Министерство образования и науки Российской Федерации
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

В. А. Труфанов

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие

Благовещенск
Издательство АмГУ
2010

ББК 22.172
Т80

Рекомендовано
учебно-методическим советом университета

Рецензент:

Кушнирук Н.Н., ст. препод. каф. «Математический анализ и моделирование» , канд. физ.-мат. наук.

Труфанов, В. А.
Т80 Математическая статистика: учебно-методическое пособие / В. А. Труфанов. — Благовещенск: Изд-во АмГУ, 2010. — 32 с.

Настоящее пособие написано в соответствии с программой курсов «Математическая статистика» для судентов специальности 010501 «Прикладная математика и информатика», 010101 «Математика».

Предполагается использование пособия для практических занятий и самостоятельного решения задач по математической статистике. Кроме задач в него включены необходимые теоретические сведения, методические указания и решения наиболее типичных задач.

Тематика занятий и их количество (8 занятий) соответствует рабочей программе курса. Все задачи занятий разбиты на группы. В группу А входят задачи для аудиторной работы, в группу С — задачи для самостоятельного решения, в группу Д — дополнительные задачи и задачи повышенной трудности.

Пособие может быть использовано студентами для овладения техникой решения задач, преподавателями — в качестве задачника по данному курсу.

ББК 22.172

Основные обозначения и сокращения

с. в. — случайная величина или случайные величины;

$\vec{X} = \vec{X}_n = (X_1, \ldots, X_n)$ — выборка, состоящая из независимых одинаково распределенных *с.в.* {наблюдений} $X_i,\ i = 1, \ldots, n$;

$\vec{X}$ — выборочный вектор;

распределение $\mathcal{F}$ — закон распределения общего вида, который характеризуется функцией распределения, плотностью или таблицей, набором числовых характеристик – $\mathrm{E}\xi,\ \mathrm{D}\xi$ и т. д.;

$m_\xi = \mathrm{E}\xi$ — математическое ожидание (среднее) *с.в.* ξ;

$\sigma^2 = \sigma_\xi^2 = \mathrm{D}\xi$ — дисперсия *с.в.* ξ;

$X_i \in \mathcal{F}$ — *с.в.* X_i имеет распределение $\mathcal{F}$, что может означать, например, $\mathrm{P}\{X_i < y\} = F(y)$ при всех y , если F – функция распределения;

$\mathbf{I}(A)$ — индикатор события A и является случайной функцией вида

$$\mathbf{I}(A) = \begin{cases} 1, & \text{если произошло событие } A, \\ 0, & \text{иначе;} \end{cases}$$

■ — символ для обозначения окончания решения задачи.

Введение

Математическая статистика, как и теория вероятностей, занимается изучением математических моделей случайных явлений и экспериментов. В то же время задачи математической статистики являются как бы обратными к задачам теории вероятностей. В теории вероятностей после задания математической модели того или иного случайного явления требуется рассчитать различные вероятностные характеристики этого явления в рамках данной модели. В математической статистике, исходя из уже имеющихся результатов эксперимента, называемые статистическими данными и состовляющие выборку, требуется определить или найти те или иные параметры, ее определяющие. То есть устанавливает определенные практические выводы на основе изучения этих данных.

Таким образом, изучение закономерностей количественного и качественного характера определенных признаков в массовых явлениях составляет основную цель математической статистики.

Математической статистика — абстрактная наука. Она изучает математическую сторону работы с числовыми данными независимо от конкретной отраслевой специфики. Ее методы применимы для обработки данных наблюдений и экспериментов любой природы, поэтому используются во всех конкретных естественных и гумантарных науках, экономике, технике, медицине и т. д., т. е. во всех отраслевых статистиках.

В математической статистике рассматриваются следующие основные задачи.

1. Оценки неизвестных параметров по результатам эксперимента (выборки). При этом требуется найти такую функцию (статистику) от результатов эксперимента, которая могла бы служить «достаточно хорошей» оценкой неизвестного истинного (теоретического) значения параметра.

2. Интервальное оценивание или, другими словами, оценивание па-

раметров с помощью доверительных интервалов. При этом требуется построить интервал со случайными границами, зависящими от результатов выборки таким образом, чтобы построенный случайный интервал накрывал неизвестное истинное значение параметра с заданной достаточно высокой вероятностью.

3. Проверка статистических гипотез о свойствах изучаемого случайного явления. В этом случае требуется на основе выборки проверить то или иное предположение или гипотезу.

При решении указанных задач используются типичные схемы исследований, которые деляться на две части.

Сначала путем наблюдений и экспериментов собираются, регистрируются статистические данные (выборка). Затем они упорядочиваются, представляются в компактной, наглядной или функциональной форме. Вычисляются различного рода средние величины, характеризующие выборку. Эта часть математической статистики называется *описательной статистикой.*

Вторая часть исследования состоит в получении на основе найденных сведений о выборке достаточно обоснованных выводов о свойствах исследуемого случайного явления. Эта часть исследования обеспечивается статистическими методами, состовляющими *статистику выводов.*

Занятие 1

Тема. **Эмпирическое распределение. Выборка и вариационный ряд.**

Цель. Ознакомиться с выборочными характеристиками, выраженные через экспериментальные данные, с приемами нахождения и построения эмпирической функции распределения.

Исходными статистическими данными являются совокупности n независимых наблюдений в случайном эксперименте, связанном со случайной величиной ξ — реализации статистической выборки $\vec{X} = (X_1, \dots, X_n)$. Совокупность значений $X_1, \dots, X_n$ интерпретируются как значения, принятые n независимыми величинами $\xi_1, \dots, \xi_n$, имеющие один и тот же закон распределения $\mathcal{F}$.

С. в. $Y = g(X_1, \dots, X_n)$, где $g(X_1, \dots, X_n)$ — произвольная измеримая по Борелю функция из $\mathbf{R}^n$ в $\mathbf{R}$, называется *статистикой.*

Рассмотрим важные для приложений примеры статистик. Выборочные моменты:

$\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ — выборочное среднее;

$\overline{X^k} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k$ — выборочный момент k-го порядка;

$\overline{g(X)} = \frac{1}{n}\sum\limits_{i=1}^{n} g(X_i)$ — вообще выборочный момент для произвольной функции $g: \mathbf{R} \to \mathbf{R}$.

Выборочная дисперсия определяется величинами

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ и } S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

При вычислении выборочной дисперсии часто удобнее использовать формулу

$$S^2 = \overline{X^2} - (\overline{X})^2. \tag{1}$$

С понятием вариационного ряда связаны другие примеры статистик. Элементы выборки $X_1, \dots, X_n$, расположенные в порядке возрастания и обозначенные $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$, называются *вариационным* *рядом. Порядковой статистикой* $X_{(k)}, k = \overline{1,n}$ называется k-й член вариационного ряда.

Для дискретных распределений в вариационном ряде могут повторяться некоторые значения, для непрерывных распределений все члены ряда — различные цифры.

Порядковые статистики можно определить в виде

$$X_{(1)} = \min\{X_1, \dots, X_n\}, \dots$$
$$\dots, X_{(k)} = \min\{X_1, \dots, X_n\} \backslash \{X_1, \dots, X_{(k-1)}\}, \dots$$
$$\dots, X_{(n)} = \min\{X_1, \dots, X_n\} \backslash \{X_1, \dots, X_{(n-1)}\} = \max\{X_1, \dots, X_n\}.$$

$X_{(1)}$ и $X_{(n)}$ называются *экстремальными* порядковыми статистиками, $X_{(k)}, k \neq 1, k \neq n$ — *промежуточной* порядковой статистикой.

Средний элемент вариационного ряда называется *выборочной медианой* и обозначается M_e^*:

$$\mathrm{M}_e^* = \begin{cases} X_{(l+1)} & \text{при } n = 2\,l + 1, \\ (X_{(l)} + X_{(l+1)})/2 & \text{при } n = 2\,l. \end{cases}$$

Эмпирической функцией распределения F_n^*, построенной на основе выборки $\vec{X}$, называется функция

$$F_n^*(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}\{X_i < y\}.$$

Эмпирическую функцию распределения можно также записать с помощью вариационного ряда

$$F_n^*(y) = \begin{cases} 0, & y \le X_{(1)}, \\ k/n, & X_{(k)} < y \le X_{(k+1)},\ k = \overline{1, n-1}, \\ 1, & y > X_{(n)}. \end{cases}$$

Основное свойство эмпирической функции распределения определяется следующим утвеждением

Т е о р е м а (Гливенко-Кантелли). Пусть $F(y)$ — общая функция распределения элементов выборки. Тогда $F_n^*(y)$ сходится к $F(y)$ почти наверное равномерно по y при $n \to \infty$, т. е.

$$\sup_{y \in R} |F_n^*(y) - F(y)| \xrightarrow{\text{п.н.}} 0, \quad n \to \infty.$$

Другими словами, при увеличении объема выборки эмпирическая функция распределения сходится к теоретической (истинной).

A1

1. Доказать формулу (1).

2. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения с параметрами a и σ^2.

а). Вычислить среднее значение и дисперсию статистики $\overline{X}$. Какое распределение имеет $\overline{X}$?

б). Вычислить среднее значение статистик S^2 и S_0^2.

О т в е т. а) $a,\ \sigma^2/n,\ N_{a,\ \sigma^2/n}$; б) $\sigma^2(n-1)/n,\ \sigma^2$.

Р е ш е н и е. б) для S^2.

$$\mathbf{E}(S^2) = \mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big] = \{\text{ в силу формулы (1)}\} =$$

$$= \mathbf{E}\big[\,\overline{X^2} - (\,\overline{X}\,)^2\big] = \mathbf{E}\Big[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2\Big] =$$

$$= \frac{1}{n}\mathbf{E}\Big(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i<j} X_i X_j\Big) =$$

$$= \{\text{ в силу независимости и одинакового распределения } X_i\} =$$

$$= \frac{1}{n}\Big[\frac{n-1}{n}\sum_{i=1}^{n}\mathbf{E}X_1^2 - \frac{2}{n}\sum_{i<j}(\,\mathbf{E}X_1)^2\Big] =$$

$$= \frac{1}{n}\Big[\,\frac{n-1}{n}\,n\,\mathbf{E}X_1^2 - \frac{2}{n}\,\frac{n\,(n-1)}{2}\,(\mathbf{E}X_1)^2\Big] = \frac{n-1}{n}(\,\mathbf{E}X_1^2 - (\mathbf{E}X_1)^2) =$$

$$= \frac{n-1}{n}\,\sigma^2. \qquad \blacksquare$$

3. Построить график эмпирической функции распределения, соответствующей выборке объема n из распределения Бернулли B_p. Использовать выборочное среднее $\overline{X}$. Доказать непосредственно, что выполнена теорема Гливенко-Кантелли:

$$\sup_{y\in R}|F_n^*(y) - F(y)| \xrightarrow{\mathrm{p}} 0,\ \text{ при } n\to\infty.$$

4. Найти по функции распределения $F(y)$ величины X_1 функцию распределения первой и последней порядковой статистики:

а) $X_{(1)} = \min\{X_1, \ldots, X_n\}$, б) $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Выписать выражения для плотности этих порядковых статистик через функцию распределения и плотность величины X_1.

О т в е т. а) $F_{X_{(1)}}(y) = 1 - (1 - F_{X_1}(y))^n$, б) $F_{X_{(n)}}(y) = F^n_{X_1}(y)$

Р е ш е н и е. а)

$$F_{X_{(1)}}(y) = \mathbf{P}\{X_{(1)} < y\} = 1 - \mathbf{P}\{X_{(1)} \geq y\} = 1 - \mathbf{P}\{X_1 \geq y, X_2 \geq y, \ldots$$
$$\ldots, X_n \geq y\} = 1 - \prod_{k=1}^{n}(1 - \mathbf{P}\{X_k < y\}) = 1 - (1 - F_{X_1}(y))^n. \qquad \blacksquare$$

5. Доказать, что функция распределения k-й порядковой статистики $X_{(k)}$ имеет вид

$$\mathbf{P}\{X_{(k)} < y\} = \mathbf{P}\{\text{хотя бы } k \text{ элементов выборки } < y\} =$$
$$= \sum_{i=k}^{n} C_n^i F(y)^i (1 - F(y))^{n-i},$$

где $F(y)$ — функция распределения величины X_1.

6. Пусть F_n^* — эмпирическая функция распределения, построенная по выборке $X_1, \ldots, X_n$, а G_n^* — эмпирическая функция распределения, построенная по выборке $Y_1, \ldots, Y_n$, где $Y_i = aX_i + b$, $a > 0$ и b — два фиксированных действительных числа.

Доказать, что при всех y имеет место равенство $G_n^*(y) = F_n^*\big((y-b)/a\big)$.

С1

1. Пусть выборка $X_1, \ldots, X_n$ имеет распределение $\mathcal{F}$ с функцией распределения F. Доказать, что для любых y и $k = 0, \ldots, n$ справедливо равенство

$$\mathbf{P}\big(F_n^*(y) = k/n\big) = C_n^k F^k(y)(1 - F(y))^{n-k}.$$

2. Для выборки из биномиального распределения с параметрами p и m найти

$$\mathbf{P}\{F_n^*(y+0) - F_n^*(y) = k/n\}.$$

О т в е т. $C_n^k (C_m^y p^y (1-p)^{m-y})^k (1 - C_m^y p^y (1-p)^{m-y})^{n-k}$ при $y \in \{0, \ldots, m\}$; 0 — иначе.

3. Найти вероятность $\mathbf{P}\{X_{(k)} < y, X_{(k+1)} \geq y\}$ в терминах общей функции распределения элементов выборки.
О т в е т. $C_n^k F^k(y)(1-F(y))^{n-k}$.

4. Пусть F_n^* — эмприческая функция распределеня, построенная по выборке $X_1, \ldots, X_n$, G_n^* — эмприческая функция распределеня, построенная по выборке $Y_1, \ldots, Y_n$, где $Y_i = G(X_i)$ и G — монотонно возрастающая непрерывная функция. Доказать, что при всех y и v справедливо равенство

$$\mathbf{P}\{F_n^*(y) < v\} = \mathbf{P}\{G_n^*(G(y)) < v\}.$$

5. Чему равна вероятность $\mathbf{P}\{F_n^*(y) < F_n^*(z)\}$?
О т в е т. $1-(F(z)-F(y))^n$, если $y < z$.

6. Пусть выборка $X_1, \ldots, X_n$ имеет распределение $\mathcal{F}$, у которого функция распределения $F(x)$ непрерывна. Какому распределению будет соответствовать выборка $Y_1, \ldots, Y_n$, где $Y_i = F(X_i)$?
О т в е т. $U_{0,1}$.

Д1

1. Пусть выборка $X_1, \ldots, X_n$ имеет распределение $\mathcal{F}$. Доказать, что для каждого $y > 0$

$$\mathbf{P}\{\sup_t |F_n^*(t) - F(t)| > y\} \leq \mathbf{P}\{\sup_{0 \leq t \leq 1} |G_n^*(t) - t| > y\}, \qquad (2)$$

где $G_n^*(t)$ — эмпирическая функция распределения, построенная по выборке, объема n из распределения $U_{0,1}$.

2. Доказать, что если в условиях предыдущей задачи функция $F(t)$ непрерывна, то неравенство (2) превращается в равенство (это свойство используется при построении критерия Колмогорова и называется непараметричностью этого критерия).

3. Пусть $\vec{X}_n$ и $\vec{Y}_n$ — две независимые выборки объема n из одного и того же непрерывного распределения, а F_n^* и G_n^* соответственно эмпирические функции распределения, построенные по этим выборкам. Доказать, что для любого $t \in (0,1]$ справедливо равенство

$$\mathbf{P}\{\sup_{y \in R} |F_n^*(y) - G_n^*(y)| < t\} = \mathbf{P}\{\sup_{0 \leq k \leq 2n} |S_k| < tn | S_{2n} = 0\},$$

где $S_k = \xi_1 + \ldots + \xi_k$, причем случайные слагаемые ξ_i независимы и $\mathbf{P}\{\xi_i = 1\} = \mathbf{P}\{\xi_i = -1\} = 1/2$.

4. Пусть $X_1, \dots, X_n$ — выборка из распределения на множестве целых чисел; положим $p_k = \mathbf{P}\{X_1 = k\}$. Обозначим через $\nu_k(n)$ число элементов выборки, равных k. Доказать, что

$$\sup_{A \subseteq Z} |P_n^*(A) - \mathbf{P}\{X_1 \in A\}| = \frac{1}{2} \sum_{k=-\infty}^{\infty} |\frac{\nu_k(n)}{n} - p_k|.$$

Занятие 2

Тема. **Методы построения оценок: метод моментов.**

Цель. Отработать методику построения оценки метода моментов (ОММ) для параметров распределений.

Пусть известен закон распределения, а параметры $\theta_1, \dots \theta_m$ в него входящие, неизвестны. Возникает задача их статистического оценивания. Обозначим $m_k = \mathrm{E}X_1^k$ теоретический (истинный) момент k-го порядка. $\overline{X^k} = \frac{1}{n}\sum_{i=1}^n X_i^k$ — выборочный (эмпирический) момент k-го порядка.

Если момент m_k существует, то в силу ЗБЧ $\overline{X^k} \xrightarrow{p} m_k$ при $n \to \infty$. Поэтому для выборки $X_1, \dots, X_n$ достаточно большего объема можно утверждать, что $\overline{X^k} \approx m_k$, т. е. выборочный момент k-го порядка близок к теоретическим моментам m_k. На этом соображении и основывается метод моментов.

Можно записать систему из m уравнений $m_k = \overline{X^k}, k = \overline{1, m}$, т. е. теоретические и выборочные моменты одинакового порядка приравниваются. Получим систему m уравнений с неизвестными величинами $\theta_1, \dots \theta_m$. Компоненты решения $\theta^* = (\theta_1^*, \dots \theta_m^*)$ этой системы называются оценками метода моментов.

A2

1. Испoлзуя метод моментов, оценить значение α по выборке из показательного распределения E_α.

О т в е т. $\alpha^* = 1/\overline{X}$.

2. Пусть $X_1, \dots, X_n$ — выборка объема n из нормального распределения $N_{a,1}$, причем $a \geq 0$. Найти ОММ a^* неизвестного параметра a.

О т в е т. $a^* = \max\{0, \overline{X}\}$.

3. Используя метод моментов, оценить параметр сдвига $\theta \in R$ распределения с плотностью

$$f_\theta(x) = e^{-(x-\theta)} I_{\{x \geq \theta\}}.$$

О т в е т. $\theta^* = \overline{X} - 1$.

4. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения N_{a,σ^2}, где $a \in R, \sigma > 0$; и оба параметра a и σ^2 неизвестны. Найти ОММ a^*, σ^{2^*} неизвестных параметров.

О т в е т. $a^* = \overline{X}, \quad \sigma^{2^*} = S^2$.

5. Используя метод моментов, оценить значения параметров a, b по выборке из равномерного распределения $U_{a,b}$.

О т в е т. $a^* = \overline{X} - S\sqrt{3}, \quad b^* = \overline{X} + S\sqrt{3}$.

6. Пусть элементы выборки $X_i, \ i = \overline{1,n}$ имеют гамма-распределение с двумя неизвестными параметрами λ и α; соответствующая плотность имеет вид $f(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} I_{\{x>0\}} / \Gamma(\alpha)$.

У к а з а н и е. Момент k-го порядка $\mathrm{E}X_1^k$ гамма-распределения равен

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} \int\limits_0^\infty x^{k+\alpha-1} e^{-\lambda x} dx = \frac{\Gamma(\alpha+k)}{\lambda^k \Gamma(\alpha)} = \frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{\lambda^k}.$$

О т в е т. $\lambda^* = \overline{X}/S^2, \quad \alpha^* = \overline{X}^2/S^2$.

C2

1. Распределение Парето с параметром $\alpha > 1$ применяется для описания величины доходов населения выше фиксированного уровня x_0. Плотность распределения Парето задается формулой

$$f_\alpha(x) = \alpha\, x_0^\alpha\, x^{-\alpha-1} I_{\{x > x_0\}}.$$

Найти ОММ α^* неизвестного параметра α.

О т в е т. $\alpha^* = \overline{X}/(\overline{X} - x_0)$.

2. Найти ОММ параметра θ равномерного распределения $U_{0,\theta}$.

О т в е т. $\theta^* = 2\overline{X}$.

3. Найти ОММ параметра θ равномерного распределения $U_{-\theta,\theta}, \ \theta > 0$ с помощью второго момента.

О т в е т. $\theta^* = \sqrt{3\overline{X^2}}$.

4. Число ξ столкновений с молекулами газа в камере Вильсона частиц, получающихся при распаде ядра урана в результате бомбардировки его нейтронами, имеет распределение

$$\mathrm{P}\{\xi = k\} = \frac{1}{2}(\frac{\theta_1^k}{k!}e^{-\theta_1} + \frac{\theta_2^k}{k!}e^{-\theta_2}),\ k = 0, 1, 2, \ldots,\ 0 < \theta_1 < \theta_2$$

(«двойное» распределение Пуассона). Найти ОММ параметров $\theta_1,\ \theta_2$.

О т в е т. $\theta^*_{1,2} = \overline{X} \mp \sqrt{S^2 - \overline{X}}$.

Д2

1. Пусть $X_1, \ldots, X_n$ — выборка объема n из распределения с плотностью $f_\theta(x) = \frac{\theta}{2}e^{-\theta|x|}$, $\theta > 0$ — неизвестно. Найти ОММ θ^* с помощью второго момента.

О т в е т. $\theta^* = \sqrt{2/\overline{X^2}}$.

2. Для логнормальной модели с плотностью

$$f_{\mu,\sigma}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(\ln x - \mu)^2\} I_{\{x>0\}}$$

найти ОММ параметров μ и σ.

О т в е т. $\mu^* = 2\ln\overline{X} - (\ln\overline{X^2})/2,\ \sigma^* = (\ln\overline{X^2} - 2\ln\overline{X})^{1/2}$.

Занятие 3

Тема. **Методы построения оценок: метод максимальноо правдоподобия.**

Цель. Отработать методику построения оценки метода максимального правдоподобия для параметров распределений.

Одним из наиболее универсальных методов оценивания неизвестных параметров распределения чвляется *метод максимального правдоподобия*, в соответствии с которым в качестве оценки $\widehat{\theta} = \widehat{\theta}(X_1, \ldots X_n)$ параметра θ по выборке $\vec{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{F}_\theta$ выбирается значение параметра, максимизирующее *функцию правдоподобия*

$$\Psi(\vec{X}, \theta) = f_\theta(X_1) \ldots f_\theta(X_n),$$

где $f_\theta(x)$ плотность распределения для непрерывной с. в. и вероятность $f_\theta(x) = \mathrm{P}\{\xi = x\}$ — в дискретном случае.

Таким образом,

$$\Psi(\vec{X}, \widehat{\theta}) \geq \Psi(\vec{X}, \theta),\ \theta \in \Theta,\ \text{или}\ \Psi(\vec{X}, \widehat{\theta}) = \sup_{\theta \in \Theta} \Psi(\vec{X}, \theta). \tag{3}$$

Статистика $\widehat{\theta}$, удовлетворяющая условиям (3), называется *оценкой максимального правдоподобия* (ОМП) параметра θ.

Технически (т. к. Ψ состоит из произведений) удобнее искать $\max \ln \Psi$ (точка $\widehat{\theta}$, дающая максимум $\ln \Psi$, дает и максимум Ψ). Если максимум $\Psi(\vec{X}, \theta)$ достиается во внутренней точке множества Θ и функция $f_\theta(x)$ дифференцируема по θ, то $\widehat{\theta}$ удовлетворяет *уравнению правдоподобия*

$$\frac{\partial \ln \Psi(\vec{X}, \theta)}{\partial \theta} = 0. \tag{4}$$

Для случая векторного параметра $\theta = (\theta_1, \ldots, \theta_m)$ это уравнение заменяется системой уравнений

$$\frac{\partial \ln \Psi(\vec{X}, \theta)}{\partial \theta_j} = 0,\ j = 1, \ldots, m. \tag{5}$$

Практически ОМП $\widehat{\theta}$ обычно находят, решая уравнения (4) (или (5)).

А3

1. Получить ОМП $\widehat{\alpha}$ неизвестного параметра α для показательного распределения E_α выборки $\vec{X}_n$.

О т в е т. $\widehat{\alpha} = 1/\overline{X}$.

2. Получить ОМП $\widehat{a}$ и $\widehat{\sigma}^2$ неизвестных параметров a, σ^2 для нормального распределения N_{a,σ^2} выборки $\vec{X}_n$.

О т в е т. $\widehat{a} = \overline{X}$, $\widehat{\sigma}^2 = S^2$.

3. С. в. ξ, характеризующая срок службы элементов электронной аппаратуры, имеют *распределение Релея*: $F_\theta(x) = 1 - e^{-x^2/\theta},\ x \geq 0$. Показать, что ОМП θ равна $\widehat{\theta} = \overline{X^2}$.

4. Найти оценку параметра $\alpha > 1$ распределения Парето из задачи 1 раздела С2 по методу максимального правдоподобия.

О т в е т. $\widehat{\alpha} = (\frac{1}{n}\sum_{i=1}^{n} \ln X_i - \ln x_0)^{-1}$.

5. Пусть $\vec{X}$ — выборка объема n из равномерного распределения $U_{\theta,\theta+5}$, где $\theta \in R$. Найти ОМП $\widehat{\theta}$ параметра θ.

О т в е т. Любая точка ОМП $\widehat{\theta} \in [X_n - 5, X_1]$.

6. С. в. X_i имеют функцию распределения

$$F(\frac{x-\mu}{\sigma}),\ \text{где}\ F(x) = (1 - e^{-x})I_{\{x\geq 0\}}\ \text{(модель сдвига-масштаба}$$

показательного закона). Найти ОМП параметров μ и σ.

О т в е т. $\widehat{\mu} = X_{(1)}$, $\widehat{\sigma} = \overline{X} - X_{(1)}$.

С3

1. Пусть $\vec{X_n}$ — выборка из равномерного распределения $U_{\theta,\theta+1}$. Показать, что любое $\widehat{\theta} \in [X_n - 1, X_1]$ максимизирует функцию правдоподобия.

2. Доказать, что ОМП $\widehat{\theta}$ параметра сдвига θ *двухпараметрического экспоненциального распределения*: $F_{\theta,b}(x) = (1 - e^{-(x-\theta)/b})I_{\{x\geq a\}}$ $(b > 0)$ совпадает с $X_{(1)}$.

3. Найти ОМП параметра λ закона Пуассона.

О т в е т. $\widehat{\lambda} = \overline{X}$.

4. Оценить по методу максимального правдоподобия неизвестные параметры θ_1 и θ_2 нормальной модели N_{θ_1,θ_2^2}.

О т в е т. $\widehat{\theta} = (\overline{X}, S)$.

5. Пусть $\vec{X_n}$ — выборка из нормального распределения $N_{\theta,2\theta}$. Убедится в том, что ОМП $\widehat{\theta} = \sqrt{1 + \overline{X^2}} - 1$.

6. С. в. X_i имеют плотность $f_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$ (*сдвиг распределения Лапласа*). Как устроено множество, на котором фунция правдоподобия максимальна для а) четного, б) нечетного размера выборки?

О т в е т. а) в качестве решения годится любое θ из $[X_{(k)}, X_{(k+1)}]$. б) ОМП будет выборочная медиана M_e^*.

Занятие 4

Тема. **Свойства оценок: несмещенность, состоятельность, асимптотическая нормальность.**

Цель. Отработать методику проверки свойств оценок.

В силу многообразия оценок, применяемых для оценивания одной и той же неизвестной величины, возникает задача выбора из них лучшей, в определенном смысле. Поэтому к оценкам предъявляется ряд требований.

Пусть оцениваемый параметр θ исвестного распределения $\mathcal{F}_\theta$ имеет оценку $\theta^* = \theta^*(X_1, \dots, X_n)$. *Смещением* оценки θ^* называется величина $b(\theta) = \mathrm{E}_\theta \theta^* - \theta$. Оценка θ^*, для которой $b(\theta) = 0$, называется *несмещенной*. Оценка θ^* называется *асимптотически несмещенной*, если $b(\theta) = b_\theta(n) \to 0$ при $n \to \infty$.

Свойство несмещенности не достаточно для того, чтобы оценка хорошо приближала неизвестный параметр. Необходимо, чтобы погрешность приближения стремилась к нулю с увеличением размера выборки. Это свойство называется *состоятельностью*: $\theta^* \xrightarrow{\mathrm{p}} \theta$ при $n \to \infty$.

Оценка θ^* называется *асимптотически нормальной оценкой* (АНО) параметра θ *с коэффициентом* $\sigma^2(\theta)$ (или *асимптотической дисперсией*), если $\sqrt{n}(\theta^* - \theta) \Longrightarrow N_{0,\, \sigma^2(\theta)}$.

Асимптотической дисперсия $\sigma^2(\theta)$ характеризует точность АНО.

А4

1. Доказать, что значение выборочной функции распределения $F_n^*(x)$ в каждой точке x является несмещенной и состоятельной оценкой для теоретической функции распределения $F(x)$.

2. Пусть $\widehat{g}$ — несмещенная оценка g с конечной положительной дисперсией $\mathrm{D}\widehat{g}$. Будет ли статистика $\widehat{g}^{\,2}$ несмещенная оценкой для g^2?

3. По n независимым наблюдениям $X_1, \dots, X_n$ над с. в. ξ требуется оценить неизвестную дисперсию $\sigma^2 = \mathrm{D}\xi$, когда среднее $a = \mathrm{E}\xi$ известно. Убедиться, что несмещенной и состоятельной оценкой для σ^2 является статистика $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$.

4. Для выборки из равномерного распределения $U_{0,\theta}$. Проверить состоятельность и несмещенность оценки $X_{(n)}$ параметра θ.

О т в е т. Смещенная, состоятельная, асимптотически несмещенная.

5. Пусть дана выборка из распределения с конечной дисперсией. Доказать, что статистика $\overline{X}$ является АНО для $\theta = \mathrm{E}X_1$. Найти коэффициент аимптотической нормальности.

О т в е т. $\sigma^2 = \mathrm{D}X_1$.

6. Доказать, что выборочная дисперсия S^2 при условии конечности $\mathrm{E}X_1^4$ является АНО дисперсии. Вычислить коэффициент аимптотиче-

ской нормальности.

О т в е т. $\sigma^2 = \mathrm{E}(X_1 - \overline{X})^4 - \sigma^4$.

С4

1. Пусть с. в. ξ принимает значения $1, 2, \ldots, N$ с некоторыми вероятностями $p_1, p_2, \ldots, p_N$ соотвестсвенно. Над ξ проведено n независимых наблюдений, и пусть ν_r означает число тех из них, когда наблюдался исход «r» ($r = 1, \ldots, N$, $\nu_1, \ldots, \nu_N$). Доказать, что относительная частота $\nu_r^* = \nu_r/n$ является несмещенной и состоятельной оценкой для вероятности p_r.

У к а з а н и я. Воспользоваться тем, что с. в. ν_r имеет биноминальное распределение B_{n,p_r}; применить неравенство Чебышёва.

2. Для оценивания неизвестной дисперсии $\sigma^2 = \mathrm{D}\xi$ при неизвестном среднем используется выборочная дисперсия S^2. Является ли эта оценка несмещенной и состоятельной? Построить несмещенную оценку.

О т в е т. Смещение $b(\sigma^2) = -\sigma^2/n$, несмещенная оценка имеет вид $S_0^2 = \frac{n}{n-1}S^2$; обе оценки состоятельны.

3. Пусть $T_n = g + b_n$, $\mathrm{D}T_n = \delta_n$, где b_n, $\delta_n \to 0$ при $n \to \infty$. Доказать, что T_n — состоятельная оценка g.

4. Пусть $\mathrm{E}g^2(X_1) < \infty$. Доказать, что статистика $\overline{g^2}$ является АНО для параметра $\theta = \mathrm{E}g^2(X_1)$. Найти коэффициент асимптотической нормальности.

О т в е т. $\mathrm{D}g(X_1)$.

5. Пусть θ_n^* — АНО для параметра θ с коэффициентом σ^2. Пусть $\theta \neq 0$. Доказать, что θ_n^2 — АНО для θ^2. Найти коэффициент асимптотической нормальности.

О т в е т. $4\sigma^2\theta^2$.

Д4

1. Пусть требуется оценить функцию $\tau(p) = 1/p$ по выборке из распределения Бернулли B_p. Показать, что в данном случае не существует несмещенных оценок.

2. Пусть выборка имеет распределения Бернулли B_p. Рассматривается класс оценок вида $p^* = (n\overline{X} + \alpha)/(n + \beta)$, где $\alpha \geq 0$, $\beta \geq 0$. Вычислить смещение и среднеквадратическую оценку вида p^*. Показать, что при $\alpha = \frac{1}{2}\sqrt{n}$, $\beta = \sqrt{n}$ ошибка не зависит от p.

Занятие 5

Тема. **Сравнение оценок: среднеквадратический, асимптотический подходы. Эффективность.**

Цель. Отработать методику и методы сравнения оценок.

Для оценивания неизвестного параметра θ наблюдаемой с. в. ξ можно использовать, вообще говоря, различные оценки, поэтому, чтобы выбрать лучшую из них, надо использовать какой-нибудь критерий сравнения *качества* (*точности*) оценок.Обычно в качестве меры точности оценки $\theta^* = \theta^*(X_1, \ldots, X_n)$ используется *среднеквадратическое отклонение* $\mathrm{E}(\theta^* - \theta)^2$, и из двух оценок θ более точной (иногда говорят *более эффективной*) считается оценка с меньшей среднеквадратической ошибкой.

Для несмещенных оценок $\mathrm{E}(\theta^* - \theta)^2 = \mathrm{D}\theta^*$, т.е. их мерой точности является дисперсия.

Оценка θ^* называется *эффективной в классе со смещением* $b_n(\theta)$, если она не хуже в среднеквадратическом смысле любой другой оценки с тем же смещением $b_n(\theta)$.

Наряду со среднеквадратическим часто применяется *асимптотический подход* к сравнению оценок. Он удобен для для сравнения АНО, когда объем выборки очень велик.

Пусть θ_1^* — АНО с коэффициентом $\sigma_1^2(\theta)$, θ_2^* — АНО с коэффициентом $\sigma_2^2(\theta)$, то θ_1^* лучше, чем θ_2^* в смысле асимптотического подхода, если для любого $\theta \in \Theta$ $\sigma_1^2(\theta) \geq \sigma_2^2(\theta)$ и хотя бы для одного θ это неравенство строгое.

А5

1. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного распределения $U_{0,\theta}$, где $\theta > 0$. Сравнить оценки $\theta^* = 2\overline{X}$ и $\widehat{\theta} = X_{(n)}$ в среднеквадратичном.

О т в е т. $X_{(n)}$ лучше, чем $2\overline{X}$.

2. Пусть θ^* — оценка θ и $b = \mathrm{E}\theta^* - \theta$ — ее смещение. Доказать формулу $\mathrm{E}(\theta^* - \theta)^2 = \mathrm{E}\theta^* - b^2$.

3. По выборке $\overrightarrow{X}$ из равномерного распределения $U_{0,\theta}$ с неизвестным параметром $\theta > 0$, который требуется оценить. Убедиться в том, что

оценки $\theta_1^* = 2\overline{X}$, $\theta_2^* = \frac{n+1}{n} X_{(n)}$ и $\theta_3^* = (n+1)X_{(1)}$ — несмещенные, но наиболее точная из них θ_2^*.

4. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения с известным средним a и с неизвестной дисперсией σ^2. Сравнить в среднеквадратическом смысле оценки параметра σ^2

$$S_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \text{ и } S_1^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - a)^2.$$

5. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения с известным средним a и с неизвестной дисперсией σ^2. При помощи асимптотического подхода сравнить следующие оценки параметра σ^2:

$$\frac{\pi}{2} (\frac{1}{n} \sum_{i=1}^{n} |X_i - a|)^2 \text{ и } \frac{1}{n} \sum_{i=1}^{n} (X_i - a)^2.$$

C5

1. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения с известным средним a и с неизвестной дисперсией σ^2. Рассмотрим три оценки неизвестного параметра: выборочные дисперсии S^2, $S_0^2 = (n/(n-1))S^2$ и $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$ (две последние из этих оценок являются несмещенными). Доказать, что по среднеквадратическому подходу оценка S^2 является лучшей из трех оценок, но в классе несмещенных оценок S_1^2 лучше, чем S_0^2.

Указание. Учесть, что $\mathrm{E}X_1^2 = \theta^2$, $\mathrm{E}X_1^4 = 3\theta^4$.

2. Рассматривается задача оценивания дисперсии σ^2 из выборки нормального распределения N_{a,σ^2} (оба параметра неизвестны). Рассмотрим класс оценок вида $\theta_\lambda^* = \lambda S_0^2$, где S_0^2 — несмещенная оценка σ^2: а) доказать, что среднеквадратическая ошибка оценки θ_λ^* имеет вид

$$\mathrm{E}(\theta_\lambda^* - \sigma^2)^2 = \varphi(\lambda)\theta_2^4, \text{ где } \varphi(\lambda) = \frac{2}{n-1}\lambda^2 + (\lambda - 1)^2;$$

б) убедиться в том, что при $\frac{n-3}{n+1} < \lambda < 1$ оценка θ_λ^* точнее, чем оценка S_0^2.

3. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения со средним $a > 0$ и с единичной дисперсией. Сравнить оценки $\overline{X}$ и $\max(0, \overline{X})$ параметра a в среднеквадратическом смысле.

О т в е т. Вторая оценка лучше.

4. Сравнить оценки $\frac{n+1}{n}X_{(n)}$ и $X_{(n)}$ в среднеквадратическом смысле.

5. Пусть $X_1, \dots, X_n$ — выборка из нормального распределения со средним a и с дисперсией σ^2. При помощи асимптотического подхода сравнить выборочное среднее и выборочную медиану как оценки параметра a.

О т в е т. Среднее лучше.

Занятие 6

Тема. **Сравнение оценок: эффективные оценки. Неравенство Рао-Крамера.**

Цель. Ознакомление с приемом нахождения эффективной оценки по критерию Рао-Крамера.

Пусть имеем семейство распределений $\mathcal{F}_\theta$, зависящие от неизвестного параметра θ, значения которого принадлежат некоторому множеству Θ.

Определим функцию

$$I(\theta) = \mathrm{E}(\frac{\partial}{\partial\theta}\ln f_\theta(X_1))^2,$$

называемую *информацией Фишера.*

Если θ^* — произвольная несмещенная оценка и функция плотности $f_\theta(x)$ почти для всех $x \in R$ непрерывно дифференцируема по $\theta \in \Theta$ (эти условия определяют свойство *регулярности* распределения), то для многих моделей (регулярных) имеет место *неравенство Рао-Крамера* для ее дисперсии

$$\mathrm{D}\theta^* \geq \frac{1}{n\,I(\theta)}.$$

Для оценок, имеющих смещение $b(\theta)$, нижней границей служит $(1 + b'(\theta))^2/(n\,I(\theta))$. Оценка θ^*, достигающая нижней границы в неравенстве Рао-Крамера, называется эффективной в классе K_b.

Таким образом, эффективная оценка является оптимальной (т. е. с минимальной дисперсией).

А6

1. Пусть $X_1, \dots, X_n$ — выборка из нормального распределения N_{a,σ^2}, где a $\in R$, $\sigma > 0$. Проверить, является ли оценка $a^* = \overline{X} \in K_0$ эффективной.

О т в е т. Да.

2. Для выборки $X_1, \dots, X_n$ из распределения Пуассона Π_λ оценка $\lambda^* = \overline{X} \in K_0$ эффективна. Показать.

3. Пусть $X_1, \dots, X_n$ — выборка из нормального распределения N_{0,σ^2}, где $\sigma > 0$. Проверить, является ли оценка $\sigma^{2^*} = \overline{X^2} \in K_0$ эффективной.

О т в е т. Да.

4. Пусть в схеме Бернулли $B_{1,\theta}$ ($\theta \in (0,1)$ — неизвестно) проведено n испытаний, результаты которых представлены выборкой $X_1, \dots, X_n$. Требуется по этим данным оценить параметр θ.

О т в е т. Среднее арифмитическое наблюдений является оптимальной оценкой неизвестного параметра в классе несмещенных оценок.

5. Доказать, что для неизвестного параметра θ геометрического распределения G_θ информация Фишера имеет вид $I(\theta) = (\theta(1-\theta)^2)^{-1}$, а эффективной оценкой функции $f(\theta) = \theta/(1-\theta)$ является статистика $\theta^* = \overline{X}$. Вычислить дисперсию оценки θ^*.

О т в е т. $\mathrm{D}\theta^* = \theta/(n(1-\theta)^2)$.

С6

1. Пусть $X_1, \dots, X_n$ — выборка из биномиального распределения $B_{n,\theta}$, и требуется оценить неизвестный параметр θ. Показать, что эффективной оценкой θ является статистика $\theta^* = \overline{X}/n$.

2. Над пуассоновской с. в. с неизвестным параметром θ произведено n независимых наблюдений $X_1, \dots, X_n$. Показать, что статистика $\theta^* = \overline{X}$ является эффективной оценкой θ.

3. Пусть $X_1, \dots, X_n$ — выборка из показательного распределения $E_{1/\theta}$ с неизвестным параметром θ. Доказать, что $\overline{X}$ — эффективная оценка θ.

4. Пусть $X_1, \dots, X_n$ — выборка из смещенного показательного распределения с плотностью $f_\beta(y) = e^{\beta - y} I_{\{y \geq \beta\}}$. Найти эффективную несмещенную оценку для параметра $\beta \in R$.

Занятие 7

Тема. **Доверительное оценивание.**

Цель. Отработать методику построения доверительных интервалов.

Пусть имеем $X_1, \dots, X_n$ выборку из распределений $\mathcal{F}_\theta$, $\theta \in \Theta \subset R$ и ε — заданное положительное число. Если найдутся две статистики $\theta^{\pm} = \theta^{\pm}(\overrightarrow{X}, \varepsilon)$ такие, что при всех $\theta \in \Theta$

$$\mathrm{P}_\theta(\theta^- < \theta < \theta^+) \geq 1 - \varepsilon, \tag{6}$$

то (θ^-, θ^+) называется *доверительным интервалом* (ДИ) уровня $1 - \varepsilon$. Если при всех θ в (6) достигается равенство, то (θ^-, θ^+) условимся называть точным ДИ уровня $1 - \varepsilon$.

Случайный интервал (θ_n^-, θ_n^+) называется *аимптотическим доверительным интервалом* уровня $1 - \varepsilon$, если при всех θ

$$\lim_{n\to\infty} \mathrm{P}_\theta\{(\theta_n^-, \theta_n^+) \ni \theta\} \geq 1 - \varepsilon.$$

Из двух ДИ, построенных в одной и той же ситуации, предпочтительный тот, у которого меньше длина.

А7

1. Построить точный ДИ для параметра θ равномерного распределения $U_{0,\theta}$ для полученной выборки.

О т в е т. $\theta \in (X_{(n)}, X_{(n)}/\sqrt[n]{\varepsilon})$.

2. Пусть $X_1, \dots, X_n$ — выборка из распределения Пуассона Π_λ, где $\lambda > 0$. Требуется построить асимптотический ДИ для параметра λ уровня доверия $1 - \varepsilon$.

О т в е т.

$$(\overline{X} - \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}}}{\sqrt{n}}, \overline{X} + \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}}}{\sqrt{n}}).$$

3. Пусть $X_1, \dots, X_n$ — выборка из распределения Бернулли B_p. Построить ДИ для p уровня доверия $1 - \varepsilon$.

У к а з а н и е. Воспльзоваться неравенством Чебышёва.

О т в е т. $(\overline{X} - 1/(2\sqrt{n\,\varepsilon}), \overline{X} + 1/(2\sqrt{n\,\varepsilon}))$.

4. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения $N_{\theta,1}$. С помощью оценки $\overline{X}$ построить точный ДИ для θ уровня доверия $1-\varepsilon$.

5. Элементы X_i выборки имеют функцию распределения $F(x-\theta)$. Пусть $F(0) = 0$, а плотность $f(x)$ такова, что $c = f(0) > 0$. а) Найти предельный закон закон для величины $n(X_{(1)} - \theta)$. б) На его основе постройте асимптотический ДИ для параметра сдвига θ.

О т в е т. а) $\mathrm{P}(n(X_{(1)} - \theta) > x) \longrightarrow \exp\{-f(0)x\}$ при $n \to \infty$.

б) $(X_{(1)},\ X_{(1)} - \ln\varepsilon/(f(0)x)$.

С7

1. Построить точный ДИ для параметра $\theta > 0$ равномерного распределения $U_{\theta,2\theta}$ для полученной выборки.

О т в е т. $\theta \in (X_{(n)}/2, X_{(n)}/(1+\sqrt[n]{\varepsilon}))$.

2. Пусть $X_1, \ldots, X_n$ — выборка из распределения Бернулли B_p. Построить асимптотический ДИ для p уровня доверия $1-\varepsilon$.

3. Рассмотрим модель сдвига из задачи 5 раздела А7, но потребуем теперь, чтобы $F(0) = 1/2$ и F была непрерывной. Найти коэффициент доверия (надежности) интервала, образованного парой порядковых статистик $(X_{(k)}, X_{(l)})$, где $k < l$. Вычислить его значение для $k = 2, l = 5, n = 6$.

О т в е т. Коэффициент доверия $2^{-n}\sum_{i=k}^{l-1} C_n^i$. Для $k = 2, l = 5, n = 6$ получаем значение коэффициент доверия $\approx 0,78$.

Занятие 8

Тема. **Статистическая проверка гипотез: наиболее мощный критерий.**

Цель. Ознакомиться с методикой построения наиболее мощного критерия — критерия Неймана-Пирсона.

В практических задачах часто требуется по эмпирическим данным проверить то или иное предположение относительно каких-нибудь свойств закона распределения наблюдаемых с. в. Все эти предположения о виде или свойствах распределения наблюдаемых с. в. называют *статистическими гипотизами*, или просто *гипотезами*, которые обозначаются $H, H_0, H_1, \ldots$.

Если предположение однозначно фиксирует закон распределения с. в., то соответствующая гипотеза называется *простой*, в противном случае называется *сложной*.

Если закон распределения известен с точностью до некоторого параметра θ (θ может быть вектором), то всякое предположение о значении параметра называют *параметрической гипотезой*.

Рассмотрим с. в. ξ и случайную выборку $\vec{X} = (X_1, \dots, X_n)$, полученную в реультате n независимых испытаний. Требуется проверить некоторую гипотезу H_0 о законе распределения ξ. Будем говорить, что задан некоторый статистический критерий $\delta(\vec{X})$ для проверки гипотезы H_0, если сформулировано правило, согласно которому принимается решение: согласуются ли наблюдаемые значения с гипотезой (обычно ее называют основной или нулевой гипотезой) или она должна быть отвергнута, как противоречащая статистическим данным.

Для построения критерия обычно разбивают выборочное пространство R^n на две части $\mathrm{R}^n = \mathrm{S} \cup (\mathrm{R}^n \backslash \mathrm{S})$ так, что

$$\delta(\vec{X}) = \begin{cases} H_0, & \text{если } \vec{X} \in \mathrm{R}^n \backslash \mathrm{S}, \\ H_1, & \text{если } \vec{X} \in \mathrm{S}. \end{cases}$$

В соотвествии с этим гипотеза H_0 принимается, если конкретная реализация выборки будет принадлежать $\mathrm{R}^n \backslash \mathrm{S}$, и отвергаться, если она будет принадлежать S.

Область S, в которой принимается вторая (альтернативная) гипотеза, назыают *критической областью*.

Для каждого критерия возможны ошибки двух родов:

- Гипотеза H_0 справедлива, но она отвергнута (реализация выборки $\vec{X} \in \mathrm{S}$), такая ошибка называется *ошибкой первого рода*. Вероятность ошибки первого рода обозначается α. По определению она равна:

  $$\alpha = \alpha(\delta) = \mathrm{P}_{H_0}(\delta(\vec{X}) \neq H_0) = \mathrm{P}_{H_0}(\delta(\vec{X}) = H_1) = \mathrm{P}_{H_0}(\vec{X} \in S).$$

  Вероятность ошибки первого рода $\alpha = \alpha(\delta)$ называют размером или уровнем значимости критерия δ.

- Гипотеза H_0 ложна, но она принята (реализация выборки. $\vec{X}$ принадлежит области $\mathrm{R}^n \backslash \mathrm{S}$). Такая ошибка называется *ошибкой второго рода*. Вероятность ошибки второго рода обозначается β. *Мощностью критерия* δ называют величину $1 - \beta$. Мощность критерия равна

$$1-\beta(\delta)=1-\mathrm{P}_{H_1}(\delta(\vec{X})\neq H_1)=\mathrm{P}_{H_1}(\delta(\vec{X})=H_1)=\mathrm{P}_{H_1}(\vec{X}\in S).$$

Эта величина характеризует вероятность принятия правильного решения (отклонение H_0) в ситуации, когда H_0 ложна.

Из двух критериев с одной и той же ошибкой первого рода α лучшим считается тот, мощность которого при альтернативах больше (обладающий наименьшей вероятностью ошибки второго рода β).

Введем при $\varepsilon\in[0,1]$ класс критериев $K_\varepsilon=(\delta(\vec{X}):\alpha(\delta)\leq\varepsilon)$. Критерий $\delta_0\in K_\varepsilon$ называют *наиболее мощным критерием* (НМК) размера ε, если $\beta(\delta_0)\leq\beta(\delta)$, для любого другого критерия $\delta\in K_\varepsilon$.

Обычно последствия указанных ошибок неравнозначны. Что опаснее — зависит от конкретной постановки задачи и содержания основной гипотезы.

Поскольку исключить ошибки α и β невозможно, то естественно стремиться минимизировать потери от этих ошибок. Конечно, желательно уменьшить α и β одновременно, но уменьшение одной из них, как правило, влечет за собой увеличение другой. В большинстве случаев одновременное уменьшение α и β связано с увеличением объема выборки n.

Рассморим построение НМК для различных двух простых гипотез о распределении элементов выборки X_i: $H_0=\{X_i$ имеют распределение $\mathcal{F}_0\}$ и $H_1=\{X_i$ имеют распределение $\mathcal{F}_1\}$.

Пусть $f_0(y)$ — плотность распределения $\mathcal{F}_0$, $f_1(y)$ — плотность распределения $\mathcal{F}_1$, а

$$f_0(\vec{X})=f_0(X_1,\ldots,X_n)=\prod_{i=1}^{n}f_0(X_i),\, f_1(\vec{X})=f_1(X_1,\ldots,X_n)=\prod_{i=1}^{n}f_1(X_i)$$

— соответствующие функции правдоподобия. Предполагается, что распределения $\mathcal{F}_0$ и $\mathcal{F}_1$ либо оба дискретны, либо оба непрерывны.

Назовем *отношением правдоподобия* частное

$$T(\vec{X})=T(X_1,\ldots,X_n)=\frac{f_1(X_1,\ldots,X_n)}{f_0(X_1,\ldots,X_n)}.$$

Это отношение позволяет разбивать выборочное пространство R^n на два подмножества $\mathrm{R}^n\backslash\mathrm{S}$ и S. Если величина $T(\vec{X})$ «велика», то отнесем $\vec{X}$

к множеству S — множеству отвержения гипотезы H_0 (и принятия H_1), если же эта величина «мала», то следует $\vec{X}$ отнести к множеству $\mathrm{R}^n \backslash \mathrm{S}$ — множеству принятия гипотезы H_0 (и отвержения H_1).

Т е о р е м а (лемма Неймана-Пирсона). НМК для проверки простой гипотезы H_0 при простой альтернативе H_1 с вероятностью ошибки первого рода α существует и задается критической областью

$$\mathrm{S} = \{\vec{X} : T(\vec{X}) = \frac{f_1(X_1, \dots, X_n)}{f_0(X_1, \dots, X_n)} \geq c\},$$

где число c должно быть выбрано так, чтобы размер критерия равнялся ε (для любого $0 < \varepsilon \leq \mathrm{P}_{H_0}(f_1(\vec{X}) > 0)$): $\alpha = \varepsilon$.

Построенный критерий носит название *критерия отношения правдоподобия* или *критерия Неймана-Пирсона.* Также предполагается, что

функция $R(c) = \mathrm{P}_{H_0}(T(\vec{X}) \geq c)$ непрерывна по c при $c > 0$.

Функция $R(c)$ есть хвост функции распределения с. в. $T(\vec{X})$:

$$R(c) = 1 - \mathrm{P}_{H_0}(T(\vec{X}) < c).$$

Ее непрерывность означает, что величина $\mathrm{P}_{H_0}(T(\vec{X}) = c)$ равна нулю для любого $c > 0$. Это предположение избавляет нас от необходимости рассматривать рандомизированные критерии.

А8

1. Имеется выборка $X_1, \dots, X_n$ из нормального распределения с неизвестным средним θ и дисперсией σ^2.

Требуется: а) построить наиболее мощный размера ε критерий для проверки гипотезы $H_0 = \{\theta = a_1\}$ против альтернативы $H_1 = \{\theta = a_2\}$, для определенности предполагаем, что $a_1 < a_2$, и найти ошибку второго рода β;

б) при заданных значениях α и β найти минимальный объем необходимого материала $n^* = n(\alpha, \beta)$ для того, чтобы ошибочные заключения могли быть сделаны с вероятностями, не большими, чем α и β;

в) вычислить ошибку второго рода β по выборке объема $n = 25$ при значениях $\sigma^2 = 25, a_1 = 0$, при альтернативах $a_2 = 1, a_2 = 5$. Ошибку первого рода α взять равной $0,05$.

г) пусть $\sigma^2 = 25, a_1 = 0, a_2 = 1, \alpha = 0,05$. Сколько необходимо провести наблюдений, чтобы ошибка второго рода β не превосходила $0,1$?

О т в е т. а) критическая область S определяется неравенством

$$\overline{X} \geq \frac{\frac{1}{2}(a_2^2 - a_1^2) - \frac{1}{n}\sigma^2 \ln c}{a_2 - a_1}.$$

НМК размера ε имеет вид

$$\delta(\vec{X}) = H_1 \text{ при } \overline{X} > a_1 + \frac{\tau_{1-\varepsilon}\sigma}{\sqrt{n}} = \{ \text{ обозначим } \} = c_1,$$

где $\tau_{1-\varepsilon}$ — квантиль уровня $1-\varepsilon$ стандартного нормального распределения.

Вероятность ошибки второго рода равна

$$\beta(\delta) = \mathrm{P}_{H_1}(\overline{X} < c_1) = \Phi_{0,1}(\sqrt{n}\ \frac{c_1 - a_2}{\sigma}).$$

б) $n^* = [\sigma^2(\tau_\alpha + \tau_\beta)/(a_1 - a_2)^2] + 1$, где $[x]$ — целая часть x.

в) При $a_2 = 1, \beta = 0,741$. При $a_2 = 5, \beta = 0,001$.

г) $n^* = 9$.

2. Имеется выборка $X_1, \ldots, X_n$ из нормального распределения N_{a,θ^2} с известным средним a и неизвестной дисперсией θ^2. Построить НМК размера ε для проверки гипотезы $H_0 = \{\theta^2 = \sigma_0^2\}$ против альтернативы $H_1 = \{\theta^2 = \sigma_1^2\}$, где $\sigma_0^2 < \sigma_1^2$.

О т в е т. Критическая область критерия Неймана-Пирсона определяется неравенством

$$\sum_{i=1}^{n}(\frac{X_i - a}{\sigma_0})^2 \geq \frac{2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\ln(\frac{\sigma_1}{\sigma_0}c) = c_1.$$

c_1 находится из уравнения, в котором размер критерия равен ε:

$$\alpha(\delta) = \mathrm{P}_{H_0}(\sum_{i=1}^{n}(\frac{X_i - a}{\sigma_0})^2 > c_1) = 1 - H_n(c_1) = \varepsilon,$$

где $H_n(x)$ — функция χ^2-распределения с n степенями свободы.

3. Имеем схему Бернулли с числом испытаний n и неизвестной вероятностью успеха p. Построить НМК размера ε для проверки гипотезы $H_0 = \{p = p_0\}$ против альтернативы $H_1 = \{p = p_1\}$, где $p_0 < p_1$.

C8

1. Имеется выборка $X_1, \dots, X_n$ из нормального распределения с неизвестным средним a и единичной дисперсией. Построить НМК размера ε для проверки гипотезы $H_0 = \{a = a_1\}$ против альтернативы $H_1 = \{a = a_2\}$, где $a_1 < a_2$.

О т в е т. НМК размера ε имеет вид

$$\delta(\vec{X}) = H_1 \text{ при } \overline{X} > a_1 + \frac{\tau_{1-\varepsilon}}{\sqrt{n}}.$$

2. Имеется выборка $X_1, \dots, X_n$ из нормального распределения со средним 0 и дисперсией σ^2. Построить НМК размера ε для проверки гипотезы $H_0 = \{\sigma = \sigma_1\}$ против альтернативы $H_1 = \{\sigma = \sigma_2\}$, где $\sigma_1 < \sigma_2$.

О т в е т. НМК размера ε имеет вид

$$\delta(\vec{X}) = H_1 \text{ при } \overline{X^2} > \frac{h_{1-\varepsilon}\sigma_1^2}{n},$$

где $h_{1-\varepsilon}$ — квантиль χ^2-распределения с n степенями свободы уровня $1-\varepsilon$.

3. По выборке $X_1, \dots, X_n$ из пуассоноского распределения Π_λ построить критерий Неймана-Пирсона для проверки гипотезы $H_0 = \{\lambda = \lambda_0\}$ против альтернативы $H_1 = \{\lambda = \lambda_1\}$, $(0 < \lambda_0 < \lambda_1)$.

Литературный список

[1] *Боровков А. А.* Математическая статистика.— Новосибирск: Наука; Изд-во Института математики, 1997.

[2] *Ван дер Варден Б.* Математическая статистика. — М.: Иностр. лит., 1960.

[3] *Емельянов Г. В., Скитович В. П.* Задачник по теории вероятностей и математической статистике. — Л.: Изд-во Ленинградского ун-та, 1967.

[4] *Ивченко Г. И., Медведев Ю. И.* Математическая статистика.— М.: Высшая школа, 1984.

[5] *Ивченко Г. И., Медведев Ю. И., Чистяков А. В* Сборник задач по математической статистике.— М.: Высшая школа, 1989.

[6] *Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф.* Справочник по теории вероятностей и математической статистике. — М.: Наука, 1985.

[7] *Розанов Ю. А.* Теория вероятностей, случайные процессы и математическая статистика. — М.: Наука, 1985.

[8] *Севостьянов Б. А.* Курс теории вероятностей и математической статистики. — М.: Наука, 1976.

[9] *Феллер В.* Введение в теорию вероятностей и ее приложения: Перевод с англ. 3-е изд. — М.: Мир, 1984.— Т. 2.

[10] Математическая статистика: Учеб. для вузов/ В. Б. Горяинов, И. В. Павлов и др.; под ред. В. С. Зарубина, А. П. Крищенко. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2001.

[11] *Климов Г. П., Кузъмин А. Д.* Вероятность, процессы, статистика. — М.: Изд-во МГУ, 1985.

[12] *Коршунов Д. А., Чернова Н. И.* Сборник задач и упражнений по математической статистике.— Новосибирск: Изд-во Института математики, 2001.

[13] Сборник задач по теории вероятностей, математической статистике и теории случайных функций/ под общей ред. А. А. Свешникова. — СПб.: Изд-во "Лань", 2007.

Содержание

Виктор Александрович Труфанов,
доцент кафедры МАиМ АмГУ;

Математическая статистика. Учебное пособие.

Изд-во АмГУ. Подписано к печати 12.12.2010. Формат 60×84/16.
Усл. печ. л. 1,91. Тираж 50. Заказ 184.
Отпечатано в типографии АмГУ.