Математическая статистика

Михайлов Максим

20 октября 2021 г.

Оглавление

Лекция 1

6 сентября

1 Организационные вопросы

Большая часть баллов зарабатывается индивидуальными заданиями, выполняемыми в Excel — 30 баллов. Тест с большим числом вопросов — 20 или 25 баллов.

2 Введение

Теория вероятности состоит в следующем: исследуется случайная величина с заданным распределением. Математическая статистика занимается обратным — даны данные, нужно приближенно найти числовые характеристики случайной величины и с некоторой уверенностью найти вид распределения. Матстатистика также исследует связанность случайных величин, их корреляцию. В идеале есть цель построить модель, которая по значениям одних случайных величин предсказывает другие.

Пусть проводится некоторое количество экспериментов, в ходе которых появились некоторые данные.

Определение. **Генеральная совокупность** — набор всех исходов проведенных экспериментов.

В реальности наблюдается некоторая выборка генеральной совокупности, ибо рассматривать всю совокупность нецелесообразно.

Определение. **Выборочная совокупность** — исходы наблюдаемых экспериментов.

Определение. Выборка называется **репрезентативной**, если её распределение совпадает с распределением генеральной совокупности.

Выборка может быть нерепрезентативной, как в примере с ошибкой выжившего. Мы считаем, что таких ошибок у нас нет и все выборки репрезентативны, ибо исправление этих ошибок — задача конкретной области, в которой используется матстатистика.

Определение (после опыта). Пусть проведено n наблюдаемых независимых экспериментов, в которых случайная величина приняла значение $X_1, X_2 \dots X_n$. Набор[1] этих данных называется **выборкой объема** n.

Определение (до опыта). **Выборкой объема** n называется набор из n независимых одинаково распределенных случайных величин.

Пусть имеется выборка в смысле "после опыта" объема n. Её можно интерпретировать как следующую дискретную случайную величину:

X_i	X_1	X_2	$\dots$	X_n	$\sum$
p_i	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$	1

Средневыборочное:

$$\overline{X} \coloneqq \sum_{i=1}^{n} \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Выборочная дисперсия:

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

2.1 Выборочная функция распределения

$$F_n^*(z) \coloneqq \frac{1}{n} \sum_{i=1}^{n} I(X_i < z) = \frac{\text{число } X_i \in (-\infty, z)}{n}$$

Примечание. I — индикатор:

$$I(X_i < z) = \begin{cases} 1, & X_i < z \\ 0, & X_i \geq z \end{cases}$$

Теорема 1.

$$\forall x \in \mathbb{R} \quad F_n^*(z) \xrightarrow[n \to \infty]{P} F(z)$$

Доказательство. Заметим, что

$$\mathbb{E} I(X_1 < z) = 1 \cdot P(X_1 < z) + 0 \cdot P(X_1 \geq z) = P(X_1 < z) = F(z)$$

[1] Или вектор.

, где $F(z)$ — функция распределения X_1. Заметим, что $F(z) \le 1 < \infty$, следовательно применим ЗБЧ Хинчина:

$$F_n^*(z) = \frac{\sum_{i=1}^n I(X_i < z)}{n} \xrightarrow{P} \mathbb{E}I(X_1 < z) = F(z)$$

□

Примечание. На самом деле имеется даже равномерная сходимость по вероятности — это теорема Гливенко-Кантелли:

$$\sup_{z\in\mathbb{R}} |F_n^*(z) - F(z)| \xrightarrow[n\to\infty]{P} 0$$

3 Первоначальная обработка статданных

Если отсортировать данные, то получим **вариационный ряд**: $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$. Если учесть повторяющиеся экземпляры, то получим **частотный вариационный ряд**:

$X_{(i)}$	$X_{(1)}$	$X_{(2)}$	$\dots$	$X_{(k)}$	$\sum$
n_i	n_1	n_2	$\dots$	n_k	n
p_i^*	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\dots$	$\frac{n_k}{n}$	1

Определение. $h \coloneqq X_{\max} - X_{\min}$ — **размах выборки**

Допустим, что разбили интервал $(X_{\min}, X_{\max})$ на k интервалов, чаще всего одинаковой длины.[2] Тогда $l_i = \frac{h}{k}$ — длина каждого интервала и интервальный ряд можно заменить интервальным вариационным рядом.

i	l_1	l_2	$\dots$	l_k	$\sum$
m_i	m_1	m_2	$\dots$	m_k	n
$\frac{m_i}{n}$	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\dots$	$\frac{m_k}{n}$	1

m_i — число попавших в i-тый интервал данных.

По такой таблице можно построить **гистограмму**. На координатной плоскости построим прямоугольники с основаниями l_i и высотами $\frac{m_i}{nl_i}$. В результате получаем ступенчатую фигуру площади 1, которая и называется гистограммой.

Теорема 2. При $n \to \infty, k(n) \to \infty$, причем $\frac{k(n)}{n} \to 0$, гистограмма будет стремиться к плотности распределения:

$$\frac{m_i}{n} \xrightarrow{P} P(X_i \in l_i) = \int_{l_i} f(x)dx$$

[2] Применяются и другие разбиения, например равнонаполненное.

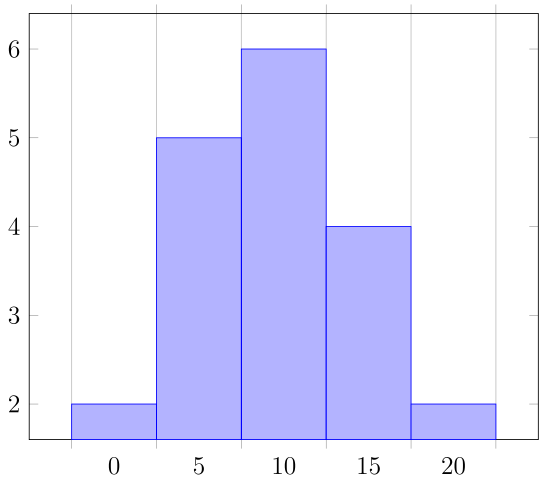

Рис. 1.1: Пример гистограммы

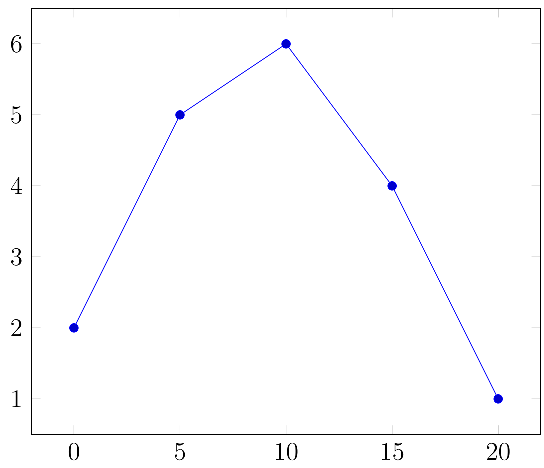

Рис. 1.2: Пример полигона

Чаще всего число интервалов берется по формуле Стёрджесса: $k \approx 1 + \log_2 n$. Иногда $k \approx \sqrt[3]{n}$.

Примечание. Иногда выборка изображается в виде **полигона**: отображаются точки, соответствующие серединам интервалов и ставим точки на высоте $\frac{m_i}{n}$.

Лекция 2

13 сентября

4 Точечные оценки

Пусть имеется выборка объема n: $X = \begin{pmatrix} X_1 & \dots & X_n \end{pmatrix}$

Определение. **Статистикой** называется измеримая функция $\theta^* = \theta^*(X_1, \dots, X_n)$.

Пусть требуется найти значение параметра θ случайной величины X по данной выборке. Оценку будем считать с помощью некоторой статистики θ^*.

4.1 Свойства статистических оценок

4.1.1 Состоятельность

Определение. Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется **состоятельной оценкой** параметра θ, если:

$$\theta^* \xrightarrow[n\to\infty]{P} \theta$$

4.1.2 Несмещённость

Определение. Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется **несмещенной оценкой** параметра θ, если

$$\mathbb{E}\theta^* = \theta$$

Примечание. То есть с равной вероятностью можем ошибиться как в меньшую, так и в большую сторону. Нет систематической ошибки.

Определение. Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется **асимптотически несмещенной оценкой** параметра θ, если

$$\mathbb{E}\theta^* \xrightarrow[n\to\infty]{} \theta$$

Примечание. То есть при достаточно большом объеме выборки ошибка исчезает, но при малом она может существовать.

4.1.3 Эффективность

Определение. Оценка θ_1^* **не хуже** оценки θ_2^*, если

$$\mathbb{E}(\theta_1^* - \theta)^2 \le \mathbb{E}(\theta_2^* - \theta)^2$$

или, если оценки несмещенные,

$$\mathbb{D}\theta_1^* \le \mathbb{D}\theta_2^*$$

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

Теорема 3. Не существует эффективной оценки в классе всех возможных оценок.

Теорема 4. В классе несмещённых оценок существует эффективная оценка.

4.2 Точечные оценки моментов

Определение. **Выборочным средним** $\overline{X_B}$ называется величина

$$\boxed{\overline{X_B} = \frac{1}{n}\sum_{i=1}^{n} X_i}$$

Определение. **Выборочной дисперсией** $\mathbb{D}_B$ называется величина

$$\boxed{\mathbb{D}_B = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_B})^2}$$

Определение. **Исправленной выборочной дисперсией** S^2 называется величина

$$\boxed{S^2 = \frac{n}{n-1}\mathbb{D}_B}$$

или

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

Определение. **Выборочным средним квадратическим отклонением** называется величина

$$\boxed{\sigma_B = \sqrt{\mathbb{D}_B}}$$

Определение. **Исправленным выборочным средним квадратическим отклонением** называется величина

$$\boxed{S = \sqrt{S^2}}$$

Определение. **Выборочным k-тым моментом** называется величина

$$\boxed{\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k}$$

Определение. **Модой** M_0^* вариационного ряда называется варианта с наибольшей частотой:

$$\boxed{M_0^* = X_i : n_i = \max_{1 \le j < n} n_j}$$

Определение. **Медианой** M_e^* вариационного ряда называется значение варианты в середине ряда:

1. Если $n = 2k - 1$, то $M_e^* = X_k$
2. Если $n = 2k$, то $M_e^* = \frac{X_k + X_{k+1}}{2}$

Величина	Команда в Excel	
	Русский	Английский
$\overline{X_B}$	СРЗНАЧ	AVERAGE
$\mathbb{D}_B$	ДИСПР	VARP
S^2	ДИСП	VAR
σ_n	СТАНДОТКЛОНП	STDEVP
S	СТАНДОТКЛОН	STDEV
M_0^*	МОДА	MODE
M_e^*	МЕДИАНА	MEDIAN

Теорема 5. Выборочное среднее $\overline{X_B}$ является несмещенной состоятельной оценкой для математического ожидания, то есть:

1. $\mathbb{E}\overline{X_B} = \mathbb{E}X = a$ — несмещенность
2. $\overline{X_B} \xrightarrow[n\to\infty]{P} \mathbb{E}X$ — состоятельность

Доказательство.

1. $$\mathbb{E}\overline{X} = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i = \frac{1}{n}\cdot n\mathbb{E}X_i = \mathbb{E}X$$

2. $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to\infty]{P} \mathbb{E}X$$

   Это верно по закону больших чисел.

$\square$

Теорема 6. Выборочный k-тый момент является несмещенной состоятельной оценкой для теоретического k-того момента, то есть:

1. $\mathbb{E}\overline{X^k} = X^k$
2. $\overline{X^k} \xrightarrow{P} \mathbb{E}X^k$

Доказательство. Следует из предыдущей теоремы, если в качестве случайной величины взять X^k. $\square$

Теорема 7.

- $\mathbb{D}_B$ — смещённая состоятельная оценка дисперсии
- S^2 — несмещённая состоятельная оценка дисперсии

Доказательство.

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

$$\begin{aligned}
\mathbb{E}\mathbb{D}_B &= \\
\mathbb{E}(\overline{X^2} - (\overline{X})^2) &= \\
\mathbb{E}\overline{X^2} - \mathbb{E}(\overline{X})^2 &= \\
\mathbb{E}X^2 - \mathbb{E}(\overline{X})^2 &= \\
\mathbb{D}\overline{X} &= \\
\mathbb{E}(\overline{X})^2 - (\mathbb{E}\overline{X})^2 &= \\
\mathbb{E}X^2 - (\mathbb{D}\overline{X} + (\mathbb{E}\overline{X})^2) &= \\
\mathbb{E}X^2 - (\mathbb{E}X)^2 - \mathbb{D}\overline{X} &= \\
(\mathbb{E}X^2 - (\mathbb{E}X)^2) - \mathbb{D}\overline{X} &= \\
\mathbb{D}X - \mathbb{D}\overline{X} &= \\
\mathbb{D}X - \mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \\
\mathbb{D}X - \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{D}X_i &=
\end{aligned}$$

$$\mathbb{D}X - \frac{1}{n^2} \cdot n\mathbb{D}X =$$
$$\mathbb{D}X - \frac{1}{n}\mathbb{D}X =$$
$$\frac{n-1}{n}\mathbb{D}X \neq \mathbb{D}X$$

$$\mathbb{E}S^2 = \mathbb{E}\left(\frac{n}{n-1}\mathbb{D}_B\right) = \frac{n}{n-1} \cdot \frac{n-1}{n}\mathbb{D}X = \mathbb{D}X$$
$$\mathbb{D}_B = \overline{X^2} - (\overline{X})^2 \xrightarrow{P} \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{D}X$$
$$S^2 = \frac{n}{n-1}\mathbb{D}_B \xrightarrow{P} \underbrace{\frac{n}{n-1}}_{\to 1}\mathbb{D}X$$

□

Примечание. $\mathbb{D}_B$ — асимптотически несмещённая оценка, т.к. при $n \to \infty$, $\frac{n-1}{n} \to 1$. Таким образом, при большой[1] выборке можно игнорировать смещённость.

4.3 Метод моментов

Изобретен Карлом Пирсоном.

Пусть имеется выборка $(X_1 \dots X_n)$ неизвестного распределения, при этом известен тип[2] распределения. Пусть этот тип определяется k неизвестными параметрами $\theta_1 \dots \theta_k$. Теоретическое распределение задает теоретические k-тые моменты. Например, если распределение непрерывное, то оно задается плотностью $f(X, \theta_1 \dots \theta_k)$ и $m_k = \int_{-\infty}^{+\infty} X^k f(x, \theta_1 \dots \theta_k)dx = h_k(\theta_1 \dots \theta_k)$. Метод моментов состоит в следующем: вычисляем выборочные моменты и подставляем их в эти равенства вместо теоретических. В результате получаем систему уравнений:

$$\begin{cases} \overline{X} = h_1(\theta_1 \dots \theta_k) \\ \overline{X^2} = h_2(\theta_1 \dots \theta_k) \\ \vdots \\ \overline{X^k} = h_k(\theta_1 \dots \theta_k) \end{cases}$$

Решив эту систему, мы получим оценки на $\theta_1 \dots \theta_k$. Эти оценки будут состоятельными[3], но смещёнными.

Пример. Пусть $X \in U(a, b), a < b$. Обработав статданные, получили оценки первого и второго момента: $\overline{X} = 2.25; \overline{X^2} = 6.75$

[1] $n \geq 100$, например.
[2] Нормальное, показательное и т.д.
[3] Если не придумывать специально плохие примеры

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

$$\mathbb{E}X = \int_a^b xf(x)dx = \int_a^b \frac{x}{b-a} = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \boxed{\frac{a+b}{2}}$$

$$\mathbb{E}X^2 = \int_a^b x^2 \cdot \frac{1}{b-a}dx = \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \boxed{\frac{a^2 + ab + b^2}{3}}$$

$$\begin{cases} 2.25 = \frac{a+b}{2} \\ 6.75 = \frac{a^2+ab+b^2}{3} \end{cases}$$

$$\begin{cases} a + b = 4.5 \\ a^2 + ab + b^2 = 20.25 \end{cases}$$

$$\begin{cases} a + b = 4.5 \\ ab = 0 \end{cases}$$

$$\begin{cases} a = 0 \\ b = 4.5 \end{cases}$$

$\square$

Лекция 3

20 сентября

4.4 Метод максимального правдоподобия

Метод максимального правдоподобия состоит в том, чтобы подобрать параметры таким образом, чтобы вероятность получения данной выборки была наибольшей. Если распределение дискретное, то вероятность выборки

$$P_\theta(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = P_\theta(X_1 = x_1)P_\theta(X_2 = x_2)\dots P_\theta(X_n = x_n)$$

Для непрерывной величины аналогично.

Поэтому исследуем такую функцию:

Определение. Функцией правдоподобия называется функция $L(\overline{X}, \theta)$, зависящая от выборки и неизвестных параметров, равная:

- В случае дискретного распределения:

$$P_\theta(X_1 = x_1)P_\theta(X_2 = x_2)\dots P_\theta(X_n = x_n)$$

- В случае абсолютно непрерывного распределения:

$$f_\theta(x_1)f_\theta(x_2)\dots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

Эту функцию неудобно исследовать, поэтому мы используем следующую функцию:

Определение. Логарифмическая функция правдоподобия:

$$M(\overline{X}, \theta) = \ln L(\overline{X}, \theta)$$

Т.к. логарифм — строго возрастающая функция, экстремумы обычной и логарифмической функций правдоподобия совпадают.

Определение. Оценкой максимального правдоподобия $\hat{\theta}$ называется значение θ, при котором функция правдоподобия достигает наибольшего значения.

Пример. Пусть $X_1 \dots X_n$ — выборка неизвестного распределения Пуассона с параметром λ: $X \in \Pi_\lambda, \lambda > 0$

$$P(X = x_i) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

$$L(\overline{X}, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{n \cdot \overline{X}}}{\prod_{i=1}^{n} x_i!} e^{-n\lambda}$$

$$\ln L(\overline{X}, \lambda) = n \cdot \overline{X} \cdot \ln \lambda - \ln \prod_{i=1}^{n} x_i! - n\lambda$$

$$\frac{\partial \ln L(\overline{X}, \lambda)}{\partial \lambda} = \frac{n\overline{X}}{\lambda} - n$$

Приравняем производную к нулю, чтобы найти точки экстремума:

$$\frac{n\overline{X}}{\lambda} - n = 0 \Rightarrow \lambda = \overline{X}$$

Таким образом $\hat{\theta} = \overline{X}$ — ОМП.

Пример. Пусть $X_1 \dots X_n$ — выборка неизвестного нормального распределения: $X \in N(a, \sigma^2), a \in \mathbb{R}, \sigma > 0$

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

$$L(\overline{X}, a, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-a)^2}{2\sigma^2}} = \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\frac{\sum(x_i-a)^2}{2\sigma^2}}$$

$$\ln L(\overline{X}, a, \sigma^2) = n \ln \sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum (x_i - a)^2$$

Не дописано

Пример. Пусть $X_1 \dots X_n$ — выборка равномерного распределения вида $U(0, \theta)$

1. Метод моментов.

$$\mathbb{E} = \frac{a+b}{2} = \frac{\theta}{2} \Rightarrow \theta = 2\overline{X}$$

2. Метод максимального правдоподобия.

$$f_\theta(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & x > \theta \end{cases}$$

$$L(\overline{X}, \theta) = \prod_{i=1}^{n} f_\theta(x_i) = \begin{cases} 0, & \theta < \max x_i = X_{(n)} \\ \frac{1}{\theta^n}, & \theta \geq X_{(n)} \end{cases}$$

L достигает наибольшего значения при $\theta = X_{(n)}$.

Сравним полученные оценки.

1. $\theta^* = 2\overline{X}$ — несмещённая оценка, т.к. $\mathbb{E}\theta^* = \mathbb{E}2\overline{X} = 2\mathbb{E}\overline{X} = \theta$

$$\mathbb{E}(\theta^* - \theta) = \mathbb{D}(\theta^*) = \mathbb{D}2\overline{X} = 4\frac{1}{n}\mathbb{D}X = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

2. Изучим случайную величину $X_{(n)}$. Её функция распределения это

$$F_{X_{(n)}}(x) = P(X_{(n)} < x) = P(X_1 < x)\dots P(X_n < x) = (F_X(x))^n$$

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{\theta}, & x > 0 \end{cases} \Rightarrow F_{X_{(n)}}(x) = \begin{cases} 0, & x < 0 \\ x^{\frac{n}{\theta^n}}, & x \geq 0 \end{cases} \Rightarrow f_{X_{(n)}}(x) = \begin{cases} 0, & x < 0 \\ \frac{nx^{n-1}}{\theta^n}, & x \geq 0 \end{cases}$$

$$\mathbb{E}X_{(n)} = \int_0^\theta x\frac{nx^{n-1}}{\theta^n}dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{\theta^n}\frac{x^{n+1}}{n+1}\Big|_0^\theta = \frac{n\theta}{n+1}$$

Таким образом, оценка смещённая, но асимптотически несмещенная.

Заменим эту оценку на несмещённую оценку $\tilde{\theta} = \frac{n+1}{n}\hat{\theta} = \frac{n+1}{n}X_{(n)}$ — сходятся к θ с одинаковой скоростью.

$$\mathbb{E}\tilde{\theta}^2 =$$
$$\mathbb{E}\left(\frac{n+1}{n}X_{(n)}\right)^2 =$$
$$\frac{(n+1)^2}{n^2}\mathbb{E}X_{(n)}^2 =$$
$$\frac{(n+1)^2}{n^2}\int_0^\theta x^2\frac{nx^{n-1}}{\theta^n}dx =$$
$$\frac{(n+1)^2}{n^2}\frac{n}{\theta^n}\frac{x^{n+2}}{n+2}\Big|_0^\theta = \frac{(n+1)^2}{n(n+2)}\theta^2$$

$$\mathbb{D}\tilde{\theta}^2 = \mathbb{E}\tilde{\theta}^2 - \mathbb{E}^2\tilde{\theta} = \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \theta^2\left(\frac{n^2+2n+1-n^2-2n}{n^2+2n}\right) = \frac{\theta^2}{n(n+2)}$$

Итак, сравним оценки.

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

Таким образом, оценка с помощью метода максимального правдоподобия лучше, её дисперсия стремится к нулю со скоростью $\frac{1}{n^2}$, а дисперсия первой оценки — со скоростью $\frac{1}{n}$. $\tilde{\theta} \to \theta$ со скоростью $\frac{1}{n}$, а $\theta^* \to \theta$ со скоростью $\frac{1}{\sqrt{n}}$

Следствие 7.1. Оценка математического ожидания $\overline{X} = 2\theta$ не будет эффективной оценкой, т.к. можно показать, что в данном случае эффективной оценкой будет

$$\boxed{\mathbb{E}X = \frac{n+1}{n} \cdot \max\{X_1 \dots X_n\}}$$

Примечание. ОМП состоятельны, часто эффективны, но могут быть смещенными.

5 Неравенство Рао-Крамера

Пусть известно, что случайная величина $X \in \mathcal{F}_\theta$ — семейству распределений с θ.

Определение. **Носителем** семейства распределений $\mathcal{F}_\theta$ называется множество $C \subset \mathbb{R}$, такое что $\forall\theta \;\; P(X \in C) = 1$.

Обозначение.

$$f_\theta(x) = \begin{cases} f_\theta(x), & \text{если распределение абсолютно непрерывное} \\ P_\theta(X = x), & \text{если распределение дискретное} \end{cases}$$

Определение. **Информацией Фишера** называется величина

$$I(\theta) = \mathbb{E}\left(\frac{\partial \ln f_\theta(x)}{\partial\theta}\right)^2$$

, если она существует.

Определение. Семейство распределений $\mathcal{F}_\theta$ называется **регулярным**, если:

1. Существует носитель C семейства $\mathcal{F}_\theta$, такой что $\forall x \in C$ функция $\ln f_\theta(x)$ непрерывно дифференцируема по θ.

2. $I(\theta)$ существует и непрерывна по θ.

Теорема 8 (неравенство Рао-Крамера). Пусть $X_1 \dots X_n$ — выборка объема n из регулярного семейства распределений $\mathcal{F}_\theta$, θ^* — несмещенная оценка параметра θ, дисперсия которой ограничена на любом компакте в области θ.

Тогда

$$\boxed{\mathbb{D}\theta^* \geq \frac{1}{nI(\theta)}}$$

Следствие 8.1. Если при условиях выше $\mathbb{D}\theta^* = \frac{1}{nI(\theta)}$, то θ^* — эффективная оценка. Это не всегда достижимо.

Пример. Пусть $X_1 \dots X_n$ — выборка нормального распределения $N(a, \sigma^2), a \in \mathbb{R}, \sigma^2 > 0$. Проверим эффективность оценки $a^* = \overline{X}$.

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

Рассмотрим носитель $C = \mathbb{R}$.

$$\ln f(x) = -\ln\sigma - \frac{1}{2}\ln(2\pi) - \frac{(x-a)^2}{2\sigma^2}$$

$$\frac{\partial \ln f(x)}{\partial a} = \frac{1}{2\sigma^2} 2(x-a) = \frac{x-a}{\sigma}$$

Производная непрерывна по a $\forall a \in \mathbb{R}$

$$I(a) = \mathbb{E}\left(\frac{\partial \ln f(x)}{\partial a}\right)^2 = \mathbb{E}\left(\frac{x-a}{\sigma}\right)^2 = \frac{1}{\sigma^4}\mathbb{E}(X - \mathbb{E}X) = \frac{1}{\sigma^4}\mathbb{D}X = \frac{1}{\sigma^4}\sigma^2 = \frac{1}{\sigma^2}$$

Сравним обе части неравенства Рао-Крамера:

$$\mathbb{D}a^* = \mathbb{D}\overline{X} = \frac{1}{n}\mathbb{D}X = \frac{1}{n}\sigma^2 = \frac{\sigma^2}{n}$$

$$\mathbb{D}a^* = \frac{\sigma^2}{n} \stackrel{?}{=} \frac{1}{nI(a)} = \frac{1}{n\frac{1}{\sigma^2}} = \frac{\sigma^2}{n}$$

Таким образом, оценка эффективна.

Примечание. Исправленная дисперсия S^2 также является эффективной оценкой.

Определение. BLUE[1]-оценка — лучшая оценка из оценок вида $\theta^* = \alpha_1 X_1 + \dots + \alpha_n X_n$.

[1] Best linear unbiased estimate.

Лекция 4

18 сентября

6 Распределения в матстатистике

6.0 Нормальное распределение

$X \in N(a, \sigma^2)$:

$$\mathbb{E}X = a, \mathbb{D}X = \sigma$$

$N(0, 1)$ — стандартное нормальное распределение

6.1 Гамма-распределение

$X \in \Gamma_{\alpha,\lambda}$, если её плотность равна:

$$f(x) = \begin{cases} 0, & x \le 0 \\ \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, & x > 0 \end{cases}$$

Свойства.

1. $\mathbb{E}\xi = \frac{\lambda}{\alpha}, \mathbb{D}\xi = \frac{\lambda}{\alpha^2}$
2. Если $\xi_1 \in \Gamma_{\alpha,\lambda_1}, \xi_2 \in \Gamma_{\alpha,\lambda_2}$, то $\xi_1 + \xi_2 \in \Gamma_{\alpha,\lambda_1+\lambda_2}$
3. $\Gamma_{\alpha,1} = E_\alpha$ — показательное распределение.
4. Если $X_i \in E_\alpha$, то $\sum_{i=1}^n X_i \in \Gamma_{\alpha,n}$
5. Если $X \in N(0, 1)$, то $X^2 \in \Gamma_{\frac{1}{2},\frac{1}{2}}$

Примечание. Гамма-распределение возникает в матстатистике как распределение квадрата стандартно нормально распределенной величины. Обобщим эту идею:

6.2 Распределение “хи-квадрат”

Определение. Распределение **хи-квадрат** с k степенями свободы называется распределение суммы k квадратов независимых стандартных нормальных величин.

$$\chi_k^2 = X_1^2 + X_2^2 + \dots + X_k^2, \quad X_i \in N(0,1)$$

Обозначение. $\chi^2 \in H_k$

Свойства.

1. $\chi_k^2 \in \Gamma_{\frac{1}{2}, \frac{k}{2}}$
2. $\chi_n^2 + \chi_m^2 = \chi_{n+m}^2$ — по определению
3. $\mathbb{E}\chi_k^2 = \frac{\lambda}{\alpha} = \frac{\frac{k}{2}}{\frac{1}{2}} = k, \mathbb{D}\chi_k^2 = \frac{\lambda}{\alpha^2} = \frac{\frac{k}{2}}{\left(\frac{1}{2}\right)^2} = 2k$

6.3 Распределение Стьюдента

Определение. Пусть случайные величины $X_0, X_1 \dots X_k$ — независимы и имеют стандартное нормальное распределение. Распределением **Стьюдента** с k степеней свободы называется распределение случайной величины

$$t_k = \frac{X_0}{\sqrt{\frac{1}{k}(X_1^2 + \dots + X_k^2)}} = \frac{X_0}{\sqrt{\frac{1}{k}\chi_k^2}}$$

Свойства.

1. $\mathbb{E}t_k = 0$
2. $\mathbb{D}t_k = \frac{k}{k-2}$

6.4 Распределение Фишера-Снедекора

Определение. Распределение $F_{m,n}$ называется распределением **Фишера-Снидекора** *(или F-**распределением**)* со степенями свободы m и n называется распределение случайной величины

$$f_{m,n} = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$$

, где χ_n^2 и χ_m^2 — независимые случайные величины с распределением χ^2.

Свойства.

1. $\mathbb{E}f_{m,n} = \frac{n}{n-2}$
2. $\mathbb{D}f_{m,n} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$

3. $F_{m,n}(x) = P(f_{m,n} < X) = P\left(\frac{1}{f_{m,n}} > \frac{1}{X}\right) = P\left(f_{m,n} > \frac{1}{X}\right) = 1 - F_{n,m}\left(\frac{1}{X}\right)$

При $n, k, m \to \infty$ эти распределения слабо сходятся к нормальному. При $n > 30$ они достаточно близки.

7 Линейные преобразования нормальных выборок

Пусть $\vec{X} = (X_1 \dots X_n)$, где $X_i \in N(0,1)$ и независимы. Будем рассматривать линейные комбинации этого вектора. Пусть A — невырожденная матрица размера $n \times n$. Рассмотрим случайный вектор $\vec{Y} = A\vec{X}$, где координаты случайного вектора $Y_i = a_{i1}X_1 + \dots + a_{in}X_n$. Будем исследовать, что из себя представляют Y_i и их совместное распределение.

Примечание. Если $\eta = a\xi + b$, то $f_\eta(\xi) = \frac{1}{|a|} f_\xi\left(\frac{\xi - b}{a}\right)$

Теорема 9. Пусть случайный вектор $\vec{X}$ имеет плотность распределения $f_{\vec{X}}(\vec{x})$ и A невырожденная матрица.

Тогда случайный вектор $\vec{Y} = A\vec{X} + \vec{b}$ имеет плотность

$$f_{\vec{Y}}(\vec{y}) = \frac{1}{|\det A|} \cdot f_{\vec{X}}(A^{-1}(\vec{y} - \vec{b}))$$

Примечание. $f_{\vec{X}}(\vec{x})$ — плотность $\vec{X}$, если $P(\vec{x} \in B) = \int \dots \int_B f_{\vec{X}}(\vec{x}) d\vec{x}$

Доказательство.

$$\begin{aligned} P(\vec{y} \in B) &= P(A\vec{x} + \vec{b} \in B) \\ &= P(\vec{x} \in A^{-1}(\vec{y} - \vec{b})) \\ &= \int \dots \int_{A^{-1}(B - \vec{b})} f_{\vec{x}}(x) d\vec{x} \end{aligned}$$

Сделаем замену $\vec{y} = A\vec{x} + \vec{b}$. Тогда $A^{-1}(B - \vec{b})$ перейдёт в B, $\vec{x}$ перейдёт в $A^{-1}(\vec{y} - \vec{b})$, $\vec{y} \in B$, $d\vec{x}$ перейдёт $|J| d\vec{y}$, где $J = |A^{-1}| = |A|^{-1}$

Итого:

$$= \int \dots \int_B f(A^{-1}(\vec{y} - \vec{b})) \cdot \frac{1}{|\det A|} d\vec{y} \Rightarrow f_{\vec{Y}}(\vec{y}) = \frac{1}{|\det A|} f_{\vec{X}}(A^{-1}(\vec{y} - \vec{b}))$$

□

Определение. $A = C$ — **ортогональна**, т.е. $C^T = C^{-1}$, $|\det C| = 1$

Теорема 10. Пусть дан случайный вектор $\vec{X} = (X_1 \dots X_n)$, где $\forall i \quad X_i \in N(0,1)$ и X_i независимы, а C — ортогональная матрица.

Тогда координаты случайного вектора $\vec{Y} = C\vec{X}$ независимы и также имеют стандартное нормальное распределение.

Доказательство. Т.к. координаты $X_i \in N(0,1)$ и независимы, то плотность $\vec{X}$:

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} f_i(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(X_1^2+X_2^2+\dots+X_n^2)} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}||\vec{X}||^2}$$

По предыдущей теореме:

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(C^T\vec{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}||C^T\vec{y}||^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}||\vec{y}||^2} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2} = \prod_{i=1}^{n} f_i(y_i)$$

Следовательно, $Y_i \in N(0,1)$ и независимы. □

Лемма 1 (Фишера). Пусть случайный вектор $\vec{X}$ состоит из независимых стандартных нормальных случайных величин, $\vec{Y} = C\vec{X}$, где C — ортогональная матрица. Тогда $\forall k :$ $1 \le k \le n-1$ случайная величина

$$T(\vec{X}) = \left(\sum_{i=1}^{n} X_i^2\right) - Y_1^2 - \dots - Y_k^2$$

не зависит от случайного вектора $Y_1 \dots Y_k$ и имеет распределение H_{n-k}

Доказательство. Т.к. C ортогональна:

$$||\vec{Y}||^2 = ||C\vec{X}||^2 = ||\vec{X}||^2 = X_1^2 + \dots + X_n^2 = Y_1^2 + \dots + Y_n^2$$

Отсюда

$$T(\vec{X}) = \sum_{i=1}^{n} Y_i^2 - Y_1^2 - \dots - Y_k^2 = Y_{k+1}^2 + \dots + Y_n^2$$

$Y_{k+1} \dots Y_n$ — независимы, имеют стандартное нормальное распределение и $T(\vec{X}) \in H_{n-k}$

$T(\vec{X})$ не зависит от $Y_1 \dots Y_k$, т.к. $Y_{k+1} \dots Y_n$ по предыдущей лемме от них не зависит. □

Теорема 11 (основная).

- $X_1 \dots X_k$ независимы и имеют нормальное распределение с параметрами a и σ^2
- $\overline{X}$ — выборочное среднее

- S^2 — исправленное выборочное среднее

Тогда имеют место следующие распределения:

1. $$\sqrt{n} \cdot \frac{\overline{X} - a}{\sigma} \in N(0,1)$$

2. $$\sum_{i=1}^{n} \left(\frac{X_i - a}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$$

3. $$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{n\sigma^{2*}}{\sigma^2} \in H_n$$

4. $$\sqrt{n} \cdot \frac{\overline{X} - a}{S} \in T_{n-1}$$

5. $\overline{X}$ и S^2 — независимые случайные величины

Доказательство.

1. $$X_i \in N(a, \sigma^2) \Rightarrow \sum_{i=1}^{n} X_i \in N(na, n\sigma^2) \Rightarrow \overline{X} \in N\left(a, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\sqrt{n}}{\sigma}(\overline{X} - a) \in N(0,1)$$

2. Верно, т.к. $\frac{X_i - a}{\sigma} \in N(0,1)$

3. $$\begin{aligned} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 &= \sum_{i=1}^{n} \left(\frac{X_i - a}{\sigma} - \frac{\overline{X} - a}{\sigma} \right)^2 \\ &= \sum_{i=1}^{n} (z_i - \overline{z})^2 \end{aligned}$$

   , где
   $$z_i = \frac{X_i - a}{\sigma} \in N(0,1), \overline{z} = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{\sum x_i - na}{\sigma n} = \frac{\overline{X} - a}{\sigma}$$
   Поэтому можем считать, что $X_i \in N(0,1)$. Применим лемму Фишера.
   $$T(\vec{X}) = \sum_{i=1}^{n} \left(X_i - \overline{X} \right) = n(\overline{X^2} - (\overline{X})^2) = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2 = \sum_{i=1}^{n} X_i^2 - Y_1$$

, где

$$Y_1 = n(\overline{X})^2 = \sqrt{n}\overline{X} = \frac{1}{\sqrt{n}}X_1 + \dots + \frac{1}{\sqrt{n}}X_n$$

Так как длина строки $\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}$ равна 1, поэтому эту строку можем дополнить до ортогональной матрицы C. Тогда Y_1 — первая координата случайного вектора $\vec{Y} = C\vec{X}$ и по лемме Фишера $T(\vec{X}) \in H_{n-1}$

5. $T(\vec{X}) = \frac{(n-1)S^2}{\sigma^2}$ не зависит от $Y_1 = \sqrt{n}\overline{X} \Rightarrow S^2$ и $\overline{X}$ независимы.

4. $$\sqrt{n}\frac{\overline{X}-a}{S} = \sqrt{n}\frac{\overline{X}-a}{\sigma} \cdot \frac{1}{\sqrt{\frac{(n-1)S^2}{\sigma^2} \cdot \frac{1}{n-1}}} = \frac{X_0}{\sqrt{\chi^2_{n-1}(n-1)}} \in T_{n-1} \text{ , т.к.:}$$

   $X_0 \in N(0,1)$ по пункту 1, $\frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$ по пункту 3 и X_0 не зависит от $\frac{(n-1)S^2}{\sigma^2}$ по пунтку 5.

□

Примечание. Эта часть была рассказана на пратике 29 сентября.

7.1 Многомерные нормальные распределения

Определение. Пусть случайный вектор $\vec{\xi} = (\xi_1 \dots \xi_n)$ имеет в средних $\vec{a} = (\mathbb{E}\xi_1 \dots \mathbb{E}\xi_n)$, K — симметричная положительно определенная метрица.

Вектор $\vec{\xi}$ имеет **многомерное нормальное** распределение с параметрами $\vec{a}$ и K, если его плотность:

$$f_{\vec{\xi}}(\vec{x}) = \frac{1}{\sqrt{2\pi}^n\sqrt{\det K}}e^{-\frac{1}{2}((\vec{x}-\vec{a})^T K^{-1}(\overline{x}-\overline{a}))}$$

Примечание. $(\vec{x}-\vec{a})^T K^{-1}(\vec{x}-\vec{a})$ — положительно определенная квадратичная форма от $(x_1 \dots x_n)$

Свойства.

1. Пусть $\vec{\eta}$ состоит из независимых стандартных нормальных величин, B — невырожденная матрица. Тогда $\vec{\xi} = B\vec{\eta} + \vec{a}$ имеет многомерное нормальное распределение с параметрами $\vec{a}, K = B^T B$

2. Пусть $\vec{\xi}$ имеет многомерное нормальное распределение с параметрами $\vec{a}$ и K. Тогда $\eta = B^{-1}(\vec{\xi}-\vec{a})$, где $B = \sqrt{K}$[1], состоит из независимы стандартных нормальных величин.

3. $K = \text{cov}(\xi_i, \xi_j)$

[1] B существует по задаче 3 из 4-ой практики.

4. Пусть $\vec{\xi}$ имеет многомерное нормальное распределение спараметрами $\vec{a}$ и K. Координаты $\vec{\xi}$ независимы тогда и только тогда, когда они не кореллированы, т.е. K — диагональная.

   Следствие 11.1. Если ξ, η — нормальные случайные величины и вектор (ξ, η) имеет ненулевую плотность, то ξ и η независимы тогда и только тогда, когда они не кореллированы, т.е. $r_{\xi,\eta} = 0$.

Теорема 12 (многомерная центральная предельная теорема). Среднее арифметическое независимых одинаково распределенных случайных векторов слабо сходится к многомерному нормальному распределению.

Лекция 5

4 октября

8 Квантили распределений

Для простоты предполагаем, что все распределения непрерывные.

Определение (1). Число t_γ называется **квантилем**[1] уровня γ, если $F(t_\gamma) = \gamma$.

С точки зрения геометрии $P(X \in \text{ область слева от } t_\gamma) = \gamma$.

Примечание.

- Медиана — квантиль уровня $\frac{1}{2}$
- Квартили — квантили уровня $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$
- Децили — квантили уровня $\frac{1}{10}, \frac{2}{10}, \dots$

Примечание. Квантиль t_γ — значение обратной функции распределения: $t_\gamma = F^{-1}(\gamma)$

Определение (2 (альтернативное)). Число t_α называется **квантилем уровня значимости** α, если $F(t_\alpha) = 1 - \alpha$.

Примечание. $\alpha = 1 - \gamma$

8.1 Квантили основных распределений в Excel

1. НОРМ.СТ.ОБР.

$$F_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dx$$

Тогда НОРМ.СТ.ОБР.$(x+0.5)$ — обратная функция функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$

[1] Или квантилью.

2. (a) СТЬЮДЕНТ.ОБР. — обратная к функции распределения Стьюдента стандартной величины.

$$t_k = \frac{X_0}{\sqrt{\frac{1}{k}\chi_k^2}}$$

   (b) СТЬЮДЕНТ.ОБР.2Х

   Возвращает t_α, такое что $P(|X| > t_\alpha) = \alpha$. Отсюда $P(|X| < t_\alpha) = 1 - \alpha$ и применяем СТЬЮДЕНТ.ОБР.2Х$(1-\alpha, k)$

3. (a) ХИ2.ОБР. — возвращает квантиль t_γ в первом смысле для распределения χ^2.

   (b) ХИ2.ОБР.ПХ — возвращает квантиль t_α

4. (a) F.ОБР. — возвращает квантиль t_γ F-распределения

   (b) F.ОБР.ПХ — возвращает квантиль t_α F-распределения

8.2 Интервальные оценки

Недостаток точных оценок в том, что мы не знаем, насколько точная наша оценка.

Пусть требуется дать оценку неизвестного параметра θ.

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

Примечание. Если θ — параметр дискретного распределения, то будет правильней написать $P(\theta_\gamma^- < \theta < \theta_\gamma^+) \geq \gamma$.

Примечание. Здесь случайные величины — интервальные оценки, а не θ. Поэтому более культурно говорить так: интервал $(\theta_\gamma^-, \theta_\gamma^+)$ накрывает неизвестный параметр θ с вероятностью γ.[2]

Примечание. В экономике γ берется 0.95, но можно брать и меньше — 0.9. Для чего-либо важного берется 0.99 или даже 0.999. Уровень надёжности выбирается в зависимости от решаемой задачи. Стандартные уровни: $0.9, 0.95, 0.99, 0.999$.

8.2.1 Интервальные оценки для нормального распеределения

Пусть $X = (X_1 \dots X_n)$ из $N(a, \sigma^2)$.

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

   По пункту 1 теоремы 11:

$$P\left(-t_\gamma < \sqrt{n}\frac{\overline{X} - a}{\sigma} < t_\gamma\right) = P\left(\left|\sqrt{n}\frac{\overline{X} - a}{\sigma}\right| < t_\gamma\right) = 2\Phi(t_\gamma) = \gamma$$

[2] А не “θ попадает в интервал $(\theta_\gamma^-, \theta_\gamma^+)$ с вероятностью γ”

, где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$. Тогда t_γ — значение обратной к Φ в точке $\frac{\gamma}{2}$. ???

Осталось решить неравенство относительно a.

$$-t_\gamma < \sqrt{n}\frac{\overline{X} - a}{\sigma} < t_\gamma$$
$$-t_\gamma \cdot \frac{\sigma}{\sqrt{n}} < a - \overline{X} < t_\gamma \cdot \frac{\sigma}{\sqrt{n}}$$
$$\overline{X} - t_\gamma \cdot \frac{\sigma}{\sqrt{n}} < a < \overline{X} + t_\gamma \cdot \frac{\sigma}{\sqrt{n}}$$

Итак получили доверительный интервал для параметра a: $\left(\overline{X} - t_\gamma \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + t_\gamma \cdot \frac{\sigma}{\sqrt{n}}\right)$

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

   По пункту 4 теоремы 11:

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

   $F_{T_{n-1}}(t_\gamma) = \frac{1+\gamma}{2}$, т.е. t_γ — квантиль распределения Стьюдента T_{n-1} уровня $\frac{1+\gamma}{2}$.

   Примечание. Если ξ — симметрично, то $P(|\xi| < t) = 2F(t) - 1$

   Доказательство.

$$P(|\xi| < t) = 2P(0 < \xi < t) = 2(F(t) - F(0)) = 2F(t) - 1$$

□

$$-t_\gamma < \sqrt{n}\frac{\overline{X} - a}{S} < t_\gamma$$
$$\overline{X} - t_\gamma \cdot \frac{S}{\sqrt{n}} < a < \overline{X} + t_\gamma \cdot \frac{S}{\sqrt{n}}$$

3. Доверительный интервал для параметра σ^2 при ???.

   По пункту 2 теоремы 11 $\frac{(n-1)S^2}{\sigma^2} \in H_{n-1}$. Пусть χ_1^2 и χ_2^2 — квантили распределения H_{n-1} уровней $1 - \frac{\gamma}{2}$ и $1 + \frac{\gamma}{2}$. Тогда:

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

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

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

Итак, доверительный интервал для параметра σ^2 надежности γ есть $\left(\frac{(n-1)S^2}{\chi_2^2}, \frac{(n-1)S^2}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 — квантили уровней $1-\frac{\gamma}{2}$ и $1+\frac{\gamma}{2}$. Следовательно, доверительный интервал для σ это $\left(\frac{\sqrt{n-1}S}{\chi_2}, \frac{\sqrt{n-1}S}{\chi_1}\right)$.

Этот интервал почти всегда не симметричен, можно его сделать симметричным, но мы этого делать не будем.

4. Доверительный интервал для парамтера σ^2 при известном параметре σ^{2*}

   По пункту 3 теоремы $\frac{n\sigma^{2*}}{\sigma^2} \in H_n$, где $\sigma^{2*} = \frac{1}{n}\sum_{i=1}^n (X_i - a)^2$. Пусть χ_1^2 и χ_2^2 — квантили распределения H_n уровней $1-\frac{\gamma}{2}$ и $1+\frac{\gamma}{2}$ соответственно.

$$P\left(\chi_1^2 < \frac{n\sigma^{2*}}{\sigma^2} < \sigma_2^2\right) = F_{H_n}(\chi_2^2) - F_{H_n}(\chi_1^2) = \gamma$$

$$\chi_1^2 < \frac{n\sigma^{2*}}{\sigma^2} < \chi_2^2$$
$$\frac{n\sigma^{2*}}{\chi_2^2} < \sigma^2 < \frac{n\sigma^{2*}}{\chi_1^2}$$

   Итак, доверительный интервал для σ^2 надежности γ это $\left(\frac{n\sigma^{2*}}{\chi_2^2}, \frac{n\sigma^{2*}}{\chi_1^2}\right)$, где χ_1^2 и χ_2^2 — квантили H_n уровней $1-\frac{\gamma}{2}$ и $1+\frac{\gamma}{2}$, $\sigma^{2*} = \frac{1}{n}\sum_{i=1}^n (X_i - a)^2$.

Для других распределений при малых объемах выборки нужно выводить формулы для каждой задачи. При больших объемах благодаря ЦПТ можно делать вид, что распределение нормальное.

Пример. $X \in N(a, \sigma^2)$, причём известно, что $\sigma = 3$. В результате обработки выборки объема $n = 36$ нашли $\overline{X} = 4.1$. Найти доверительный интервал параметра a надежности $\gamma = 0.95$.

Решение. $t_\gamma : 2\Phi(t_\gamma) = 0.95, \Phi(t_\gamma) = 0.475, t_\gamma = 1.96$

$$\overline{X} - t_\gamma \cdot \frac{\sigma}{\sqrt{n}} < a < \overline{X} + t_\gamma \cdot \frac{\sigma}{\sqrt{n}}$$

$$4.1 - 1.96 \cdot \frac{3}{\sqrt{36}} < a < 4.1 + 1.96 \cdot \frac{3}{\sqrt{36}}$$
$$4.1 - 0.98 < a < 4.1 + 0.98$$

$$3.12 < a < 5.08$$

Ответ: $(3.12, 5.08)$ $\square$

Пример. $X \in N(a, \sigma^2)$. В результате обработки выборки объема $n = 25$ нашли $\overline{X} = 42.32, S = 6.4$. Найти доверительный интервал надежности $\gamma = 0.95$.

Решение. По таблице двустороннего распределения Стьюдента T_{n-1} $t_\gamma = 2.064$

$$\overline{X} - t_\gamma \cdot \frac{S}{\sqrt{n}} < a < \overline{X} + t_\gamma \cdot \frac{S}{\sqrt{n}}$$

$$42.32 - 2.064 \cdot \frac{6.4}{\sqrt{25}} < a < 42.32 + 2.064 \cdot \frac{6.4}{\sqrt{25}}$$

$$42.32 - 2.642 < a < 42.32 + 2.642$$

$$39.678 < a < 44.962$$

Ответ: $(39.678, 44.962)$ $\square$

Лекция 6

11 октября

9 Гипотезы

Определение. **Гипотезой** H называется предположение о свойствах случайной величины.

Определение. Гипотеза называется **простой**, если она однозначно определяет распределение, т.е. $H : \mathcal{F} = \mathcal{F}_1$, где $\mathcal{F}_1$ — распределение известного типа с известными параметрами.

Определение. Все остальные гипотезы называются **сложными**, т.к. они являются объединением конечного или бесконечного числа простых гипотез.

Определение (основная модель гипотез). Гипотеза $H_1 = \overline{H_0}$ — конкурирующая *(альтернативная)* гипотеза, состоящая в том, что основная гипотеза H_0 неверна.

Примечание. С помощью статистических методов нельзя доказать гипотезу, можно только сказать, что она верна с некоторой уверенностью.

Основная гипотеза H_0 принимается или отклоняется с помощью статистики критерия K:

$$K(X_0 \dots X_n) \to \mathbb{R} = S \cup^{1} \overline{S} \to (H_0, H_1)$$

$$\begin{cases} H_0, & \text{если } K \in \overline{S} \\ H_1, & \text{если } K \in S \end{cases}$$

Определение. Если точка находится на границе областей S и $\overline{S}$, она называется **критической**.

Определение. **Ошибка I рода** состоит в том, что нулевая гипотеза отвергается, когда она верна.

[1] Объединение на самом деле дизъюнктно.

Определение. **Ошибка II рода** состоит в том, что отвергается альтернативная, когда она верна.

Определение. α — вероятность ошибки II рода, β — вероятность ошибки I рода/

Пример. H_0 — деталь годная, H_1 — деталь бракованная.

Ошибка I рода — признать годную деталь бракованной.

Ошибка II рода — признать бракованную деталь годной.

Примечание. При росте выборки вероятности ошибок уменьшаются, при уменьшении вероятности одной ошибки другая вероятность увеличивается.

9.1 Способы сравнения критериев

Пусть имеются критерии K_1 и $K_2, \alpha_1, \beta_1, \alpha_2, \beta_2$ — вероятности ошибок при соответствующих критериях, h_1 — потери в результате ошибке I рода, h_2 — потери в результате ошибки II рода.

Тогда рассмотрим способы сравнения критериев:

1. Минимакс: K_1 не хуже, чем K_2, если $\max(\alpha_1 h_1, \beta_1 h_2) \leq \max(\alpha_2 h_1, \beta_2 h_2)$
2. Критерий называется **баесовским**, если $U = \alpha k_1 + \beta k_2$ минимально.
3. Пусть ε — допустимый уровень ошибки I рода. Обозначим $K_\varepsilon := \{K_i \mid \alpha_i \leq \varepsilon\}$.

   Определение. Критерий $K \in K_\varepsilon$ называется **наиболее мощным** критерием уровня ε, если $\beta \leq \beta_i \ \ \forall i$.

9.2 Критерий согласия

Определение. Критерий K называется **критерием асимптотического уровня** ε, если вероятность ошибки первого рода α стремится к ε при $n \to \infty$.

Определение. Критерий K для проверки гипотезы H_0 против альтернативы $H_1 = \overline{H_0}$ называется состоятельным, если вероятность ошибки II рода $\beta \to 0$ при $n \to \infty$.

Определение. Критерием **согласия** уровня ε называются состоятельные критерии асимптотического уровня ε.

9.3 Построение критериев согласия

В качестве критериев согласия берётся статистика $K(X_1 \ldots X_n)$ со свойствами:

1. Если H_0 верна, то $K(X_1 \ldots X_n) \rightrightarrows Z$ — известное распределение с известными параметрами.
2. Если H_0 не верна, то $K(X_1 \ldots X_n) \xrightarrow{P} \infty$

Для заданного уровня значимость ε находим константу t_k, такую что $P(|Z| \geq t_k) = \alpha$. В результате получаем критерий согласия уровня значимости $\alpha = \varepsilon$:

$$\begin{cases} H_0, & |K| < t_k \\ H_1, & |K| \geq t_k \end{cases}$$

Теорема 13. Этот критерий является критерием согласия.

Доказательство.

1. K — критерий асимптотического уровня:

   Пусть H_0 верна. Тогда по построению $K \rightrightarrows Z$, т.е. $F_K(x) \to F_Z(x)$ и

$$\begin{aligned} \alpha &= P(|K| \geq t_k \mid H_0) \\ &= 1 - P(|K| < t_k) \\ &= 1 - (F_K(t_k) - F_K(-t_k)) \\ &\xrightarrow[n\to\infty]{} 1 - (F_Z(t_k) - F_Z(-t_k)) \\ &= P(|Z| \geq t_k) \\ &= \varepsilon \end{aligned}$$

2. K — состоятельный критерий:

   Пусть H_1 верна. Тогда $K(X_1 \dots X_n) \xrightarrow{P} \infty$, т.е.

$$\forall C \;\; P(|K| \geq C \mid H_1) \xrightarrow[n\to\infty]{P} 1 \Rightarrow \beta = P(|K| < t_k \mid H_1) \xrightarrow[n\to\infty]{P} 0$$

□

Упражнение. Гипотеза о среднем нормальной совокупности с известной дисперсией.

Пусть имеется выборка $(X_1 \dots X_n) \in X \in N(a, \sigma^2)$, причём второй параметр известен.[2]

$H_0 : a = a_0, H_1 : a \neq a_0$.

В качестве статистики критерия возьмём $\sqrt{n} \cdot \frac{\overline{X} - a_0}{\sigma}$. Проверим, что оно имеет требуемые свойства:

1. Если H_0 верна, т.е. $a = a_0$, то $\sqrt{n}\frac{\overline{X}-a_0}{\sigma} = \sqrt{n}\frac{\overline{X}-a}{\sigma} \in N(0,1)$

[2] Например, мы измеряем что-то инструментом заданной точности.

2. Если H_0 неверно, т.е. $a \neq a_0$, то $|K| \to \infty$:

$$|K| = \left|\sqrt{n}\frac{\overline{X}-a_0}{\sigma}\right| = \underbrace{\sqrt{n}}_{\to\infty}\left|\underbrace{\frac{\overline{X}-a}{\sigma}}_{\in N(0,1)} + \underbrace{\frac{a-a_0}{\sigma}}_{\neq 0}\right| \xrightarrow[n\to\infty]{P} \infty$$

Таким образом, этот критерий — критерий согласия. Для уровня значимости $\alpha = \varepsilon$ выберем C, такую что $\varepsilon = P(|K| \geq C) \Rightarrow P(|K| < C) = 1-\varepsilon \Rightarrow 2\Phi(C) = 1-\varepsilon \Rightarrow 2\Phi(C) = \frac{1-\varepsilon}{2}$

Итого:

$$\begin{cases} H_0, & |K| = \left|\sqrt{n}\frac{\overline{X}-a_0}{\sigma}\right| < C \\ H_1, & |K| = \left|\sqrt{n}\frac{\overline{X}-a_0}{\sigma}\right| \geq C \end{cases}$$

Заметим, что если мы решим это неравенство, то получим доверительный интервал для параметра a нормального распределения при известном σ.

Примечание. Аналогично можно проверять для неизвестного σ, тогда в критерии σ заменится на S.

9.4 Доверительные интервалы как критерии гипотез о параметрах распределения

Пусть имеется выборка $(X_1 \dots X_n)$ случайной величины $X \in \mathcal{F}_\theta$, где $\mathcal{F}_\theta$ — распределение известного типа с неизвестным параметром θ. Проверяется гипотеза: $H_0 : \theta = \theta_0$ против $H_1 : \theta \neq \theta_0$. Пусть для θ построен доверительный интервал (θ^-, θ^+) надежности γ. Тогда следующий критерий является критерием согласия уровня $\alpha = 1-\gamma$:

$$\begin{cases} H_0, & \theta_0 \in (\theta^-, \theta^+) \\ H_1, & \theta_0 \notin (\theta^-, \theta^+) \end{cases}$$

Доказательство.

$$\alpha = P(\theta_0 \notin (\theta^-, \theta^+) \mid X \in \mathcal{F}_\theta) = 1 - P(\theta_0 \in (\theta^-, \theta^+) \mid X \in \mathcal{F}_\theta) = 1-\gamma = \alpha$$

Доказывать состоятельность критерия нужно в каждом случае отдельно. □

Пример. По выборке объема $n = 36$ из нормальной совокупности с известным $\sigma = 1.44$ найдено выборочное среднее $\overline{X} = 21.36$. Проверить гипотезу $H_0 : a = 21$ против $H_1 : a \neq 21$ при уровне значимости $\alpha = 0.05$.

$$K = \sqrt{n}\frac{\overline{X}-a_0}{\sigma} = \sqrt{36}\frac{21.6-21}{1.44} = 2.5$$

$$\Phi(t_k) = \frac{1-\alpha}{2} = 0.475$$
$$t_k = 1.96$$

Т.к. $|K| = 2.5 > 1.96$, гипотеза отклоняется.

Лекция ещё не дописана.

Примечание. Следующий материал был рассказан на практике 13 октября.

9.5 Распределение Коши

Пусть дан источник некоторого излучения в точке $(0, 1)$, который равномерно посылает лучи во все стороны.

Случайная величина ξ — точка пересечения луча с осью OX.

Найти $F_\xi(x), f_\xi(x), \mathbb{E}\xi$.

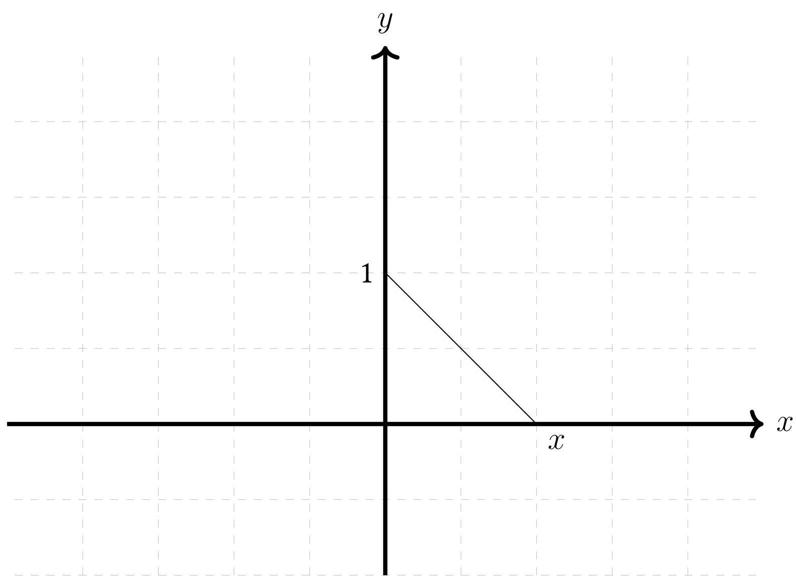

Рис. 6.1: Источник

$$F_\xi(x) = P(\xi < x) = P(\xi < 0) + P(0 < \xi < x) = 0.5 + \frac{1}{\pi}\operatorname{arctg} x$$
$$f_\xi(x) = F'_\xi(x) = \frac{1}{\pi}\frac{1}{x^2+1}$$
$$\mathbb{E}\xi = \int_{-\infty}^{+\infty} x f_\xi(x) = \int_{-\infty}^{+\infty} x\frac{1}{\pi}\frac{1}{x^2+1}dx = \frac{1}{2\pi}\ln\left(1+x^2\right)\Big|_{-\infty}^{+\infty}, \nexists$$

Пусть теперь источник сдвинут на θ по оси x. Тогда $f_\xi(x) = \frac{1}{\pi(1+(x-\theta)^2)}$. Попробуем оценить θ. $\overline{X}$ не работает, т.к. оно убежит на бесконечность: $\mathbb{E}\frac{S_n}{n} = \mathbb{E}X$. Оценим с помощью медианы. По симметрии $\theta = \text{Me}\xi$.

$$\text{Me}^* = \begin{cases} X_{(k+1)}, & n = 2k+1 \\ \frac{X_{(k)}+X_{(k+1)}}{2}, & n = 2k \end{cases}$$

Теорема 14. Если $f(\text{Me}) \neq 0$, то $\text{Me}^* \xrightarrow{P} \text{Me}$, причём сходится со скоростью $\frac{1}{\sqrt{n}}$.

В целом при большом числе выбросов медиана помогает. Например, оценивать зарплату нужно по медиане, а не по среднему.

У медианы также есть свои недостатки: она сходится медленнее, чем выборочное среднее — эффективность обычно ниже на 20-30%, но бывают и случаи хуже.

Есть и другие оценки, например **усечённое среднее**. Выкидываются наименьшие и наибольшие k точек и считается выборочное среднее:

$$\frac{\sum\limits_{i=k+1}^{n-k} X_{(i)}}{n-2k}$$

Несложно заметить, что это нечто промежуточное между выборочным средним и медианой — если $k = 0$, то получаем выборочное среднее, если $k = \frac{n-1}{2}$, то получаем медиану.

Другой пример: составим по исходной выборке выборку объема $\frac{n(n-1)}{2}$, состоящую из $\frac{X_i+X_j}{2}, 1 \leq i, j \leq n$. **Среднее Уолша** — медиана этой выборки. У этой оценки эффективность падает на $\approx 12\%$ относительно выборочного среднего.

Упражнение 1. Дано n призывников с вероятностью болезни $p = 0.01$. Разбиваем призывников на группы по k человек в группе. Считаем, что $n \,\vdots\, k$, т.е. групп $\frac{n}{k}$. В каждой группе:

- Если суммарный результат отрицательный, то 1 анализ.
- Иначе $k+1$ анализ.

Найти оптимиальное значение k и среднее значение числа анализов.

Решение. ξ_i — число анализов в i-той группе.

$$P(\xi_i = 1) = (1-p)^k \quad P(\xi_i = k+1) = 1-(1-p)^k$$
$$\mathbb{E}\xi_i = (1-p)^k + (k+1)(1-(1-p)^k) = k+1-k(1-p)^k$$

$$\xi = \frac{n}{k} \cdot \xi_i$$

$$\mathbb{E}\xi = n\left(1 + \frac{1}{k} - (1-p)^k\right) = f(k)$$

Т.к. p мало, пусть оно $p \to 0$. $(1-p)^k \sim 1 - pk$.

$$f(k) \sim n\left(\frac{1}{k} + pk\right)$$

$$f'(k) = n\left(-\frac{1}{k^2} + p\right) = 0$$

$$k = \frac{1}{\sqrt{p}} = 10$$

$$\mathbb{E}\xi \approx n\left(\frac{1}{10} + 0.01 \cdot 10\right) = 0.2n$$

$\square$