Содержание

Лекция 1.

Теория вероятности изучает характеристику случайных величин, тогда как математическая статистика решает обратную задачу
Допустим, что у нас есть случайная величина, по ней мы можем найти матожидание, моменты и оценить, какое распределение имеет случайная величина.

Выборки

Def. Выборка - набор данных, полученных в ходе экспериментов. Тогда количество экспериментов n - объем Выборки
Def. Генеральной совокупностью называются все результаты проведенных экспериментов
Def. Выборочной совокупностью называются наблюдаемые данные экспериментов
Не все данные экспериментов мы можем наблюдать, например, выборы, тогда опросы голосовавших - выборочная совокупность, а результаты выборов - генеральная. Очевидно, что выборочная и генеральная совокупности могут иметь различные распределения.
Def. Выборка называется **репрезентативной**, если ее распределение близко к распределению генеральной совокупностью
Пример - ошибка выжившего. Во время Второй Мировой стал вопрос, в каких местах стоит бронировать корпус самолета. Самолеты возвращались с пулевыми отверстиям, и интуитивно казалось, что стоит бронировать те места, которые больше всего пострадали. Однако не были учтены те самолеты, которые не вернулись, а те, которые выжили, выжили благодаря тому, что были прострелены в нелетальных местах, поэтому было принято решение бронировать фюзеляж в менее пострадавших местах
В дальнейшем считаем, что все выборки репрезентативны

Def. 1. Выборкой объема n называется набор из n экспериментаных данных $\vec{X} = (x_1, x_2, \dots, x_n)$ (апостериорное определение)
Def. 2. Выборкой объема n называется набор из n независимых одинаково распределенных случайных величин $\vec{X} = (X_1, X_2, \dots, X_n)$ (априорное определение)

Выборочные характеристики

Можно выборку рассматривать как дискретную случайную величину с одинаковыми вероятностями $p_i = \dfrac{1}{n}$ и вычислить для нее математическое ожидание, дисперсию и функцию распределения

Def. Выборочным средним $\overline{x}$ называется величина $\overline{x} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$

Def. Выборочной дисперсией D^* называется величина $D^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2$ (или $D^* = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{x}^2$)

По закону больших чисел выборочное среднее будет сходиться к матожиданию

Def. Исправленной дисперсией называется величина $S^2 = \frac{n}{n-1}D^* = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$

Def. Выборочной функцией распределения $F^*(x)$ называется функция $F^*(x) = \frac{\text{число данных } x_i < x}{n}$

Th. Выборочная функция распределения поточечно сходится к теоретической функции распределения:

$$\forall y \in \mathbb{R} F^*(y) \xrightarrow{p} F(y)$$

$F(y) = P(X < y)$

$F_y^* = \frac{1}{n}\sum_{i=1}^{n} I(X_i < y) \xrightarrow[\text{по ЗБЧ}]{p} EI(X_i < y) = P(X_i < y) = P(X_1 < y) = F_{X_1}(y)$

Усилим теорему

Th. Гливенко-Кантелли. $\sup\limits_{x \in \mathbb{R}} |F^*(x) - F(x)| \xrightarrow{p} 0$

Th. Колмогорова. $\sqrt{n}\sup\limits_{x \in \mathbb{R}} |F^*(x) - F(x)| \rightrightarrows K$ - распределение Колмогорова с функцией распределения $F_K(x) = \sum\limits_{j=-\infty}^{\infty}(-1)^j e^{-2j^2x^2},\ x \in [0; \infty)$

Начальная обработка статданных

1. Ранжирование данных - упорядочиваем выборки по возрастанию. В результате получаем вариационный ряд $\vec{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$

   $X_{(1)} = \min X_i; \quad X_{(n)} = \max X_i$

   $X_{(i)} = i$-ая порядковая статистика
2. Объединим повторяющиеся данные - получаем т.н. частотный вариационный ряд

X_i	$X_{(1)}$	$\ldots$	$X_{(r)}$	$\sum$
n_i	n_1	$\ldots$	n_r	n

   Иногда часть данных отбрасывается сверху и снизу (по 5, по 10, по 5% и так далее), чтобы сделать выборку репрезентативной

   Тогда $\overline{x} = \frac{1}{n}\sum X_i n_i,\ D^* = \frac{1}{n}\sum (X_i - \overline{x})^2 n_i$

3. Чтобы уменьшить количество вычислений или сделать гистограмму, делают интервальный вариационный ряд: разбиваем данные на интервалы и считаем, сколько данных n_i попало в интервал.

   Тогда n_i - частота интервала A_i

   Есть два основные способа разбиения на интервалы:

   (a) Интервалы одинаковой длины

   (b) Равнонаполненные интервалы (в каждом интервале примерно одинаковое количество данных)

   Число интервалов K такое, что $\dfrac{K(n)}{n} \longrightarrow 0$ и $K(n) \underset{n\to\infty}{\longrightarrow} \infty$

   Обычно применяют формулу Стерджесса $K \approx 1 + \log_2 n$ или $K \approx \sqrt[3]{n}$

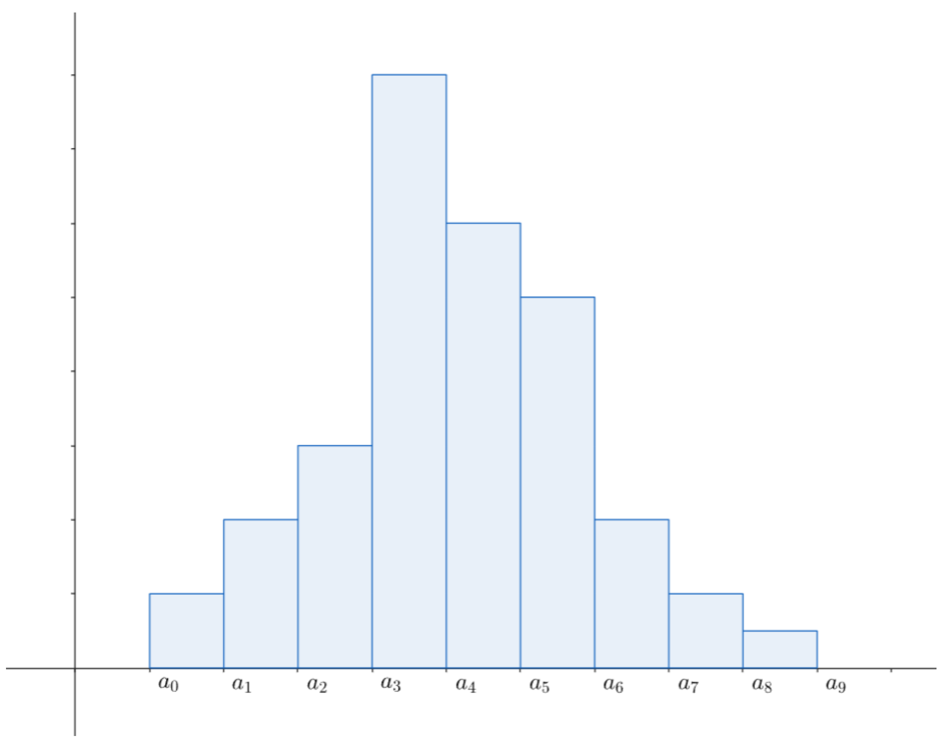
   Пусть получили интервальный вариационный ряд

интервалы	$[a_0; a_1)$	$[a_1; a_2)$	...	$[a_{K-1}; a_K]$	$\sum$
частоты	n_1	n_2	...	n_K	n

Геометрическая интерпретация данных

- Гистограмма

  Строится ступенчатая фигура из прямоугольников, основание i-ого прямоугольника - интервал, высота прямоугольника - $\dfrac{n_i}{nl_i}$, где l_i - длина интервала

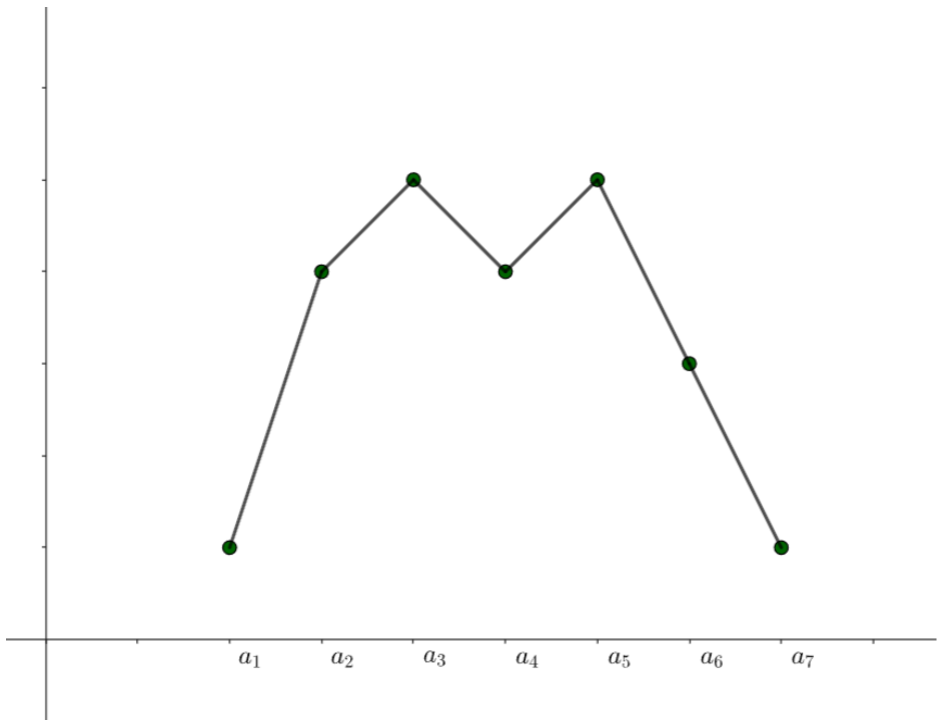


Визуально можно сделать гипотезу, как ведет себя распределение.

Th. Гистограмма поточечно сходится к теоретической плотности

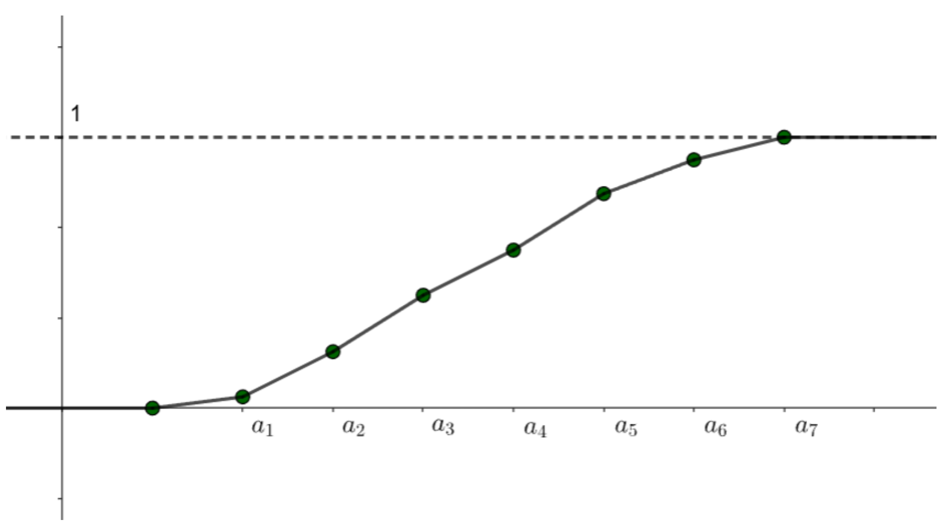
- Полигон

  На оси абсцисс отмечаем значения частотного вариационного ряда, по оси ординат - их частоты. Получившиеся точки соединяем отрезками



- Выборочная функция распределения

  На основе таблицы строится график функции распределения



  Она может быть ступенчатой, ломаной или соединена по усмотрению

Лекция 2.

Точечная оценка

Пусть имеется выборка $\vec{X} = (X_1, X_2, \ldots, X_n)$ объемом n

Пусть требуется найти приближенную оценку θ^* неизвестного параметра θ

Находим ее при помощи некоторой функции обработки данных $\theta^* = \theta^*(X_1, \ldots, X_n)$

Def. Такая функция называется статистикой

Def. А оценка θ^* называется точечной оценкой

Свойство точечных оценок

1. Состоятельность

   Def. Статистика $\theta^* = \theta^*(X_1, \ldots, X_n)$ неизвестного параметра называется состоятельной, если $\theta^* \xrightarrow{p} \theta$ при $n \to \infty$

2. Несмещенность

   Def. Оценка θ^* параметра θ называется несмещенной, если математическое ожидание $E\theta^* = \theta$

   Nota. Оценка θ^* называется асимптотически несмещенной, если $E\theta^* \xrightarrow{p} \theta$ при $n \to \infty$

3. Эффективность

   Def. Оценка θ_1^* не хуже θ_2^*, если $E(\theta_1^* - \theta)^2 \leq E(\theta_2^* - \theta)^2$. Или, если θ_1^* и θ_2^* несмещенные, то $D\theta_1^* \leq D\theta_2^*$

   Def. Оценка θ^* называется эффективной, если она не хуже всех остальных оценок

   Nota. Не существует эффективной оценки в классе всех возможных оценок

   Th. В классе несмещенных оценок существует эффективная оценка

4. Асимптотическая нормальность

   Def. Оценка θ^* параметра θ называется асимптотически нормальной, если $\sqrt{n}(\theta^* - \theta) \rightrightarrows N(0, \sigma^2(\theta))$ при $n \to \infty$

Точечные оценки моментов

Def. Выборочным средним $\overline{x}$ называется величина $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

Def. Выборочной дисперсией D^* называется величина $D^* = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{x})^2$

Def. Исправленной дисперсией S^2 называется величина $S^2 = \frac{n}{n-1}D^* = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \overline{x})^2$

Def. Выборочным средним квадратическим отклонением называется величина $\sigma^* = \sqrt{D^*}$

Def. Исправленным средним квадратическим отклонением называется величина $S = \sqrt{S^2}$

Def. Выборочным k-ым моментом называется величина $\overline{x^k} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k$

Def. Модой Mo^* называется варианта x_k с наибольшей частотой $n_k = \max\limits_i(n_1, n_2, \dots, n_m)$

Def. Выборочной медианой Me^* называется варианта x_i в середине вариационного ряда $\begin{cases} \text{Me}^* = X_{(k)}, & \text{если } n = 2k-1 \\ \frac{X_{(k)}+X_{(k+1)}}{2}, & \text{если } n = 2k \end{cases}$

Th. $\overline{x}$ - состоятельная несмещенная оценка теоретического матожидания $\mathring{A}X = a$

1) $E\overline{x} = a$

2) $\overline{x} \xrightarrow{p} a$ при $n \to \infty$

1) $E\overline{x} = E\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n}\sum\limits_{i=1}^{n} EX_i = \frac{1}{n}nEX_1 = EX_1 = a$

2) $\overline{x} = \frac{\overline{x}_1 + \dots + \overline{x}_n}{n} \xrightarrow{p} a$ согласно Закону Больших Чисел

Nota. Если второй момент конечен, то $\overline{x}$ - асимптотически нормальная оценка. По ЦПТ $\frac{S_n - nEX_1}{\sqrt{n}\sqrt{DX_1}} = \sqrt{n}\frac{\overline{x} - EX_1}{\sqrt{DX_1}} \rightrightarrows N(0,1)$ или $\sqrt{n}(\overline{x} - EX_1) \rightrightarrows N(0; DX_1)$

Th. Выборочный k-ый момент является состоятельной несмещенной оценкой теоретического k-ого момента

1) $\overline{EX^k} = EX^k$

2) $\overline{X^k} \xrightarrow{p} X^k$

Это следует из предыдущей теоремы, если взять X^k вместо X

Th. Выборочной дисперсией D^* и S^2 являются состоятельными оценками теоретической дисперсией, при этом D^* - смещенная оценка, а S^2 - несмещенная оценка

Заметим, что $D^* = \overline{X^2} - \overline{X}^2$
$ED^* = E(\overline{X^2} - \overline{X}^2) = E\overline{X^2} - E(\overline{X}^2) = EX^2 - E(\overline{X}^2)$
Так как $D\overline{X} = E(\overline{X}^2) - (E\overline{X})^2$, то $EX^2 - E(\overline{X}^2) = EX^2 - ((E\overline{X})^2 + D\overline{X}) = (EX^2 - EX) - D\overline{X} = DX - D\overline{X} = DX - D\left(\frac{X_1 + \dots + X_n}{n}\right) = DX - \frac{1}{n^2}\sum_{i=1}^{n} DX_i = DX - \frac{1}{n^2}nDX_1 = DX - \frac{1}{n}DX = \frac{n-1}{n}DX$,
то есть D^* - смещенная вниз оценка
$ES^2 = E(\frac{n}{n-1}D^*) = \frac{n}{n-1}\frac{n-1}{n}DX = DX \Longrightarrow S^2$ - несмещенная вниз оценка
2. $D^* = \overline{X^2} - \overline{X}^2 \xrightarrow{p} EX^2 - (EX)^2 = DX$ - состоятельная оценка
$S^2 = \frac{n}{n-1}D^* \xrightarrow{p} DX$

Nota. Отсюда видим, что выборочная дисперсия - асимптотически несмещенная оценка. Поэтому при большом (обычно не меньше 100) объеме выборке можно считать обычную выборочную дисперсию

Метод моментов (Пирсона)

Постановка задачи: пусть имеется выборка объема n неизвестного распределения, но известного типа, которое задается k параметрами: $\theta = (\theta_1, \theta_2, \dots, \theta_k)$. Требуется дать оценки данным неизвестным параметрам
Идея метода состоит в том, что сначала находим оценки k моментов, а затем с помощью теоретических формул из теории вероятности даем оценки этих параметров
Пусть $\vec{X}$ - выборка из абсолютно непрерывного распределения F_θ с плотностью известного типа, которая задается k параметрами $f_\theta(x, \theta_1, \dots, \theta_k)$
Тогда теоретические моменты находим по формуле $m_i = \int_{-\infty}^{\infty} x^i f_\theta(x, \theta_1, \dots, \theta_k)dx = h_i(\theta_1, \dots, \theta_n)$
Получаем систему из k уравнений с k неизвестными. В эти уравнения подставляем найденные оценки моментов и, решая получившуюся систему уравнений, находим нужные оценки параметров

$$\begin{cases} \overline{x} = h_1(\theta_1^*, \dots, \theta_n^*) \\ \overline{x^2} = h_2(\theta_1^*, \dots, \theta_n^*) \\ \dots \\ \overline{x^k} = h_k(\theta_1^*, \dots, \theta_n^*) \end{cases}$$

Nota. Оценки по методу моментов как правило состоятельные, но часто смещенные

Ex. Пусть $X \in U(a, b)$. Обработав статданные, нашли оценки первого и второго моментов: $\overline{x} = 2.25; \overline{x^2} = 6.75$

Найти оценки параметров a^*, b^*

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

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

Получаем:
$$\begin{cases} \overline{x} = \frac{a^*+b^*}{2} \\ \overline{x^2} = \frac{a^{*2}+a^*b^*+b^{*2}}{3} \end{cases} \iff \begin{cases} \frac{a^*+b^*}{2} = 4.5 \\ a^{*2}+a^*b^*+b^{*2} = 20.25 \end{cases} \iff \begin{cases} \frac{a^*+b^*}{2} = 4.5 \\ a^*b^* = 0 \end{cases} \iff \begin{cases} a^* = 0 \\ b^* = 4.5 \end{cases}$$

Лекция 3.

Метод максимального правдоподобия

Пусть имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ из распределения известного типа, определяемого неизвестными параметрами $\theta = (\theta_1, \ldots, \theta_n)$

Идея метода состоит в следующем: подбираем параметры таким образом, чтобы вероятность получения данной выборки при случайном эксперименте была наибольшей.

Если распределение дискретное, то $P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)\ldots P(X_n = x_n)$

Def. Функцией правдоподобия $L(\vec{X}, \theta)$ называется функция $L(\vec{X}, \theta) = P(X_1 = x_1)\ldots P(X_n = x_n) = \prod\limits_{i=1}^{n} P(X_i = x_i)$ при дискретном распределении

и $L(\vec{X}, \theta) = f_\theta(x_1)\ldots f_\theta(x_n) = \prod\limits_{i=1}^{n} f_\theta(x_i)$ в абсолютно непрерывном распределении

Def. Логарифмической функцией правдоподобия называется функция $\ln L(\vec{X}, \theta)$

Nota. Так как $y = \ln x$ возвростающая функция, точки максимума совпадают, а такую функцию правдоподобия становится легче дифференцировать

Def. Оценкой максимального правдоподобия $\hat{\theta}$ называется значение θ, при котором функция правдоподобия $L(\vec{X}, \theta)$ достигает наибольшего значения (при фиксированных значениях выборки)

Ex. 1. Пусть $\vec{X} = (X_1, \ldots, X_n)$ - выборка из распределения Пуассона Π_λ с неизвестным $\lambda > 0$

Мет. Для распределения Пуассона $P(X = x_i) = \dfrac{\lambda^{x_i}}{x_i!}e^{-\lambda}$

Получаем функцию максимального правдоподобия $L(\vec{X}, \lambda) = \prod\limits_{i=1}^{n} \dfrac{\lambda^{x_i}}{x_i!}e^{-\lambda} = \dfrac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}e^{-n\lambda} = \dfrac{\lambda^{n\overline{x}}}{\prod_{i=1}^n x_i!}e^{-n\lambda}$

$\ln L(\vec{X}, \lambda) = n\overline{x}\ln\lambda - \ln\prod\limits_{i=1}^{n} x_i! - n\lambda$

$\dfrac{\partial \ln L}{\partial \lambda} = \dfrac{n\overline{x}}{\lambda} - n = 0 \Longrightarrow \hat{\lambda} = \overline{x}$ - оценка максимального правдоподобия

Убедимся, что этот экстремум - максимум: $\dfrac{\partial^2 \ln L}{\partial \lambda^2} = -\dfrac{n\overline{x}}{\lambda} < 0 \Longrightarrow \hat{\lambda} = \overline{x}$ - точка максимума

Ex. 2. Пусть $(X_1, \ldots, X_n)$ из $N(a, \sigma^2)$

$f_{a,\sigma^2}(x) = \dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$

$L(\vec{X}, a, \sigma^2) = \prod\limits_{i=1}^{n} \dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-a)^2}{2\sigma^2}} = \dfrac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}e^{-\frac{\sum_{i=1}^n(x_i-a)^2}{2\sigma^2}}$

$\ln L(\vec{X}, a, \sigma^2) = -n\ln\sigma - \dfrac{n}{2}\ln 2\pi - \dfrac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i - a)^2$

$\dfrac{\partial \ln L}{\partial a} = -\dfrac{1}{2\sigma^2}\sum\limits_{i=1}^{n} -2(x_i - a) = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(x_i - a) = \dfrac{n\overline{x} - na}{\sigma^2}$

$\dfrac{\partial \ln L}{\partial \sigma} = -\dfrac{n}{\sigma} - \sum\limits_{i=1}^{n}(x_i - a)^2\dfrac{1}{2}\cdot(-2)\cdot\sigma^{-3} = \dfrac{1}{\sigma^3}\sum\limits_{i=1}^{n}(x_i - a)^2 - \dfrac{n}{\sigma}$

$\begin{cases} \frac{n\overline{x}-na}{\sigma^2} = 0 \\ \frac{1}{\sigma^3}\sum_{i=1}^n(x_i - a)^2 - \frac{n}{\sigma} = 0 \end{cases} \Longrightarrow \begin{cases} \hat{a} = \overline{x} \\ \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n(x_i - a)^2 = D^* \end{cases}$

Ex. 3. Пусть $(X_1, \ldots, X_n)$ из $U(0, \theta)$. Найти оценку θ этого распределения.

Воспользуемся методом моментов:

$EX = \dfrac{a+b}{2} = \dfrac{\theta}{2} \Longrightarrow \overline{x} = \dfrac{\theta^*}{2} \Longrightarrow \theta^* = 2\overline{x}$

Воспользуемся методом максимального правдоподобия:

$f_\theta = \begin{cases} 0, & x < 0 \\ \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & x > \theta \end{cases}$

$X_{(n)} = \max\limits_i(X_1, \ldots, X_n)$

$L(\vec{X}, \theta) = \prod\limits_{i=1}^{n} f_\theta(x_i) = \begin{cases} 0, & \text{если } \theta < X_{(n)} \\ \frac{1}{\theta^n}, & \text{если } \theta \ge X_{(n)} \end{cases}$

$L(\vec{X}, \theta)$ достигает наибольшего значения при наименьшем значении θ^n, то есть при $\hat{\theta} = X_{(n)}$

Сравним оценки:

$\theta^* = 2\overline{x}$ - несмещенная оценка, так как $E\theta^* = 2E\overline{x} = 2EX = \theta$

$E(\theta^* - \theta)^2 = D\theta^* = D2\overline{x} = 4D\overline{x} = 4\dfrac{D\overline{x}}{n} = \dfrac{4}{n}\dfrac{\theta^2}{12} = \dfrac{\theta^2}{3n}$

Изучим распределение $X_{(n)}$: $F_{X_{(n)}}(x) = P(X_{(n)} < x) = P(X_1 < x, X_2 < x, \ldots, X_n < x) = P(X_1 < x)\ldots P(X_n < x) = F_{X_1}(x)\ldots F_{X_n}(x) = F^n_{(x_1)}(x)$

$$F_{X_1}(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{\theta}, & 0 \le x \le \theta \\ 1, & x > \theta \end{cases} \Longrightarrow F_{X(n)}(x) = \begin{cases} 0, & x < 0 \\ \frac{x^n}{\theta^n}, & 0 \le x \le \theta \\ 1, & x > \theta \end{cases} \Longrightarrow f_{X(n)}(x) = \begin{cases} 0, & x < 0 \\ n\frac{x^{n-1}}{\theta^n}, & 0 \le x \le \theta \\ 1, & x > \theta \end{cases}$$

$EX_{(n)} = \displaystyle\int_0^\theta x \cdot \frac{nx^{n-1}}{\theta^n}dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{nx^{x+1}}{\theta^n(n+1)}\Big|_0^\theta = \frac{n\theta}{n+1}$ - смещенная вниз оценка

$\tilde{\theta} = \dfrac{n+1}{n}X_{(n)}$ - несмещенная оценка (будем считать, что эффективность не изменилась)

$E\tilde{\theta}^2 = E(\dfrac{n+1}{n}X_{(n)})^2 = \dfrac{(x+1)^2}{n^2}EX_{(n)} = \dfrac{(n+1)^2}{n^2}\displaystyle\int_0^\theta x^2\frac{nx^{n-1}}{\theta^n}dx = \frac{(n+1)^2\theta^2}{n(n+2)}$

$D\tilde{\theta} = E\tilde{\theta}^2 - (E\tilde{\theta})^2 = \dfrac{\theta^2}{n(n+2)}$

$D\tilde{\theta} = \dfrac{\theta^2}{n(n+2)} < \dfrac{\theta^2}{3n} = D\theta^*$

Таким образом, оценка по методу правдоподобия сходится быстрее, чем оценка по методу моментов, поэтому она лучше

Отсюда следует, что при равномерном распределении выборочное среднее не является эффективной оценкой для математического ожидания; вместо нее половина максимального элемента выборки будет лучше

Nota. Эффективной здесь будет несмещенная оценка $\dfrac{n+1}{2n}X_{(n)}$

В общем случае для $U(a,b)$ будет такая эффективная оценка матожидания - $\dfrac{X_{(1)} + X_{(n)}}{2}$, длины интервала - $\dfrac{n+1}{n-1}(X_{(n)} - X_{(1)})$

Nota. При методе максимального правдоподобия обычно получаем состоятельные и эффективные оценки, но часто смещенные

Неравенство Рао-Крамера

Пусть $X \in F_\theta$ - семейство распределений с параметром $\theta \in \mathbb{R}$

Def. Носителем семейства распределений F_θ называется множество $C \subset \mathbb{R}$ такое, что $P(X \in C) = 1\ \forall X \in F_\theta$

$$f_\theta(x) = \begin{cases} \text{плотность } f_\theta(x) \text{ при непрерывном распределении} \\ P_\theta(X = x) \text{ при дискретном распределении} \end{cases}$$

Def. Информацией Фишера $I(\theta)$ семейства распределений F_θ называется величина $I(\theta) =$

$E\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)^2$ при условии, что она существует

Def. Семейство распределений F_θ называется регулярным, если:

- существует носитель C семейства F_θ такой, что $\forall x \in C$ функция $\ln f_\theta(x)$ непрерывно дифференцируема по θ
- информация Фишера $I(\theta)$ существует и непрерывна по θ

Th. Пусть $(X_1, \ldots, X_n)$ - выборка объема n из регулярного семейства F_θ,
$\theta^* = \theta^*(X_1, \ldots, X_n)$ - несмещенная оценка параметра θ, дисперсия которой $D\theta^*$ ограничена в любой замкнутой ограниченной области параметра θ
Тогда $\boxed{D\theta^* \geq \frac{1}{nI(\theta)}}$

Следствие: если при данных условиях получили $D\theta^* = \frac{1}{nI(\theta)}$, то оценка θ^* является эффективной (то есть дальше улучшать уже некуда)

Ex. Пусть $(X_1, \ldots, X_n)$ из $N(a, \sigma^2)$ (то есть $F_a = N(a, \sigma^2)$, σ^2 зафиксируем)
Проверим эффективность $a^* = \overline{x}$
Плотность $f_a(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$, носитель - вся прямая $\mathbb{R}$
$\ln f_a(x) = -\ln\sigma - \frac{1}{2}\ln 2\pi - \frac{(x-a)^2}{2\sigma^2}, \qquad a \in (-\infty, \infty)$
$\frac{\partial}{\partial a}\ln f_a(x) = \frac{1}{2\sigma^2}\cdot 2(x-a) = \frac{x-a}{\sigma^2}$ - непрерывна для всех $a \in \mathbb{R}$
$I(a) = E\left(\frac{\partial}{\partial a}\ln f_a(X)\right)^2 = E\left(\frac{X-a}{\sigma^2}\right)^2 = \frac{1}{\sigma^4}E(X-a)^2 = \frac{E(X-EX)^2}{\sigma^4} = \frac{DX}{\sigma^4} = \frac{1}{\sigma^2}$ - непрерывна по a
Из этого следует, что $N(a, \sigma^2)$ - регулярное семейство относительно параметра a
$Da^* = D\overline{x} = \frac{DX}{n} = \frac{\sigma^2}{n}$ - ограничена по параметру a
По неравенству Рао-Крамера $Da^* = \frac{\sigma^2}{n} = \frac{1}{nI(a)} = \frac{1}{n}\sigma^2$; из этого следует, что a^* - эффективная оценка параметра a

Nota. Аналогично можно показать, что S^2 - несмещенная эффективная оценка для параметра σ^2

Лекция 4.

Основные распределения математической статистики

Def. Случайная величина имеет нормальное распределение $\xi \in N(a, \sigma^2)$ с параметрами a и σ^2, если ее плотность имеет вид $f_\xi(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$
На практике нормальное распределение встречается чаще всего в силу ЦПТ

Def. Распределение $N(0,1)$ с параметрами $a = 0, \sigma^2 = 1$ называется стандартным нормальным распределением. Его плотность равна $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. В дальнейшем такую случайную величину будем называть стандартной нормалью
Свойства

1. $a = E\xi \qquad \sigma^2 = D\xi$
2. Линейность: $\xi \in N(a, \sigma^2)$, то $\eta = b\xi + \gamma \in N(ab + \gamma, b^2\sigma^2)$
3. Стандартизация: Если $\xi \in N(a, \sigma^2)$, то $\eta = \dfrac{\xi - a}{\sigma} \in N(0,1)$
4. Устойчивость относительно суммирования: если $\xi_1 \in N(a_1, \sigma_1^2)$, $\xi_2 \in N(a_2, \sigma_2^2)$, независимы то $\xi_1 + \xi_2 \in N(a_1 + a_2, \sigma_1^2 + \sigma_2^2)$

Распределение «хи-квадрат»

Def. Распределение «хи-квадрат» H_n со степенями свободы n называется распределение суммы квадратов независимых стандартных нормальных величин: $\chi_n^2 = X_1^2 + X_2^2 + \dots + X_n^2$, где $X \in N(0,1)$ и независимы
Свойства

1. $E\chi_n^2 = n$

   Так как $\forall i \; X_i \in N(0,1)$, то $EX_i^2 = DX_i^2 + (EX_i)^2 = 1 \Longrightarrow E(X_i^2 + \dots X_n^2) = \sum\limits_{i=1}^{n} X_i^2 = n$

2. Устойчивость относительно суммирования: если $X \in H_n$, $Y \in H_m$, независимы, то $X + Y \in H_{n+m}$ (по определению)
3. $\dfrac{\chi_k^2}{k} \xrightarrow[k\to\infty]{p} 1$ (по Закону Больших Чисел)

Распределение Стьюдента

Def. Пусть $X_0, X_1, \dots, X_k$ - независимые стандартные нормальные величины. Распределением Стьюдента T_k с k степенями свободы называется распределение случайной величины $t_k = \dfrac{X_0}{\sqrt{\frac{1}{k}(X_1^2 + \dots + X_k^2)}} = \dfrac{X_0}{\sqrt{\frac{\chi_k^2}{k}}}$

<u>Свойства</u>

1. $Et_k = 0$ - в силу симметрии
2. $t_k \rightrightarrows N(0,1)$ (на практике при $k \geq 100$ распределение Стьюдента можно считать стандартным нормальным)

Распределение Фишера-Снедекера

Def. Распределением Фишера-Снедекера $F_{n,m}$ (другое название - F-распределение) со степенями свободы n и m называется распределение случайной величины $f_{n,m} = \dfrac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}}$, где χ_n^2 и χ_m^2 - независимые случайные величины с распределением «хи-квадрат»

<u>Свойства</u>

1. $Ef_{n,m} = \dfrac{n}{n-2}$
2. $f_{n,m} \xrightarrow[n,m\to\infty]{p} 1$

Математическое ожидание и дисперсия случайного вектора

Пусть $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ - случайный вектор, где случайная величина X_i - компонента (координата) случайного вектора

Def. Математическим ожидание случайного вектора называется вектор с координатами из математических ожиданий компонент: $E\vec{X} = \begin{pmatrix} EX_1 \\ \vdots \\ EX_n \end{pmatrix}$

Def. Дисперсией случайного вектора (или матрицей ковариаций) случайного вектора $\vec{X}$ называется матрица $D\vec{X} = E(\vec{X} - E\vec{X})(\vec{X} - E\vec{X})^T$, состоящая из элементов $d_{ij} = \text{cov}(X_i, X_j)$

Nota. На главной диагонали стоят дисперсии компонент: $d_{ii} = DX_i$

Nota. $D\vec{X}$ - симметричная положительно определенная матрица

<u>Свойства</u>

1. $E(A\vec{X}) = AE\vec{X}$
2. $E(\vec{X}+\vec{B}) = E\vec{X}+\vec{B}$, где $\vec{B}$ - вектор чисел
3. $D(A\vec{X}) = A\cdot D\vec{X}\cdot A^T$
4. $D(\vec{X}+\vec{B}) = D\vec{X}$

Многомерное нормальное распределение

Def. Пусть случайный вектор $\vec{\xi} = \begin{pmatrix}\xi_1\\ \vdots\\ \xi_n\end{pmatrix}$ имеет вектор средних $\vec{a} = E\vec{\xi}$, K - симметричная положительно определенная матрица. Вектор $\vec{\xi}$ имеет нормальное распределение в $\mathbb{R}^n$ с параметрами $\vec{a}$ и K, если его плотность $f_{\vec{\xi}}(\vec{X}) = \dfrac{1}{\left(\sqrt{2\pi}\right)^n\sqrt{\det K}}e^{-\frac{1}{2}(\vec{X}-\vec{a})^TK^{-1}(\vec{X}-\vec{a})}$

Свойства

1. Матрица $K = D\vec{\xi} = \left(\operatorname{cov}(\xi_i,\xi_j)\right)$ - матрица ковариаций
2. При $\vec{a} = \vec{0}$ и $K = E$ имеем вектор из независимых стандартных нормальных величин

При $\vec{a} = \vec{0}$ и $K = E$: $f_{\vec{\xi}}(X_1,\dots,X_n) = \dfrac{1}{\left(\sqrt{2\pi}\right)^n}e^{-\frac{1}{2}\begin{pmatrix}X_1 & \dots & X_n\end{pmatrix}E\begin{pmatrix}X_1 & \dots & X_n\end{pmatrix}^T} = \dfrac{1}{\left(\sqrt{2\pi}\right)^n}e^{-\frac{1}{2}(X_1^2+\dots+X_n^2)} = \dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}X_1^2}+\dots+\dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}X_n^2}$

Так как плотность распалась на произведение плотностей стандартного нормального распределение, то все компоненты имеют стандартное нормальное распределение

Далее вектором из независимых стандартных нормальных величин для краткости будем называть стандартным нормальным вектором

3. $\sqsupset\vec{X}$ - стандартный нормальный вектор, B - невырожденная матрица, тогда вектор $\vec{Y} = B\vec{X}+\vec{a}$ имеет многомерное нормальное распределение с параметрами $\vec{a}$ и $K = BB^T$
4. $\sqsupset\vec{Y}\in N(\vec{a},K)$. Тогда вектор $\vec{X} = B^{-1}(\vec{Y}-\vec{a})$ - стандартный нормальный вектор, где $B = \sqrt{K}$

   Следствие. Эквивалентное определение: Многомерное нормальное распределение - это то, которое получается из стандартного нормального вектора при помощи невырожденного преобразования и сдвиг
5. $\sqsupset\vec{X}$ - стандартный нормальный вектор, C - ортогональная матрица. Тогда $\vec{Y} = C\vec{X}$ - стандартный нормальный вектор

Так как C - ортогональная, то $C^T = C^{-1}$. Тогда по третьему свойству $K = CC^T = E$, а по второму свойству $\vec{Y}$ - стандартный нормальный вектор

6. $\sqsupset$ случайный вектор $\xi \in N(\vec{a}, K)$. Тогда его координаты независимы тогда и только тогда, когда они не коррелированы (то есть матрица ковариаций K диагональная)
Следствие. Если плотность совместного распределения случайных величин ξ и η ненулевая, то они независимы тогда и только тогда, когда их коэффициент корреляции равен нулю

Многомерная центральная предельная теорема

Th. Среднее арифметическое независимых одинаково распределенных случайных векторов слабо сходится к многомерному нормальному распределению

Лемма Фишера

Пусть вектор $\vec{X}$ - стандартный нормальный вектор, C - ортогональная матрица, $\vec{Y} = C\vec{X}$. Тогда $\forall 1 \le k \le n-1$ случайная величина $T(\vec{X}) = \sum\limits_{i=1}^{n} X_i^2 - Y_1^2 - Y_2^2 - \ldots Y_k^2$ не зависит от $Y_1, Y_2, \ldots, Y_k$ и имеет распределение «хи-квадрат» со степенями свободы $n-k$

Так как C - ортогональное преобразование, то $\|\vec{X}\| = \|\vec{Y}\|$, то есть $\sum\limits_{i=1}^{n} X_i^2 = \sum\limits_{i=1}^{n} Y_i^2 \Longrightarrow$

$T(\vec{X}) = \sum\limits_{i=1}^{n} X_i^2 - Y_1^2 - Y_2^2 - \ldots Y_k^2 = Y_{k+1}^2 + \cdots + Y_n^2$

Согласно свойству 5 $Y_i \in N(0,1)$ и независимы, то по определению «хи-квадрат» $T(\vec{X}) \in H_{n-k}$ и не зависит от $Y_1, \ldots, Y_k$

Основная теорема

Эта теорема также известна как основное следствие леммы Фишера

Th. Пусть $(X_1, \dots, X_n)$ - выборка из нормального распределения $N(a, \sigma^2)$, $\overline{x}$ - выборочное среднее, S^2 - исправленная дисперсия.
Тогда справедливы следующие высказывания:

1. $\sqrt{n}\dfrac{\overline{x} - a}{\sigma} \in N(0, 1)$
2. $\displaystyle\sum_{i=1}^{n} \frac{(X_i - a)^2}{\sigma^2} \in H_n$
3. $\displaystyle\sum_{i=1}^{n} \frac{(X_i - \overline{x})^2}{\sigma^2} = \frac{nD^*}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$
4. $\sqrt{n}\dfrac{\overline{x} - a}{S} \in T_{n-1}$
5. $\overline{x}$ и S^2 независимы

1. Так как $X_i \in N(a, \sigma^2)$, то $\displaystyle\sum_{i=1}^{n} X_i \in N(na, n\sigma^2) \Longrightarrow \overline{x} \in N\left(a, \frac{\sigma^2}{n}\right) \Longrightarrow \overline{x} - a \in N\left(0, \frac{\sigma^2}{n}\right) \Longrightarrow$ $\dfrac{\sqrt{n}}{\sigma}(\overline{x} - a) \in N(0, 1)$
2. Так как $X_i \in N(a, \sigma^2)$, то $\dfrac{X_i - a}{\sigma} \in N(0, 1)$ и $\displaystyle\sum_{i=1}^{n} \frac{(X_i - a)^2}{\sigma^2} \in H_n$ по определению
3. $\displaystyle\sum_{i=1}^{n} \frac{(X_i - \overline{x})^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - a}{\sigma} - \frac{\overline{x} - a}{\sigma}\right)^2 = \sum_{i=1}^{n} (z_i - \overline{z})^2$, где $z_i = \dfrac{X_i - a}{\sigma} \in N(0, 1)$, $\overline{z} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} z_i = \dfrac{\sum_{i=1}^{n} X_i - na}{\sigma} = \dfrac{\overline{x} - a}{\sigma}$

   Поэтому можно считать, что изначально $X_i \in N(0, 1)$

   $T(\vec{X}) = \displaystyle\sum_{i=1}^{n} (X_i - \overline{x})^2 = nD^* = n(\overline{x^2} - \overline{x}^2) = \sum_{i=1}^{n} X_i^2 - n\overline{x}^2 = \sum_{i=1}^{n} X_i^2 - Y_1^2$, где $Y_1 = \sqrt{n}\overline{x} = \dfrac{X_1}{\sqrt{n}} + \dots + \dfrac{X_n}{\sqrt{n}}$

   Строчка $\left(\dfrac{1}{\sqrt{n}}, \dots, \dfrac{1}{\sqrt{n}}\right)$ имеет длину 1, поэтому ее можно дополнить до ортогональной матрицы C, тогда Y_1 - первая компонента $\vec{Y} = C\vec{X}$, и согласно лемме Фишера $T(\vec{X}) = \displaystyle\sum_{i=1}^{n} (X_i - \overline{X})^2 \in H_{n-1}$
5. Согласно лемме Фишера $T(\vec{X}) = (n-1)S^2$ не зависит от $Y_1 = \sqrt{n}\overline{x} \Longrightarrow S^2$ и $\overline{x}$ - независимы
4. $\sqrt{n}\dfrac{\overline{x} - a}{S} = \sqrt{n}\dfrac{\overline{x} - a}{\sigma} \cdot \dfrac{1}{\sqrt{\frac{S^2(n-1)}{\sigma^2} \cdot \frac{1}{n-1}}} = \dfrac{\sqrt{n}\frac{\overline{x} - a}{\sigma}}{\frac{\chi^2_{n-1}}{n-1}}$

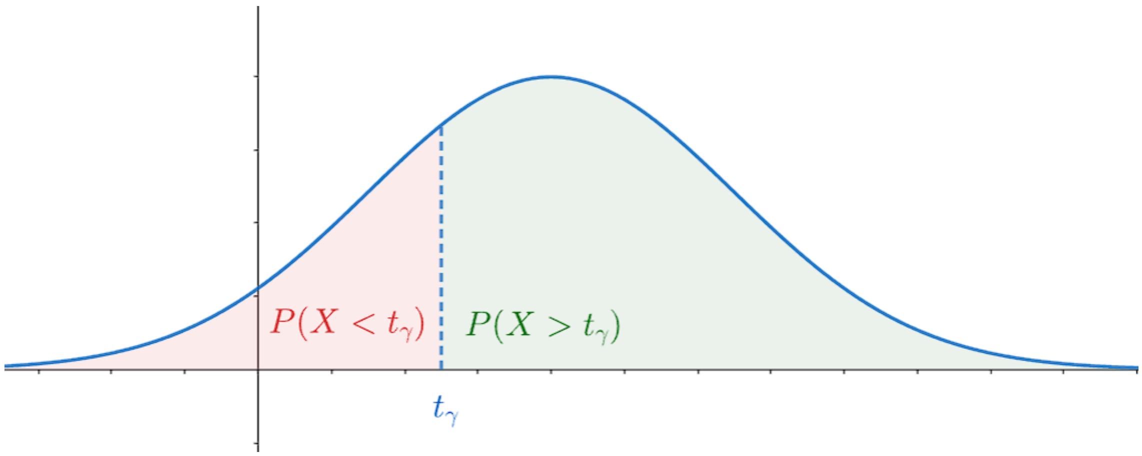
   Так как по пятому пункту числитель и знаменатель независимы, по определению получаем распределение Стьюдента

Лекция 5.

Квантильное распределение

Предполагаем, что распределение абсолютно непрерывное и $F(x)$ - функция распределения

Def. 1. Число t_γ называется квантилем распределения уровня γ, если значения функции распределения $F(t_\gamma) = \gamma$ или $P(X < t_\gamma) = \gamma$ $(t_\gamma = F^{-1}(\gamma))$



Ex. Медиана - квантиль уровня $\dfrac{1}{2}$

Def. 2. Число t_α называется квантилем уровня значимости α, если $P(X > t_\alpha) = \alpha$ или $F(t_\alpha) = 1-\alpha$
Ясно, что $\gamma = 1-\alpha$

Интервальные оценки

Недостатки точечных оценок - неизвестно насколько они далеки от реального значения параметра и насколько им можно доверять. Особенно это заметно при малых выборках. Поэтому мы указываем интервал, в котором лежит этот параметр с заданной вероятностью (надежностью) γ. Такие оценки называются интервальными (доверительными)

Def. Интервал $(\theta_\gamma^-; \theta_\gamma^+)$ называется доверительным интервалом параметра θ надежности γ, если вероятность $P(\theta_\gamma^- < \theta < \theta_\gamma^+) = \gamma$

Nota. Если имеем дискретную случайную величину, то $P(\theta_\gamma^- < \theta < \theta_\gamma^+) \geq \gamma$
Nota. Так как параметр θ - константа, то бесмысленно говорить о его попадании в интервал. Правильно: доверительный интервал накрывает параметр θ с вероятностью γ

Nota. 1. $\alpha = 1-\gamma$ называется уровнем значимости доверительного интервала

Nota. 2. Обычно пытаются строить симметричный доверительный интервал относительно несмещенной оценки θ^*

Nota. 3. Возникает вопрос, какой уровень γ выбрать для исследования. Стандартные уровени надежности γ: $0.9,\ 0.95,\ 0.99,\ 0.999$. Самый мейнстримный - 0.95. В малых выборках используют 0.9

Вспомним основную теорему:

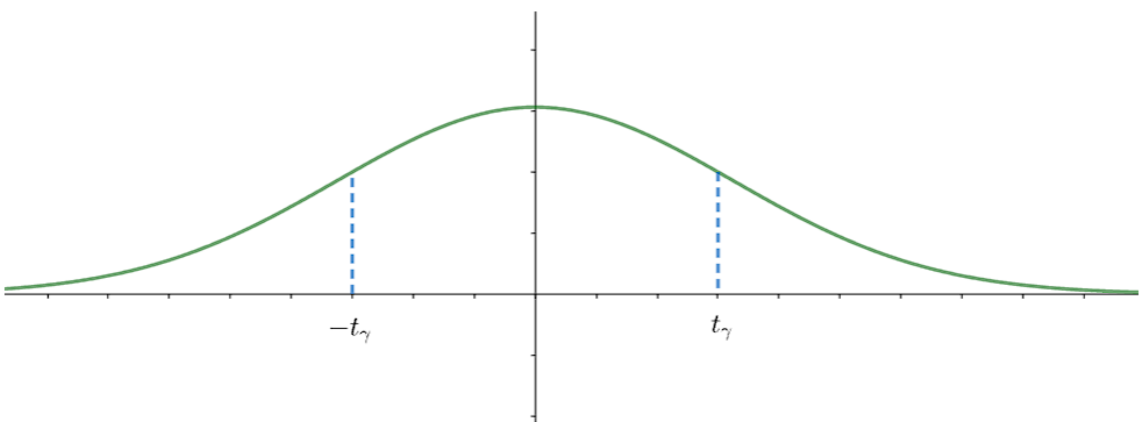
$\sqsupset (X_1, \dots, X_n)$ - выборка объема n из $N(\alpha, \sigma^2)$

$\overline{x}$ - выборочное среднее, S^2 - исправленная дисперсия, D^* - выборочная дисперсия

Тогда:

1. $\sqrt{n}\dfrac{\overline{x} - a}{\sigma} \in N(0, 1)$
2. $\sum\limits_{i=1}^{n}\left(\dfrac{X_i - a}{\sigma}\right)^2 = \dfrac{n\tilde{\sigma^2}}{\sigma^2} \in H_n$, где $n\tilde{\sigma^2} = \sum\limits_{i=1}^{n}(X_i - a)^2$
3. $\sum\limits_{i=1}^{n}\left(\dfrac{X_i - \overline{x}}{\sigma}\right)^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{nD^*}{\sigma^2} \in H_{n-1}$
4. $\sqrt{n}\dfrac{\overline{x} - a}{S} \in T_{n-1}$
5. $\overline{x}$ и S^2 - независимы

Nota. Если $F(x)$ - функция симметричного относительно $x = 0$ распределения, то $P(|X| < t) = 2F(t) - 1$



Доверительные интервалы для параметров нормального распределения

Пусть $\vec{X} = (X_1, \dots, X_n)$ - выборка объема n из $N(a, \sigma^2)$. Хотим найти интервалы для параметров a и σ^2

I. Доверительный интервал для параметра a при известном значении σ^2

По пункту 1 из теоремы $\sqrt{n}\dfrac{\overline{x}-a}{\sigma} \in N(0,1)$

$P\left(-t_\gamma < \sqrt{n}\dfrac{\overline{x}-a}{\sigma} < t_\gamma\right) = P\left(\left|\sqrt{n}\dfrac{\overline{x}-a}{\sigma}\right| < t_\gamma\right) = 2F_0(t_\gamma) - 1 = \gamma$

$F_0(t_\gamma) = \dfrac{1+\gamma}{2} \Longrightarrow t_\gamma$ - квантиль уровня $\dfrac{1+\gamma}{2}$ для $N(0,1)$, где $F_0(x) = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}dz$

Решая неравенство, получаем $-t_\gamma < \sqrt{n}\dfrac{\overline{x}-a}{\sigma} < t_\gamma$

$-t_\gamma\dfrac{\sigma}{\sqrt{n}} < \overline{x} - a < t_\gamma\dfrac{\sigma}{\sqrt{n}}$

$\overline{x} - t_\gamma\dfrac{\sigma}{\sqrt{n}} < a < \overline{x} + t_\gamma\dfrac{\sigma}{\sqrt{n}}$ - симметричный интервал относительно $\overline{x}$

Доверительный интервал надежности γ: $\left(\overline{x} - t_\gamma\dfrac{\sigma}{\sqrt{n}}, \overline{x} + t_\gamma\dfrac{\sigma}{\sqrt{n}}\right)$, где t_γ - квантиль $N(0,1)$ уровня $\dfrac{1+\gamma}{2}$

II. Доверительный интервал для параметра a при неизвестном σ^2

Из пункта 4 из теоремы $\sqrt{n}\dfrac{\overline{x}-a}{S} \in T_{n-1}$

$P\left(-t_\gamma < \sqrt{n}\dfrac{\overline{x}-a}{S} < t_\gamma\right) = P\left(\left|\sqrt{n}\dfrac{\overline{x}-a}{S}\right| < t_\gamma\right) = 2F_{T_{n-1}}(t_\gamma) = \gamma$

$F_{T_{n-1}}(t_\gamma) = \dfrac{1+\gamma}{2} \Longrightarrow t_\gamma$ - квантиль T_{n-1} уровня $\dfrac{1+\gamma}{2}$

Аналогично с примером выше получаем интервал $\left(\overline{x} - t_\gamma\dfrac{S}{\sqrt{n}}, \overline{x} + t_\gamma\dfrac{S}{\sqrt{n}}\right)$, где t_γ - квантиль T_{n-1} уровня $\dfrac{1+\gamma}{2}$

III. Доверительный интервал для параметра σ^2 при неизвестном a

По пункту 3 из теоремы $\displaystyle\sum_{i=1}^{n}\left(\dfrac{X_i - \overline{x}}{\sigma}\right)^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{nD^*}{\sigma^2} \in H_{n-1}$

Пусть χ_1^2 и χ_2^2 - квантили H_{n-1} уровней $\dfrac{1-\gamma}{2}$ и $\dfrac{1+\gamma}{2}$

Тогда $P\left(\chi_1^2 < \dfrac{(n-1)S^2}{\sigma^2} < \chi_2^2\right) = F_{H_{n-1}}(\chi_1^2) - F_{H_{n-1}}(\chi_2^2) = \dfrac{1-\gamma}{2} - \dfrac{1+\gamma}{2} = \gamma$

$\chi_1^2 < \dfrac{(n-1)S^2}{\sigma^2} < \chi_2^2$

$\dfrac{1}{\chi_2^2} < \dfrac{\sigma^2}{(n-1)S^2} < \dfrac{1}{\chi_1^2}$

$\dfrac{(n-1)S^2}{\chi_2^2} < \sigma^2 < \dfrac{(n-1)S^2}{\chi_1^2}$ или $\dfrac{nD^*}{\chi_2^2} < \sigma^2 < \dfrac{nD^*}{\chi_1^2}$

Получаем интервал $\left(\dfrac{(n-1)S^2}{\chi_2^2}, \dfrac{(n-1)S^2}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 - квантили H_{n-1} уровней $\dfrac{1-\gamma}{2}$ и $\dfrac{1+\gamma}{2}$

Nota. Данный интервал не симметричен относительно неизвестного параметра σ^2

IV. Доверительный интервал для параметра σ^2 при известном a

По пункту 2 из теоремы $\sum\limits_{i=1}^{n}\left(\dfrac{X_i-a}{\sigma}\right)^2 = \dfrac{n\tilde{\sigma^2}}{\sigma^2} \in H_{n-1}$

Пусть χ_1^2 и χ_2^2 - квантили H_n уровней $\dfrac{1-\gamma}{2}$ и $\dfrac{1+\gamma}{2}$

Тогда $P\left(\chi_1^2 < \dfrac{n\tilde{\sigma^2}}{\sigma^2} < \chi_2^2\right) = F_{H_n}(\chi_1^2) - F_{H_n}(\chi_2^2) = \dfrac{1-\gamma}{2} - \dfrac{1+\gamma}{2} = \gamma$

Аналогично получаем интервал $\left(\dfrac{n\tilde{\sigma^2}}{\chi_2^2}, \dfrac{n\tilde{\sigma^2}}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 - квантили H_n уровней $\dfrac{1-\gamma}{2}$ и $\dfrac{1+\gamma}{2}$, $n\tilde{\sigma^2} = \sum\limits_{i=1}^{n}(X_i-a)^2$

Nota. $\tilde{\sigma^2} - D^* = \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-a)^2 - \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{x})^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i^2 - 2aX_i + a^2 - X_i^2 + 2\overline{x}X_i - \overline{x}^2) = \dfrac{1}{n}(na^2 - 2an\overline{x} + 2\overline{x}\cdot n\overline{x} - n\overline{x}^2) = a^2 - 2a\overline{x} + \overline{x}^2 = (a-\overline{x})^2 \Longrightarrow \tilde{\sigma^2} = D^* + (a-\overline{x})^2$

Получаем $\left(\dfrac{n(D^*+(a-\overline{x})^2)}{\chi_2^2}, \dfrac{n(D^*+(a-\overline{x})^2)}{\chi_1^2}\right)$

Асимптотические доверительные интервалы

Def. Интервал $(\theta_\gamma^-, \theta_\gamma^+)$ называется асимптотическим доверительным интервалом параметра θ уровня γ, если $P(\theta_\gamma^- < \theta < \theta_\gamma^+) \underset{n\to\infty}{\longrightarrow} \gamma$

Ex. Доверительный интервал вероятности события A

Пусть $p = P(A), q = 1-p, n$ - число испытаний или объем выборки $(X_1, \dots, X_n)$, где $X_i \in \{0, 1\}$

$p^* = \dfrac{n_A}{n} = \overline{x}$ - оценка p

Согласно Центральной предельной теореме $\sqrt{n}\dfrac{p^*-p}{DX_1} = \sqrt{n}\dfrac{p^*-p}{\sqrt{pq}} \rightrightarrows N(0,1)$

Так как $p^* \xrightarrow{p} p$, то $\sqrt{n}\dfrac{p^*-p}{\sqrt{p^*(1-p^*)}} = \sqrt{n}\underbrace{\dfrac{p^*-p}{\sqrt{p(1-p)}}}_{\rightrightarrows N(0,1)}\underbrace{\dfrac{\sqrt{p(1-p)}}{\sqrt{p^*(1-p^*)}}}_{\xrightarrow{p} 1} \rightrightarrows N(0,1)$

$P\left(\left|\sqrt{n}\dfrac{p^*-p}{\sqrt{p^*(1-p^*)}}\right| < t_\gamma\right) \underset{n\to\infty}{\longrightarrow} 2F_0(t_\gamma) - 1 = \gamma$

$F_0(t_\gamma) = \dfrac{1+\gamma}{2}$, t_γ - квантиль $N(0,1)$ уровня $\dfrac{1+\gamma}{2}$

Получаем $\left|\sqrt{n}\dfrac{p^*-p}{\sqrt{p^*(1-p^*)}}\right| < t_\gamma$

$|p^*-p| < t_\gamma\dfrac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}$

Итак, $\left(-t_\gamma \frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n}}, t_\gamma \frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n}}\right)$, где t_γ - квантиль $N(0,1)$ уровня $\frac{1+\gamma}{2}$

Лекция 6.

Проверка статистических гипотез

$\exists \vec{X} = (X_1, \dots, X_n)$ из некоторого распределения F

Def. Гипотезой H называется предположение о распределении наблюдаемой случайной величины.
Доказать какое-то утверждение с помощью методов матстатистики невозможно - можно лишь с какой-то долей уверенности утверждать

Def. Гипотеза называется простой, если она однозначно определяет распределение: $H : F = F_1$, где F_1 - распределение известного типа с известными параметрами
В противном случае гипотеза называется сложной - она является объединением конечного или бесконечного числа гипотез
Например, «величина X принадлежит нормальному распределению» - сложная гипотеза, а «величина X принадлежит нормальному распределению с матожиданием $a = 1$ и дисперсией $\sigma^2 = 1$» - простая
В общем случае работаем со схемой из двух или более гипотез. В ходе проверки принимается ровна одна из них. Мы ограничемся самой простой схемой из 2 гипотез: H_0 - основная (нулевая) гипотеза, $H_1 = \overline{H_0}$ - альтернативная (конкурирующая) гипотеза, состоящая в том, что основная гипотеза неверна
Основная гипотеза H_0 принимается или отклоняется при помощи статистики критерия K
$K(X_1, \dots, X_n) \longrightarrow \mathbb{R} = \overline{S} \cup S \longrightarrow (H_0, H_1)$

$$\begin{cases} H_0, & \text{если } K(X_1, \dots, X_n) \in \overline{S} \\ H_1, & \text{если } K(X_1, \dots, X_n) \in S \end{cases}$$

Вместо «гипотеза доказана» лучше употреблять «гипотеза принимается/отвергается»
Область S называется критической областью, а точка $t_{\text{кр}}$ на границе областей называется критической

Def. Ошибка первого рода состоит в том, что H_0 отклоняется, хотя она верна. Аналогично, ошибка второго рода состоит в том, что H_1 отклоняется, хотя она верна.
Def. Вероятность α ошибки первого рода называется уровнем значимости критерия. Вероятность ошибки второго рода обозначаем β. Мощностью критерия называется вероятность $1 - \beta$ (вероятность недопущения ошибки второго рода)

Ясно, что критерий будет тем лучше, чем меньше вероятности ошибок α и β. При увеличении объема выборки уменьшаются обе вероятности. При фиксированном объему попытки уменьшить одну вероятность увеличат другую
Одним из способов является фиксация одной вероятности (принято α) и уменьшение другой

Построение критериев согласия

Def. Говорят, что критерий K является критерием асимптотического уровня ε, если вероятность ошибки первого рода $\alpha \xrightarrow[n\to\infty]{} \varepsilon$
Def. Критерий K для проверки гипотезы H_0 называется состоятельным, если вероятность ошибки второго рода $\beta \xrightarrow[n\to\infty]{} 0$
Def. Критерием согласия уровня ε называем состоятельный критерий асимптотического уровня ε
Обычно критерий согласия строится по следующей схеме: берется статистика $K(X_1, \dots, K_n)$, обладающая свойствами:

1. Если H_0 верна, то $K(X_1, \dots, X_n) \rightrightarrows Z$, где Z - известное распределение
2. Если H_0 неверна, то есть верна H_1, то $K(X_1, \dots, X_n) \xrightarrow[n\to\infty]{p} \infty$ (достаточно сильно отклоняться от распределения Z)

Построенный таким образом критерий является критерием согласия, то есть обладает свойствами

1. критерия асимптотического уровня
2. состоятельного критерия

Пусть $t_{\text{кр}}$ - критическая точка такая, что $P(|Z| > t_{\text{кр}}) = \varepsilon$ - заданный уровень ошибки первого рода
$\begin{cases} H_0, & \text{если } |K| < t_{\text{кр}} \\ H_1, & \text{если } |K| \geq t_{\text{кр}} \end{cases}$

1. Тогда $\alpha = P(|K| \geq t_{\text{кр}} \mid H_0) = 1 - P(|K| < t_{\text{кр}} \mid H_0) = 1 - (F_K(t_{\text{кр}}) - F_K(-t_{\text{кр}})) \xrightarrow[n\to\infty]{\substack{\text{т.к. при верной } H_0 \\ F_K(x) \xrightarrow[n\to\infty]{} F_Z(x)}} 1 - (F_Z(t_{\text{кр}}) - F_Z(-t_{\text{кр}})) = P(|Z| \geq t_{\text{кр}}) = \varepsilon$
2. Если H_1 верна, то $|K| \xrightarrow{p} \infty$, то есть $\forall C\ P(|K| > C \mid H_1) \xrightarrow{p} 1 \implies \beta = P(|K| < C \mid H_1) \xrightarrow{p} 0$

Гипотеза о среднем нормальной совокупности при известной дисперсии

$\exists \vec{X} = (X_1, \dots, X_n)$ из $N(a, \sigma^2)$, причем σ^2 известен.

Проверяется гипотеза, что $H_0 : a = a_0$, против $H_1 : a \neq a_0$ для уровня значимости α

1. По пункту 1 теоремы, если $H_0 : a = a_0$ верна, то $K = \sqrt{n}\dfrac{\overline{x} - a_0}{\sigma} = \sqrt{n}\dfrac{\overline{x} - a}{\sigma} \in N(0,1)$
2. Если верна $H_1 : a \neq a_0$, то $|K| = \sqrt{n}\left|\dfrac{\overline{x} - a_0}{\sigma}\right| = \sqrt{n}\left|\dfrac{\overline{x} - a}{\sigma} + \dfrac{a - a_0}{\sigma}\right| =$

$$= \left| \underbrace{\sqrt{n}\frac{\overline{x} - a}{\sigma}}_{\substack{\in N(0,1),\text{ограничен} \\ \text{по вероятности}}} + \underbrace{\sqrt{n}}_{\to\infty} \underbrace{\frac{a - a_0}{\sigma}}_{\text{const}} \right| \xrightarrow{p} \infty$$

Для уровня значимости α находим $t_{\text{кр}}$ такую, что $\alpha = P(|K| \geq t_{\text{кр}} \mid H_0) = P(|Z| \geq t_{\text{кр}}) \Longrightarrow P(|Z| < t_{\text{кр}}) = 2F_0(t_{\text{кр}}) - 1 = 1 - \alpha$

$F_0(t_{\text{кр}}) = 1 - \dfrac{\alpha}{2}$ - то есть $t_{\text{кр}}$ - квантиль стандартного нормального распределения уровня $1 - \dfrac{\alpha}{2}$

$$\begin{cases} H_0, & \text{если } |K| < t_{\text{кр}} \\ H_1, & \text{если } |K| \geq t_{\text{кр}} \end{cases}$$

Гипотеза о среднем нормальной совокупности при неизвестной дисперсии

1. По пункту 4 основной теоремы, если $H_0 : a = a_0$ верна, то $K = \sqrt{n}\dfrac{\overline{x} - a_0}{S} = \sqrt{n}\dfrac{\overline{x} - a}{S} \in T_{n-1}$
2. Если верна $H_1 : a \neq a_0$, то $|K| = \sqrt{n}\left|\dfrac{\overline{x} - a_0}{S}\right| = \sqrt{n}\left|\dfrac{\overline{x} - a}{S} + \dfrac{a - a_0}{S}\right| =$

$$= \left| \underbrace{\sqrt{n}\frac{\overline{x} - a}{S}}_{\substack{\in T_{n-1},\text{ограничен} \\ \text{по вероятности}}} + \underbrace{\sqrt{n}}_{\to\infty} \underbrace{\frac{a - a_0}{S}}_{\text{const}} \right| \xrightarrow{p} \infty$$

Аналогично получаем $t_{\text{кр}}$ - квантиль распределения T_{n-1} уровня $1 - \dfrac{\alpha}{2}$

Доверительные интервалы как критерии гипотез по параматрам распределения

$\sqsupset (X_1, \ldots, X_n)$ из F_θ, где F_θ - распределение известного типа с неизвестным параметром θ

Проверяется гипотеза, что $H_0 : \theta = \theta_0$, против $H_1 : \theta \neq \theta_0$

Допустим, что для θ построен доверительный интервал $(\theta_\gamma^-, \theta_\gamma^+)$, то есть $P(\theta_\gamma^- < \theta < \theta_\gamma^+) = \gamma$.

Тогда критерий $\begin{cases} H_0, & \text{если } \theta_0 \in (\theta_\gamma^-, \theta_\gamma^+) \\ H_1, & \text{если } \theta_0 \notin (\theta_\gamma^-, \theta_\gamma^+) \end{cases}$ будет уровня $\alpha = 1 - \gamma$

$\alpha = P(\theta_0 \notin (\theta_\gamma^-, \theta_\gamma^+) \mid H_0) = 1 - P(\theta_0 \in (\theta_\gamma^-, \theta_\gamma^+) \mid X \in F_{\theta_0}) = 1 - \gamma$

Поэтому доверительные интервалы можно использовать для проверки гипотез

Но почему в схеме $\begin{cases} H_0 \,:\, a = \overline{x} \\ H_1 \,:\, a \neq \overline{x} \end{cases}$ основная гипотеза всегда верна, тогда как выборочно среднее на практике почти всегда не равняется матожиданию. Потому что ...

> А вот нефиг такие гипотезы вообще выдвигать
>
> © Блаженов А. В.

Критерий вероятности появления события

$\sqsupset P(A) = p$ - вероятность успеха при одном испытании. При достаточно большом количестве испытаний n событие A появилось m раз. Проверяется $H_0 \,:\, p = p_0$ против $H_1 \,:\, p \neq p_0$
В качестве статистики критерия возьмем величину $K = \dfrac{m - np_0}{\sqrt{np_0q_0}}$

1. Если H_0 верна, то $K = \dfrac{m - np}{\sqrt{npq}} \rightrightarrows N(0,1)$ по ЦПТ
2. Lab.

Из тех же соображений $t_{\text{кр}}$ - квантиль $N(0,1)$ уровня $1 - \dfrac{\alpha}{2}$

$$\begin{cases} H_0 \,:\, p = p_0, & \text{если } |K| < t_{\text{кр}} \\ H_1 \,:\, p \neq p_0 & \text{если } |K| \geq t_{\text{кр}} \end{cases}$$

Ex. При посеве 4000 семян 970 всходов оказались рецессивного цвета, а 3030 - доминантного. Проверим гипотезу $H_0 : p = \dfrac{1}{4}$ - Мендель прав, против $H_1 : p \neq \dfrac{1}{4}$ - Мендель не прав, для уровня значимости - 0.05

$$K = \frac{m - np_0}{\sqrt{np_0q_0}} = \frac{970 - 4000 \cdot \frac{1}{4}}{\sqrt{4000 \frac{1}{4} \frac{3}{4}}} \approx -1.095$$

Так как $|K| = 1.095 < 1.96 = t_{\text{кр}}$, то $H_0 \,:\, p = \dfrac{1}{4}$ верна

Лекция 7.

Критерии для проверки гипотез о распределении

Простая параметрическая гипотеза

Пусть имеется выборка $(X_1, \ldots, X_n)$ объема n из неизвестного распределения $\mathcal{F}$. Проверяется простая гипотеза $H_0 : \mathcal{F} = \mathcal{F}_1$ против $H_1 : \mathcal{F} \neq \mathcal{F}_1$, где $\mathcal{F}_1$ - распределение известного типа с известными нами параметрами $\theta = (\theta_1, \ldots, \theta_m)$

I. **Критерий Колмогорова**

Если $\mathcal{F}_1$ - абсолютно непрерывное распределение с функцией распределения $F(x)$, то применим критерий

$\exists K = \sqrt{n}\sup\limits_x |F^*(x) - F(x)|$, где $F^*(x)$ - выборочная функция распределения

То есть используем теорему Колмогорова: если $H_0 : \mathcal{F} = \mathcal{F}_1$, то $K = \sqrt{n}\sup\limits_x |F^*(x) - F(x)| \rightrightarrows \mathcal{K}$ - распределение Колмогорова с функцией распределения $F_{\mathcal{K}}(x) = \sum\limits_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}$

Для уровня значимости α находим квантиль t_α такой, что $P(\xi \geq t_\alpha) = \alpha$, где $\xi \in \mathcal{K}$

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{F}_1, & \text{если } K < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{F}_1, & \text{если } K \geq t_\alpha \end{cases}$$

II. **Критерий «хи-квадрат» Пирсона**

Пусть выборка разбита на k интервалов $A_1, A_2, \ldots, A_k$, $A_i = [a_{i-1}, a_1)$

n_i - соответствующая частота интервала

При распределении $\mathcal{F}_1$ теоретические вероятности попадания в эти интервалы $p_i = F_{\mathcal{F}_1}(a_i) - F_{\mathcal{F}_1}(a_{i-1})$. Тогда $n_i' = p_i \cdot n$ - теоретические частоты

В качестве статистики критерия выберем $\chi^2_{\text{набл.}} = \sum\limits_{i=1}^{k} \dfrac{(n_i - n_i')^2}{n_i'}$

> **Th. Пирсона.** Если $H_0 : \mathcal{F} = \mathcal{F}_1$ верна, то $\chi^2_{\text{набл.}} \rightrightarrows \chi^2_{k-1}$ - распределение «хи-квадрат» с $k - 1$ степенями свободы

Критерий: $\exists t_\alpha$ - квантиль χ^2_{k-1} уровня α

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} \geq t_\alpha \end{cases}$$

Nota. Часто обозначают $t_\alpha = \chi^2_{\text{теор.}}$

Nota. При этом частота каждого интервала должна быть не меньше 5, а объем выборки - не меньше 50. Число интервалов лучше брать по формуле Стерджесса

Сложная параметрическая гипотеза

Здесь мы будем проверять гипотезу $H_0 : \mathcal{F} \in \mathcal{F}_\theta$ против $H_1 : \mathcal{F} \notin \mathcal{F}_\theta$, где $\mathcal{F}_\theta$ - распределение известного типа с неизвестными параметрами

III. **Критерий «хи-квадрат» Фишера**

Пусть выборка разбита на k интервалов $A_1, A_2, \ldots, A_k$, $A_i = [a_{i-1}, a_1)$

n_i - соответствующая частота интервала A_i

Def. Оценка максимального правдоподобия по частотам называется значения неизвестных параметров, при которых вероятность появления таких частот явялется максимальной

Пусть $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ - оценка максимального правдоподобия по частотам неизвестных параметров. Тогда теоретические вероятности попадания в интервал считаем по формуле $p_i = F_{\mathcal{F}_\theta}(a_i) - F_{\mathcal{F}_\theta}(a_{i-1})$, теоретическая частота - $n'_i = np_i$

В качестве статистики критерия берется функция $\chi^2_{\text{набл.}} = \sum\limits_{i=1}^{k} \dfrac{(n_i - n'_i)^2}{n'_i}$

Th. Фишера.

Если $H_0 : \mathcal{F} \in \mathcal{F}_\theta$ верна, то $\chi^2_{\text{набл.}} = \sum\limits_{i=1}^{k} \dfrac{(n_i - n'_i)^2}{n'_i} \rightrightarrows \chi^2_{k-m-1}$ - распределение «хи-квадрат», где m - число параметров неизвестного распределения

$\sqsupset t_\alpha$ - квантиль χ^2_{k-m-1} уровня α

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{F}_1, & \text{если } \chi^2_{\text{набл.}} \geq t_\alpha \end{cases}$$

Nota. Часто в качестве оценки неизвестных параметров берется просто оценка максимального правдоподобия

Ex. Имеется выборка в виде вариационного ряда $(5.2; \dots ; 22.8)$, $n = 120$, при разбиении на $k = 8$ интервалов получили

A_i	$[5.2, 7.4)$	$[7.4, 9.6)$	$[9.6, 11.8)$	$[11.8, 14)$	$[14, 16.2)$	$[16.2, 18.4)$	$[18.4, 20.6)$	$[20.6, 22.8)$	$\sum$
n_i	12	17	14	13	18	14	13	19	120

Проверить гипотезу о равномерном распределении $H_0 : \mathcal{F} \in U(a, b)$ против $H_1 : \mathcal{F} \notin U(a, b)$ при уровне значимости $\alpha = 0.05$

Дадим оценку параметров методом максимального правдоподобия: $\hat{a} = 5.2 \quad \hat{b} = 22.8$

Теоретическая вероятность будет $p'_i = \dfrac{1}{8}$, теоретическая частота - $n_i = 15$

$$\chi^2_{\text{набл.}} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i} = \frac{(12-15)^2}{15} + \frac{(17-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(13-15)^2}{15} + \frac{(18-15)^2}{15} + \frac{(14-15)^2}{15} + \frac{(19-15)^2}{15} = 3.2$$

При $\alpha = 0.05$ и $S = k - m - 1 = 5$ квантиль χ^2_S уровня α равен $t_\alpha = 11.07$

Так как $\chi^2_{\text{набл.}} < t_\alpha$, нулевая гипотеза о равномерном распределении принимается

Критерии для проверки однородности

IV. **Критерий Колмогорова-Смирнова**

Пусть имеются 2 независимых выборки $(X_1, \dots, X_n)$ и $(Y_1, \dots, Y_m)$ объемов n и m из неизвестных непрерывных распределений $\mathcal{F}$ и $\mathcal{J}$

Проверяется $H_0 : \mathcal{F} = \mathcal{J}$ (данные однородны) против $H_1 : \mathcal{F} \neq \mathcal{J}$

В качестве статистики критерия берется функция $K = \sqrt{\frac{nm}{n+m}} \sup\limits_x |F^*(x) - G^*(x)|$, где F^* и G^* - соответствующие выборочные функции распределения

Th. Колмогорова-Смирнова.
Если $H_0 : \mathcal{F} = \mathcal{J}$ верна, то $K \rightrightarrows \mathcal{K}$ - распределение Колмогорова

Критерий: t_α - квантиль $\mathcal{K}$ уровня значимости α

$$\begin{cases} H_0 : \mathcal{F} = \mathcal{J}, & \text{если } K < t_\alpha \\ H_1 : \mathcal{F} \neq \mathcal{J}, & \text{если } K \geq t_\alpha \end{cases}$$

Проверки однородности выборок из нормальных совокупностей

V. **Критерий Фишера**

Пусть имеют две независимые выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ объемов n и m из нормальных распределений $X \in N(a_1, \sigma_1^2)$ и $Y \in N(a_2, \sigma_2^2)$

Проверяется $H_0 : \sigma_1 = \sigma_2$ против $H_1 : \sigma_1 \neq \sigma_2$

В качестве статистики критерия берется функция $K = \frac{S_X^2}{S_Y^2}$, где S_X^2, S_Y^2 - соответствующие исправленные дисперсии, причем $S_X^2 \geq S_Y^2$

Th. Если $H_0 : \sigma_1 = \sigma_2$ верна, то $K = \frac{S_X^2}{S_Y^2} \in F(n-1, m-1)$ - распределение Фишера-Снедекера

По пункту 3 основной теоремы $\frac{(n-1)S^2}{\sigma^2} \in \chi^2_{n-1}$. Если H_0 верна, то $K = \frac{S_X^2}{S_Y^2} = \frac{(n-1)S_X^2 \sigma_2^2}{\sigma_1^2 (m-1) S_Y^2} \frac{(m-1)}{(n-1)} = \frac{\frac{\chi^2_{n-1}}{n-1}}{\frac{\chi^2_{m-1}}{m-1}} \in F(n-1, m-1)$

t_α - квантиль $F(n-1, m-1)$ уровня α

$$\begin{cases} H_0 : \sigma_1 = \sigma_2, & \text{если } K < t_\alpha \\ H_1 : \sigma_1 \neq \sigma_2, & \text{если } K \geq t_\alpha \end{cases}$$

Nota. Здесь критерий согласия работает чуть иным образом: при верной альтернативной гипотезе $K = \frac{S_X^2}{S_Y^2} \xrightarrow{p} \frac{\sigma_1^2}{\sigma_2^2} > 1$

Если нулевая гипотеза отклоняется, то отклоняется общая гипотеза об однородности. А если основная гипотеза принимается, то применяем критерий Стьюдента

VI. **Критерий Стьюдента**

Пусть $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ из нормальных распределений $X \in N(a_1, \sigma^2)$ и $Y \in N(a_2, \sigma^2)$

Проверяется $H_0 : a_1 = a_2$ против $H_1 : a_1 \neq a_2$

Th. $\sqrt{\frac{nm}{n+m}} \frac{(\overline{x} - a_1) - (\overline{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \in T_{n+m-2}$

По пункту 5 основной теоремы считаем, что числитель и знаменатель независимы

$$\sqrt{\frac{nm}{n+m}} \frac{(\overline{x} - a_1) - (\overline{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} = \sqrt{\frac{nm}{n+m}} \frac{\frac{\overline{x} - a_1}{\sigma} - \frac{\overline{y} - a_2}{\sigma}}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2(n+m-2)}}}$$

По пункту 1 основной теоремы $\sqrt{n}\frac{\overline{x} - a_1}{\sigma}, \sqrt{m}\frac{\overline{y} - a_2}{\sigma} \in N(0, 1) \Longrightarrow$

$$\frac{\overline{x} - a_1}{\sigma} \in N\left(0, \frac{1}{n}\right), \frac{\overline{y} - a_2}{\sigma} \in N\left(0, \frac{1}{m}\right) \Longrightarrow$$

$$\frac{\overline{x} - a_1}{\sigma} - \frac{\overline{y} - a_2}{\sigma} \in N\left(0, \sqrt{\frac{n+m}{nm}}\right) \Longrightarrow$$

$$\sqrt{\frac{nm}{n+m}} \left(\frac{\overline{x} - a_1}{\sigma} - \frac{\overline{y} - a_2}{\sigma}\right) \in N(0, 1)$$

По пункту 3 основной теоремы $\frac{(n-1)S_X^2}{\sigma^2} \in \chi^2_{n-1}, \frac{(m-1)S_Y^2}{\sigma^2} \in \chi^2_{m-1} \Longrightarrow$

$$\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2(n+m-2)} \in \frac{\chi^2_{n+m-2}}{m+n-2}$$

Из этого $\sqrt{\frac{nm}{n+m}} \frac{(\overline{x} - a_1) - (\overline{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \in \frac{N(0,1)}{\frac{\chi^2_{n+m-2}}{n+m-2}} = T_{n+m-2}$

В качестве статистики возьмем $\sqrt{\frac{nm}{n+m}} \frac{(\overline{x} - a_1) - (\overline{y} - a_2)}{\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}}$, по теореме при $a_1 = a_2$ получаем, что $K \in T_{n+m-2}$

Если верна альтернативная гипотеза, то $K \longrightarrow \infty$

Критерий: t_α - квантиль $|T_{n+m-2}|$ уровня α

$$\begin{cases} H_0 : a_1 = a_2, & \text{если } K < t_\alpha \\ H_1 : a_1 \neq a_2, & \text{если } K \geq t_\alpha \end{cases}$$

Nota. Если при обоих критериях согласились с нулевой гипотезой, то соглашаемся с гипотезой об однородности выборок

Nota. Критерий хорошо работает, если выборки из нормальных распределении (или очень близких к ним)

Лекция 8.

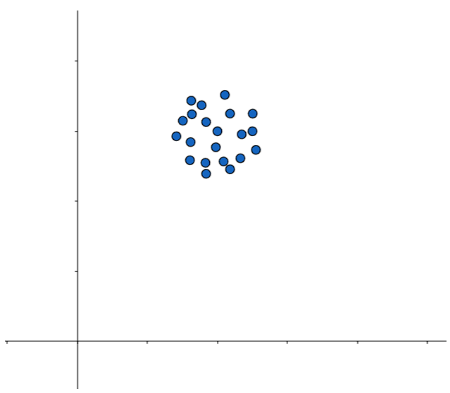
Статистическая зависимость

Def. Зависимость называется статистической, если изменение одной случайной величины вызывает изменение распределения другой. Если при этом изменяется среднее значение другой случайной величины, то такая зависимость называется коррелеционной. Если при увеличении одной случайной величины среднее значение другой также увеличивается, то говорят, что имеет место прямая корреляция. Аналогично, если уменьшается, то – обратная

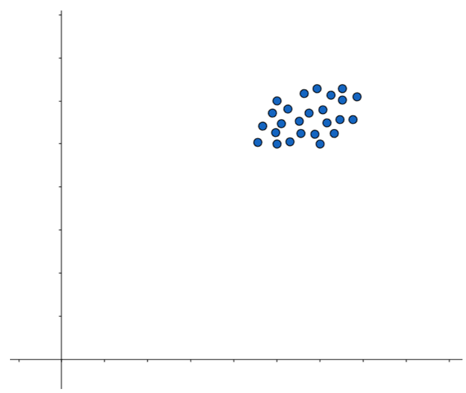
Корреляционное облако

Пусть в ходе n экспериментов появились значения двух случайных величин X и Y:
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Нанеся точки на координатную плоскость, получаем корреляционное облако, о виде которого можно делать предположения о наличии/отсутствии связи

Ex.



Здесь, возможно, нет зависимости



Здесь можно предположить прямую корреляцию

Корреляционная таблица

Пусть даны данные X и Y при n экспериментов. Эти данные удобно представить в виде коррелеционной таблицы: по вертикали отмечают различные значения x, а по горизонтали - y, в клетках таблицы отмечаются частота появления n_{xy}

Ex. $n = 50$

X\\Y	10	20	30	40	n_x	$\overline{y}_x$
2	7	3	0	0	10	13
4	3	10	10	2	25	4.4
6	0	2	10	3	15	30.67
n_y	10	15	20	5	Σ 50	

По диагонали таблицы можно предположить, что корреляция есть

Имеет смысл вычислить условное среднее по формуле $\overline{y}(x) = \dfrac{1}{n_x}\sum n_{xy}y_i$. Так как в нашем примере условные средние растут с ростом x, то имеет место прямая корреляция

Nota. Если данных много или X и Y – непрерывные случайные величины, то лучше составить интервальную корреляционную таблицу: разбить случайные величины на интервалы, по вертикали отметить интервалы $[a_{i-1}, a_i)$ случайной величины X, по горизонтали – $[b_{j-1}, b_j)$ случайной величины Y, в клетках отметить частоты $n_{ij} : [a_{i-1}, a_i) \times [b_{j-1}, b_j)$. В дальнейшем интервалы можно заменить их серединами

Критерий «хи-квадрат» для проверки независимости

Пусть выборка $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ представлена в виде интервальной корреляционной таблицы. Случайная величина X разбита на k интервалов, а Y – на m интервалов

Обозначим $v_{i\cdot}$ – частота i-ого интервала $[a_{i-1}, a_i)$ случайной величины X, $v_{\cdot j}$ – частота j-ого интервала $[b_{j-1}, b_j)$ случайной величины Y, v_{ij} – число точек в $[a_{i-1}, a_i) \times [b_{j-1}, b_j)$

X\\Y	$[b_0, b_1)$	$[b_1, b_2)$	...	$[b_{m-1}, b_m)$	$v_{i\cdot}$
$[a_0, a_1)$	v_{11}	v_{12}	...	v_{1m}	$v_{1\cdot}$
		...			
$[a_{k-1}, a_k)$	v_{k1}	v_{k2}	...	v_{km}	$v_{k\cdot}$
$v_{\cdot j}$	10	15	20	5	Σn

Проверяется основная гипотеза $H_0 : X$ и Y независимы против $H_1 = \overline{H_0} : X$ и Y зависимы

Если H_0 верна, то $p_{ij} = P(X \in [a_{i-1}, a_i), Y \in [b_{j-1}, b_j)) = P(X \in [a_{i-1}, a_i)) \cdot P(Y \in [b_{j-1}, b_j))$

Тогда по закону больших чисел $\dfrac{v_{i\cdot}}{n} \xrightarrow{p} p_{i\cdot}$, $\dfrac{v_{\cdot j}}{n} \xrightarrow{p} p_{\cdot j}$

Поэтому основанием для отклонения основной гипотезы будет заметная разница между величинами $\dfrac{v_{i\cdot}}{n}\dfrac{v_{\cdot j}}{n}$ и $\dfrac{v_{ij}}{n}$ или v_{ij} и $\dfrac{1}{n}v_{i\cdot}v_{\cdot j}$

В качестве статистики берется $K = n\sum\limits_{i,j}\dfrac{\left(v_{ij} - \frac{1}{n}v_{i\cdot}v_{\cdot j}\right)^2}{v_{i\cdot}v_{\cdot j}}$

> **Th.** Если H_0 верна, то $K \rightrightarrows H_{(k-1)(m-1)}$

Пусть t_α – квантиль $H_{(k-1)(m-1)}$ уровня α, тогда

$$\begin{cases} H_0 : X \text{ и } Y \text{ независимы, если } K < t_\alpha \\ H_0 : X \text{ и } Y \text{ зависимы, если } K \geq t_\alpha \end{cases}$$

Nota. Для работы критерия необходимо, что бы частота в каждой клетке была больше 5, а объем выборки был достаточно большой

Однофакторный дисперсионный анализ

Предположим, что на случайную величину X (результат) может влиять фактор Z (необязательно, что Z – случайная величина, эксперимент может быть управляемым)
Пусть при различных «k уровней» фактора Z получено k независимых выборок случайной величины X: $X^{(1)} = (X_1^{(1)}, \dots, X_{n_1}^{(1)}), \dots, X^{(k)} = (X_1^{(k)}, \dots, X_{n_k}^{(k)})$
Всего было получено $n = \sum\limits_{i=1}^{k} n_i$ значений

Nota. В общем говоря, распределение этих выборок отличается, поэтому эти выборки разных случайных величин

Общая, внутригрупповая и межгрупповая дисперсии

Для каждой выборки вычислим выборочное среднее и дисперсию: $\overline{x}^{(j)} = \dfrac{1}{n_j}\sum\limits_{i=1}^{n_j} X_i^{(j)}$, $D^{(j)} = \dfrac{1}{n_j}\sum\limits_{i=1}^{n_j}(X_i^{(j)} - \overline{x}^{(j)})^2$
Объединим все выборки в общую и также вычислим выборочнее среднее и дисперсию:
$\overline{x} = \dfrac{1}{n}\sum\limits_{i,j} x_i^{(j)} = \dfrac{1}{n}\sum\limits_{j=1}^{k} n_j \cdot \overline{x}^{(j)}$ - общее среднее
$D_\text{О} = \dfrac{1}{n}\sum\limits_{i,j}(X_i^{(j)} - \overline{x})^2$ - общая дисперсия

Def. Внутригрупповой (или остаточной) дисперсией называется среднее групповых дисперсий: $D_\text{В} = \dfrac{1}{n}\sum\limits_{j=1}^{k} n_j D^{(j)}$

Def. Межгрупповой (или факторной) дисперсией называется величина $D_\text{М} = \dfrac{1}{n}\sum\limits_{j=1}^{k} n_j(\overline{x} - \overline{x}^{(j)})^2$

Th. О разложении дисперсии. Общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсией: $D_\text{О} = D_\text{В} + D_\text{М}$

Смысл: внутригрупповая дисперсия показывает средний (случайный) разброс внутри выборок, межгрупповая - насколько отличаются среднее при различных уровнях фактора, то есть именно эта величина отражает влияния фактора

Вывод по наличии корреляции можно сделать, если доля D_{M} достаточно велика

Проверка гипотезы о влиянии фактора

Предполагаем, что X имеет нормальное распределение и фактор Z может влиять только на ее математическое ожидание, но на дисперсию и тип распределения, поэтому можно считать, что данные независимых k выборок при различных уровнях фактора Z также имеют нормальное распределение с одинаковой дисперсией: $X^{(j)} \in N(a_j, \sigma^2)$

Проверяется основная гипотеза $H_0 : a_1 = a_2 = \cdots = a_k$ (фактор не оказывает влияния) против $H_1 = \overline{H_0}$: есть влияние

По пункту 3 основной теоремы $\sum\limits_{i=1}^{n} \left(\dfrac{x_i - \overline{x}}{\sigma} \right)^2 = \dfrac{nD^*}{\sigma^2} \in H_{n-1}$

Из этого $\dfrac{n_j D^{(j)}}{\sigma^2} \in H_{n_j - 1} \quad \forall\ 1 \leq j \leq k$

Так как распределение «хи-квадрат» устойчиво относительно суммирования, то $\sum\limits_{j=1}^{k} \dfrac{n_j D^{(j)}}{\sigma^2} \in H_{n-k}$, так как $(n_1 - 1) + \cdots + (n_k - 1) = n - k$

Пусть основная гипотеза верна, тогда все данные можно считать выборкой одной случайной величины и по пункту 3 $\dfrac{nD_{\text{O}}}{\sigma^2} \in H_{n-1}$

Согласно теореме о разложении дисперсии $D_{\text{O}} = D_{\text{B}} + D_{\text{M}}$, тогда $\dfrac{nD_{\text{O}}}{\sigma^2} = \dfrac{nD_{\text{B}}}{\sigma^2} + \dfrac{nD_{\text{M}}}{\sigma^2}$

Так как $\dfrac{nD_{\text{O}}}{\sigma^2} \in H_{n-1}$, $\dfrac{nD_{\text{B}}}{\sigma^2} \in H_{n-k}$, то $\dfrac{nD_{\text{M}}}{\sigma^2} \in H_{k-1}$

Тогда при верной основной гипотезе получим, что $\dfrac{nD_{\text{M}}}{\sigma^2} \dfrac{\sigma^2 (n-k)}{nD_{\text{B}}} = \dfrac{n-k}{n-1} \dfrac{D_{\text{M}}}{D_{\text{B}}} \in F(k-1, n-k)$ – распределение Фишера-Снедекера со степенями $k-1$ и $n-k$

В качестве статистики берется $K = \dfrac{n-k}{n-1} \dfrac{D_{\text{M}}}{D_{\text{B}}}$, в качестве критической точки t_α – квантиль $F(k-1, n-k)$ уровня α

$$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_k \text{ (фактор оказывает влияние), если } K < t_\alpha \\ H_0 : \text{фактор влияния не оказывает, если } K \geq t_\alpha \end{cases}$$

Лекция 9.

Исследование статистической корреляции

Математическая модель регрессии

Пусть случайная величина X зависит от случайной величины Z (необязательно случайной)

Def. Регрессией X на Z называется функция $f(z) = E(X|Z = z)$. Она показывает зависимость среднего значения X от значения Z

Уравнение $x = f(z)$ называется уравнением регрессии, а график этой функции - линия регрессии

Пусть при n экспериментах при значениях $Z_1, Z_2, \dots, Z_n$ фактора Z наблюдались значения $X_1, X_2, \dots, X_n$ случайной величины X

Обозначим через ε_i разницу между экспериментальным и теоретическими значениями случайной величины X, то есть $\varepsilon_i = X_i - f(z_i)$

ε - это случайный член модели или так называемая теоретическая ошибка

Nota. Обычно можно считать, что ε_i независимы друг от друга и имеет нормальное распределение с $a = 0$, так как $E\varepsilon_i = E(X_i - f(Z_i)) = E(X|Z = Z_i) - E(X|Z = Z_i) = 0$

Цель: нам нужно по экспериментальным данным $(z_1, x_1), \dots, (z_n, x_n)$ как можно лучше оценить функцию $f(z)$

Nota. При этом предполагая (часто из теории), что $f(z)$ - функция определенного вида, но параметры которой неизвестны. Если нет, то начинаем подбирать модели самого простого вида. В противном случае, наилучшим решением была бы кривая, проходящая через все точки

Метод наименьших квадратов

Пусть известен из теории вид функции $f(z)$. Метод наименьших квадратов состоит в выборе параметров $f(z)$ таким образом, чтобы минимизировать сумму квадратов ошибок $\sum\limits_{i=1}^{n} \varepsilon_i^2 = \sum\limits_{i=1}^{n} (X_i - f(Z_i))^2 \to \min$

Def. Пусть θ - набор неизвестных параметров функции $f(z)$. Оценка $\hat{\theta}$ параметра θ, при которой достигается минимум $\sum\limits_{i=1}^{n} \varepsilon_i^2$, называется оценкой метода наименьших квадратов (или ОМНК)

Линейная парная регрессия

Пусть имеется теоретическая модель линейной регрессии

$f(z) = \alpha + \beta z + \varepsilon$ - теоретическая модель, где ε - теоретическая ошибка отражающая влияние невключенных в модель факторов, возможной нелинейности, ошибок измерения и просто случая

Пусть $(z_1, x_1), \dots, (z_n, x_n)$ - экспериментальные данные. По ним методом наименьших квадратов строим экспериментальную модель линейной регрессии $f(z) = a + bz$, где a и b - ОМНК параметров α и β

$\hat{\varepsilon}_i = X_i - f(Z_i) = X_i - (a + bZ_i)$ - экспериментальная ошибка

Найдем ОМНК параметров α и β

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}(X_i-(a+bZ_i))$$
$$\frac{\partial}{\partial a}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}-2(X_i-a-bZ_i)=-2\sum_{i=1}^{n}X_i+2\sum_{i=1}^{n}a+2b\sum_{i=1}^{n}Z_i=2(n\overline{x}-na-bn\overline{z})$$
$$\frac{\partial}{\partial b}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}-2Z_i(X_i-a-bZ_i)=-2\sum_{i=1}^{n}X_iZ_i+2\sum_{i=1}^{n}aZ_i+2b\sum_{i=1}^{n}Z_i^2=2(n\overline{xz}-an\overline{z}-bn\overline{z^3})$$
$$\begin{cases}-2(n\overline{x}-na-nb\overline{z})=0\\ -2(n\overline{zx}-na\overline{z}-nb\overline{z^2})=0\end{cases}\iff\begin{cases}a+b\overline{z}=\overline{x}\\ a\overline{z}+b\overline{z^2}=\overline{zx}\end{cases}$$

Получили систему линейных уравнений. Будем называть ее нормальной системой. При решении получаем:

$$\begin{cases}a=\overline{x}-b\overline{z}\\ (\overline{x}b\overline{z})\overline{z}+b\overline{z^2}=\overline{zx}\end{cases}\iff\begin{cases}a=\overline{x}-b\overline{z}\\ b=\frac{\overline{zx}-\overline{z}\,\overline{x}}{\hat{\sigma}_z^2}\end{cases}\text{ - ОМНК}$$

Запишем уравнение линейной регрессии в удобном виде: $\overline{x}_z=f(z)=E(X|Z=z)$

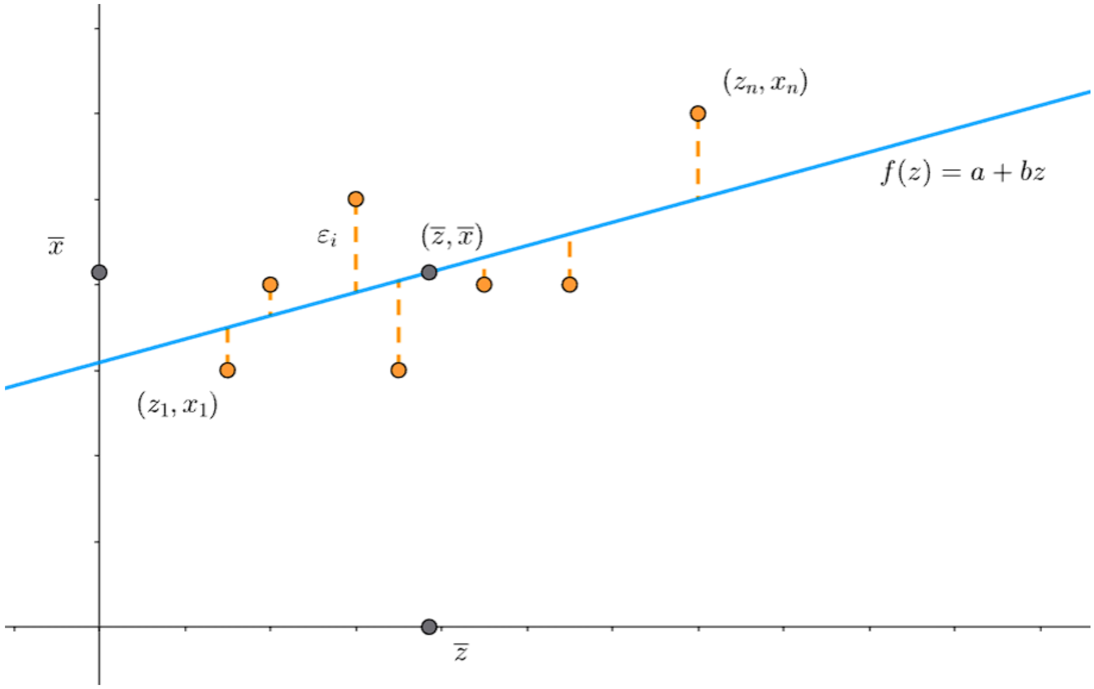
$$\overline{x}_z=a+bz$$
$$\overline{x}_z=\overline{x}-b\overline{z}+bz$$
$$\overline{x}_z-\overline{x}=\frac{\overline{zx}-\overline{x}\,\overline{z}}{\hat{\sigma}_z^2}(z-\overline{z})$$
$$\overline{x}_z-\overline{x}=\frac{\hat{\sigma}_x}{\hat{\sigma}_z}\frac{\overline{zx}-\overline{x}\,\overline{z}}{\hat{\sigma}_z\hat{\sigma}_x}(z-\overline{z})=\frac{\hat{\sigma}_x}{\hat{\sigma}_z}\hat{r}(z-\overline{z}),$$ где $\hat{r}$ - выборочный коэффициент линейной корреляции

Или $\dfrac{\overline{x}_z-\overline{x}}{\hat{\sigma}_x}=\hat{r}\dfrac{z-\overline{z}}{\hat{\sigma}_z}$ - выборочное уравнение линейной регрессии

Nota. Прямая регрессии проходит через точку из выборочных средних

Nota. При $n\to\infty$ $\overline{x}\longrightarrow EX,\overline{z}\longrightarrow EZ,\hat{\sigma}_x\longrightarrow\sigma_x,\hat{\sigma}_z\longrightarrow\sigma_z,\overline{x}_z\longrightarrow E(X|Z=z),\hat{r}\longrightarrow r$, получаем $\dfrac{E(X|Z=z)-EX}{\sigma_x}=r\dfrac{z-EZ}{\sigma_z}$ - теоретическое уравнение линейной регрессии

Геометрический смысл линии регрессии



Суть МНК: находим такую прямую, чтобы сумма квадратов длин этих отрезков (по сути отклонений) была минимальна (или дисперсия экспериментальных данных относительно прямой была минимальна)

Выборочный коэффициент линейной корреляции

Def. $\hat{r} = \dfrac{\overline{zx} - \overline{x}\,\overline{z}}{\hat{\sigma}_z \hat{\sigma}_x}$ называется выборочным коэффициентов линейной корреляции. Ясно, что она будет точечной оценкой теоретического коэффициента линейной корреляции. Также $\hat{r}$ является несмещенной оценкой

Поэтому выборочный коэффициент корреляции характеризует силу линейной связи. Знак коэффициента показывает направления корреляции (прямая или обратная)

Силу связи можно примерно оценить от шкале Чеддока:

Количественная мера $\hat{r}$	Качественная мера
0.1 – 0.3	Слабая
0.3 – 0.5	Умеренная
0.5 – 0.7	Заметная
0.7 – 0.9	Высокая
> 0.9	Весьма высокая

Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть (Z, X) распределена нормально. По выборке объема n вычислен выборочный коэффициент корреляции $\hat{r}$, а r - теоретический коэффициент корреляции

Проверяется $H_0 : r = 0$ (выборочный коэффициент корреляции статистически незначим) против $H_1 : r \neq 0$ (коэффициент статистически значим)

Если H_0 верна, то $K = \dfrac{\hat{r}\sqrt{n-2}}{\sqrt{n-\hat{r}^2}} \in T_{n-2}$ - распределение Стьюдента с степенью $n-2$

Получаем критерий. Пусть t_α - квантиль $|T_{n-2}|$ (двухстороннее распределение Стьюдента) уровня α

$$\begin{cases} H_0 : r = 0, & \text{если } |K| < t_\alpha \\ H_1 : r \neq 0, & \text{если } |K| \geq t_\alpha \end{cases}$$

Надо понимать, что корреляция - более тонкое понятие, чем зависимость

А термин *регрессия* получил свое название чисто исторически: статистик Гальтон в 1886 году исследовал зависимость роста детей от роста родителей

$E(\text{Р}_{\text{сына}} | Z_{\text{отца}} = Z_1, Z_{\text{матери}} = Z_1) = 0.27Z_1 + 0.2Z_2 + \text{const}$

$E(\text{Р}_{\text{дочери}} | Z_{\text{отца}} = Z_1, Z_{\text{матери}} = Z_1) = \dfrac{1}{1.08}\text{Р}_{\text{сына}}$

Дальше он заметил, что при у самых высоких родителей рост детей был меньше относительно них (скатывался к среднему, происходил регресс)

Позже исследовали экономические результаты фирм, показатели спортсменов, которые после успешного сезона уменьшались, после чего появлялось куча теорий. Сейчас все это объясняется простым случаем

Выборочное корреляционное отношение

Выборочный коэффициент корреляции характеризует только силу линейной связи. Следующий подход основан на однофакторном дисперсионном анализе

Пусть есть k выборок случайной величины X при k различных уровнях фактора Z. Вычислены общая, внутригрупповая и межгрупповая дисперсии. По теореме $D_{\text{О}} = D_{\text{М}} + D_{\text{В}}$

Def. Выборочным корреляционным отношением X на Z называется величина $\eta_{X,Z} = \sqrt{\dfrac{D_{\text{М}}}{D_{\text{О}}}}$

Свойства:

1. $0 \leq \eta_{X,Z} \leq 1$ $(D_{\text{М}}, D_{\text{О}} \geq 0)$
2. Если $\eta = 1$, то $D_{\text{М}} = D_{\text{О}} \Longrightarrow D_{\text{В}} = 0$, имеем функциональную зависимость X от Z
3. Если $\eta = 0$, то $D_{\text{М}} = 0 \Longrightarrow$ корреляция отсутствует

4. $\eta \geq |\hat{r}|$
5. Если $\eta = |\hat{r}|$, то все точки экспериментальных данных лежат на прямой линейной регрессии (то есть данная линейная модель является идеальной)