

Математическая статистика

Михайлов Максим

4 ноября 2021 г.

Оглавление

Лекция 1	6 сентября	4
1	Организационные вопросы	4
2	Введение	4
2.1	Выборочная функция распределения	5
3	Первоначальная обработка статданных	6
Лекция 2	13 сентября	8
4	Точечные оценки	8
4.1	Свойства статистических оценок	8
4.1.1	Состоятельность	8
4.1.2	Несмещённость	8
4.1.3	Эффективность	9
4.2	Точечные оценки моментов	9
4.3	Метод моментов	12
Лекция 3	20 сентября	14
4.4	Метод максимального правдоподобия	14
5	Неравенство Рао-Крамера	17
Лекция 4	18 сентября	19
6	Распределения в матстатистике	19
6.0	Нормальное распределение	19
6.1	Гамма-распределение	19
6.2	Распределение “хи-квадрат”	20
6.3	Распределение Стьюдента	20
6.4	Распределение Фишера-Снедекора	20
7	Линейные преобразования нормальных выборок	21
7.1	Многомерные нормальные распределения	24
Лекция 5	4 октября	26
8	Квантили распределений	26
8.1	Квантили основных распределений в Excel	26
8.2	Интервальные оценки	27
8.2.1	Интервальные оценки для нормального распределения	27
Лекция 6	11 октября	31
9	Гипотезы	31
9.1	Способы сравнения критериев	32
9.2	Критерий согласия	32

9.3	Построение критериев согласия	32
9.4	Доверительные интервалы как критерии гипотез о параметрах распределения	34
9.5	Распределение Коши	35
Лекция 7	18 октября	38
9.6	Критерии для проверки гипотез о распределении	38
9.6.1	Критерий χ^2 для параметрической гипотезы	38
9.6.2	Критерий χ^2	40
9.6.3	Критерий Колмогорова	40
9.7	Критерии для проверки однородности	40
9.7.1	Критерий Колмогорова-Смирнова	41
9.7.2	Критерий Фишера	41
9.7.3	Критерий Стьюдента	42
Лекция 8	25 октября	43
10	Статистическая зависимость	43
10.1	Корреляционное облако	43
10.2	Корреляционная таблица	43
10.3	Критерий χ^2 для проверки независимости	44
10.4	Однофакторный дисперсионный анализ	46
10.4.1	Общая, межгрупповая и внутригрупповая дисперсия	46
10.4.2	Проверка гипотезы о влиянии фактора	47

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 $X_1, X_2 \dots X_n$. Набор¹ 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:

$$\bar{X} := \sum_{i=1}^n \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^n X_i$$

Выборочная дисперсия:

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

2.1 Выборочная функция распределения

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

Примечание. I — индикатор:

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

Теорема 1.

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

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 $F(z) \leq 1 < \infty$, 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:

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

3 Первоначальная обработка статданных

Если отсортировать данные, то получим **вариационный ряд**: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. 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. $h := X_{\max} - X_{\min}$ — **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 $n \rightarrow \infty, k(n) \rightarrow \infty$, причем $\frac{k(n)}{n} \rightarrow 0$, 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 : $X = (X_1 \dots X_n)$

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:

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

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

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

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

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

или, если оценки несмещенные,

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

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

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

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

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

Определение. Исправленной выборочной дисперсией S^2 называется величина

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

или

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

Определение. Выборочным средним квадратическим отклонением называется величина

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

Определение. Исправленным выборочным средним квадратическим отклонением называется величина

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

Определение. Выборочным k -тым моментом называется величина

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

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

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

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. $\overline{X_B} \xrightarrow[n \rightarrow \infty]{P} \mathbb{E}X$ — 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.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow[n \rightarrow \infty]{P} \mathbb{E}X$$

Это верно по закону больших чисел.

□

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

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

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

Теорема 7.

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

Доказательство.

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

$$\mathbb{E}\mathbb{D}_B =$$

$$\mathbb{E}(\bar{X}^2 - (\bar{X})^2) =$$

$$\mathbb{E}\bar{X}^2 - \mathbb{E}(\bar{X})^2 =$$

$$\mathbb{E}X^2 - \mathbb{E}(\bar{X})^2 =$$

$$\mathbb{D}\bar{X} =$$

$$\mathbb{E}(\bar{X})^2 - (\mathbb{E}\bar{X})^2 =$$

$$\mathbb{E}X^2 - (\mathbb{D}\bar{X} + (\mathbb{E}\bar{X})^2) =$$

$$\mathbb{E}X^2 - (\mathbb{E}X)^2 - \mathbb{D}\bar{X} =$$

$$(\mathbb{E}X^2 - (\mathbb{E}X)^2) - \mathbb{D}\bar{X} =$$

$$\mathbb{D}X - \mathbb{D}\bar{X} =$$

$$\mathbb{D}X - \mathbb{D}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) =$$

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

$$\begin{aligned}
\mathbb{D}X - \frac{1}{n^2} \cdot n\mathbb{D}X &= \\
\mathbb{D}X - \frac{1}{n}\mathbb{D}X &= \\
\frac{n-1}{n}\mathbb{D}X &\neq \mathbb{D}X \\
\mathbb{E}S^2 = \mathbb{E}\left(\frac{n}{n-1}\mathbb{D}_B\right) &= \frac{n}{n-1} \cdot \frac{n-1}{n}\mathbb{D}X = \mathbb{D}X \\
\mathbb{D}_B = \overline{X^2} - (\overline{X})^2 &\xrightarrow{P} \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{D}X \\
S^2 = \frac{n}{n-1}\mathbb{D}_B &\xrightarrow{P} \underbrace{\frac{n}{n-1}}_{\rightarrow 1} \mathbb{D}X
\end{aligned}$$

□

Примечание. \mathbb{D}_B — асимптотически несмещённая оценка, т.к. при $n \rightarrow \infty$, $\frac{n-1}{n} \rightarrow 1$. Таким образом, при большой¹ выборке можно игнорировать смещённость.

4.3 Метод моментов

Изобретен Карлом Пирсоном.

Пусть имеется выборка $(X_1 \dots X_n)$ неизвестного распределения, при этом известен тип² 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 состоятельными³, 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 $f(x) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x > b \end{cases}$

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

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

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

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

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

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

□

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 $L(\bar{X}, \theta)$, 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:

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

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 : $X \in \Pi_\lambda, \lambda > 0$

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

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

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

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

Приравняем производную к нулю, чтобы найти точки экстремума:

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

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

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

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

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

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

Не дописано

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

1. Метод моментов.

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

2. Метод максимального правдоподобия.

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

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

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

Сравним полученные оценки.

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

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

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

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

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

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

Таким образом, оценка смещённая, но асимптотически несмещённая.

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.

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

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 $\frac{1}{n}$. $\tilde{\theta} \rightarrow \theta$ со скоростью $\frac{1}{n}$, а $\theta^* \rightarrow \theta$ со скоростью $\frac{1}{\sqrt{n}}$

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

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

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}_θ 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}_θ называется **регулярным**, если:

1. Существует носитель C семейства \mathcal{F}_θ , 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}_θ , θ^* — несмещенная оценка параметра θ , дисперсия которой ограничена на любом компакте в области θ .

Тогда

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

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 $a^* = \bar{X}$.

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

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 $a \ \forall a \in \mathbb{R}$

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

Сравним обе части неравенства Рао-Крамера:

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

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

Таким образом, оценка эффективна.

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

Определение. BLUE¹-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:

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

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. $\mathbb{E}\chi_k^2 = \frac{\lambda}{\alpha} = \frac{\frac{k}{2}}{\frac{1}{2}} = k, \mathbb{D}\chi_k^2 = \frac{\lambda}{\alpha^2} = \frac{\frac{k}{2}}{(\frac{1}{2})^2} = 2k$

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-Снедекора (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 $n, k, m \rightarrow \infty$ 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 $\forall i \ X_i \in N(0, 1)$ 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. \square

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 $\forall k : 1 \leq k \leq n - 1$ 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. \square

Теорема 11 (основная).

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

- S^2 — исправленное выборочное среднее

Тогда имеют место следующие распределения:

1.

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

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.

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

4.

$$\sqrt{n} \cdot \frac{\bar{X} - a}{S} \in T_{n-1}$$

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

Доказательство.

1.

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

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

3.

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

, где

$$z_i = \frac{X_i - a}{\sigma} \in N(0, 1), \bar{z} = \frac{\sum_{i=1}^n z_i}{n} = \frac{\sum x_i - na}{\sigma n} = \frac{\bar{X} - a}{\sigma}$$

Поэтому можем считать, что $X_i \in N(0, 1)$. Применим лемму Фишера.

$$T(\vec{X}) = \sum_{i=1}^n (X_i - \bar{X})^2 = n(\bar{X}^2 - (\bar{X})^2) = \sum_{i=1}^n X_i^2 - n(\bar{X})^2 = \sum_{i=1}^n X_i^2 - Y_1$$

, где

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

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 $Y_1 = \sqrt{n}\bar{X} \Rightarrow S^2$ и \bar{X} независимы.

4.

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

$X_0 \in N(0, 1)$ 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 пункту 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:

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

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 $\vec{\eta} = B^{-1}(\vec{\xi} - \vec{a})$, где $B = \sqrt{K}$ ¹, состоит из независимых 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 с параметрами \vec{a} и K . Координаты $\vec{\xi}$ независимы тогда и только тогда, когда они не коррелированы, 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 коррелированы, 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 квантилем¹ 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.

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

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.2X

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

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 распределения

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:

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

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 .

$$\begin{aligned} -t_\gamma &< \sqrt{n} \frac{\bar{X} - a}{\sigma} < t_\gamma \\ -t_\gamma \cdot \frac{\sigma}{\sqrt{n}} &< a - \bar{X} < t_\gamma \cdot \frac{\sigma}{\sqrt{n}} \\ \bar{X} - t_\gamma \cdot \frac{\sigma}{\sqrt{n}} &< a < \bar{X} + t_\gamma \cdot \frac{\sigma}{\sqrt{n}} \end{aligned}$$

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

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

По пункту 4 теоремы 11:

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

$F_{T_{n-1}}(t_\gamma) = \frac{1+\gamma}{2}$, 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.

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

□

$$\begin{aligned} -t_\gamma &< \sqrt{n} \frac{\bar{X} - a}{S} < t_\gamma \\ \bar{X} - t_\gamma \cdot \frac{S}{\sqrt{n}} &< a < \bar{X} + t_\gamma \cdot \frac{S}{\sqrt{n}} \end{aligned}$$

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 параметра 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.

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

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

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

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 $\bar{X} = 4.1$. Найти доверительный интервал параметра a надежности $\gamma = 0.95$.

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

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

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

$$4.1 - 0.98 < a < 4.1 + 0.98$$

$$3.12 < a < 5.08$$

Ответ: (3.12, 5.08)

□

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

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

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

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

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

$$39.678 < a < 44.962$$

Ответ: (39.678, 44.962)

□

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 :

$$K(X_0 \dots X_n) \rightarrow \mathbb{R} = S \cup {}^1\overline{S} \rightarrow (H_0, H_1) \\ \begin{cases} H_0, & \text{если } K \in \overline{S} \\ H_1, & \text{если } K \in S \end{cases}$$

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 $\beta \leq \beta_i \quad \forall i$.

9.2 Критерий согласия

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

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

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

9.3 Построение критериев согласия

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

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

Для заданного уровня значимости 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 $K \Rightarrow Z$, т.е. $F_K(x) \rightarrow F_Z(x)$ и

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

2. K — состоятельный критерий:

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

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

□

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 $\sqrt{n} \cdot \frac{\bar{X} - a_0}{\sigma}$. Проверим, что оно имеет требуемые свойства:

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

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

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

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

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:

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

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}_θ — 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 $\bar{X} = 21.36$. Проверить гипотезу $H_0 : a = 21$ против $H_1 : a \neq 21$ при уровне значимости $\alpha = 0.05$.

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

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

$$t_k = 1.96$$

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

В математических пакетах могут не сравнивать с критической точкой, а считать статистику и искать вероятность.

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

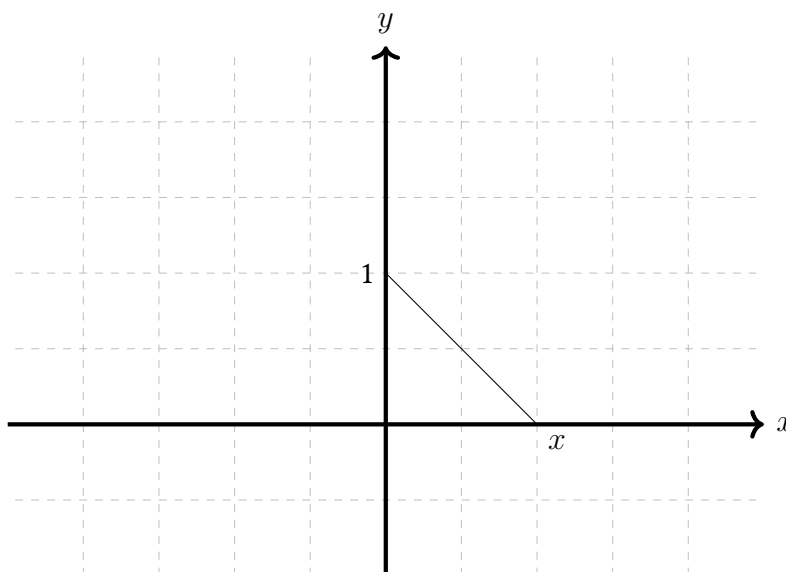


Рис. 6.1: Источник

$$F_\xi(x) = P(\xi < x) = P(\xi < 0) + P(0 < \xi < x) = 0.5 + \frac{1}{\pi} \arctg x$$

$$f_\xi(x) = F'_\xi(x) = \frac{1}{\pi} \frac{1}{x^2 + 1}$$

$$\mathbb{E}\xi = \int_{-\infty}^{+\infty} x f_\xi(x) dx = \int_{-\infty}^{+\infty} x \frac{1}{\pi} \frac{1}{x^2 + 1} dx = \frac{1}{2\pi} \ln(1 + x^2) \Big|_{-\infty}^{+\infty}, \#$$

Пусть теперь источник сдвинут на θ по оси x . Тогда $f_\xi(x) = \frac{1}{\pi(1+(x-\theta)^2)}$. Попробуем оценить θ . \bar{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:

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

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 $n:k$, т.е. групп $\frac{n}{k}$. В каждой группе:

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

Найти оптимальное значение k и среднее значение числа анализов.

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

$$\begin{aligned} P(\xi_i = 1) &= (1 - p)^k & P(\xi_i = k + 1) &= 1 - (1 - p)^k \\ \mathbb{E} \xi_i &= (1 - p)^k + (k + 1)(1 - (1 - p)^k) = k + 1 - k(1 - p)^k \end{aligned}$$

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

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

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

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

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

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

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

□

Лекция 7

18 октября

На прошлой лекции мы обсуждали проверку статистических гипотез, эта лекция будет посвящена основному набору оных.

9.6 Критерии для проверки гипотез о распределении

9.6.1 Критерий χ^2 для параметрической гипотезы

Этот критерий самый популярный.

Пусть дана выборка $(X_1 \dots X_n)$ неизвестного распределения \mathcal{F} . Проверяется основная сложная гипотеза $H_0 : \mathcal{F} \in \mathcal{F}_\theta$, т.е. \mathcal{F} принадлежит классу распределений \mathcal{F}_θ , параметризованное набором из m параметров: $\theta = (\theta_1 \dots \theta_m)$.

Пусть $\hat{\theta} = (\hat{\theta}_1 \dots \hat{\theta}_m)$ — оценка этих параметров методом максимального правдоподобия. Пусть выборка разбита на k интервалов $A_1 \dots A_k$, где $A_i = [a_{i-1}, a_i)$. Пусть n_i — соответствующие экспериментальные частоты попадания в интервал A_i , p_i — соответствующие теоретические вероятности попадания в эти интервалы при распределении $\mathcal{F}_{\hat{\theta}}$

Примечание. $p_i = \mathcal{F}_{\hat{\theta}}(a_i) - \mathcal{F}_{\hat{\theta}}(a_{i-1})$

Тогда $n'_i = np_i$ — теоретические частоты попадания в A_i .

В качестве статистики критерия берётся:

$$K = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i} = \sum_{i=1}^k \frac{n_i^2}{n'_i} - n$$

Теорема 15 (Фишера). Если гипотеза $H_0 : \mathcal{F} \in \mathcal{F}_\theta$ верна, то

$$K = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i} \in H_{k-m-1}$$

, т.е. K имеет распределение χ^2 с $k - m - 1$ степенями свободы, где k — число интервалов и m — число параметров, задающих распределение.

Доказательство. Использует многомерное нормальное распределение. \square

Критерий используется следующим образом: для заданного уровня значимости α находим критическую точку t_k , такую что $P(\chi_{k-m-1}^2 \geq t_k) = \alpha$. Тогда критерий имеет вид:

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

Примечание. $t_k = \text{ХИ2.ОБР.ПХ}(\alpha, k - m - 1)$

Примечание. Частота интервалов должна быть ≥ 5 . Если нет, то объединяем соседние интервалы.

Примечание. Желательно выборку разбить на большое число равнонаполненных интервалов.

Пример. Имеется выборка в виде частотного вариационного ряда объёма $n = 120$: (5.2 ... 82.8). При разбиении её на 8 интервалов получили интервальный ряд:

A_i	n	n_i
[5.2; 7.4)	12	15
[7.4; 9.6)	17	15
[9.6; 11.8)	14	15
[11.8; 14)	13	15
[14; 16.2)	18	15
[16.2; 18.4)	14	15
[18.4; 20.6)	13	15
[20.6; 22.6)	11	15
Σ	120	120

Проверим гипотезу о равномерности распределения при уровне значимости $\alpha = 0.05$: $H_0 : \mathcal{F} \in U(a; b), H_1 : \mathcal{F} \notin U(a; b)$

$\hat{a} = X_{\min} = 5.2, \hat{b} = X_{\max} = 82.8, n_i = \frac{120}{8} = 15$ — теоретические частоты.

$$\chi_{\text{набл}}^2 = \sum_{i=1}^8 \frac{(n_i - n'_i)^2}{n'_i} = 3.2$$

$\alpha = 0.05$, число степеней свободы: $k - m - 1 = 8 - 2 - 1 = 5, t_k(0.05; 5) = 11.07, \chi_{\text{набл}}^2 = 3.2 < 11.07$, гипотеза о равномерном распределении принимается.

9.6.2 Критерий χ^2

Проверяется основная (простая) гипотеза $H_0 : \mathcal{F} = \mathcal{F}_\theta$, где \mathcal{F}_θ — распределение известного типа с известными параметрами, против $H_1 : \mathcal{F} \neq \mathcal{F}_\theta$. В качестве статистики берётся та же самая функция

$$K = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i}$$

Теорема 16 (Парона??). Если гипотеза H_0 верна, то

$$K = \sum_{i=1}^k \frac{(n_i - n'_i)^2}{n'_i} \in H_{k-1}$$

9.6.3 Критерий Колмогорова

Приведён по историческим причинам.

Пусть имеется выборка $(X_1 \dots X_n)$ неизвестного распределения \mathcal{F} . Проверяется простая гипотеза $H_0 : \mathcal{F} = \mathcal{F}_1$ против $H_1 : \mathcal{F} \neq \mathcal{F}_1$. Пусть $F_1(x)$ — непрерывная функция распределения \mathcal{F}_1 . Тогда применяем критерий:

$$K = \sqrt{n} \sup_x |F^*(x) - F_1(x)|$$

, где $F^*(x)$ — выборочная функция распределения.

Теорема 17. Если гипотеза H_0 верна, то

$$K = \sqrt{n} \sup_x |F^*(x) - F_1(x)| \xrightarrow{n \rightarrow \infty} \mathcal{K}$$

, где \mathcal{K} — распределение Колмогорова с функцией распределения $F_{\mathcal{K}}(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 x^2}$.

Для уровня значимости находим t_k и дальше как обычно.

В некоторых статистических пакетах это распределение есть, в Excel — нет. Исторически оно не распространилось.

Недостаток этого критерия в том, что он не применим в дискретном случае.

9.7 Критерии для проверки однородности

Мы хотим узнать, случайна ли эта выборка, или её кто-то неправильно собрал данные.

9.7.1 Критерий Колмогорова-Смирнова

Также используется редко.

Пусть имеются две независимых выборки $(X_1 \dots X_n)$ и $(Y_1 \dots Y_m)$ объёмов n и m соответственно неизвестных непрерывных распределений \mathcal{F} и \mathcal{G} . Проверяется гипотеза $H_0 : \mathcal{F} = \mathcal{G}$ против гипотезы $H_1 : \mathcal{F} \neq \mathcal{G}$. В качестве статистики берётся:

$$K = \sqrt{\frac{nm}{n+m}} \sup_x |F^*(x) - G^*(x)|$$

, где $F^*(x)$ и $G^*(x)$ — соответствующие выборочные функции распределения.

Теорема 18. Если гипотеза H_0 верна, то

$$K = \sqrt{\frac{nm}{n+m}} \sup_x |F^*(x) - G^*(x)| \xrightarrow[n \rightarrow \infty]{m \rightarrow \infty} \mathcal{K}$$

Примечание. Чаще всего в случае нормальных распределений используются критерии Фишера и Стьюдента. Сначала применяем критерий Фишера и если он не отвергает основную гипотезу, то применяем критерий Стьюдента.

Ещё часто применяется ранговый критерий Уилкоксона-Манна-Уитни. Мы его не рассмотрим, но общая идея в следующем: рассматривается только одна выборка и если выборка составлялась не случайно, то порядок возрастания/убывания нарушен.

9.7.2 Критерий Фишера

Пусть имеются две независимых выборки $(X_1 \dots X_n)$ и $(Y_1 \dots Y_m)$ объёмов n и m соответственно из нормальных распределений $N(a_1, \sigma_1^2)$ и $N(a_2, \sigma_2^2)$. Проверяется гипотеза $H_0 : \sigma_1 = \sigma_2$ против гипотезы $H_1 : \sigma_1 \neq \sigma_2$. В качестве статистики берётся:

$$K = \frac{S_x^2}{S_y^2}$$

, где S_x^2, S_y^2 — соответствующие исправленные дисперсии, причём $S_x^2 \geq S_y^2$

Теорема 19. Если H_0 верна, то $\frac{S_x^2}{S_y^2} \in F(n-1, m-1)$ — распределение Фишера с $n-1, m-1$ степенями свободы.

Доказательство. По пункту 3 основной теоремы $\frac{(n-1)S_x^2}{\sigma^2} \in H_{n-1}$ или $\frac{S_x^2}{\sigma^2} \in \frac{\chi_{n-1}^2}{n-1}$.

При $\sigma_1 = \sigma_2 = \sigma$:

$$\frac{S_x^2}{S_y^2} = \frac{S_x^2}{\sigma^2} \cdot \frac{\sigma^2}{S_y^2} = \frac{\chi_{n-1}^2}{n-1} \cdot \frac{m-1}{\chi_{m-1}^2} \stackrel{\text{def}}{=} F(n-1, m-1)$$

□

Критерий по статистике очевиден.

Примечание. При $H_1 : \sigma_1 \neq \sigma_2$, т.е. $\sigma_1 > \sigma_2$, $K = \frac{S_x^2}{S_y^2} = \frac{\sigma_1^2}{\sigma_2^2} > 1$

При H_0 выполнено $K \rightarrow 1$.

9.7.3 Критерий Стьюдента

Пусть имеются две независимых выборки $(X_1 \dots X_n)$ и $(Y_1 \dots Y_m)$ объёмов n и m соответственно из нормальных распределений $N(a_1, \sigma^2)$ и $N(a_2, \sigma^2)$ с одинаковой дисперсией σ^2 . Проверяется гипотеза $H_0 : a_1 = a_2$ против гипотезы $H_1 : a_1 \neq a_2$.

Теорема 20. Случайная величина

$$\sqrt{\frac{nm}{n+m}} \frac{(\bar{X} - a_1) - (\bar{Y} - a_2)}{\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}} \in T_{n+m-2}$$

, где T_{n+m-2} — распределение Стьюдента с $n+m-2$ степенями свободы. Это не зависит от того, верна гипотеза или нет.

В качестве статистики берётся:

$$K = \sqrt{\frac{nm}{n+m}} \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}}$$

Из теоремы видим, что если H_0 верна, то $K \in T_{n+m-2}$, если нет, то $K \xrightarrow[n \rightarrow \infty]{m \rightarrow \infty} \infty$.

Критерий: пусть t_k — квантиль распределения Стьюдента $|T_{n+m-2}|$ уровня значимости α .

$$\begin{cases} H_0 : a_1 = a_2, & K < t_k \\ H_1 : a_1 \neq a_2, & K \geq t_k \end{cases}$$

Существует масса других критериев, но все они строятся похожим образом.

Лекция 8

25 октября

10 Статистическая зависимость

Определение. Функциональная зависимость имеет место, когда две величины связаны жесткими законами природы.

Определение. Зависимость называется **статистической**, если изменение одной величины влияет на распределение другой. Если при этом изменяется среднее значение¹ другой случайной величины, то зависимость называется **корреляционной**. Если среднее значение *увеличивается* при увеличении первой случайной величины, то корреляция *прямая*, а если *уменьшается* — *обратная*.

10.1 Корреляционное облако

Пусть в ходе экспериментов получились значения случайных величин X и $Y : (X_i, Y_i), 1 \leq i \leq n$. Нанося эти точки на координатную плоскость XOY , получим корреляционное облако. По его виду можно сделать предположение о зависимости.

Пример. X, Y имеют нормальное распределение с одинаковыми параметрами.

- Если корреляционное облако имеет форму круга, то величины *независимы*.
- Если корреляционное облако имеет форму эллипса с большой осью параллельной прямой вида $y = kx + b, k > 0$, то скорее всего *зависимость прямая*.

10.2 Корреляционная таблица

Экспериментальные данные представляем в виде таблицы:

¹ Математическое ожидание.

$X_i \backslash Y_i$	Y_1	Y_2	\dots	Y_m
X_1	n_{11}	n_{12}	\dots	n_{1m}
X_2	n_{21}	n_{22}	\dots	n_{2m}
\vdots	\vdots	\vdots	\ddots	\vdots
X_n	n_{n1}	n_{n2}	\dots	n_{nm}

Пример.

$X_i \backslash Y_i$	10	20	30	40	n_x	$\overline{y_x}$
2	7	3	0	0	10	13
4	3	10	10	2	25	24.4
6	0	2	10	3	15	30.67
n_y	10	15	20	5	50	

- n_x — частота значения x
- n_y — частота значения y
- $\overline{y_x}$ — условное среднее случайной величины y :

$$\overline{y_x} = \frac{1}{n_x} \sum_i n_{xy_i} y_i$$

В нашем примере условное матожидание $\overline{y_x}$ увеличивается при увеличении x , следовательно скорее всего есть прямая корреляция.

Примечание. При большом числе данных удобнее составлять интервальную корреляционную таблицу и заменить интервалы на среднее.

10.3 Критерий χ^2 для проверки независимости

Пусть выборка $(X_1, Y_1) \dots (X_n, Y_n)$ случайных величин X и Y представлена в виде интервальной корреляционной таблицы. Случайная величина X при этом разбита на k интервалов, а Y на m интервалов. Обозначим $v_{i.}$ = число значений случайной величины X , попавших в i -тый интервал $[a_{i-1}, a_i)$, $1 \leq i \leq k$. Обозначим $v_{.j}$ = число значений случайной величины Y , попавших в j -тый интервал $[b_{j-1}, b_j)$, $1 \leq j \leq m$. Обозначим v_{ij} = число точек (X, Y) , попавших в $[a_{i-1}, a_i) \times [b_{j-1}, b_j)$.

$X_i \backslash Y_i$	$[b_0; b_1)$	$[b_1; b_2)$	\dots	$[b_{m-1}; b_m)$	$v_i = \sum_{j=1}^n v_{ij}$
$[a_0; a_1)$	v_{11}	v_{12}	\dots	v_{1m}	$v_{1.}$
$[a_1; a_2)$	v_{21}	v_{22}	\dots	v_{2m}	$v_{2.}$
\vdots	\vdots	\vdots	\ddots	\vdots	
$[a_{n-1}; a_n)$	v_{n1}	v_{n2}	\dots	v_{nm}	$v_{n.}$
$v_j = \sum_{i=1}^n v_{ij}$	$v_{.1}$	$v_{.2}$	\dots	$v_{.m}$	

По этой таблице проверяется основная гипотеза $H_0 : X$ и Y независимы против $H_1 : X$ и Y зависимы.

Вспомним определение независимых случайных величин:

$$P(X \in \mathfrak{B}_1, Y \in \mathfrak{B}_2) = P(X \in \mathfrak{B}_1) \cdot P(Y \in \mathfrak{B}_2)$$

Согласно этому определению, если гипотеза H_0 верна, то вероятность попадания пары (X, Y) в любой прямоугольник равна произведению теоретических вероятностей попасть случайным величинам в эти интервалы.

$$p_{ij} = P(X \in [a_{i-1}; a_i), Y \in [b_{j-1}; b_j)) = P(X \in [a_{i-1}; a_i)) \cdot P(Y \in [b_{j-1}; b_j)) = p_i \cdot p_j$$

По закону больших чисел при $n \rightarrow \infty$:

$$\frac{v_{i.}}{n} \xrightarrow{P} p_i \quad \frac{v_{.j}}{n} \xrightarrow{P} p_j \quad \frac{v_{ij}}{n} \xrightarrow{P} p_{ij}$$

Поэтому основанием для отклонения гипотезы служит заметная разница между величинами между $\frac{v_{ij}}{n}$ и $\frac{v_i}{n} \cdot \frac{v_j}{n}$, т.е. между v_{ij} и $\frac{v_i v_j}{n}$.

В качестве статистики критерия берётся функция:

$$K = n \sum_{i,j} \frac{\left(v_{ij} - \frac{v_i v_j}{n}\right)^2}{v_i \cdot v_j}$$

Теорема 21. Если гипотеза H_0 верна, то $K \Rightarrow \chi^2_{(k-1)(m-1)}$

Получили критерий согласия: t_k — квантиль распределения $H_{(k-1)(m-1)}$

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

10.4 Однофакторный дисперсионный анализ

Предположим, что на случайную величину X (*признак-результат*) может влиять фактор Z (*признак-фактор*). Z — не обязательно случайная величина.

Пример. Хотим проверить, как влияет температура на разложение ????. Проводим измерения при разных температурах, регулируя термостат. Тогда температура — не случайная величина, она управляема.

Пусть при различных k уровнях фактора Z получены k независимых выборок случайной величины X : $X^{(1)} = (X_1^{(1)} \dots X_{n_1}^{(1)}) \dots X^{(k)} = (X_1^{(k)} \dots X_{n_k}^{(k)})$. В общем случае размеры выборок могут быть различны.

10.4.1 Общая, межгрупповая и внутригрупповая дисперсия

$$\bar{X}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^{(j)} \quad \mathbb{D}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_i^{(j)} - \bar{X}^{(j)})^2$$

Объединив все данные в одну общую выборку, получим выборку объёма $n = n_1 + \dots + n_k$. Вычислим общее выборочное среднее как

$$\bar{X} = \frac{1}{n} \sum_{i,j} X_i^{(j)} = \frac{1}{n} \sum_{j=1}^k \bar{X}^{(j)} n_j$$

и общую выборочную дисперсию:

$$D_o = \frac{1}{n} \sum_{i,j} (X_i^{(j)} - \bar{X})^2$$

Определение. Внутригрупповой (*остаточной*) дисперсией называется среднее² групповых дисперсий:

$$D_b = \frac{1}{n} \sum_{j=1}^k \mathbb{D}^{(j)} n_j$$

Определение. Межгрупповой (*факторной*) дисперсией или дисперсией выборочных средних называется величина

$$D_m = \frac{1}{n} \sum_{j=1}^k (\bar{X}^{(j)} - \bar{X})^2 n_j$$

Теорема 22 (о разложении дисперсий). Общая дисперсия равна сумме межгрупповой и внутригрупповой дисперсий:

$$D_o = D_b + D_m$$

² взвешенное

Доказательство. Неинтересное, алгебраическое. □

Смысл:

- Внутригрупповая дисперсия показывает средний разброс внутри выборок.
- Межгрупповая дисперсия показывает, насколько отличны выборочные средние при различных уровнях фактора. Таким образом, её величина в общей сумме отражает влияние фактора.

10.4.2 Проверка гипотезы о влиянии фактора

Предположим, что случайная величина X имеет нормальное распределение и фактор Z может влиять только на математическое ожидание, но не на дисперсию и тип распределения.

Может показаться, что ограничение слишком строгое, но в реальной жизни это условие выполняется часто.

Поэтому можно считать, что данные независимые выборки при разных уровнях Z также имеют нормальное распределение с одинаковым параметром σ^2 :

$$X_i^{(j)} \in N(a_i, \sigma^2)$$

Проверяется основная гипотеза $H_0 : a_1 = a_2 = \dots = a_k$, т.е. фактор Z не влияет на X . $H_1 : Z$ влияет на X . По пункту 3 основной теоремы:

$$\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{nD_{\text{в}}}{S^2} \in H_{n-1}$$

Отсюда для каждой из k выборок:

$$\frac{n_j \mathbb{D}^{(j)}}{\sigma} \in H_{n_j-1}, 1 \leq j \leq k$$

Т.к. распределение χ^2 устойчиво по суммированию, то получаем:

$$\sum_{j=1}^k \frac{n_j \mathbb{D}^{(j)}}{\sigma} = \frac{\sum_{j=1}^k n_j \mathbb{D}^{(j)}}{\sigma^2} = \frac{nD_{\text{в}}}{\sigma^2} \in H_{n-k}$$

, т.к. $\sum_{j=1}^k (n_j - 1) = n - k$.

Все это выполнено вне зависимости от того, верна H_0 или нет. Пусть H_0 верна, тогда все выборки можно считать одной выборкой объёма n и по тому же свойству:

$$\frac{nD_o}{\sigma^2} \in H_{n-1}$$

На записи все эти формулы видны примерно как “Ыаыаыаацпы”

Согласно теореме о разложении дисперсии:

$$\underbrace{\frac{nD_o}{\sigma^2}}_{\in H_{n-1}} = \frac{nD_B}{\sigma^2} + \underbrace{\frac{nD_M}{\sigma^2}}_{\in H_{n-k}}$$

Следовательно, $\frac{nD_B}{\sigma^2} \in H_{k-1}$ ³. В итоге при верной гипотезе H_0 мы получили $\frac{nD_B}{\sigma^2} \in H_{k-1}$, а $\frac{nD_M}{\sigma^2} \in H_{n-k}$. Тогда:

$$\frac{\frac{nD_B}{\sigma^2(k-1)}}{\frac{nD_M}{\sigma^2(n-k)}} = \frac{\frac{D_B}{k-1}}{\frac{D_M}{n-k}} \in F(k-1, n-k)$$

В результате имеем критерий $K = \frac{n-k}{n-1} \frac{D_B}{D_M}$, находим t_k — квантиль $F(k-1, n-k)$ уровня значимости α , искомое очевидно строится.

³ Это неочевидный факт, но мы его доказывать не будем.