

БРС

| Вид учебной деятельности | Максимальное количество баллов |
|--------------------------|--------------------------------|
| 8 лекций                 | 10                             |
| 15 практических занятий  | 10                             |
| РГЗ – часть 1            | 20                             |
| РГЗ – часть 2            | 20                             |
| Экзамен                  | 40                             |

## Список литературы

1. Постовалов С. Н. Математическая статистика. Конспект лекций: учебное пособие / С.Н. Постовалов, Е.В. Чимитова, В.С. Карманов; Новосиб. гос. техн. ун-т. - Новосибирск, 2014. – 138 с.
2. Математическая статистика: методические указания к расчетно-графическому заданию для 4 курса ФПМИ.
3. Математическая статистика: методические указания по курсу "Теория вероятностей и математическая статистика" для 3 курса ФПМИ.
4. Гмурман В.Е. Теория вероятностей и математическая статистика: учебное пособие для вузов / В. Е. Гмурман. - М., 2008. – 478 с.
5. Ивченко Г.И., Медведев Ю.И. Математическая статистика: Учебник. — М.: Книжный дом «ЛИБРОКОМ», 2014. — 352 с.
6. Ивченко Г.И., Медведев Ю.И., Чистяков А.В. Сборник задач по математической статистике. М.: Высш. шк., 1989. – 255 с.
7. Коршунов Д.А., Чернова Н.И. Сборник задач и упражнений по математической статистике: Учебное пособие. 2004. — 128 с.
8. Чернова Н.И. Математическая статистика: Учеб. пособие / Новосиб. гос. ун-т. Новосибирск, 2007. 148 с.
9. Шуленин В.П. Математическая статистика. В 3х частях. 2012 г.

### Соглашения в обозначениях

Случайные величины  $\xi, \eta$  и  $X, Y$ .

Переменные, реализации случайных величин  $x, y$ .

Функция распределения случайной величины  $F(x)$  или  $F_\xi(x)$ .

Функция плотности распределения случайной величины  $f(x)$  или  $f_\xi(x)$ .

### 1.1. Выборка. Выборочный метод

Пусть  $X_n = \{X_1, \dots, X_n\}$  – выборка объема  $n$ , полученная в результате наблюдения случайной величины  $\xi$  с функций распределения  $F(x)$ .

Будем считать, что

- наблюдения  $X_1, \dots, X_n$  независимы и имеют одну и ту же функцию распределения;
- множество возможных значений  $\xi$  образуют генеральную совокупность  $L(\xi)$ , из которой берется выборка  $X_n$ ;
- $F_\xi \in F = \{F(x, \theta), \theta \in \Theta\}$  – параметрическая статистическая модель. Параметр  $\theta$  может быть как скалярным, так и векторным.

### ***Определение***

Статистикой называется любая измеримая функция от выборки, которая, в свою очередь, также является случайной величиной.

### Примеры статистик

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ — выборочное среднее,}$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ — выборочная дисперсия,}$$

$$S_0^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ — несмещенная выборочная дисперсия,}$$

$$\overline{X^k} = \frac{1}{n} \sum_{i=1}^n X_i^k \text{ — выборочный момент порядка } k$$

## 1.2. Порядковые статистики и вариационный ряд

Упорядочим все наблюдения в выборке и произведем их перенумерацию:  
 $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \dots \leq X_{(n)}$  – вариационный ряд.

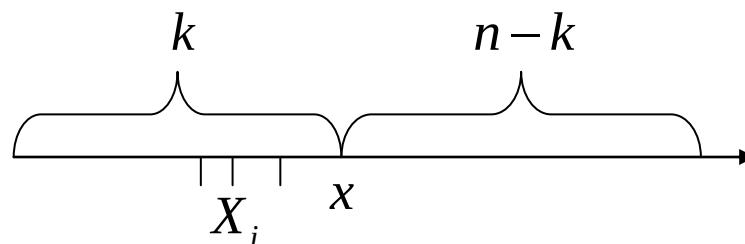
### *Определение*

Величина  $X_{(i)}$  называется  $i$ -ой порядковой статистикой.

Найдем распределение  $i$ -ой порядковой статистики.

Введем вспомогательную случайную функцию:  $\mu_n(x)$  – количество наблюдений, не превосходящих  $x$ .

Найдем  $P\{\mu_n(x) = k\}$ . Событие  $\mu_n(x) = k$  означает, что в интервал  $(-\infty, x]$  попало  $k$  наблюдений, а в  $(x, +\infty)$  попало  $(n - k)$  наблюдений.



Число способов, которыми можно выбрать  $k$  элементов из  $n$  равно  $C_n^k$ .

В результате получаем:  $P\{\mu_n(x) = k\} = C_n^k F^k(x) (1 - F(x))^{n-k}$

$$P\{X_{(i)} < x\} = P\{\mu_n(x) \geq i\} = P\{\mu_n(x) = i \vee \mu_n(x) = i+1 \vee \dots \vee \mu_n(x) = n\} =$$

$$= \sum_{k=i}^n P\{\mu_n(x) = k\} = \sum_{k=i}^n C_n^k F^k(x) (1 - F(x))^{n-k}.$$

### 1.3. Эмпирическая функция распределения

Функция  $F_n(x) = \frac{\mu_n(x)}{n}$  называется эмпирической функцией распределения. По определению эмпирическая функция распределения является случайной функцией;  $\forall x \in R, F_n(x)$  – дискретная случайная величина, принимающая значения

$$0 = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, \frac{n}{n} = 1$$

при этом  $P\left\{F_n(x) = \frac{k}{n}\right\} = P\{\mu_n(x) = k\} = C_n^k F^k(x)(1-F(x))^{n-k}$ .

Если все  $X_i$  (наблюдения в выборке) различны, то

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ \frac{k}{n}, & X_{(k)} \leq x < X_{(k+1)}, k = 1, \dots, n-1 \\ 1, & x \geq X_{(n)} \end{cases}$$

или  $F_n(x) = \frac{1}{n} \sum_{i=1}^n h(x - X_i)$ , где  $h(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$  – функция Хевисайда (единичного скачка).

**Теорема 1.1.**

Пусть  $F_n(x)$  – эмпирическая функция распределения случайной величины  $\xi \in F_\xi$ .

Тогда  $\forall |x| < \infty, \forall \varepsilon > 0 \lim_{n \rightarrow \infty} P\{|F_n(x) - F(x)| < \varepsilon\} = 1$

*Доказательство:*

З.Б.Ч.

$\eta_i$  – независимы, одинаково распределенные,  $M\eta_i = a$  то  $\frac{1}{n} \sum \eta_i \xrightarrow{P} a$ , при  $n \rightarrow \infty$

$$\left( P\left\{ \frac{1}{n} \sum \eta_i - a \right\} < \varepsilon \rightarrow 1; \forall \varepsilon > 0, n \rightarrow \infty \right)$$

Введем случайную величину  $\eta_i = h(x - X_i) = \begin{cases} 0, X_i < x \\ 1, X_i \geq x \end{cases}$ , найдем ее математическое

ожидание  $M\eta_i = Mh(x - X_i) = 1 \cdot P\{X_i \leq x\} + 0 \cdot P\{X_i > x\} = F(x)$  подставим в З.Б.Ч. получим условия теоремы. ◀

Таким образом, при  $n \rightarrow \infty$  эмпирическая функция распределения  $F_n(x)$  является оценкой теоретической функции распределения  $F(x)$ .

Введем статистику  $D_n = \sup_{|x| < \infty} |F_n(x) - F(x)|$

**Теорема 1.2. (Гливенко - Кантелли)**

$$P\left\{\lim_{n \rightarrow \infty} D_n = 0\right\} = 1$$

**Теорема 1.3. (Колмогоров)**

$$\lim_{n \rightarrow \infty} P\left\{\sqrt{n}D_n < t\right\} = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}$$

$K(t)$  – распределение Колмогорова.

$$P\left\{D_n < \frac{t}{\sqrt{n}}\right\} = K(t)$$

## 1.4. Непараметрическое оценивание плотности распределения

### 1.4.1. Гистограмма

Разобьем область определения на  $k$  интервалов:

$$-\infty = t_0 < t_1 < t_2 < \dots < t_k = \infty,$$

$n_i$  — количество наблюдений на  $[t_i, t_{i+1})$ ,  $\sum_{i=1}^k n_i = n$ .

Высота столбца гистограммы  $\frac{n_i}{n(t_i - t_{i-1})}$ :

$$\sum_{i=1}^k \frac{n_i}{n(t_i - t_{i-1})} (t_i - t_{i-1}) = \frac{\sum_i n_i}{n} = \frac{n}{n} = 1$$

Гистограмма довольно грубый способ оценивания плотности распределения, связанный с неопределенностью выбора  $k$ , границ интервалов, потерей информации при группировании.

### 1.4.2. Ядерные оценки плотности (оценки Розенבלата-Парзена)

Пусть  $g(t)$  – колоколообразная (ядерная) функция, удовлетворяющая условиям:

$$g(t) = g(-t), \int_{-\infty}^{+\infty} g(t) dt = 1, \int_{-\infty}^{+\infty} t^2 g(t) dt = 1, \int_{-\infty}^{+\infty} t^m g(t) dt < \infty; 0 \leq m < \infty$$

и  $\lim_{n \rightarrow \infty} \lambda_n = 0; \lim_{n \rightarrow \infty} n\lambda_n = \infty$  тогда функцию плотности можно оценить:

$$\hat{f}_n(x) = \frac{1}{n\lambda_n} \sum_{i=1}^n g\left(\frac{x - X_i}{\lambda_n}\right),$$

$\lambda_n$  – параметр размытости ядерной функции.

Основное преимущество ядерных оценок состоит в том, что они непрерывны, в отличие от эмпирической функции распределения и гистограмм.

### 1.4.3. Цензурированная выборка. Оценка Каплана-Мейера

*Цензурированная справа* выборка может быть представлена в виде

$$(X_1, \delta_1), (X_2, \delta_2), \dots, (X_n, \delta_n),$$

где  $X_i = \min(T_i, C_i)$  – значение наблюдения,

$T_i$  – момент наступления системного события,

$C_i$  – момент цензурирования,

$$\delta_i = \begin{cases} 1, & \text{если } T_i < C_i, X_i = T_i \\ 0, & \text{если } T_i \geq C_i, X_i = C_i \end{cases} \text{ – индикатор цензурирования, который равен единице,}$$

если  $i$ -е наблюдение полное, нулю – если цензурированное.

Выборка называется *случайно цензурированной*, если  $T_i$  и  $C_i$  представляют собой независимые случайные величины, причем  $T_i$  принадлежит закону распределения вероятностей с функцией  $F(x)$ , а  $C_i$  – закону  $F_C(x)$ .

Обозначим через  $a_1 < a_2 < \dots < a_k = \tau$ ,  $k \leq n$  моменты времени, в которые были зафиксированы системные события ( $X_i, \delta_i = 1$ ).

Тогда *оценку Каплана-Мейера* можно вычислить по формуле:

$$\hat{F}_n(x) = 1 - \prod_{a_i \leq x} \left( 1 - \frac{d_i}{r_i} \right),$$

где  $d_i = \sum_{t_j = a_i} \delta_j$ ,  $r_i$  — количество наблюдений, для которых  $t_j \geq a_i$ ,  $j = 1, 2, \dots, n$ .