Математическая статистика

> ## 1
>
> date: 2022-09-05
> subject: Математическая статистика
> number: 1
> type: lection
>
> ---
>
> Баллы:
>
> - баллы: задание + теор тест + посещение
> - очно: 45 + 20 + 20
> - дист: 45 + 40 + 0
> - экзамен: 25
> - распределение: очно/дистанционно по формату — старосты #todo
>
> - [ ] сделать опрос дист/очно
>
> Необходимо знать:
>
> - законы больших чисел
> - типы распределений
>
> ---
>
> Задача Математической Статистики: вы знаете только частично о том, что вы изучаете. Возникает чёрный ящик. На основе экспериментальных данных нужно дать оценки на числовые характеристики распределения. Отвечает не на теоретические вопросы, а на практические с некоторой вероятностью.
> Возможно построение модели, которая с некоторой надёжностью предсказывает распределение.
>
> Работаем с экспериментальными данными.
>
> **Генеральная совокупность** — все результаты данной серии экспериментов (или экспериментальных значений случайно

величины).

Если эксперимент чистый, независимый, то эти данные должны в точности соответствовать случайной величине. Но нюанс - вы не можете посмотреть всевозможные результаты экспериментов.

Выборочная совокупность — имеющиеся у нас данные (выборка из генеральной совокупности, возможно неполная).

Выборочная совокупность может не отражать реальное поведение случайно величины. *байка про самолёты-бомбардировщики в вмв*. (ошибка выжившего)

Репрезентативная выборка — выборка, имеющая то же самое распределение, что и теоретическая.

В дальнейшем предполагаем, что все выборки репрезентативные.

Выборка объёма n — набор экспериментальных данных $(x_1, \ldots, x_n)$.
Выборка объёма n — набор $(X_1, \ldots, X_n)$ независимых одинаково распределённых случайных величин.

Note: поэтому $EX_i = EX_2, DX_i = Dx_2$ и используем обозначения EX_1, DX_1 и т.д.

Выборочные характеристики

Выборку можно рассматривать как дискретную случайную величину

$$\begin{array}{c|c|c|c} X & x_1 & \ldots & x_n \\ \hline p^* & \frac{1}{n} & \ldots & \frac{1}{n} \end{array}$$

Среднее выборочное — число

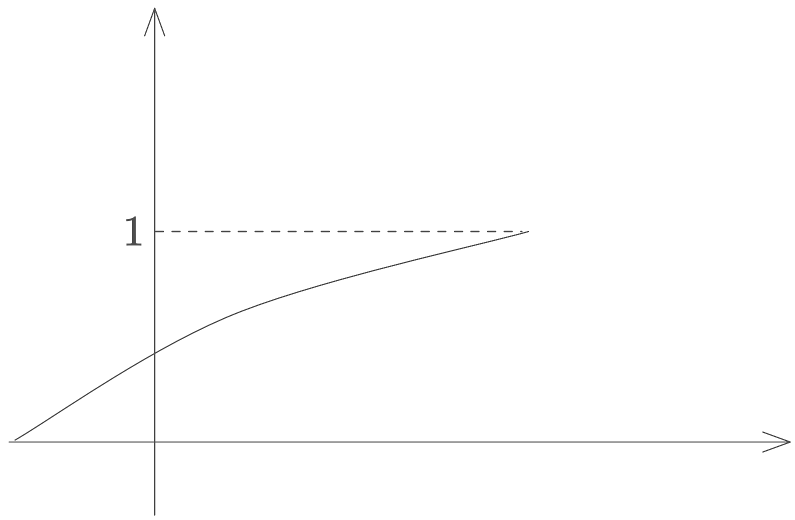
$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

Выборочная дисперсия — число

$$D_B = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Выборочная функция распределения —

$$F^*(y) = \frac{\text{число данных } x_i \in (-\infty, y)}{n}$$



Теорема Гливенко-Кантелли:
$\overline{X} = (x_1, \dots, x_n)$ объёма n, $F^*(y), F(y)$ – выборочная и теоретическая функции распределения
Тогда $\sup_{y\in\mathbb{R}} |F^*(y) - F(y)| \to 0$ при $p \to \infty$

Начальная обработка статистических данных

Данные могут быть неоднородные.
Пусть однородные, тогда её можно **ранжировать** (упорядочить данные по возрастанию). В результате получаем **вариационный ряд**

- $X_{(1)}, X_{(2)}, \dots, X_{(n)}$

i-ая порядковая статистика — $X_{(i)}$.

Статистика – некоторая процедура, обрабатывающая статистические данные.

Данные могут повторятся. Например считать оценки (2,3,4,5) среди большого количества людей.

Note: Если мы объединяем повторяющиеся результаты с учётом числа повторов, то получаем частотный вариационный ряд. (надо ли это делать или нет — смотрят по ситуации)

Другой нюанс: всё таки могут быть искажения в выборке. Бывает, что отбрасывается часть первых и часть последних (уже в вариационном ряду).

Если данных много и они не повторяются разумно разбить выборку на интервалы и составить так называем **интервально-вариационный ряд**.

Интервалы:

- равной длины (гистограммы, выдвижение гипотезы о типе распределения)
- равнонаполненные (проверка гипотез о типе распределения)

Число интервалов обычно (не всегда) берётся по формуле

$$K \approx 1 + \log_2 n$$

(иногда $\sqrt[3]{n}$ или что-то зависящее не от n, а от системы, например система оценивая итмо)

В результате получаем K интервалов $[a_{i-1}, a_i)$.
ν_i — число данных попавших в i-ый интервал.
$\frac{\nu_i}{n}$ — относительная частота. (оценка теоретической вероятности попадания случайно величины в данный интервал)

Чтобы получить из этого дискретную величине можно взять середины каждого интервала, которым будет соответствовать вероятность $\frac{\nu_i}{n}$.

$\overline{x} = \frac{1}{n} \sum c_i \nu_i \quad c_i = \frac{a_{i-1} + a_i}{2}$

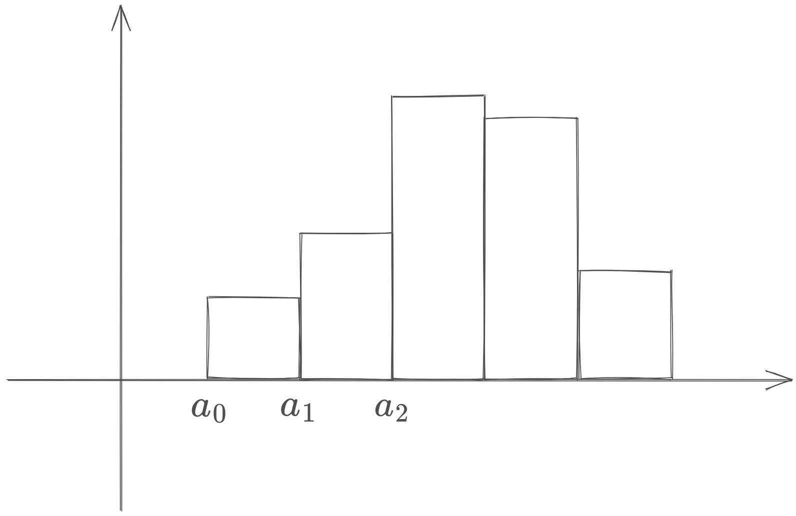
$D_B = \frac{1}{n} \sum (c_i - \overline{x})^2 \cdot \nu_i$

Геометрическая интерпретация данных

Обычно удобнее рисовать гистограммы.

Гистограмма — набор прямоугольников для каждого интервала. Основание $[a_{i-1}, a_i)$ длины $l = a_i - a_{i-1}$, а высоту берём пропорционально частоте, причём, чтобы суммарная площадь равнялась 1. высота: $\frac{\nu_i}{nl}$

Гистограмма является приближением плотности распределения (если она непрерывная) и по её виду можно выдвинуть гипотезу о типе распределения. Именно поэтому для этих целей лучше брать интервалы одинаковой длины.



Теорема:
Если число интервалов $k(n) \to \infty$ и при этом $\frac{k(n)}{n} \to 0$, то гистограмма по вероятности поточечно сходится к теоретической плотности.

Полигон — кусочно-линейная функция соединяющая точки вида (x_i, ν_i)

1

> **Математическая статистика с 2022-09-05**
>
> Сделать то, что мы делали на практике с ценами акций
>
> - разбить на промежутки длины 1 + log_2(n)
> - посчитать частоты
> - посчитать частоты/n
> - центры отрезков
> - функция распределения
> - гистограмму + полигон

2

Точечная оценка

Пусть имеется выборка объёма n: $\vec{x}(x_1, x_2, \ldots, x_n)$

Статистикой называется (измеримая - Борелевская) функция

$$\Theta^* = \Theta^*(X_1, \ldots, X_n)$$

Пусть требуется найти приближённую оценку неизвестного параметра Θ по выборке $(x_1, \ldots, x_n)$

Оценка считается при помощи некоторой статистики

Свойства статистических оценок

- состоятельность. При увеличении объёма данных повышается точность.

Статистика
$\Theta^* = \Theta^*(X_1, \ldots, X_n)$
неизвестного параметра Θ называется **состоятельной**, если
$\Theta^* \to \Theta \quad n \to \infty$

- Статистика
  $\Theta^* = \Theta^*(X_1, \ldots, X_n)$
  неизвестного параметра Θ называется **несмещённой**,если

  $$E\Theta^* = \Theta$$

  ... асимптотически несмещённой, если вместо = стоит $\to, n \to \infty$

Оценка $\Theta_1^* = \Theta_1^*(X_1, \ldots, X_n)$ не хуже оценки $\Theta_2^* = \Theta_2^*(X_1, \ldots, X_n)$, если

$$E(\Theta_1^* - \Theta*)^2 \leq E(\Theta_2^* - \Theta*)$$

Если это несмещённый оценки, то можно эквивалентно записать

$$DQ_1^* \leq DQ_2^*$$

Оценка Θ^* называется **эффективной**, если она не хуже всех остальных оценок

В классе всех возможных оценок не существует эффективной оценки.

Теорема: В классе несмещённых оценок *существует* эффективная оценка причём *единственная*.

Точные оценки моментов

> **Среднее выборочное** — число
>
> $$\overline{x} = \frac{x_1 + \dots + x_n}{n}$$
>
> **Выборочная дисперсия** — число
>
> $$D_B = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Исправленное выборочное среднее — число

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

Если выборка задана в виде частотного вариационного ряда

- $$\begin{array}{c|c|c|c} x_i & x_1 & \dots & x_k \\ \hline n_i & n_1 & \dots & n_k \end{array}$$

  , то удобнее использовать несколько другие формулы $X = \frac{1}{n}\sum x_i n_i$ и $D_B = \frac{1}{n}\sum(X_i - X)^2 n_i$

Выборочным средним квадратическим отклонением называется величина $\sigma^* = \sqrt{D_B}$, а исправленным $S = \sqrt{S^2}$

Выборочный k-ый момент — величина

$$X^k = \frac{1}{n}\sum X_i^k$$

Мода Mo^* вариационного ряда — варианта с наибольшей частотой. $Mo^* = x_i$, где $n_i = max(n_1, \ldots, n_k)$

Медиана Me^* вариационного ряда — значение варианты в середине ряда:

- $n = 2k - 1$ - нечётный, то медина это x_k
- $n = 2k$ - чётный, то $\frac{x_k + x_{k+1}}{2}$

Note: соответствующие функции в Excel: СРЗНАЧ, ДИСП.Г, ДИСП.В, СТАНДОТКЛОН.{Г,В}, МЕДИАНА, МОДА.ОДН

Теорема 1:
Выборочное среднее является несмещённой состоятельной оценкой для математического ожидания:

- Несмещённость $E\overline{X} = EX = a$
- Состоятельность $\overline{X} \to^p_{n\to\infty} EX = a$
  Доказательство:
- $E\overline{X} = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum EX_i = \frac{1}{n}\cdot nEX = EX$
- $\overline{X} = \frac{\overline{X_1} + \cdots + \overline{X_n}}{n} \to^p_{n\to\infty} EX$ по закону больших чисел Хинчина.

Теорема 2:
Выборочный k-ый момент $\overline{X^k}$ является несмещённой состоятельной оценкой теоретического k-ого момента M_k

- Несмещённость $E\overline{X^k} = m_k$
- Состоятельность $\overline{X^k} \to^p_{n\to\infty} m_k$
  Доказательство: Это следствие пред. теоремы если в качестве случайно величины взять X^k

Теорема 3:
Выборочные дисперсии являются состоятельными оценками для дисперсии. При этом D_B это смещённая вниз оценка, а S^2 несмещённая оценка.
Доказательство:

- Смещённость. Заметим, что $D_B = \frac{1}{n}\sum(X_i - \overline{X}) = \overline{X^2} - \overline{X}^2$.
  $D\overline{X} = E\overline{X}^2 - (E\overline{X})^2$

Будем смотреть

$$\begin{aligned}ED_B &= E(\overline{X^2} - \overline{X}^2) = E\overline{X^2} - E\overline{X}^2 = \\ &= E\overline{X^2} - ((E\overline{X})^2 + D\overline{X}) = EX^2 - (EX)^2 - D\overline{X} = \\ &= DX - D\overline{X} = DX - D\left(\frac{1}{n}\sum X_i\right) = \\ &= DX - \frac{1}{n^2}\sum DX_i = DX - \frac{1}{n^2}\cdot nDX = \\ &= Dx - \frac{1}{n}DX = \frac{n-1}{n}DX\end{aligned}$$

$ES^2 = E\left(\frac{n}{n-1}D_B\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}DX = DX$

- Состоятельность:
  $D_B = \overline{X^2} - \overline{X}^2 \to^p_{n\to\infty} EX^2 - (EX)^2 = DX$
  $S^2 = \frac{n}{n-1}D_B \to DX$

Отсюда видим, что D_B асимптотически несмещённая оценка, т.к. $\frac{n}{n-1} \to 1, n \to \infty$, поэтому при больших объёмах выборки (начиная со 100) обычно считают просто выборочную дисперсию. Если объём выборки мал (30-50), то обязательно заменять её на исправленную дисперсию.

Метод моментов (Пирсона)

Пусть имеет выборка неизвестного распределения. При этом из теории знаем тип распределения и что оно определяется набором параметром $\vec{\Theta} = (\Theta_1, \ldots, \Theta_n)$. Цель — дать оценку данных параметров.

Если знаем параметры распределения и его тип, то теоретические k-ые моменты можем вычислить по известным формулам. Например: Если распределение абсолютно непрерывное с плотностью $f(x, \Theta_1, \ldots, \Theta_n)$, то

$$m_i = \int_{-\infty}^{\infty} X^i f(X, \Theta_1, \ldots, \Theta_k)\,dx = h_i(\Theta_1, \ldots, \Theta_k)$$

Суть метода: в этих уравнениях теоретические моменты заменяем их оценками, после чего находим неизвестные k параметров, решая систему из k уравнений.

$$\begin{cases} \overline{X} = h_1(\Theta_1, \ldots, \Theta_k) \\ \overline{X^2} = h_2(\Theta_1, \ldots, \Theta_k) \\ \ldots \\ \overline{X^k} = h_k(\Theta_1, \ldots, \Theta_k) \end{cases}$$

Такие оценки как правило состоятельные, но смещённые.

Пример: Пусть $X \in U(a;b), a < b$ - равномерное распределение. При обработке стат. данных получили оценки первого и второго моментов. $\overline{X} = 2.25; \overline{X^2} = 6.75$. Дать оценки параметрам a, b.
Плотность

$$f(x) = \begin{cases} 0 & , x < a \\ \frac{1}{b-a} & , a \le x \le b \\ 0 & , x > b \end{cases}$$

$EX = \int_a^b X \frac{1}{b-a}\, dx = \frac{1}{b-a} \cdot \frac{b^2-a^2}{2} = \frac{a+b}{2}$
$EX^2 = \int_a^b X^2 \frac{1}{b-a}\, dx = \frac{a^2+ab+b^2}{3}$

$$\begin{cases} \overline{X} = 2.25 = \frac{a^*+b^*}{2} \\ \overline{X}^2 = 6.75 = \frac{a^{*2}+a^*b^*+b^{*2}}{3} \end{cases} \iff \begin{cases} a^* + b^* = 4.5 \\ a^*b^* = 0 \end{cases}$$

Литература:

- Как не ошибаться
- Чернова
- Логутин - Наглядная математическая статистикаю

> ## 2
>
> Известно, что данное распределение является показательным. При обработке данных получили среднее выборочное $\overline{X} = 0.86$. Дать оценку неизвестному параметру α. Является ли эта оценка а) состоятельной б) несмещённой. Если она смещена, то куда?
>
> $$f_\alpha(x) = \begin{cases} 0 & , x < 0 \\ \alpha e^{-\alpha x} & , x \ge 0 \end{cases}$$
>
> $EX = \frac{1}{\alpha} \implies \alpha = \frac{1}{EX} \implies \alpha^* = \frac{1}{\overline{X}} = \frac{1}{0.86} \approx 1.162$
> Состоятельность:

$\alpha^* = \frac{1}{\overline{X}} \to^p_{n\to\infty} \frac{1}{EX} = \alpha$

Несмещённость:

$E\alpha^* = E\frac{1}{\overline{X}} \leq \frac{1}{E\overline{X}} = \frac{1}{EX} = \alpha$ – оценка является смещённой вниз

(по неравенству Йенсена)

Математическая статистика с 2022-09-12

Посмотреть экстремумыфункций нескольких переменных, дифференциалы второго порядка, ..