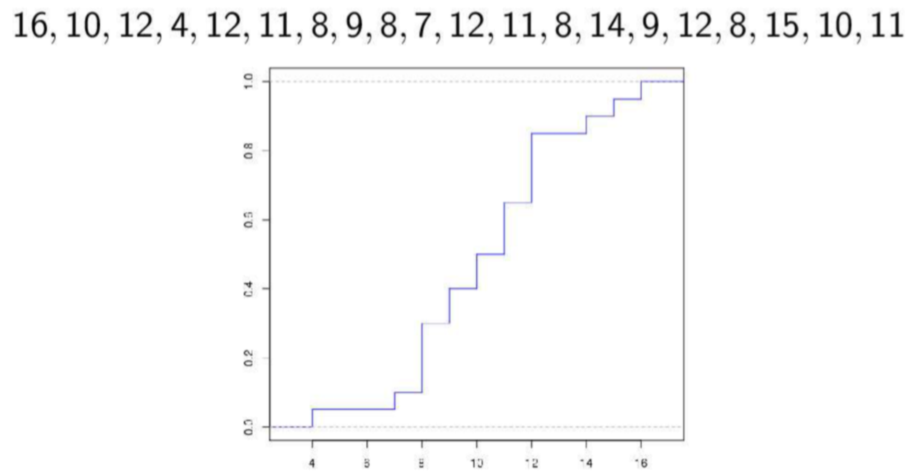
1.1.2 Выборочная (кумулятивная) функция распределения

Как и у любой случайной величины, у случайно величины, порождающей выборку, есть функция распределения. Однако, зная исключительно выборку, функцию распределения мы не получим, зато можем получить кумулятивную (выборочную) функцию распределения, которая будет аппрокисимировать реальную функцию распределения:

Определение 1.13. Выборочной функцией распределения $F(x)$ называется

$$F(x) = \frac{1}{n}\#\{X_i \leq x\}$$

то есть доля элементов выборки, не превосходящих x. По нашему определению, eCDF $F(x)$ имеет ступенчатый вид, где высота каждой ступеньки $\frac{1}{n}$, а сами они расположены в точках X_i :



1.2 Оценки параметров

Анализируемые методами математической статистики данные обычно рассматриваются как реализация выборки из некоторого распределения, известного с точностью до параметра (или нескольких параметров). При таком подходе для определения распределения, наиболее подходящего для описания данных, достаточно уметь оценивать значение параметра по реализации. В этой главе будет рассказано, как сравнивать различные оценки по точности.

1.2.1 Что такое оценки

Давайте поймем в общем случае, о чем идет речь.

Пусть есть некоторая выборка, которая порождена случайной величиной X. Считаем известным, что эта величина лежит в параметрическом семействе распределений $X \sim F_\theta(x)$, $\theta \in \Theta$ (например, $X \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, то есть мы знаем, что выборка была порождена нормальной случайной величиной, но нужно оценить, какое же у нее было среднее. Надо отметить, что зачастую в двупараметрических распределениях мы не знаем оба параметра, а не один, что в целом не должно мешать построению оценки). В реальности есть некоторое $\theta_0 \in \Theta$, которое соответствует настоящей случайной величине.

Задача в том, чтобы восстановить θ_0 по выборке $x_1,\ x_2,\ x_3\ \ldots\ x_n$ наиболее точно. Восстановление (или оценку) будем обозначать $\hat{\theta} = f(x_1, x_2, \ldots x_n)$. Как мы помним, это же выражение называлось статистикой.

Упражнение 4. Пусть есть выборка $x_1,\ x_2,\ \ldots\ x_n$ из точек, случайно выбранных с отрезка $[0;\theta]$ (то есть $X \sim U[0;\theta]$). Тогда $\Theta = [0; +\infty]$. Давайте приведем несколько интуитивно ясных оценок θ:

$$\hat{\theta}_1 = \max\{x_1, x_2, \ldots x_n\} = x_{(n)}$$

$$\hat{\theta}_2 = \frac{2(x_1 + x_2 + \ldots + x_n)}{n} = 2\overline{X}$$

$$\hat{\theta}_3 = 2MED(x_1, x_2, \ldots x_n)$$

Все они кажутся достаточно хорошими оценками для θ. Но надо как-то научиться сравнивать их точность.

1.2.2 Несмещенность, состоятельность

Определение 1.14. Оценка $\hat{\theta}(x_1, x_2, \ldots x_n)$ называется несмещенной, если

$$\mathbb{E}_\theta \hat{\theta}(x_1, x_2, \ldots x_n) = \theta$$

для всех θ.

Нужно разобраться с тем, что такое $\mathbb{E}_{\lnsim}$: при подсчете математического ожидания нам потребуется знание распределения (или плотность) каждого x_i, которое зависит от θ. Здесь нужно понимать, что несмещенность означает, что в среднем мы получаем θ в оценке, если считаем распределения x_i именно с этим θ (а не, например, с истинным θ_0).

Замечание 1.15. При этом важно отметить, что требование «для всех» обязательно: поскольку мы не знаем реальный θ_0. Проще всего понять это следующим образом: пусть у нас есть какая-нибудь простая оценка $\hat{\theta} f(x_1, x_2, \ldots x_n) \equiv 42$. Тогда она будет идеальной для $\theta = 42$, однако для всех остальных будет иметь смещение.

Упражнение 5. Пусть есть n приборов. Записано их время до поломки $x_1, x_2, \ldots x_n$. Пусть эта выборка порождена величиной из показательного распределения с неизвестным параметром θ:

$$F_\theta(x) = \begin{cases} 1 - e^{-\theta x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Оценить среднее время до поломки прибора. Как мы знаем, это математическое ожидание E_θ и есть эта величина, равная $\frac{1}{\theta}$.

Упражнение 6. Пусть есть выборка $x_1,\ x_2,\ \dots\ x_n$ из точек, случайно выбранных с отрезка $[0;\theta]$ (то есть $X \sim U[0;\theta]$). Возьмем оценку $\hat{\theta}_1 = \max\{x_1, x_2, \dots x_n\} = x_{(n)}$. Проверьте ее на несмещенность или поправьте так, чтобы она стала несмещенной.

Упражнение 7. Рассмотрим выборку из какого-либо распределения с двумя параметрами μ и σ, (скажем, нормального закона $\mathcal{N}(\mu, \sigma^2)$). По свойствам математического ожидания выборочное среднее X несмещенно оценивает параметр μ. В качестве оценки для неизвестной дисперсии σ^2 можно взять выборочную дисперсию

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Покажите, что такая оценка смещена (ее как раз и называют смещенная выборочная дисперсия). Как ее исправить?

Само по себе свойство несмещенности не достаточно для того, чтобы оценка хорошо приближала неизвестный параметр. Например, первый элемент X_1 выборки из закона Бернулли служит несмещенной оценкой для θ. Однако, его возможные значения 0 и 1 даже не принадлежат $\Theta = (0, 1)$. Необходимо, чтобы погрешность приближения стремилась к нулю с увеличением размера выборки. Это свойство в математической статистике называется состоятельностью.

Определение 1.16. Оценка $\hat{\theta}(x_1, x_2, \dots x_n)$ называется состоятельность, если

$$\hat{\theta_n}(x_1, x_2, \dots x_n) \xrightarrow{p} \theta$$

для всех θ.

Обычно такое свойство довольно тяжело доказать (например, можно посчитать распределение оценки). Однако есть дополнительное утверждение:

Теорема 1.17. *Если оценка $\hat{\theta}_n$ не смещена, $\mathbb{E}\hat{\theta}_n = \theta$, и дисперсия $\mathbb{D}\hat{\theta}_n \to 0$, то оценка $\hat{\theta}_n$ состоятельна.*

Упражнение 8. Пусть есть выборка из семейства $X_i \sim U[0;\theta]$. Покажите состоятельность оценки $X_{(n)}$.

Упражнение 9. Для случайных величин $X_i \sim \mathcal{N}(\theta, 1)$, докажите состоятельность оценки $\hat{\theta}_n = \overline{X}$.

1.2.3 Метод моментов

В этой главе рассматриваются несколько методов получения оценок параметров статистических моделей, в том числе — метод моментов и метод максимального правдоподобия.

Моментом $k-$го порядка случайной величины X называется величина

$$\mu_k = \mathbb{E}X^k$$

Моменты существуют не всегда. Например, у закона Коши математическое ожидание μ_1 не определено.

Положим $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$. Если момент μ_k существует, то в силу закона больших чисел

$$M_k \xrightarrow{p} \mu_k$$

На этом соображении основывается так называемый метод моментов.

Допустим, что распределение элементов выборки зависит от m неизвестных параметров $\theta_1, \dots \theta_m$. Пусть $\mathbb{E}|X|^m < \infty$ для всех $\theta \in \Theta$ (отсюда следует конечность всех моментов до m из неравенства Ляпунова). Тогда существуют все $\mu_k = \mu_k(\theta)$, $k = 1, ..., m$, и можно записать систему из m (вообще говоря, нелинейных) уравнений

$$\mu_k(\theta) = M_k \quad k = 1, 2 \dots m$$

Решение системы относительно вектора $(\hat{\theta}_1, \hat{\theta}_2, \dots \hat{\theta}_n)$ и будет оценкой по методу моментов.

Упражнение 10. Рассмотрим показательное распределение со сдвигом

$$f_\theta(x) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

Найдите оценку $\hat{\theta}$ по методу моментов.

1.2.4 Метод максимального правдоподобия

Рассмотрим другой метод получения оценок. Для начала договоримся, что величина X распределена дискретно с набором параметров θ (θ положим k-мерным вектором). Тогда вероятность того, что X реализует данное значение x есть $f(x, \theta) = P_\theta(X = x)$; в случае же выборки ее совместная вероятность реализации (в случае, как обычно, независимых случайных величин) есть:

$$f(x, \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \dots \cdot f(x_n, \theta)$$

Отметим, что функция f уже зависит от $n + k$ переменных $-$ n значений СВ и k параметров распределения. В реальной жизни нам известны (то есть фиксированы) значения выборки; получившуюся функцию только от параметров называют $L(\theta)$ $-$ функцией правдоподобия.

В таком случае легко определить верную с данной точки зрения оценку параметров θ: для нее выборка должна быть наиболее правдоподобной, то есть максимизировать свою вероятность реализации. Иными словами, $\hat{\theta}$ должно быть решением оптимизационной задачи $L(\theta) \to \max$; такие оценки называют оценками максимального правдоподобия (ОМП).

Заметим еще, что

$$\ln L(\theta) = \ln\left[f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta)\right] = \sum_{i=1}^{n} \ln f(x_i, \theta),$$

что чаще удобнее максимизировать, то есть брать производную (ввиду монотонности логарифма максимум не поменяется.

Упражнение 11. Найдите ОМП для выборки из независимых случайных величин в схеме Бернулли с вероятность θ.

Решение. Для начала рассмотрим $f(x, \theta)$ — это вероятность успеха в одном опыте. Ее можно записать таблицей, а можно и следующим более удобным образом:

$$f(k, \theta) = \theta^k (1-\theta)^{1-k},$$

где k принимает значение 0 или 1. Пусть выпишем функцию правдоподобия:

$$L(x, \theta) = \Pi_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x-i} = \theta^{x_1+x_2+\ldots+x_n} (1-\theta)^{n-(x_1+x_2+\ldots+x_n)}$$

Возьмем производную от логарифма полученного выражения:

$$\frac{d}{d\theta} \ln L(x, \theta) = \frac{d}{d\theta} \left[(x_1 + x_2 + \ldots + x_n) \ln \theta + (n - (x_1 + x_2 + \ldots + x_n)) \ln(1-\theta)\right]$$

$$\frac{d}{d\theta} \ln L(x, \theta) = \frac{x_1 + x_2 + \ldots + x_n}{\theta} - \frac{n - (x_1 + x_2 + \ldots + x_n)}{1-\theta} = 0$$

Отсюда получаем, что $\hat{\theta} = \frac{x_1+x_2+\ldots+x_n}{n} = \overline{X}$. □

В случае непрерывных величин вместо $f(x, \theta) = P_\theta(X = x)$ вероятности реализации берут плотность вероятности реализации: $f(x, \theta) = p_\theta(x)$.

Упражнение 12. Найдите ОМП показательного распределения со сдвигом:

$$f_\theta(x) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta \end{cases}$$

Можно ли здесь дифференцировать? Как полученная оценка соотносится с оценкой по методу моментов?