Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Агафонова Н.Ю.

Краткий курс лекций по математической статистике

Учебное пособие для студентов механико-математического факультета и факультета компьютерных наук и информационных технологий

Саратов, 2019

Содержание

1 ТЕОРИЯ ВЕРОЯТНОСТЕЙ

1.1 Случайные величины

Вероятностным пространством называется тройка $(\Omega, \mathcal{F}, P)$, где: Ω – это множество элементарных исходов, элементы $\omega \in \Omega$ которого называются элементарными исходами; $\mathcal{F}$ – сигма-алгебра подмножеств множества Ω. Подмножества $A \in \mathcal{F}$ называются случайными событиями; P – вероятностная мера или вероятность, то есть сигма-аддитивная конечная мера, такая что $P(\Omega) = 1$.

Определение. Случайной величиной называется действительнозначная функция $\xi(\omega) : \mathcal{F} \to \mathbb{R}$, т.е. такая что $\forall x \in \mathbb{R}$ множество

$$\{\omega : \xi(\omega) < x\} \in \mathcal{F}$$

.

Определение. Функцией распределения вероятностей случайной величины ξ называется функция $F_\xi(x)$, при каждом значении x равная вероятности того, что случайная величина ξ примет значение меньшее, чем x, то есть

$$F_\xi(x) = P(\xi < x),\ x \in \mathbb{R}.$$

Свойства функции распределения $F_\xi(x)$:

1. $0 \leqslant F_\xi(x) \leqslant 1,\ \forall x \in \mathbb{R}$;
2. $F_\xi(x)$ – неубывающая, непрерывная слева функция;
3. $\lim\limits_{x \to -\infty} F_\xi(x) = 0, \lim\limits_{x \to +\infty} F_\xi(x) = 1$;
4. $\mathsf{P}\{a \leqslant \xi < b\} = F(b) - F(a)$.

Среди случайных величин можно выделить два основных типа: дискретные и абсолютно непрерывные случайные величины.

Дискретной называется случайная величина ξ, принимающая конечное или счетное множество значений.

Дискретную случайную величину обычно задают *рядом распределения вероятностей*, т.е. таблицей вида:

ξ	x_1	x_2	$\dots$	x_k	$\dots$
p	p_1	p_2	$\dots$	p_k	$\dots$

Здесь x_k – значения сл.в, которые она принимает с вероятностями

$$p_k = P(\xi = x_k), \quad \sum_{k=1}^{\infty} p_k = 1.$$

Непрерывной называется случайная величина ξ, функция распределения которой дифференцируема, т.е. существует производная $f(x) = F'_{\xi}(x)$, называемая **плотностью распределения вероятностей** случайной величины ξ.

Случайная величина называется **абсолютно непрерывной**, если для любого $x \in \mathbb{R}$ ее функцию распределения можно представить в виде

$$F_{\xi}(x) = \int_{\infty}^{x} f(t)\, dt.$$

Определение Случайные величины ξ и η называются независимыми, если для любых $x, y \in \mathbb{R}$ справедливо равенство:

$$P\{\omega : \xi(\omega) < x, \eta(\omega) < y) = P(\omega : \xi(\omega) < x)P(\omega : \eta(\omega) < y),$$

или, в терминах функций распределения:

$$F_{\xi\eta}(x, y) = F_{\xi}(x)F_{\eta}(y).$$

Определение. Математическим ожиданием дискретной случайной величины ξ называется сумма ряда

$$M\xi = \sum_{k=1}^{\infty} x_k p_k,$$

при условии его абсолютной сходимости.

Определение. Математическим ожиданием абсолютно непрерывной случайной величины ξ с плотностью распределения $f(x)$ называется интеграл

$$M\xi = \int\limits_{-\infty}^{+\infty} xf(x)\,dx,$$

также при условии, что он абсолютно сходится.

Пусть ξ, η – произвольные случайные величины, a, b – константы.

Свойства математического ожидания:

1. $MC = C$, где $C = const$.

2. $M(a\xi + b\eta) = aM\xi + bM\eta$.

3. Если случайные величины ξ, η имеют конечные математические ожидания $M\xi < \infty$, $M\eta < \infty$, то $M(\xi \pm \eta) = M\xi \pm M\eta$.

4. Математическое ожидание произведения независимых случайных величин, имеющих конечные математические ожидания, равно произведению их математических ожиданий, т.е.

$$M\xi\eta = M\xi M\eta.$$

5. Если ξ – дискретная случайная величина, а функция $\varphi(x)$ непрерывна на $\mathbb{R}$, то для случайной величины $\eta = \varphi(x)$ справедливо равенство

$$M\eta = \sum_{k=1}^{\infty} \varphi(x_k)p_k.$$

Если ξ – непрерывная случайная величина, а функция $\varphi(x)$ является непрерывной на $\mathbb{R}$, то для случайной величины $\eta = \varphi(x)$ справедливо

$$M\eta = \int\limits_{-\infty}^{+\infty} \varphi(x)f(x)dx.$$

Определение. Дисперсией случайной величины называется математическое ожидание квадрата отклонения случайной величины от ее математического ожидания:

$$D\xi = M(\xi - M(\xi))^2.$$

Дисперсия характеризует рассеяние (разбросанность) значений случайной величины около ее математического ожидания.

Свойства дисперсии:

1. $D\xi \geq 0$.

2. Если $P\{\xi = C\} = 1$, то $D(C) = 0$, где $C = const$.

3. $D(C\xi) = C^2 D\xi$.

4. $D\xi = M\xi^2 - (M(\xi))^2$.

5. Для произвольных сл.в. ξ и η $D(\xi \pm \eta) = D\xi + D\eta \pm 2cov(\xi, \eta)$.

Определение. Ковариация двух случайных величин – это числовая характеристика зависимости случайных величин, определяемая следующим образом:

$$cov(\xi, \eta) = M[(\xi - M(\xi)(\eta - M(\eta))].$$

Свойства ковариации:

1. Если ξ, η являются независимыми случайными величинами, то

$$cov(\xi, \eta) = 0.$$

2. Ковариация случайной величины с собой равна дисперсии:

$$cov(\xi, \xi) = D\xi.$$

3. Ковариация симметрична:

$$cov(\xi, \eta) = cov(\eta, \xi).$$

4. $cov(\xi + a, \eta + b) = cov(\xi, \eta)$.

Определение. Коэффициентом корреляции $r(\xi, \eta)$ называют величину

$$r(\xi, \eta) = \frac{cov(\xi, \eta)}{\sqrt{D(\xi)}\sqrt{D(\eta)}}.$$

Коэффициент корреляции характеризует степень зависимости между величинами или, другими словами, степень близости связи между величинами к линейному закону.

Основные свойства коэффициента корреляции:

1. Если случайные величины ξ, η независимы, то $r(\xi, \eta) = 0$.

2. $|r(\xi, \eta)| \leq 1$.

3. Если случайные величины ξ, η линейно зависимы, т.е. $\eta = a\xi + b$, то $r(\xi, \eta) = \pm 1$, и наоборот.

1.2 Основные законы распределения

Распределение Бернулли $\varepsilon \sim Bern(p)$.

Закон распределения $P\{\varepsilon = k\} = p^k q^{n-k}$, где $k \in \{0, 1\}$.

Сл.в. ε принимает значение 1 в случае наступления события A ("успех"), значение 0 —в случае наступления события $\overline{A}$ ("неуспех"). Параметром распределения является вероятность успеха p.

Математическое ожидание сл.в. ε

$$M\varepsilon = p.$$

Дисперсия равна

$$D\xi = pq.$$

Биномиальное распределение $\xi \sim Bin(n, p)$.

Сл.в. ξ –количество успехов в n независимых одинаковых испытаниях Бернулли.

Параметры распределения: n – количество испытаний, p –вероятность успеха в одном испытании.

Закон распределения $P\{\xi = k\} = C_n^k p^k q^{n-k}$, $k = 0, 1, \ldots, n$.

Математическое ожидание сл.в. ξ

$$M\xi = np.$$

Дисперсия сл.в. ξ

$$D\xi = npq.$$

Пуассоновское распределение $\xi \sim Pois(\lambda)$.

Сл.в. ξ – количество успехов в n независимых одинаковых испытаниях Бернулли, где n –велико.

Параметр распределения: λ – среднее количество успехов в большом количестве испытаний n.

Закон распределения

$$P\{\xi = k\} = \frac{\lambda^k e^{-\lambda}}{k!},$$

где $k = 0, 1, \ldots$.

Математическое ожидание сл.в. ξ

$$M\xi = \lambda.$$

Дисперсия равна

$$D\xi = \lambda.$$

Равномерное распределение $\xi \sim R[a, b]$.

Функция плотности имеет вид

$$f(x) = \begin{cases} 1/(b-a), & x \notin [a, b]; \\ 0, & x \in [a, b]. \end{cases}$$

Математическое ожидание сл.в. ξ равно

$$M\xi = \frac{b+a}{2}.$$

Дисперсия равна

$$D\xi = \frac{(b-a)^2}{12}.$$

Показательное (экспоненциальное) распределение $\xi \sim \Pi(\lambda)$.

Функция плотности имеет вид

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0; \\ 0, & x \leqslant 0. \end{cases}$$

Математическое ожидание сл.в. ξ равно:

$$M\xi = \frac{1}{\lambda}.$$

Дисперсия равна

$$D\xi = \frac{1}{\lambda^2}.$$

Таким образом, вероятностный смысл параметра λ показательного распределения состоит в следующем: параметр λ есть величина, обратная математическому ожиданию сл.в. Также следует отметить характеристическое свойство показательного распределения: среднее значение сл.в совпадает со средним квадратическим отклонением.

Нормальное распределение (распределение Гаусса) $\xi \sim N(a, \sigma^2)$.

Функция плотности сл.в. $\xi \sim N(a, \sigma^2)$ имеет вид

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

Функция плотности сл.в. $\xi_0 \sim N(0, 1)$ имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big).$$

Числовые характеристики стандартной нормальной случайной величины $\xi_0 \sim N(0,1)$.

Математическое ожидание сл.в. ξ_0 равно

$$M\xi_0 = 0.$$

Дисперсия равна

$$D\xi_0 = 1.$$

Таким образом, для стандартной нормальной величины $\xi_0 \sim N(0,1)$ $M\xi = 0$, $D\xi = 1$, для нормальной величины $\xi \sim N(a, \sigma^2)$ $M\xi = M(\sigma\xi_0 + a) = a$, $D\xi = D(\sigma\xi_0 + a) = \sigma^2$. Таким образом, параметрами нормального распределения являются математическое ожидание и дисперсия.

1.3 Условные распределения. Функция регрессии

Рассмотрим двумерное распределение (ξ, η) .

Определение. Функцией распределения (ф.р.) двумерного вектора или функцией совместного распределения случайных величин ξ и η называют функцию

$$F_{\xi,\eta}(x,y) = P\{\omega : \xi(\omega) < x, \eta(\omega) < y\}.$$

Свойства ф.р. двумерного вектора:

1. Для любых x, y из $\mathbb{R}$ $0 \leqslant F_{\xi,\eta}(x,y) \leqslant 1$.

2. При фиксированном x_0 функция $F_{\xi,\eta}(x_0, y)$ и при фиксированном y_0 функция $F_{\xi,\eta}(x, y_0)$ являются неубывающими, непрерывными слева функциями соответствующей переменной.

3. $\lim\limits_{x\to\infty} F_{\xi,\eta}(x,y) = F_\eta(y)$, $\lim\limits_{y\to\infty} F_{\xi,\eta}(x,y) = F_\xi(x)$,
$\lim\limits_{x\to\infty, y\to\infty} F_{\xi,\eta}(x,y) = 1$,
$\lim\limits_{x\to-\infty, y\to-\infty} F_{\xi,\eta}(x,y) = \lim\limits_{x\to-\infty} F_{\xi,\eta}(x,y) = \lim\limits_{y\to-\infty} F_{\xi,\eta}(x,y) = 0.$

Сл. вектор называют **абсолютно непрерывным**, если его ф.р. представима в виде

$$F_{\xi,\eta}(x,y) = \int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f(x,y)\,dxdy,$$

где $f(x,y)$ – функция плотности распределения вероятностей сл.вектора (ξ,η).

Функция $f(x,y)$ обладает следующими свойствами:

1. Для любых x,y из $\mathbb{R}$ $f(x,y) \geqslant 0$.
2. $\frac{\partial^2 F_{\xi,\eta}(x,y)}{\partial x \partial y} = f(x,y)$ п.в. по $x,y \in \mathbb{R}$.
3. $\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} f(x,y)\,dxdy = 1$.
4. $\iint\limits_{S} f(x,y)\,dS = \mathsf{P}\{\xi \in S\}$.

Функции плотности распределения вероятностей сл.величин ξ и η могут быть восстановлены по функции плотности совместного распределения:

$$f_\xi(x) = \int\limits_{-\infty}^{\infty} f(x,y)\,dy, \quad f_\eta(y) = \int\limits_{-\infty}^{\infty} f(x,y)\,dx.$$

Пусть $F_{\xi,\eta}(x,y)$ – функция совместного распределения вероятностей сл.величин ξ и η.

Определение. Условное распределение сл. в. η относительно сл. в. ξ задается функцией распределения

$$F_{\eta|\xi}(x,y) = \frac{F_{\xi,\eta}(x,y)}{F_\xi(x)},$$

при $F_\xi(x) > 0$, и $F_{\eta|\xi}(x,y) = 0$, при $F_\xi(x) = 0$.

Определение. Условным математическим ожиданием сл. в. η относительно сл. в. ξ называется сл.в.

$$M_{\eta|\xi}(x,y) = \int\limits_{-\infty}^{\infty} y\,dF_{\eta|\xi}(x,y),$$

с функцией распределения $F_\xi(x)$.

Определение. Функцией регрессии сл.в. η на сл.в. ξ называется функция $g(\xi) = M_{\eta|\xi}$. Если $g(\xi) = M_{\eta|\xi} = a\xi + b$, то говорят о линейной регрессии.

Построение уравнения линейной регрессии методом наименьших квадратов (МНК).

Предположим, что в двумерном распределении (ξ, η) возможно получение информации о сл.в. ξ, т.е. мы можем наблюдать её значения, либо знаем распределение. Требуется по имеющейся информации о сл.в. ξ сделать выводы о сл.в. η.

Пусть связь между величинами выражается некоторой линейной функцией $g(x) = ax + b$. Оценим коэффициенты методом наименьших квадратов (МНК).

Обозначим η – наблюдаемое значение η, $\hat{\eta}$ – вычисленные (прогнозные) значения.

Согласно МНК требуется найти такие значения оценок параметров $\hat{a}$ и $\hat{b}$, чтобы была минимальной сумма квадратов отклонений прогнозных значений от наблюдаемых:

$$L(\hat{a}, \hat{b}) = M(\eta - \hat{\eta})^2 \to min.$$

Значит, для нахождения оценки параметров парной регрессионной модели МНК необходимо найти экстремум (минимум) функции двух аргументов.

Запишем необходимые условия экстремума:

$$\begin{cases} \frac{\partial L}{\partial \hat{a}} = -2M(\eta - \hat{a} - \hat{b}\xi) = 0, \\ \frac{\partial L}{\partial \hat{b}} = -2M\xi(\eta - \hat{a} - \hat{b}\xi) = 0, \end{cases}$$

Раскрывая скобки получим:

$$\begin{cases} M\eta - \hat{a}n - \hat{b}M\xi = 0, \\ M\xi\eta - \hat{a}M\xi - \hat{b}M\xi^2 = 0. \end{cases} \quad (1)$$

Из первого уравнения системы имеем оценку параметра a:

$$\hat{a} = M\eta - \hat{b}M\xi. \tag{2}$$

Преобразуем второе уравнение системы и подставим полученную оценку $\hat{a}$

$$M\xi\eta - \hat{a}M\xi - \hat{b}M\xi^2 = M\xi\eta - M\xi M\eta - \hat{b}(M\xi^2 - (M\xi)^2) = 0,$$

$$cov(\xi, \eta) - \hat{b}D\xi = 0.$$

Отсюда получаем оценку параметра b:

$$\hat{b} = \frac{cov(\xi, \eta)}{D_\xi} = \frac{cov(\xi, \eta)}{\sigma_\xi \sigma_\eta}\frac{\sigma_\eta}{\sigma_\xi} = r\frac{\sigma_\eta}{\sigma_\xi}. \tag{3}$$

Таким образом, решение системы уравнений имеет вид

$$\begin{cases} \hat{b} = r\frac{\sigma_\eta}{\sigma_\xi}; \\ \hat{a} = M\eta - \hat{b}M\xi. \end{cases}$$

Уравнение регрессии η на ξ имеет вид:

$$\eta = M\eta + r\frac{\sigma_\eta}{\sigma_\xi}(\xi - M\xi).$$

Уравнение регрессии ξ на η имеет вид

$$\xi = M\xi + 1/r\frac{\sigma_\eta}{\sigma_\xi}(\eta - M\eta)$$

Заметим, что каждое из уравнений имеет $M\xi$ своим решением , т.е графики проходят через точку $(M\xi, M\eta)$ – центр совместного распределения величин ξ и η.

1.4 Законы больших чисел. Сходимость последовательностей случайных величин

Пусть $\{\xi_n\}_{n=1}^{\infty}$ —последовательность случайных величин, заданных на одном и том же вероятностном пространстве $(\Omega, \mathcal{F}, P)$.

Определение. Последовательность случайных величин $\{\xi_n\}_{n=1}^{\infty}$ называется сходящейся **по вероятности** к случайной величине ξ, если для лю-

бого $\varepsilon > 0$

$$P\{|\xi_n - \xi| < \varepsilon\} \to 1, \quad n \to \infty.$$

Сходимость **по вероятности** обозначается $\{\xi_n\} \xrightarrow{p} \xi$.

Определение. Последовательность случайных величин $\{\xi_n\}_{n=1}^{\infty}$ называется сходящейся **почти наверное** к случайной величине ξ , если для любого $\varepsilon > 0$

$$P\{\omega : |\xi_n(\omega) \nrightarrow \xi(\omega)|\} \to 0, \quad n \to \infty,$$

или

$$P\{\omega : \lim_{n\to\infty} \xi_n = \xi\} = 1, \quad n \to \infty.$$

Сходимость **почти наверное** обозначается $\{\xi_n\} \xrightarrow{\text{п.н.}} \xi$

Определение. Последовательность случайных величин $\{\xi_n\}_{n=1}^{\infty}$ называется сходящейся **в среднем** к случайной величине ξ , если для любого n существуют $M\xi_n$, $M\xi$ и

$$\lim_{n\to\infty} M|\xi_n - \xi| = 0, \quad n \to \infty.$$

Сходимость **в среднем** обозначается $\{\xi_n\} \xrightarrow{\text{L}} \xi$.

Определение. Последовательность случайных величин $\{\xi_n\}_{n=1}^{\infty}$ называется сходящейся **по распределению** к случайной величине ξ , если для любого $x \in \mathbb{R}$

$$\lim_{n\to\infty} F_n(x) = F(x), \quad n \to \infty,$$

где $F_n(x) = P\{\xi_n < x\}$, $F(x) = P\{\xi < x\}$ — функции распределения случайных величин.

Сходимость **по распределению** обозначается $\{\xi_n\} \xrightarrow{\text{d}} \xi$.

Теорема.

Из сходимости почти наверное вытекает сходимость по вероятности,

$$\{\xi_n\} \xrightarrow{\text{п.н.}} \xi \Rightarrow \{\xi_n\} \xrightarrow{p} \xi.$$

Из сходимости по вероятности вытекает сходимость по распределению,

$$\{\xi_n\} \xrightarrow{p} \xi \Rightarrow \{\xi_n\} \xrightarrow{\mathrm{d}} \xi.$$

Из сходимости в среднем вытекает сходимость по вероятности,

$$\{\xi_n\} \xrightarrow{\mathrm{L}} \xi \Rightarrow \{\xi_n\} \xrightarrow{p} \xi.$$

Определение. Говорят, что последовательность случайных величин $\{\xi_i\}_{i=1}^{\infty}$ удовлетворяет **закону больших чисел** (ЗБЧ), если последовательность $\{S_n\}_{n=1}^{\infty}$, составленная из частных сумм $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$, при любом ε удовлетворяет

$$P\Big\{\Big|\frac{S_n}{n} - \frac{MS_n}{n}\Big| < \varepsilon\Big\} \to 1, \quad \text{при}\, n \to \infty.$$

Таким образом, если последовательность удовлетворяет ЗБЧ, то при достаточно большом количестве слагаемых среднее из наблюдаемых значений случайных величин и среднее из их мат.ожиданий практически равны с вероятностью 1.

В терминах сходимости сл.величин ЗБЧ означает, что

$$\frac{S_n}{n} - \frac{MS_n}{n} \xrightarrow{p} 0.$$

Теорема. (Закон больших чисел в форме Чебышёва.)
Пусть $\{\xi_i\}_{i=1}^{\infty}$ —последовательность независимых случайных величин со средними $M\xi_i = a_i < \infty$, при всех $i \in \mathbb{Z}_+$, и дисперсиями $D\xi_i \leqslant C$, при всех $i \in \mathbb{Z}_+$, где $C = const$.
Тогда последовательность $\{\xi_i\}_{i=1}^{\infty}$ удовлетворяет ЗБЧ, и при любом ε удовлетворяет неравенству

$$P\Big\{\Big|\frac{S_n}{n} - \frac{MS_n}{n}\Big| > \varepsilon\Big\} \leqslant \frac{C}{n\varepsilon^2}.$$

Напомним, что $\xi_0 \sim N(0,1)$ —стандартная нормальная случайная величина с $M\xi_0 = 0$ и $D\xi_0 = 1$. Обозначим также через $F_{\xi_0}(x)$ — функцию

распределения вероятностей случайной величины $\xi_0 \sim N(0,1)$.

Центральная предельная теорема(ЦПТ).

Пусть $\{\xi_i\}_{i=1}^{\infty}$ —последовательность независимых одинаково распределенных случайных величин с $M\xi_i = a < \infty$ и $D\xi_i = \sigma^2$, при каждом $i \in \mathbb{Z}_+$.

Тогда для любого $x \in \mathbb{R}$ при $n \to \infty$

$$P\Big\{\frac{S_n/n - a}{\sigma/\sqrt{n}} < x\Big\} \to \frac{1}{\sqrt{2\pi}} \int_{\infty}^{x} e^{-\frac{u^2}{2}}\, du.$$

Замечание.

Введем в рассмотрение сл.величины

$$Z_n = \frac{S_n - MS_n}{\sqrt{DS_n}} = \frac{S_n/n - a}{\sigma/\sqrt{n}}.$$

Тогда выражение в левой части предела есть функция распределения сл.в. Z_n, а справа записана функция распределения величины $\xi_0 \sim N(0,1)$. Следовательно, ЦПТ означает, что "суммы"большого количества независимых и одинаково распределенных сл.в. "ведут"себя по нормальному закону, а именно

$$F_{Z_n}(x) \to F_{\xi_0}(x), n \to \infty, \quad \text{или} \quad Z_n \xrightarrow{\mathrm{d}} \xi_0$$

2 МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

2.1 Вариационные ряды и их характеристики

Рассмотрим постановку задачи математической статистики: по результатам наблюдения за некоторой случайной величиной ξ требуется сделать выводы о неизвестном законе распределения этой величины $\mathcal{L}(x, \theta)$ либо о неизвестных параметрах $\theta_1, \ldots, \theta_n$ известного распределения.

Пусть ξ — случайная величина (сл.в.) с некоторой (теоретической) функцией распределения $F_\xi(x) = P\{\xi < x\}$, $x \in \mathbb{R}$.

Определение. Совокупность n независимых одинаково распределенных случайных величин $X_1, X_2, \ldots, X_n$ называется выборкой, извлеченной из распределения случайной величины ξ.

Определение. Набор n значений $x_1, x_2, \ldots, x_n$ сл. величин $X_1, X_2, \ldots, X_n$ называется реализацией выборки или выборкой из генеральной совокупности *значений* случайной величины ξ.

Под генеральной совокупностью понимается множество всех возможных значений случайной величины ξ. Объем совокупности есть количество всех ее элементов, объем выборки или выборочной совокупности обозначается n, генеральной совокупности — N.

Определение. Выборочным пространством называется множество $\mathbf{X_n} = \{\overline{\mathbf{X}}_\mathbf{n} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})\}$ всех возможных выборок объема n, извлеченных из распределения $\mathcal{L}(x, \theta)$ случайной величины ξ.

В силу того, что наблюдать все выборки из выборочного пространства мы не можем, статистические выводы о наблюдаемой случайной величине приходится делать по имеющейся выборке. *Выборочным методом* в статистике называется метод, на основании которого, за свойства генеральной совокупности принимаются свойства выборочной совокупности. Для того, чтобы полученные при таком подходе оценки были более точными, выборка должна

быть в первую очередь *репрезентативной*, т.е. отражать истинные свойства генеральной совокупности. Обеспечивается репрезентативность, как правило, с помощью методов отбора единиц наблюдения в выборку.

Первым этапом статистического анализа данных является группировка полученных в результате наблюдения за процессом или явлением данных.

Определение. Вариационным рядом называется последовательность расположенных в порядке неубывания элементов выборки $x_1^* \leqslant x_2^* \leqslant, \ldots, \leqslant x_n^*$.

Элементы вариационного ряда называются вариантами.

Определение. Точечным вариационным рядом называется таблица вида

x_i	x_1	x_2	$\ldots$	x_m
n_i	n_1	n_2	$\ldots$	n_m

,

где x_i —варианты, n_i —частоты, m —количество групп (различных значений), $n = \sum\limits_{i=1}^{m} n_i$ —объем выборки.

Для графического представления точечных вариационных рядов используется полигон частот—ломаная с вершинами в точках (x_i, n_i).

Определение. Интервальным вариационным рядом называется таблица вида

x_i	$[x_1, x_2]$	$(x_2, x_3]$	$\ldots$	$(x_m, x_{m+1}]$
n_i	n_1	n_2	$\ldots$	n_m

,

где x_i —варианты, n_i —частоты, m —количество групп (интервалов), $n = \sum\limits_{i=1}^{m} n_i$ —объем выборки.

Для графического представления интервальных вариационных рядов используется гистограмма частот — фигура, составленная из прямоугольников, одной стороной которых служат интервалы $(x_i, x_{i+1}]$, а длина второй равна n_i.

Пусть $X_1, X_2, \ldots, X_n$ выборка из распределения сл.в. ξ с теоретической функцией распределения $F_\xi(x)$.

Определение. Эмпирической функцией распределения (ЭФР) называется функция

$$\widetilde{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} e(x - X_i), \tag{4}$$

где $e(x) = 1$, при $x > 0$, $e(x) = 0$, при $x <= 0$.

Таким образом, если $X_i < x$, то $e(x) = 1$, если $X_i \geqslant x$, то $e(x) = 0$, а сумма $e(x - X_i)$ будет равна количеству элементов выборки, которые приняли значение, строго меньше некоторого $x \in \mathbb{R}$.

Пусть $x_1, x_2, \ldots, x_n$ — реализация выборки $X_1, X_2, \ldots, X_n$, т.е. наблюдавшиеся значения сл.в. ξ. Обозначим $\mu(x)$ — число элементов выборки, строго меньших $x \in \mathbb{R}$.

Тогда эмпирическая функция распределения $\widetilde{F_n}(x)$ может быть определена как

$$\widetilde{F_n}(x) = \frac{\mu(x)}{n}. \tag{5}$$

С помощью полигона и гистограммы частот можно оценить вид функции плотности наблюдаемого распределения, ЭФР является оценкой (теоретической) функции распределения. С помощью анализа графиков можно сделать предположение (выдвинуть статистическую гипотезу) о типе наблюдаемого распределения.

Числовые характеристики вариационных рядов

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{L}(x, \theta)$ — выборка из распределения сл.в. ξ.

Определение. Выборочным средним называется величина

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{6}$$

Если данные представлены в виде точечного или интервального вариационного ряда, то для вычисления используют формулу:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{m} x_i * n_i, \tag{7}$$

где m – количество групп в точечном или интервалов в интервальном вариационном ряду, n_i – частота, т.е. количество элементов выборки, принадлежащих i-той группе или i-тому интервалу, x_i – варианта для точечного ряда и середина i-того интервала для интервального ряда.

Определение. Выборочной дисперсией (смещенной) называется величина

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2. \tag{8}$$

Она характеризует среднее из квадратов отклонений наблюдаемой величины от выборочного среднего. Величина $S = \sqrt{S^2}$ называется выборочным средним квадратическим отклонением (смещенным) величин выборки от выборочного среднего.

Определение. Выборочной дисперсией (несмещенной) называется величина

$$\widetilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2. \tag{9}$$

Аналогично, величина $\widetilde{\sigma} = \sqrt{\widetilde{\sigma}^2}$ называется выборочным несмещенным средним квадратическим отклонением. Очевидно, что смещенная и несмещенная выборочные дисперсии связаны формулой

$$\widetilde{\sigma}^2 = \frac{n}{n-1}S^2.$$

Кроме того, при любом n $\widetilde{\sigma}^2 \geqslant S^2$, а $\lim\limits_{n\to\infty} S^2 = \widetilde{\sigma}^2$

Если данные представлены в виде точечного или интервального вариационного ряда, то для вычисления используют формулу:

$$S^2 = \frac{1}{n}\sum_{i=1}^{m}(x_i - \overline{x})^2 n_i. \tag{10}$$

или

$$S^2 = \left(\frac{1}{n}\sum_{i=1}^{m} x_i^2 n_i\right) - \overline{x}^2. \tag{11}$$

Здесь m – количество групп в точечном или интервалов в интервальном вариационных рядах, n_i – частота, т.е. количество элементов выборки,

принадлежащих i –той группе или i –тому интервалу, x_i – варианта для точечного ряда и середина i –того интервала для интервального ряда.

2.2 Оценки параметров распределения

Пусть $X_1, X_2, \ldots, X_n$ выборка, извлеченная из распределения $\mathcal{L}(x, \theta)$ случайной величины ξ. Пусть требуется получить оценку неизвестного значения параметра θ.

Например, известно, что рост человека распределен по нормальному закону, т.е. $\xi \sim N(a, \sigma^2)$, где a –среднее значение величины, и σ^2 — дисперсия, являются параметрами распределения. По результатам измерения роста 100 человек можно получить выборочное среднее значение, которое и будет оценкой параметра a.

При построении оценок необходимо, чтобы они удовлетворяли определенным свойствам.

Определение. Оценкой параметра θ распределения $\mathcal{L}(x, \theta)$ называется величина $\widetilde{\theta}_n = f(X_1, X_2, \ldots, X_n)$, где $f(t_1, t_2, \ldots, t_n)$ некоторая непрерывная функция.

Заметим, что $\widetilde{\theta}_n$ как функция от случайных величин также есть случайная величина.

Определение. Оценка $\widetilde{\theta}_n$ параметра θ называется **несмещенной**, если

$$M\widetilde{\theta}_n = \theta.$$

(математическое ожидание оценки равно оцениваемому параметру).

Определение. Оценка $\widetilde{\theta}_n$ параметра θ называется **состоятельной**, если при всех $\varepsilon > 0$

$$\lim_{n\to\infty} P\{|\widetilde{\theta}_n - \theta| < \varepsilon\} = 1. \tag{12}$$

(т.е. если оценка состоятельная, то при достаточно большом объеме выборки оценка параметра с высокой вероятностью практически равна параметру).

Также можно заметить, что состоятетельность оценки означает сходимость по вероятности последовательности оценок $\widetilde{\theta}_n$ к параметру θ: $\{\theta_n\} \xrightarrow{p} \theta$.

Оценка $\overline{x}$ является несмещенной и состоятельной оценкой математического ожидания $M\xi$, ЭФР $\widetilde{F_n}(x)$ —несмещенной и состоятельной оценкой функции распределения $F_\xi(x) = P\{\xi < x\}$.

Теорема.(о несмещенной и состоятельной оценке математического ожидания)

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{L}(x, \theta)$ — выборка из распределения сл.в. ξ с конечным математическим ожиданием $M\xi = a < \infty$. Выборочное среднее $\overline{x}$ является несмещенной и состоятельной оценкой математического ожидания.

Доказательство.

Докажем несмещенность $M\overline{x} = M\xi = a$. Действительно

$$M\overline{x} = M\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} M\xi = \frac{1}{n} n * M\xi = M\xi = a.$$

Докажем состоятельность, т.е. что для любого $\varepsilon > 0$ имеет место

$$\lim_{n\to\infty} P\{|\overline{x} - a| < \varepsilon\} = 1.$$

По неравенству Чебышёва для любого $\varepsilon > 0$

$$P\{|\overline{x} - a| < \varepsilon\} \geqslant \frac{D(\overline{x})}{\varepsilon^2} = \frac{1}{\varepsilon^2} D\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) =$$

$$= \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{\varepsilon^2 n^2} n \cdot D\xi = \frac{1}{\varepsilon^2 n} D\xi.$$

С другой стороны, по свойству вероятности имеем

$$P\{|\overline{x} - a| < \varepsilon\} \leqslant 1.$$

Таким образом, доказано, что при любом $\varepsilon > 0$

$$\lim_{n\to\infty} P\{|\overline{x} - a| < \varepsilon\} = 1,$$

и оценка $\overline{x}$ является состоятельной.

Теорема.(о несмещенной и состоятельной оценке функции распределения)

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{L}(x, \theta)$ — выборка из распределения сл.в. ξ с функцией распределения $F_\xi(x) = P\{\xi < x\}$. Эмпирическая функция распределения (ЭФР) $\widetilde{F_n}(x)$ является несмещенной и состоятельной оценкой функции распределения $F_\xi(x)$.

Доказательство.

Рассмотрим вначале ЭФР в виде

$$\widetilde{F_n}(x) = \frac{\mu(x)}{n},$$

где $\mu(x)$ —число элементов выборки строго меньших $x \in \mathbb{R}$. Величина $\mu(x)$ является случайной, принимает значения из множества $\{0, \ldots, n\}$ с вероятностями

$$P\{\mu(x) = k\} = C_n^k P\{X_i < x\} P\{X_i \geqslant x\} = C_n^k F_\xi(x)(1 - F_\xi(x)).$$

Таким образом, величина $\mu(x)$ распределена по биномиальному закону $Bin(n, p)$ с параметрами n и $p = F_\xi(x)$, а значит, имеет числовые характеристики $M\mu(x) = nF_\xi(x)$ и $D\xi = nF_\xi(x)(1 - F_\xi(x))$.

Для доказательства несмещенности найдем $M\widetilde{F_n}(x)$.

$$M\widetilde{F_n}(x) = M\Big(\frac{\mu(x)}{n}\Big) = \frac{1}{n} M(\mu(x)) = \frac{1}{n} nF_\xi(x) = F_\xi(x),$$

что по определению означает несмещенность оценки.

Для доказательства состоятельности снова воспользуемся неравенством Чебышёва. Для любого $\varepsilon > 0$ имеем

$$P\{|\widetilde{F_n}(x) - F_\xi(x)| < \varepsilon\} \geqslant \frac{D(\widetilde{F_n}(x))}{\varepsilon^2} = \frac{1}{\varepsilon^2 n^2} nF_\xi(x)(1 - F_\xi(x)) =$$

$$= \frac{F_\xi(x)(1-F_\xi(x))}{\varepsilon^2 n} \to 1 \quad \text{при} \quad n \to \infty.$$

Учитывая, что $P\{|\widetilde{F_n}(x) - F_\xi(x)| < \varepsilon\} \leqslant 1$, доказана состоятельность оценки $\widetilde{F_n}(x)$ для функции распределения $F_\xi(x)$.

Теорема.(о несмещенной оценке дисперсии)

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{L}(x, \theta)$ — выборка из распределения сл.в. ξ с конечным математическим ожиданием $M\xi = a < \infty$ и дисперсией $D\xi = \sigma^2$. Выборочная дисперсия

$$\widetilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$$

является несмещенной оценкой дисперсии.

Доказательство.

Заметим сначала, что так как элементы выборки предполагаются независимыми случайными величинами, то при $i \neq j$ $M(X_i X_j) = M(X_i)M(X_j) = a \cdot a = a^2$. Если же $i = j$, то $M(X_i X_i) = M(X_i^2) = \sigma^2 + a^2$.

Для доказательства несмещенности покажем, что $M(\widetilde{\sigma}^2) = \sigma^2$.

Действительно,

$$M(\widetilde{\sigma}^2) = M\Big(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2\Big) = \frac{1}{n-1}\sum_{i=1}^{n}\big(M(X_i^2) - 2M(X_i \cdot \overline{x}) + M(\overline{x} \cdot \overline{x})\big) \tag{13}$$

Рассмотрим первое слагаемое, стоящее в скобках под знаком суммы. Так как элементы выборки представляют собой одинаково распределенные случайные величины, то $M(X_1^2) = \cdots = M(X_n^2) = M\xi^2 = D\xi + (M\xi)^2 = \sigma^2 + a^2$.

Второе слагаемое (без множителя 2) запишем в виде

$$M(X_i \cdot \overline{x}) = M(X_i \cdot \frac{1}{n}\sum_{j=1}^{n} X_j) = \frac{1}{n}\sum_{j=1}^{n} M(X_i \cdot X_j) =$$

$$= \frac{1}{n}(MX_i \cdot MX_1 + MX_i \cdot MX_2 + \cdots + MX_i^2 + \cdots + MX_i \cdot MX_n) =$$

$$\frac{1}{n}(a^2 + a^2 + \cdots + a^2 + \sigma^2 + \cdots + a^2) = \sigma^2/n + a^2.$$

При вычислении третьего слагаемого одно $\overline{x}$ распишем по определению, второе $\overline{x}$ оставим в прежнем виде. Получим:

$$M(\overline{x}\cdot\overline{x}) = M(\overline{x}\cdot\frac{1}{n}\sum_{i=1}^{n}X_i) = \frac{1}{n}\sum_{i=1}^{n}M(X_i\cdot\overline{x}) = \cdots = \sigma^2/n + a^2.$$

Подставляя полученные выражения в (13) получаем требуемое равенство, и несмещенность доказана.

2.3 Эффективные оценки. Неравенство Рао-Крамера.

Рассмотрим параметрическую статистическую модель

$$\{\mathbf{X}_n; F_\xi(x,\theta) : \theta\in\Theta\},$$

где $\mathbf{X}_n$ – выборочное пространство, $F_\xi(x,\theta)$ – функция распределения наблюдаемой сл.в. ξ, известная с точностью до параметра θ, Θ –множество значений параметра θ.

Определение. Оценка $\widetilde{\theta}^*$ параметра θ называется эффективной, если $D\widetilde{\theta}^* = \inf\limits_{\widetilde{\theta}\in\Theta} D\widetilde{\theta}$.

Теорема.(о единственности эффективной оценки)

Пусть $\hat{\theta}(\overline{X}_n)$ и $\widetilde{\theta}(\overline{X}_n)$ две эффективные оценки для параметра θ параметрической модели $\{\mathbf{X}_n; F_\xi(x,\theta) : \theta\in\Theta\}$.

Тогда $\hat{\theta}(\overline{X}_n) = \widetilde{\theta}(\overline{X}_n)$, где $P\{\overline{X}_n \in \{\overline{x}_n : \hat{\theta}(\overline{x}_n) \neq \widetilde{\theta}(\overline{x}_n)\}\} = 0.$

Доказательство.

Рассмотрим оценку

$$\theta^*(\overline{X}_n) = \frac{\hat{\theta}(\overline{X}_n) + \widetilde{\theta}(\overline{X}_n)}{2}.$$

Т.к. $\hat{\theta}(\overline{X}_n)$ и $\widetilde{\theta}(\overline{X}_n)$ эффективные, то $D\hat{\theta}(\overline{X}_n) = D\widetilde{\theta}(\overline{X}_n)$, и $M\hat{\theta}(\overline{X}_n) = M\widetilde{\theta}(\overline{X}_n) = \theta.$

Найдем характеристики оценки $\theta^*(\overline{X}_n)$.

$$M\theta^*(\overline{X}_n) = M\frac{\hat{\theta}(\overline{X}_n) + \widetilde{\theta}(\overline{X}_n)}{2} = \frac{1}{2}(M\hat{\theta}(\overline{X}_n) + M\widetilde{\theta}(\overline{X}_n)) = \theta.$$

$$D\theta^*(\overline{X}_n) = D\frac{\hat{\theta}(\overline{X}_n) + \widetilde{\theta}(\overline{X}_n)}{2} = \frac{1}{4}(D\hat{\theta}(\overline{X}_n) + D\widetilde{\theta}(\overline{X}_n) + 2cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n))) =$$
$$= \frac{1}{2}D\hat{\theta}(\overline{X}_n) + \frac{1}{2}cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n)).$$

Имеем

$$|cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n))| = |M(\theta - \hat{\theta}(\overline{X}_n))(\theta - \widetilde{\theta}(\overline{X}_n))| \leqslant$$
$$\leqslant \sqrt{M(\theta - \hat{\theta}(\overline{X}_n))^2 M(\theta - \widetilde{\theta}(\overline{X}_n))^2} = \sqrt{D\hat{\theta}(\overline{X}_n)D\widetilde{\theta}(\overline{X}_n)} = D\hat{\theta}(\overline{X}_n)$$

Тогда

$$D\theta^* = |D\theta^*| = \frac{1}{2}|D\hat{\theta}(\overline{X}_n) + cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n))| \leqslant$$
$$\leqslant \frac{1}{2}|D\hat{\theta}(\overline{X}_n)| + \frac{1}{2}|cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n))| \leqslant D\hat{\theta}(\overline{X}_n).$$

Таким образом, θ^* несмещенная оценка, и ее дисперсия $D\theta^*(\overline{X}_n) = D\hat{\theta}(\overline{X}_n)$, и следовательно, оценка $\theta^*(\overline{X}_n)$ также является эффективной.

Получаем уравнение

$$D\theta^*(\overline{X}_n) = \frac{1}{2}D\hat{\theta}(\overline{X}_n) + \frac{1}{2}cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n)),$$

откуда следует $cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n)) = D\hat{\theta}(\overline{X}_n)$.

Вычислим коэффициент корреляции

$$r(\hat{\theta}, \widetilde{\theta}) = \frac{cov(\hat{\theta}(\overline{X}_n), \widetilde{\theta}(\overline{X}_n))}{\sqrt{D\hat{\theta}(\overline{X}_n)D\widetilde{\theta}(\overline{X}_n)}} = \frac{D\hat{\theta}(\overline{X}_n)}{D\hat{\theta}(\overline{X}_n)} = 1.$$

Следовательно, величины $\hat{\theta}(\overline{X}_n)$ и $\widetilde{\theta}(\overline{X}_n)$ линейно связаны

$$\hat{\theta}(\overline{X}_n) = a\widetilde{\theta}(\overline{X}_n) + b.$$

Пользуясь несмещенностью, получаем

$$M\hat{\theta}(\overline{X}_n) = aM\widetilde{\theta}(\overline{X}_n) + b$$

$$\theta = a\theta + b.$$

Отсюда следует, что $a = 1$, $b = 0$, $\hat{\theta}(\overline{X}_n) = \widetilde{\theta}(\overline{X}_n)$.

Неравенство Рао-Крамера. (Информационное неравенство)

Рассмотрим параметрическую статистическую модель

$$\{\mathbf{X}_n; F_\xi(x,\theta) : \theta \in \Theta\},$$

где $\mathbf{X}_n$ – выборочное пространство, $F_\xi(x,\theta)$ – функция распределения наблюдаемой сл.в. ξ, известная с точностью до параметра θ, Θ –множество значений параметра θ.

Обозначим $f(x,\theta)$ функцию плотности для абсолютно непрерывных сл.в. и закон распределения для дискретных сл.в.

Определение. Параметрическая модель называется **регулярной**, если она удовлетворяет условиям регулярности.

Условия регулярности.

1. параметрическое множество Θ — открытый интервал;
2. носитель распределения – множество $A = \{x : f(x,\theta) > 0\}$ не зависит от параметра;
3. для любого $\theta \in \Theta$ и $x \in A$ существует и конечна производная
$$\frac{\partial f(x,\theta)}{\partial \theta};$$
4. для любого $\theta \in \Theta$
$$M\frac{\partial \ln f(x,\theta)}{\partial \theta} = 0; \quad M\Big(\frac{\partial \ln f(x,\theta)}{\partial \theta}\Big)^2 < \infty;$$
5. в интеграле $\int f(x,\theta)dx$ допустимо дифференцирование по параметру под знаком интеграла дважды.

Теорема.(Неравенство Рао-Крамера)

Пусть параметрическая модель $\{\mathbf{X}_n; F_\xi(x,\theta) : \theta \in \Theta\}$ регулярна, и $\widetilde{\theta}(\overline{X}_n)$ несмещенная оценка параметра θ.

Тогда

$$D(\widetilde{\theta}(\overline{X}_n)) \geqslant \frac{1}{nI(\theta)},$$

где

$$I(\theta) = \Big(M\frac{\partial \ln f(x,\theta)}{\partial\theta}\Big)^2$$

– количество информации по Фишеру в одном наблюдении.

2.4 Методы построения оценок параметров распределения

К основным методам относятся **метод моментов** и **метод максимального правдоподобия**.

Начальным моментом порядка k, $k \in \mathbb{Z}_+$ распределения случайной величины ξ называется величина

$$m_k = M\xi^k, \text{ где } \quad M\xi^k < \infty.$$

Центральным моментом порядка k, $k \in \mathbb{Z}_+$ распределения случайной величины ξ называется величина

$$\mu_k = M(\xi - M\xi)^k, \quad \text{где} \quad M\xi < \infty.$$

Таким образом, $m_1 = M\xi$ – математическое ожидание сл. величины ξ, $m_2 = M\xi^2$, $\mu_1 = M(\xi - M\xi) = 0$, $\mu_2 = M(\xi - M\xi)^2 = D\xi$ – дисперсия сл.в. ξ.

Выборочным начальным моментом порядка k называется величина

$$\widetilde{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k. \tag{14}$$

Выборочным центральным моментом порядка k называется величина

$$\widetilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^k. \tag{15}$$

Таким образом, $\widetilde{m}_1 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i = \overline{x}$ – выборочное среднее (несмещенная оценка математического ожидания сл. величины ξ), $\widetilde{\mu}_2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{x})^2 = S^2$ – смещенная оценка дисперсии сл.в. ξ.

Метод моментов состоит в том, что за оценку параметров распределения принимается решение системы уравнений

$$\begin{cases} m_k = \widetilde{m}_k, \\ \mu_k = \widetilde{\mu}_k. \end{cases}$$

Количество и тип уравнений в системе выбирается в зависимости от распределения.

Пусть наблюдаемая случайная величина имеет распределение с неизвестным параметром θ . Обозначим $P(x,\theta)$ закон распределения этой величины, подразумевая, что $P(x,\theta) = P\{\xi = x\}$, если величина дискретна, $P(x,\theta) = f(x)$, если $P(x,\theta) = f(x)$ величина абсолютно непрерывна.

Функцией правдоподобия называется функция

$$L(\overline{X}_n, \theta) = \prod_{i=1}^{n} P(X_i, \theta).$$

Таким образом, для выборки из распределения дискретной случайной величины

$$L(\overline{X}_n, \theta) = P(X_1, \theta) \cdot P(X_2, \theta) \cdot \cdots \cdot P(X_n, \theta),$$

для выборки из распределения абсолютно непрерывной случайной величины

$$L(\overline{X}_n, \theta) = f(X_1, \theta) \cdot f(X_2, \theta) \cdot \cdots \cdot f(X_n, \theta).$$

Метод максимального правдоподобия (ММП) состоит в том, что за оценку параметра θ принимается точка максимума функции правдоподобия $L(x,\theta)$.

Таким образом, алгоритм ММП таков:

1. Составить функцию правдоподобия $L(x,\theta)$;

2. Взять натуральный логарифм $\ln L(x,\theta)$;

3. Найти производную $\frac{\partial \ln L(x,\theta)}{\partial \theta}$ и решение уравнения $\frac{\partial \ln L(x,\theta)}{\partial \theta} = 0$;

4. Доказать, что решение уравнения является точкой максимума

$$\frac{\partial^2 \ln L(x,\theta)}{\partial \theta^2} < 0.$$

Рассмотрим применение ММП.

Задача. Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{N}(a,\sigma)$ выборка из нормального распределения с математическим ожиданием a и дисперсией σ^2. Оценим методом максимального правдоподобия параметры нормального распределения.

Составим функцию правдоподобия

$$L(\overline{X}_n, a, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-a)^2/2\sigma^2} = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-a)^2\Big).$$

Возьмем ln

$$\ln L(\overline{X}_n, a, \sigma) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-a)^2.$$

Вычислим производные и решим соответствующие уравнения

$$\begin{cases} \dfrac{\partial \ln L(\overline{X}_n, a, \sigma)}{\partial a} = \frac{1}{\sigma^2}\sum\limits_{i=1}^{n}(X_i-a) = 0, \\ \dfrac{\partial \ln L(\overline{X}_n, a, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum\limits_{i=1}^{n}(X_i-a)^2 = 0. \end{cases}$$

Получаем:

$$\begin{cases} n\overline{x} - na = 0, \\ -n + \frac{1}{\sigma^2}\sum\limits_{i=1}^{n}(X_i-a)^2 = 0. \end{cases}$$

$$\begin{cases} a = \overline{x}, \\ -n + \frac{1}{\sigma^2}\sum\limits_{i=1}^{n}(X_i-\overline{x})^2 = 0. \end{cases}$$

Решение этой системы, таким образом, имеет вид $a = \overline{x}$, $\sigma^2 = S^2$.

Докажем, что эти значения доставляют максимум функции правдоподобия.

$$\begin{cases} \dfrac{\partial^2 \ln L(\overline{X}_n, a, \sigma)}{\partial a^2} = -\dfrac{n}{\sigma^2}|_{\sigma^2=S^2} = -\dfrac{n}{S^2} < 0, \\ \dfrac{\partial^2 \ln L(\overline{X}_n, a, \sigma)}{\partial \sigma^2} = \dfrac{n}{\sigma^2} - 3\dfrac{1}{\sigma^4}\sum\limits_{i=1}^{n}(X_i - a)^2|_{a=\overline{x},\sigma^2=S^2} = -\dfrac{2n}{S^2} < 0, \\ \dfrac{\partial^2 \ln L(\overline{X}_n, a, \sigma)}{\partial a \partial \sigma} = -\dfrac{2}{\sigma^3}\sum\limits_{i=1}^{n}(X_i - a)|_{a=\overline{x},\sigma^2=S^2} = -\dfrac{2}{\sigma^3}(n\overline{x} - n\overline{x}) = 0. \end{cases}$$

Таким образом, оценками, полученными ММП, параметров нормального распределения являются $\hat{a}_{\text{ММП}} = \overline{x}$, $\hat{\sigma^2}_{\text{ММП}} = S^2$.

2.5 Доверительные интервалы

Если по некоторым причинам невозможно построить выборку достаточного объема, то даже для несмещенных и состоятельных оценок равенство $\widetilde{\theta} \approx \theta$ не может быть выполнено. В таких случаях оценка параметра θ строится в виде интервала.

Определение. Доверительным интервалом надежности γ называется интервал со случайными концами $(\widetilde{\theta}_1, \widetilde{\theta}_2)$, который накрывает неизвестное значение параметра θ с вероятностью, не меньшей γ , т.е.

$$P\{\theta \in (\widetilde{\theta}_1, \widetilde{\theta}_2)\} \geqslant \gamma. \tag{16}$$

Вероятность γ называется также доверительной вероятностью, ее значения обычно выбирают близкими к единице: 0,9; 0,95; 0,99 и т.д.

Точность интервальной оценки существенно зависит от того, что известно о наблюдаемом распределении. Так, если нет никакой информации о типе наблюдаемого распределения, то построенная оценка будет "грубой". Если же известен закон распределения, то доверительный интервал можно построить более точно.

Прежде чем рассматривать построение доверительных интервалов, изучим некоторые важные распределения.

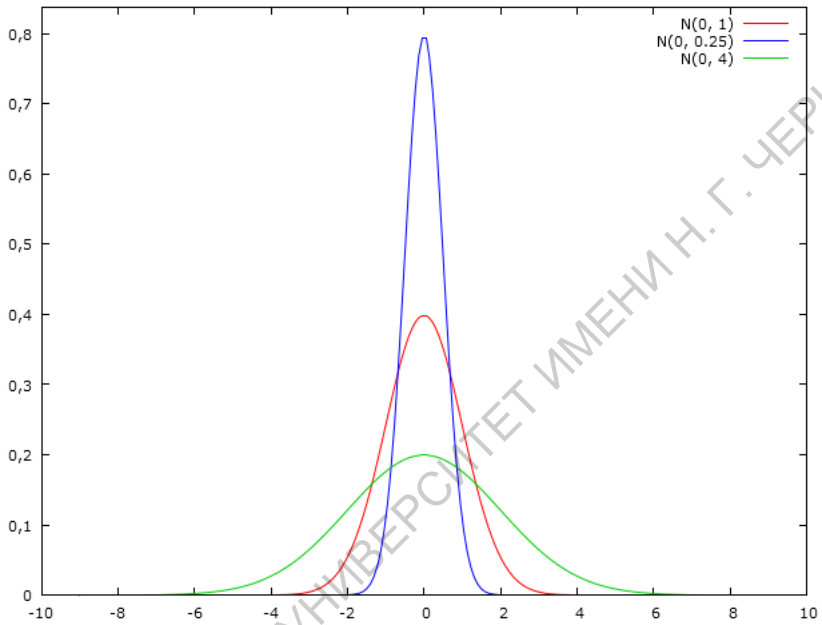
Стандартное нормальное распределение $\mathcal{N}(0, 1)$

Сл. в. $\xi_0 \sim \mathcal{N}(0, 1)$ имеет функцию плотности

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

Числовые характеристики стандартной нормальной величины $M\xi = 0$, $D\xi = \sigma^2 = 1$.

Графики функции плотности $\xi \sim \mathcal{N}(a, \sigma^2)$ представлены на рисунке



Распределение хи-квадрат $\chi^2(n)$

Рассмотрим n независимых одинаково распределенных случайных величин со стандартным нормальным распределением $\xi_1, \xi_2, \ldots, \xi_n$.

Случайной величиной с распределением $\chi^2(n)$ с n степенями свободы называется величина $\chi^2(n) = \xi_1^2 + \xi_2^2 + \cdots + \xi_n^2$.

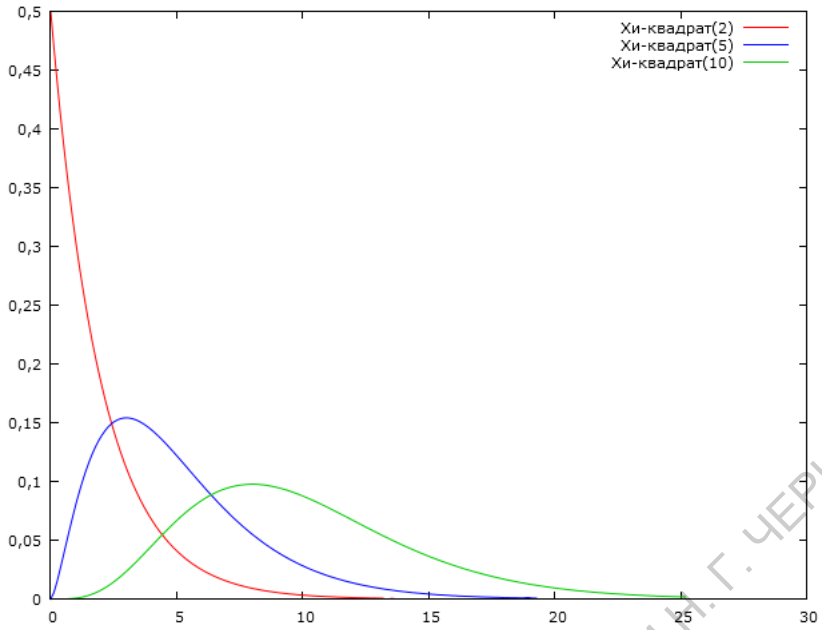
Функция плотности этой сл. в. имеет вид

$$f(x) = \frac{x^{n/2-1} e^{-x/2}}{\Gamma(n/2) 2^{n/2}}, \quad \text{при} \quad x \geqslant 0, f(x) = 0, \quad \text{при} \quad x < 0.$$

Здесь $\Gamma(\lambda) = \int\limits_0^\infty t^{\lambda-1} e^{-t} dt, \quad \lambda > 0,$ —гамма-функция.

Числовые характеристики $M\xi = n$, $D\xi = \sigma^2 = 2n$.

На рисунках приведены графики распределения хи–квадрат при различных значениях степеней свободы.



Распределение Стьюдента

Рассмотрим независимые одинаково распределенные случайных величин со стандартным нормальным распределением $\eta, \xi_1, \xi_2, \ldots, \xi_n$.

Случайной величиной с распределением Стьюдента или T – распределением с n степенями свободы называется величина

$$t(n) = \frac{\eta}{\sqrt{\chi^2(n)/n}} = \frac{\eta}{\sqrt{(\xi_1^2 + \xi_2^2 + \cdots + \xi_n^2)/n}}$$

Функция плотности этой сл. в. имеет вид

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}. \qquad (17)$$

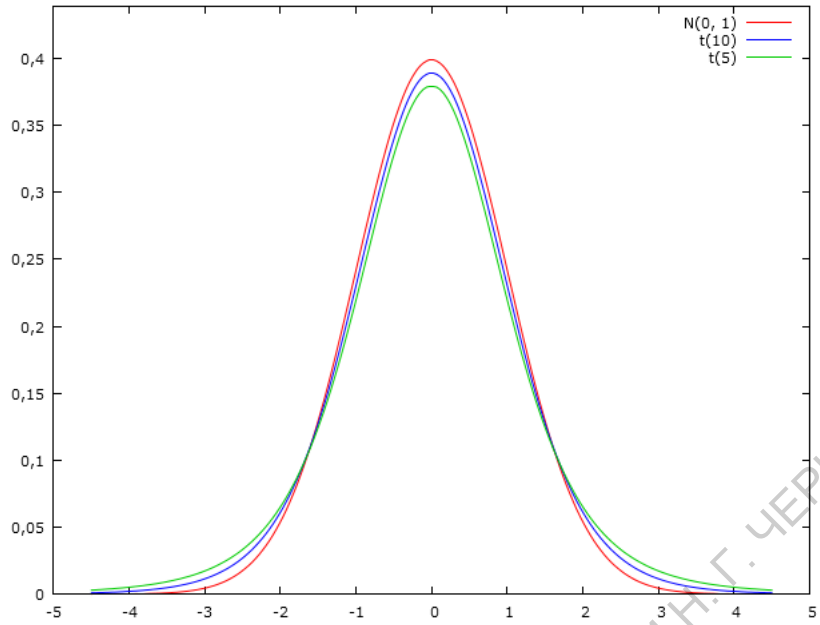
При $n = 1$ функция плотности (17) принимает вид

$$f(x) = \frac{1}{\sqrt{\pi}(1+x^2)},$$

т.е. получаем распределение Коши.

Числовые характеристики при $n > 2$ $M\xi = 0$, $D\xi = \sigma^2 = n/(n-2)$.

На рисунке приведен графики функции плотности распределения Стьюдента для числа степеней свободы $n = 5, 10$, и график функции плотности стандартного нормального распределения.



Лемма Фишера(следствие)

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{N}(a, \sigma)$ выборка из нормального распределения математическим ожиданием a и дисперсией σ^2.

Тогда сл. величины $\overline{x}$ и S^2 независимы, а сл.в

$\frac{\overline{x}-a}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ распределена по стандартному нормальному закону,

$\frac{\overline{x}-a}{\widetilde{\sigma}/\sqrt{n}} \sim T(n-1)$ распределена по закону Стьюдента,

$\frac{\sqrt{n}\,\sigma^2}{\widetilde{\sigma}^2} \sim \chi^2(n-1)$ распределена по закону хи-квадрат.

Пусть известно, что наблюдаемая выборка извлечена из нормального распределения. Рассмотрим построение доверительных интервалов для параметров нормального распределения.

1. Построение доверительного интервала для параметра a нормального распределения (при известном σ)

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{N}(a, \sigma)$ – выборка из нормального распределения, где a – неизвестное математическое ожидание, а σ^2 - известная дисперсия. Пусть задана доверительная вероятность или надежность доверительного интервала γ. Требуется построить интервал $(\widetilde{a}_1, \widetilde{a}_2)$ так, чтобы $\mathsf{P}\{a \in (\widetilde{a}_1, \widetilde{a}_2)\} \geqslant \gamma$.

Доверительный интервал имеет вид

$$\overline{x} - t\frac{\sigma}{\sqrt{n}} < a < \overline{x} + t\frac{\sigma}{\sqrt{n}}, \tag{18}$$

где t - значение аргумента, при котором $2\Phi(t) = \gamma$.

Решение:

По лемме Фишера имеем

$$\frac{\overline{x} - a}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Из курса теории вероятностей известно, что для любой сл.в. ξ,

$$\mathsf{P}\{a \leqslant \xi < b\} = F(b) - F(a),$$

а для величины $\xi_0 \sim \mathcal{N}(0,1)$, справедливо представление $F_{\xi_0}(x) = 0.5 + \Phi(x)$, при любом $x \in \mathbb{R}$, где

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_0^x e^{-t^2/2}\, dt$$

— функция Лапласа.

В силу симметричности нормального распределения относительно математического ожидания будем строить интервал симметрично относительно нуля. Для заданной доверительной вероятности γ найдем такое t, чтобы $\mathsf{P}\{-t < \xi_0 < t\} = \gamma$. Для этого воспользуемся представлением с помощью функции Лапласа:

$$\mathsf{P}\{-t < \xi_0 < t\} = F_{\xi_0}(t) - F_{\xi_0}(-t) = 0.5 + \Phi(t) - 0.5 - \Phi(-t) =$$

$$= \Phi(t) + \Phi(t) = 2\Phi(t) = \gamma.$$

Тогда с использованием найденного t получаем

$$\mathsf{P}\{-t < \xi_0 < t\} = \mathsf{P}\{-t < \frac{\overline{x} - a}{\sigma/\sqrt{n}} < t\} = \mathsf{P}\{\overline{x} - t\frac{\sigma}{\sqrt{n}} < a < \overline{x} + t\frac{\sigma}{\sqrt{n}}\}.$$

Интервал построен.

2. Построение доверительного интервала для параметра a нормального распределения (при неизвестном σ)

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{N}(a, \sigma)$ – выборка из нормального распределения, где a – неизвестное математическое ожидание, а σ^2 - неизвестная дисперсия. Пусть задана доверительная вероятность или надежность доверительного интервала γ. Требуется построить интервал $(\widetilde{a}_1, \widetilde{a}_2)$ так, чтобы $\mathsf{P}\{a \in (\widetilde{a}_1, \widetilde{a}_2)\} \geqslant \gamma$.

Решение.

По лемме Фишера имеем

$$\frac{\overline{x} - a}{\widetilde{\sigma}/\sqrt{n}} \sim T(n-1).$$

В силу симметричности распределения Стьюдента относительно математического ожидания также как и в предыдущем случае будем строить интервал симметричным относительно нуля. Воспользуемся функцией плотности распределения Стьюдента (17) и для заданной доверительной вероятности γ найдем такое t_γ, чтобы

$$\mathsf{P}\{-t_\gamma < t_n < t_\gamma\} = \int\limits_{-t_\gamma}^{t_\gamma} f_{t_n}(x)\, dx = \gamma.$$

Тогда для найденного t_γ доверительный интервал надежности γ имеет вид

$$\overline{x} - t_\gamma \frac{\widetilde{\sigma}}{\sqrt{n}} < a < \overline{x} + t_\gamma \frac{\widetilde{\sigma}}{\sqrt{n}}.$$

2.6 Проверка статистических гипотез.

Статистической гипотезой называется предположение о виде наблюдаемого неизвестного распределения либо о неизвестных параметрах известного распределения.

Выдвинутую гипотезу называют основной или нулевой гипотезой и обозначают H_0 . Предположения, противоречащие нулевой гипотезе, включают в конкурирующую или противоположную гипотезу и обозначают H_1 .

Различают простые гипотезы, состоящие из одного предположения, и сложные гипотезы. Примеры простых гипотез: H_0 : "$M\xi = 0$", H_0 : "$\xi \sim \mathcal{N}(0,1)$" . Примеры сложных гипотез: H_1 : "$M\xi > 0$" , H_1 : "$\xi \sim \mathcal{N}(a, \sigma^2)$" .

Критерием проверки гипотез называют правило, согласно которому выдвинутую гипотезу принимают или отвергают.

Критерием назовем случайную величину $K = f(X_1, X_2, \ldots, X_n)$ — функцию от элементов выборки, закон распределения которой известен, и которую используют для проверки гипотез.

Прежде чем переходить к построению критериев, введем некоторые понятия.

Очевидно, что выдвинутая гипотеза H_0 может быть как ложной, так и истинной. Если гипотеза H_0 истинна, и ее приняли, либо гипотеза H_0 ложная и ее отвергли, то ошибки не произошло. В оставшихся случаях возникает одна из двух ошибок.

Ошибка первого рода при принятии гипотезы заключается в том, что нулевая гипотеза H_0 была истинна, но ее отвергли. Вероятность ошибки первого рода обозначают α , и называют **уровнем значимости** критерия.

Итак, уровень значимости критерия означает вероятность отвергнуть верную нулевую гипотезу.

Уровень значимости выбирают близким к нулю, как правило 0.05 или 0.01 .

Ошибка второго рода при принятии гипотезы состоит в том, что нулевая гипотеза H_0 была ложной, но ее приняли.

Вероятность ошибки второго рода обозначается β.

Мощностью критерия называют величину $1-\beta$, то есть вероятность отвергнуть ложную нулевую гипотезу.

Рассмотрим выборочное пространство $\mathbf{X_n} = \{\overline{\mathbf{X}}_\mathbf{n} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})\}$ — множество выборок объема n из распределения наблюдаемой случайной величины. Разобъём множество $\mathbf{X_n}$ на два подмножества W и $\overline{W}$, $W \cup \overline{W} = \mathbf{X_n}$.

Если наблюдалась выборка $\overline{X}_n = (X_1, X_2, \ldots, X_n) \in W$, то гипотезу H_0 отвергаем, если наблюдалась выборка $\overline{X}_n = (X_1, X_2, \ldots, X_n) \in \overline{W}$, гипотезу H_0 принимаем. Множество W назовем критическим множеством, множество $\overline{W}$ — областью принятия гипотезы.

Основная задача при решении задачи принятия гипотезы сводится к тому, чтобы построить критическую область. Другими словами, требуется указать правило, согласно которому наблюдавшиеся выборки относят к критическому множеству.

Решается эта задача с использованием понятия ошибки первого рода.

Обозначим α вероятность ошибки первого рода, т.е. события

$$\{\overline{X}_n = (X_1, X_2, \ldots, X_n) \in W | H_0\}$$

— наблюдалась выборка из критической области, в то время как была верна нулевая гипотеза H_0.

Определение. Уровнем значимости критерия называется

$$\alpha = \mathsf{P}\{\overline{X}_n = (X_1, X_2, \ldots, X_n) \in W | H_0\}.$$

Определение. Мощностью критерия называется

$$1-\beta = \mathsf{P}\{\overline{X}_n = (X_1, X_2, \ldots, X_n) \in W | H_1\}.$$

Критерий $K = f(X_1, X_2, \ldots, X_n)$ как функция отображает выборочное пространство $\mathbf{X_n}$ на множество действительных чисел $\mathbb{R}$. Тогда при условии монотонности функции $f(X_1, X_2, \ldots, X_n)$ критическая область W и область принятия гипотезы $\overline{W}$ отображаются в непересекающиеся множества на оси $\mathbb{R}$ и имеют границу, которую обозначим k_α. Тогда критерий как правило принятия решения может иметь вид (левосторонняя критическая область): если $K_{\text{набл}} = f(x_1, x_2, \ldots, x_n) < k_\alpha$ гипотезу H_0 следует отвергнуть, иначе —принять.

Критерий Неймана-Пирсона

Пусть известно наблюдаемое распределение, но неизвестны его параметры. Пусть относительно параметра распределения выдвинуты две простые гипотезы $H_0 : "\theta = \theta_0"$ и $H_1 : "\theta = \theta_1"$.

Построим критерий для проверки истинности простой гипотезы H_0 против простой гипотезы H_1.

Как и ранее, $P(x, \theta)$ —закон распределения наблюдаемой случайной величины, $\overline{X}_n = (X_1, X_2, \ldots, X_n)$ –выборка.

Функцией правдоподобия называется функция

$$L(\overline{X}_n, \theta) = \prod_{i=1}^{n} P(X_i, \theta).$$

Отношением правдоподобия называется функция

$$\varphi(\overline{X}_n) = \frac{L(\overline{X}_n, \theta_1)}{L(\overline{X}_n, \theta_0)} = \frac{\prod\limits_{i=1}^{n} P(X_i, \theta_1)}{\prod\limits_{i=1}^{n} P(X_i, \theta_0)}.$$

Критерий Неймана-Пирсона состоит в том, что за границу критической области принимается такое значение C_φ, при котором

$$\mathsf{P}\{\varphi(\overline{X}_n) \geqslant C_\varphi = \alpha\}.$$

Правило принятия решения имеет вид:

если наблюдаемое значение $\varphi(\overline{x}_n) < C_\varphi$, принимают гипотезу H_0, если $\varphi(\overline{x}_n) \geqslant C_\varphi$ принимают H_1 (отвергают H_0). Здесь $\overline{x}_n = (x_1, x_2, \ldots, x_n)$ – наблюдавшаяся выборка.

Пример Построение оптимального критерия Неймана-Пирсона для проверки двух простых гипотез $H_0 : "a = a_0"$ против $H_1 : "a = a_1"$, $a_0 < a_1$, относительно параметра a нормального распределения. Дисперсия σ^2 известна.

Решение. Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{N}(a, \sigma^2)$.

Функция правдоподобия имеет вид:

$$L(\overline{X}_n, a) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-a)^2/2\sigma^2} = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - a)^2\Big).$$

Составим отношение правдоподобия

$$\varphi(\overline{X}_n) = \frac{L(\overline{X}_n, a_1)}{L(\overline{X}_n, a_0)} = \exp\Big(\frac{a_1 - a_0}{\sigma^2} n\overline{X} - \frac{n(a_1^2 - a_0^2)}{2\sigma^2}\Big).$$

Согласно критерию Неймана-Пирсона граница критической области должна удовлетворять неравенству $\varphi(\overline{x}_n) \geqslant C_\varphi$.

Подставим найденное представление отношения правдоподобия и разрешим неравенство относительно сл.величины $n\overline{X} = \sum\limits_{i=1}^{n} X_i$:

$$\sum_{i=1}^{n} X_i \geqslant \frac{\sigma^2}{a_1 - a_0}\Big(\ln C_\varphi - \frac{n(a_1^2 - a_0^2)}{2\sigma^2}\Big) = C.$$

Рассматривая в качестве критерия случайную величину $n\overline{X} = \sum\limits_{i=1}^{n} X_i$ приходим к выводу, что она имеет распределение $\mathcal{N}(na, n\sigma^2)$. В предположении, что верна нулевая гипотеза $H_0 : "a = a_0"$ имеем

$$\frac{\sum\limits_{i=1}^{n} X_i - na_0}{n\sigma^2} \sim \mathcal{N}(0, 1),$$

а, следовательно, для заданного уровня значимости α можно определить

$$\mathsf{P}\{\sum_{i=1}^{n} X_i \geqslant C\} = \mathsf{P}\Big\{\frac{\sum\limits_{i=1}^{n} X_i - na_0}{n\sigma^2} \geqslant \frac{C - na_0}{n\sigma^2}\Big\} =$$

$$= 1 - F_{\xi_0}\Big(\frac{C - na_0}{n\sigma^2}\Big) = 1 - F_{\xi_0}(u_{1-\alpha}) = \alpha.$$

Таким образом, из соотношения

$$\frac{C - na_0}{n\sigma^2} = u_{1-\alpha}$$

определяем границу критической области в виде

$$C = na_0 + u_{1-\alpha}\sigma\sqrt{n}.$$

Проверка сложных гипотез

Рассмотрим проверку сложных гипотез относительно параметров распределения.

Сложными гипотезами называются гипотезы вида $H_0 : "\theta \in \Theta_0"$, $\Theta_0 \in \mathbb{R}$, $H_0 : "a > a_0"$, $H_0 : "a \neq a_0"$, и т.п.

Пусть нулевая гипотеза имеет вид $H_0 : "\theta \in \Theta_0"$, а конкурирующая $H_1 : "\theta \in \Theta_1"$, где два $\Theta_0 \cap \Theta_1 = \emptyset$ – непересекающиеся множества из $\mathbb{R}$.

Тогда вероятности ошибок первого и второго рода можно рассматривать как заданные на соответствующих множествах функции:

$$\alpha(\theta) = \mathsf{P}\{\overline{X}_n = (X_1, X_2, \ldots, X_n) \in W | \theta \in \Theta_0\};$$

$$\beta(\theta) = \mathsf{P}\{\overline{X}_n = (X_1, X_2, \ldots, X_n) \in \overline{W} | \theta \in \Theta_1\}.$$

Определение. Размером критерия называют величину $\alpha = \max\limits_{\theta \in \Theta_0} \alpha(\theta)$.

Определение. Функцией мощности критерия называют вероятность отвергнуть нулевую гипотезу при любых возможных значения параметра:

$$m(\theta) = \mathsf{P}\{\overline{X}_n = (X_1, X_2, \ldots, X_n) \in W | \theta \in \Theta\}$$

Таким образом, на множестве Θ_0 $m(\theta) = \alpha(\theta)$, на множестве Θ_1 $m(\theta) = 1 - \beta(\theta)$.

Пусть α – некоторый фиксированный размер критерия. Предположим, что критерий максимизирует функцию мощности по всем возможным критериям на множестве Θ_1. Тогда этот критерий называют **равномерно наиболее мощным**.

То есть равномерно наиболее мощный критерий имеет наименьшую вероятность ошибки второго рода при фиксированной вероятности ошибки первого рода.

Равномерно наиболее мощные критерии можно построить только в некоторых частных случаях для проверки простых гипотез вида $"\theta = \theta_0"$.

Пример построения равномерно наиболее мощного критерия.

Пусть $X_1, X_2, \ldots, X_n \sim \mathcal{N}(a, \sigma^2)$, $H_0 : "a = a_0"$, $H_1 : "a > a_0"$, σ — известно. Воспользуемся результатом построения критической области для проверки гипотезы $H_0 : "a = a_0"$ против гипотезы $H_1 : "a = a_1"$, $a_1 > a_0$ с помощью критерия Неймана-Пирсона.

Было получено, что критическая область задается неравенством

$$W = \left\{ \overline{X_n} : \sum_{i=1}^{n} X_i \geq C_\varphi = na_0 + u_{1-\alpha}\sigma\sqrt{n} \right\}. \tag{19}$$

Здесь $u_{1-\alpha}$ значение аргумента функции распределения $F_{\xi_0}(x)$ при котором достигается вероятность $1 - \alpha$, $F(u_{1-\alpha}) = 1 - \alpha$. Величина $u_{1-\alpha}$ квантиль распределения уровня $1 - \alpha$.

Замечая, что область $a > a_0$ состоит из величин, удовлетворяющих условию $a_1 > a_0$ гипотезы $H_1 : "a = a_1"$, а также, что C_φ в (*) не зависит от a_1, приходим к выводу, что при заданном уровне значимости α для всех a из области $a > a_0$ критерий задается условием (19), а значит, является равномерно наиболее мощным.

Критерии согласия

Предыдущие критерии предназначались для проверки гипотез относительно параметров некоторого известного распределения, т.е. была известна

функция распределения наблюдаемой сл.в., но не были известны значения параметров распределения. Этот факт влиял на построение критерия (такие критерии называют также параметрическими).

Пусть теперь нет информации о типе наблюдаемого распределения, т.е. неизвестна функция распределения. Тогда по-прежнему можно выдвигать гипотезы относительно числовых характеристик распределения, но для проверки использовать другие методы (непараметрической статистики). Более распространенной в этом случае является задача проверки гипотезы о типе наблюдаемого распределения, для решения которой используют критерии согласия.

Пусть $X_1, X_2, \ldots, X_n$ – наблюдаемая выборка, и пусть на основании некоторых статистических и вероятностных фактов выдвинута гипотеза $H_0 : \xi \sim \mathcal{L}(x, \theta)$, или H_0 : "наблюдаемая сл.в. имеет распределение с ф.р. $F_0(x)$" , или "выдвинутая гипотеза согласуется с наблюдаемыми данными".

Рассмотрим два основных критерия проверки таких гипотез. Критерий Колмогорова основан на анализе различий между предполагаемой, гипотетической функцией распределения $F_0(x)$ и наблюдаемой, эмпирической функцией распределения $\widetilde{F_n}(x)$. Критерий Пирсона (критерий χ^2) основан на анализе различий между предполагаемой, гипотетической функцией плотности распределения $f_0(x)$ и наблюдаемой, эмпирической функцией плотности распределения $\widetilde{f}_n(x)$, которые задаются соответственно теоретическими и наблюдаемыми частотами.

Критерий Колмогорова.

Проверяется простая гипотеза H_0 : "$F_\xi(x) = F_0(x)$" против сложной гипотезы H_0 : "$F_\xi(x) \neq F_0(x)$", α –заданный уровень значимости.

Для проверки гипотезы используется сл.в.(статистика) $D(\overline{X}_n)$, наблюдаемые значения которой вычисляются

$$D_{\text{набл}} = D(\overline{x}_n) = \sup_{x \in \mathbb{R}} |\widetilde{F_n}(x) - F_0(x)|.$$

Если $D_{\text{набл}} >= D_{\text{крит}}$, гипотеза H_0 отклоняется, иначе, если $D_{\text{набл}} < D_{\text{крит}}$ гипотеза H_0 принимается, т.е. наблюдаемые значения согласуются с выдвинутой гипотезой. Здесь $D_{\text{крит}} = D_{1-\alpha}$ —квантиль уровня $1-\alpha$ распределения величины $D(\overline{X}_n)$.

Определение. Квантилью уровня α, $0 < \alpha < 1$ распределения $F_\xi(x)$ называется число u_α, такое что $F_\xi(u_\alpha) = \mathsf{P}\{\xi < u_\alpha\} = \alpha$.

Критерий Колмогорова А.Н. основан на предельном соотношении

$$\lim_{n\to\infty} \mathsf{P}\{\sqrt{n}D(\overline{X}_n) < x\} = F_K(x), \quad x > 0. \tag{20}$$

Здесь $F_K(x)$ – функция распределения сл.в K с распределением Колмогорова

$$F_K(x) = \begin{cases} \sum\limits_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}, & x >= 0; \\ 0, & x < 0. \end{cases} \tag{21}$$

Критерий Пирсона.

Проверяется простая гипотеза H_0 : "$F_\xi(x) = F_0(x)$" против сложной гипотезы H_0 : "$F_\xi(x) \neq F_0(x)$", α — заданный уровень значимости.

Для проверки гипотезы используется сл.в.(статистика) χ^2, наблюдаемые значения которой вычисляются следующим образом:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'}, \tag{22}$$

где m – количество групп в вариационном ряду, n_i – наблюдаемые частоты, n_i' –теоретические частоты, вычисленные в соответствии с выдвинутой гипотезой. Критическое значение $\chi^2_{\text{крит}} = \chi^2(1-\alpha, m-r-1)$ –квантиль уровня $1-\alpha$ распределения сл. в. χ^2 с $m-r-1$ степенями свободы. Здесь r – число параметров гипотетического распределения, оцениваемых по выборке. (Замечание. При использовании таблиц распределения часто пишут для простоты $\chi^2_{\text{крит}} = \chi^2(\alpha, m-r-1)$)

Если $\chi^2_{\text{набл}} < \chi^2_{\text{крит}}$, то гипотезу о согласии наблюдаемого распределения с гипотетическим принимают, в противном случае гипотезу H_0 отвергают.

2.7 Анализ связи двух величин. Парная линейная регрессионная модель.

Пусть производятся наблюдения некоторого процесса, который описывается двумя случайными величинами (ξ, η). Тогда результаты наблюдений представляют собой двумерную выборку $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Выборочные числовые характеристики в этом случае имеют вид.

Выборочные средние

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Выборочные дисперсии (смещенные)

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2, \quad S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{y})^2.$$

Величины $S_x = \sqrt{S_x^2}$ и $S_y = \sqrt{S_y^2}$ представляют собой соответственно выборочные средние квадратические отклонения (смещенные) величин ξ и η.

Выборочное среднее (совместное) есть величина

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} X_i \cdot Y_i. \tag{23}$$

Коэффициент корреляции (выборочный)

$$r_{\text{выб}} = r = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{S_x^2}\sqrt{S_y^2}}$$

(Для удобства далее опускаем пометку выб.)

Как известно, если коэффициент корреляции $r = Cov(\xi, \eta)/\sqrt{D\xi D\eta}$ равен нулю, то величины ξ и η некоррелированы, если $|r| = 1$, то величины ξ и η линейно связаны $\eta = a\xi + b$.

Следовательно, при изучении связи величин необходимо оценить отличие выборочного коэффициента корреляции от нуля. Проверим гипотезу о

значимости коэффициента корреляции, т.е. H_0 : "$r = 0$" против гипотезы H_1 : "$r \neq 0$" .

Статистика

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$$

имеет распределение Стьюдента с $n-2$ степенями свободы. Для заданного уровня значимости α критическое значение равно $t_{\text{кр}} = t(\frac{\alpha}{2}; n-2)$.

Если $t_{\text{набл}} < t_{\text{кр}}$, принимают гипотезу H_0 : "$r = 0$" , в противном случае H_0 отвергают, и делают вывод о значимом отличии коэффициента корреляции от нуля.

Если установлено, что коэффициент значимо отличается от нуля, то линейная функция должна достаточно близко описывать имеющуюся между величинами, но неизвестную связь.

Построение уравнения парной линейной регрессии.

Пусть наблюдалась выборка $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, и выборочный коэффициент значимо отличается от нуля. Найдем коэффициенты уравнения $Y = aX + b$, которое наилучшим образом аппроксимирует $Y = f(X)$ ($\eta = f(\xi)$).

Величина Y называется зависимой переменной, признаком, величина Y называется независимой переменной, фактором, регрессором.

В парной регрессионной модели зависимая переменная зависит только от одного регрессора. Оценим коэффициенты уравнения методом наименьших квадратов (МНК). Будем обозначать $\hat{Y}$ – вычисленные (прогнозные) значения.

Согласно МНК требуется найти такие значения оценок параметров $\hat{a}$ и $\hat{b}$, чтобы была минимальной сумма квадратов отклонений прогнозных значений от наблюдаемых:

$$L(\hat{a}, \hat{b}) = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (Y_t - (\hat{a} + \hat{b}X_t))^2 \to min.$$

Значит, для нахождения оценки параметров парной регрессионной модели МНК необходимо найти экстремум (минимум) функции двух аргументов.

Запишем необходимые условия экстремума:

$$\begin{cases} \frac{\partial L}{\partial \hat{a}} = -2\sum\limits_{t=1}^{n}(Y_t - \hat{a} - \hat{b}X_t) = 0, \\ \frac{\partial L}{\partial \hat{b}} = -2\sum\limits_{t=1}^{n} X_t(Y_t - \hat{a} - \hat{b}X_t) = 0, \end{cases}$$

Раскрывая скобки получим:

$$\begin{cases} \sum\limits_{t=1}^{n} Y_t - \hat{a}n - \hat{b}\sum\limits_{t=1}^{n} X_t = 0, \\ \sum\limits_{t=1}^{n} X_t Y_t - \hat{a}\sum\limits_{t=1}^{n} X_t - \hat{b}\sum\limits_{t=1}^{n} X_t^2 = 0. \end{cases} \tag{24}$$

Из первого уравнения системы имеем оценку параметра a:

$$\hat{a} = \frac{1}{n}\sum_{t=1}^{n} Y_t - \hat{b}\frac{1}{n}\sum_{t=1}^{n} X_t = \bar{y} - \hat{b}\bar{x}. \tag{25}$$

Преобразуем второе уравнение системы и подставим полученную оценку $\hat{a}$

$$\sum_{t=1}^{n} X_t Y_t - \hat{a}\sum_{t=1}^{n} X_t - \hat{b}\sum_{t=1}^{n} X_t^2 = n\overline{xy} - \hat{a}\bar{x} - n\hat{b}(S_x^2 + \bar{x}^2) = 0;$$

$$\overline{xy} - \bar{x}\bar{y} + \hat{b}\bar{x}^2 - \hat{b}S_x^2 - \hat{b}\bar{x}^2 = 0.$$

Отсюда получаем оценку параметра b:

$$\hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x S_y}\frac{S_y}{S_x} = r\frac{S_y}{S_x}, \tag{26}$$

Таким образом решение системы уравнений имеет вид

$$\begin{cases} \hat{b} = r\frac{S_y}{S_x}; \\ \hat{a} = \bar{y} - \hat{b}\bar{x}. \end{cases}$$

Уравнение регрессии Y на X имеет вид:

$$\bar{y}_x = \bar{y} + r\frac{S_y}{S_x}(x - \bar{x}).$$

Уравнение регрессии X на Y имеет вид

$$\bar{x}_y = \bar{x} + 1/r\frac{S_x}{S_y}(y - \bar{y}).$$

Заметим, что каждое из уравнений имеет $\bar{x}$ своим решением , т.е графики проходят через точку $(\bar{x}, \bar{y})$ –центр совместного распределения величин X и Y .

Теорема Гаусса-Маркова.

Пусть Y — сл.величина, значения которой подлежат прогнозу, X — неслучайная величина, фактор, от которого зависит величина Y , ε —сл. в., которая описывает все случайные воздействия и ошибки наблюдений.

Тогда общий вид спецификации парной линейной регрессионной модели:

$$Y = a + bX + \varepsilon, \tag{27}$$

где a и b – параметры модели (постоянные неизвестные коэффициенты);
X – объясняющая (независимая) переменная – регрессор (детерминированная величина);
Y – объясняемая (зависимая) переменная – отклик (случайная величина);
ε – случайное возмущение (случайная величина), которое характеризует отклонение от уравнения регрессии $f(X) = a + bX$ (теоретической линейной зависимости).

Случайные ошибки возникают по двум основным причинам: из-за ошибок выбора модели, когда истинная зависимость выражается другой функцией, из-за ошибок измерений.

Уравнения для отдельных наблюдений зависимой переменной Y имеют следующий вид (схема Гаусса-Маркова):

$$Y_t = a + bX_t + \varepsilon_t,$$

где $Y_t, X_t,\ t = 1, \ldots, n$ – выборочные данные (наблюдения);
n – объем выборки (количество наблюдений).

В регрессионных моделях ошибки $\varepsilon_t,\ t = 1, \ldots, n$ должны удовлетворять условиям Гаусса-Маркова:

1. Математическое ожидание случайных возмущений равно нулю

$$M\{\varepsilon_t\} = 0, \quad t = 1, \ldots, n \tag{28}$$

2. Дисперсия возмущений не зависит от номера (момента) наблюдений t

$$Var\{\varepsilon_t\} = const = \sigma^2 \tag{29}$$

3. Некоррелированность возмущений для различных наблюдений:

$$cov\{\varepsilon_t, \varepsilon_s\} = 0 \text{ при } t \neq s, \tag{30}$$

что означает отсутствие систематической связи между значениями случайного члена в любых наблюдениях t и s. Нарушение этого условия называется автокорреляцией возмущений.

Модель (27) при выполнении условий Гаусса-Маркова (28)– (30) является классической регрессионной моделью. Если возмущения имеют совместное нормальное распределение:

$$\varepsilon_t \sim N(0, \sigma^2),$$

то модель называется классической нормальной регрессионной моделью.

Теорема Гаусса-Маркова.

В предположениях модели (27)– (30) оценки , полученные МНК имеют наименьшую дисперсию в классе всех линейных несмещенных оценок (т.е. являются эффективными).

Анализ вариации зависимой переменной в регрессии

Оценим дисперсию ошибок σ^2. По теореме Гаусса— Маркова имеем «наилучшие» оценки коэффициентов регрессии a, b. Оценим дисперсию ошибок σ^2.

Обозначим через $\hat{Y} = \hat{a} + \hat{b}X_t$ прогноз значения Y_t в точке X_t. Остатки регрессии e_t определяются из уравнения $Y_t = \hat{Y}_t + e_t = \hat{a} + \hat{b}X_t + e_t$. Остатки e_t, так же как и ошибки ε_t, являются случайными величинами, однако остатки, в отличие от ошибок, наблюдаемые.

Оценка σ^2 связана с суммой квадратов остатков регрессии $e_t = Y_t - \hat{a} - \hat{b}X_t$. В самом деле,

$$\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n}(Y_t - \hat{a} - \hat{b}X_t)^2 = \sum_{t=1}^{n}(\overline{Y} + y_t - \hat{a} - \hat{b}\overline{X} - \hat{b}x_t)^2 =$$

$$= \sum_{t=1}^{n}(y_t - \hat{b}x_t)^2 = \sum_{t=1}^{n}(\overline{Y} + y_t - \hat{a} - \hat{b}\overline{X} - \hat{b}x_t)^2 =$$

$$= \sum_{t=1}^{n}((b - \hat{b})x_t + (\varepsilon_t - \overline{\varepsilon}))^2 =$$

$$= x_t^2(b - \hat{b})^2 + 2(b - \hat{b})\sum_{t=1}^{n} x_t(\varepsilon_t - \overline{\varepsilon}) + \sum_{t=1}^{n}(\varepsilon_t - \overline{\varepsilon})^2 =$$

$$= I + II + III$$

Вычислим математическое ожидание $M\sum\limits_{t=1}^{n} e_t^2 = M(I) + M(II) + M(III)$.

$$M(I) = M(\sum_{t=1}^{n} x_t^2(b - \hat{b})^2) = \sum_{t=1}^{n} x_t^2 V(\hat{b}) = \sum_{t=1}^{n} x_t^2 \frac{\sigma^2}{\sum\limits_{k=1}^{n} x_k^2} = \sigma^2.$$

Введем обозначение

$$\omega_t = \frac{x_t^2}{\sum\limits_{k=1}^{n} x_k^2}.$$

Используя соотношение

$$\hat{b} = \sum_{t=1}^{n} \omega_t y_t = \sum_{t=1}^{n} \omega_t(bx_t + \varepsilon_t - \overline{\varepsilon}) = b + \sum_{t=1}^{n} \omega_t \varepsilon_t$$

, получаем

$$M(II) = -2M(\sum_{t=1}^{n} t\omega_t\varepsilon_t \sum_{t=1}^{n} x_s(\varepsilon_s - \overline{\varepsilon})) =$$

$$= -2M(\sum_{t=1}^{n} t, s\omega_t x_s \varepsilon_t \varepsilon_s - \sum_{t=1}^{n} t\omega_t \varepsilon_t \overline{\varepsilon} \sum_{t=1}^{n} s x_s) =$$

$$= -2\sum_{t=1}^{n} t\omega_t x_t \sigma^2 = -2\sigma^2,$$

$$M(III) = M(\sum_{t=1}^{n} \varepsilon_t^2 - 2\overline{\varepsilon} \sum_{t=1}^{n} \varepsilon_t + n\overline{\varepsilon^2}) =$$

$$= n\sigma^2 - 2n\frac{1}{n}\sigma^2 + n\frac{1}{n}\sigma^2 =$$

$$= (n-1)\sigma^2.$$

Таким образом, $M\sum\limits_{t=1}^{n} e_t^2 = M(I) + M(II) + M(III) = \sigma^2 - 2\sigma^2 + (n-1)\sigma^2 = (n-2)\sigma^2$.

Отсюда следует, что $s^2 = \hat{\sigma^2} = \frac{1}{n-2}\sum\limits_{t=1}^{n} e_t^2$ является несмещенной оценкой дисперсии ошибок σ^2 .

Замечание. О статистических свойствах МНК-оценок параметров регрессии.

Пусть выполняется условие нормальной линейной регрессионной модели $\varepsilon \sim N(0, \sigma^2 I_n)$, т. е. ε — многомерная нормально распределенная случайная величина. Другими словами Y_t имеют совместное нормальное распределение.

Тогда МНК-оценки коэффициентов регрессии a , b также имеют совместное нормальное распределение, так как они являются линейными функциями от Y_t:

$$\hat{a} \sim N\Big(a, \sigma^2 \frac{\sum X_t^2}{n\sum x_t^2}\Big),$$

$$\hat{b} \sim N\Big(b, \sigma^2 \frac{1}{\sum x_t^2}\Big).$$

Список литературы

[1] Гихман И.И., Скороход А.В., Ядренко М.И. Теория вероятностей и математическая статистика. — Киев: Издательское объединение "Вища школа"; 1979. — 408 с.

[2] Математическая статистика: Учеб. для вузов / В.Б. Горяинов, И.В. Павлов, Г.М. Цветкова и др.; Под ред. В.С. Зарубина, А.П. Крищенко. - М.: Изд-во МГТУ им. Н.Э. Баумана, 2001. – 424 с.

[3] Боровков А.А. Математическая статистика. — Новосибирск: Наука; Издательство института математики, 1997. — 772 с.

[4] Лагутин М.Б. Наглядная математическая статистика: Учебное пособие / М.Б. Лагутин.—М.Ж БИНОМ. Лаборатория знаний, 2007.– 472 с.: ил.

[5] Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. — 3-е изд., перераб. и доп. — М.: Дело, 2000. —400 с.