

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ  
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«ОРЕНБУРГСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»**

**МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ  
ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ**

**Теория вероятностей и математическая статистика**

**Направление подготовки (специальность) Экономика**

**Профиль образовательной программы Финансы и кредит**

**Форма обучения заочная**

## СОДЕРЖАНИЕ

<b>1. Конспект лекций .....</b>	<b>3</b>
<b>1.1 Лекция № 1 Случайные события. Вероятность события.....</b>	<b>3</b>
<b>1.2 Лекция № 2 Повторные независимые испытания.....</b>	<b>6</b>
<b>1.3 Лекция № 3 Дискретная случайная величина (ДСВ).....</b>	<b>8</b>
<b>1.4 Лекция № 4 Непрерывная случайная величина (НСВ).....</b>	<b>10</b>
<b>1.5 Лекция № 5 Статистическое оценивание параметров распределения.....</b>	<b>13</b>
<b>1.6 Лекция № 7 Корреляционный анализ.....</b>	<b>21</b>
<b>1.7 Лекция № 8 Регрессионный анализ.....</b>	<b>26</b>
<b>2. Методические указания по проведению практических занятий .....</b>	<b>29</b>
<b>2.1 Практическое занятие № 1 ПЗ-1 Случайные события. Вероятность события ..</b>	<b>29</b>
<b>2.2 Практическое занятие № 2 ПЗ-2 Повторные независимые испытания.....</b>	<b>30</b>
<b>2.3 Практическое занятие № 3 ПЗ-3 Дискретная случайная величина.....</b>	<b>30</b>
<b>2.4 Практическое занятие № 4 ПЗ-4 Непрерывная случайная величина.....</b>	<b>31</b>
<b>2.5 Практическое занятие № 5 ПЗ-5 Статистическое оценивание параметров распределения.....</b>	<b>32</b>
<b>2.6 Практическое занятие № 6 ПЗ-7 Корреляционный анализ.....</b>	<b>33</b>
<b>2.7 Практическое занятие № 7 ПЗ-8 Регрессионный анализ.....</b>	<b>33</b>

# 1. КОНСПЕКТ ЛЕКЦИЙ

## 1.1 Лекция № 1 (2 часа).

Тема: «Случайные события. Вероятность события» (интерактивная форма)

### 1.1.1 Вопросы лекции:

1. Основные понятия теории вероятностей
2. Классическое определение вероятности
3. Элементы комбинаторики
4. Геометрическое определение вероятности
- 5 Статистическое определение вероятности

### 1.1.2 Краткое содержание вопросов:

#### 1. Основные понятия теории вероятностей

Теория вероятностей является основой математической статистики. **Математическая статистика** – раздел математики, изучающий математические методы сбора, систематизации, обработки и интерпретации результатов наблюдений с целью выявления статистических закономерностей.

**Теория вероятностей** – это математическая наука, изучающая закономерности случайных явлений (событий величин, функций, процессов и др.)

Предметом теории вероятностей является изучение вероятностных закономерностей массовых однородных случайных событий.

Одним из основных понятий теории вероятностей является понятие случайного события.

**Случайное событие** – это любой факт, который может либо произойти, либо не произойти при выполнении некоторого комплекса условий.

Обычно случайные события обозначаются заглавными латинскими буквами: A, B, C.

Геометрически случайные события удобно изображать с помощью диаграмм Эйлера-Венна.

Виды событий: **достоверные** ( $\Omega$ ); **невозможные** ( $\emptyset$ ); - A является частным случаем B; **равносильные**; **несовместные** (несовместимые); **совместные** (совместимые); **равновозможные**.

#### Операции над событиями

**Суммой** событий A и B называют событие **A+B**, состоящее в наступлении хотя бы одного из этих событий A или B.

**Произведением** событий A и B называют событие **A·B**, состоящее в одновременном наступлении этих событий A и B.

Событие  $\bar{A}$  называется противоположным событием (дополнением) события A, если непоявление одного события влечет появление другого. Сумма противоположных событий есть событие достоверное, а произведение невозможное:

События  $\{H_i\}_{i=1}^n$  образуют полную группу попарно несовместимых событий, если любые два из них несовместны и хотя бы одно непременно должно произойти в результате испытания:

$$\sum_{i=1}^n H_i = \Omega \text{ - полнота,}$$

$$H_i \cdot H_j = \emptyset, \forall i \neq j \text{ - попарная несовместимость.}$$

## 2. Классическое определение вероятности

**Вероятность события** — это численная мера объективной возможности его появления.

В соответствии с классическим определением:

**Вероятность  $P(A)$  события  $A$**  равняется отношению числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов, образующих полную группу:

$$P(A) = \frac{m}{n}. \quad (*)$$

При этом полагают, что:

- испытание содержит конечное число исходов;
- все исходы испытания равновозможны и несовместимы.

Такого рода опыт называют «схемой случаев» (или «схемой урн», так как любую подобную вероятностную задачу можно свести к задаче с урнами с разноцветными шарами).

Из классического определения вероятности вытекают следующие очевидные **свойства вероятности события**.

### 3. Элементы комбинаторики

**Комбинаторика** — это раздел математики, изучающий методы решения задач на подсчет числа различных комбинаций.

В комбинаторике есть два важных правила, часто применяемых при решении комбинаторных задач.

#### 1. Правило умножения комбинаторики.

#### 2. Правило сложения комбинаторики.

$m$ -элементные подмножества (комбинации) могут отличаться:

- составом элементов;
- порядком следования элементов;
- возможностью повтора элементов в подмножестве;
- объемом подмножества.

В соответствии с этим выделяют следующие виды подмножеств.

#### 1. Размещения

Число всех размещений  $A_n^m$  из  $n$  элементов по  $m$  (где  $m < n$ ), определяется по формуле:

$$A_n^m = \frac{n!}{(n-m)!}$$

#### 2. Перестановки

Число всех перестановок  $P_n$  из  $n$  элементов определяется по формуле:

$$P_n = n!$$

Заметим, что перестановки — это частный вид размещений, когда  $n = m$ :

$$P_n = A_n^n$$

#### 3. Сочетания

Число всех сочетаний  $C_n^m$  из  $n$  элементов по  $m$  (где  $m < n$ ), определяется по формуле:

$$C_n^m = \frac{n!}{m!(n-m)!}$$

Рассмотренные выше комбинации относятся к так называемому **выбору без возвращения** — входящие в их состав элементы не повторяются.

Кроме того, комбинаторика рассматривает и случаи с повторением элементов, входящих в рассматриваемые подмножества — так называемый **выбор с возвращением** — размещения, сочетания и перестановки из  $n$  элементов по  $m$ , в которых некоторые элементы (или все) могут быть одинаковыми.

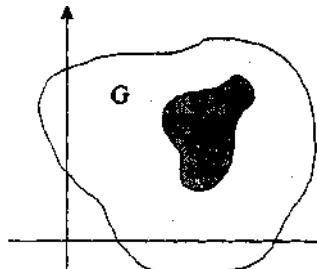
Рассматривая конкретную задачу, необходимо выяснить, каким требованиям удовлетворяют комбинации элементов. Только после этого можно использовать нужные

вычислительные формулы, комбинируя их с правилами сложения и умножения комбинаторики.

#### 4. Геометрическое определение вероятности

Классическое определение вероятности основывается на том числе всех возможных случаев конечно.

Если распределение возможных исходов испытания непрерывно и бесконечно, то при решении задач используется понятие **геометрической вероятности** — вероятности попадания точки в область (отрезок, часть плоскости и т. д.).



При определении геометрической вероятности полагают, что имеется область  $G$  и в ней меньшая область  $g$  с квадрируемой границей. На  $G$  наудачу бросает точка. Событие  $A$  — попадание точки в область  $g$ . Вероятность попадания в какую-либо часть  $G$  пропорциональна мере этой части (обозначим  $\text{mes}$ ) и не зависит от ее расположения и формы.

**Геометрической вероятностью** события  $A$  называется отношение меры области  $g$ , благоприятствующей событию  $A$  к мере всей области  $G$ :

$$P(g) = \frac{\text{mes}(g)}{\text{mes}(G)}.$$

Область, на которую распространяется геометрическая вероятность может быть:

- **одномерной;**
- **двумерной;**
- **трехмерной;**
- **$n$ -мерной.**

При этом вероятность попадания случайно взятой точки в область размерности меньшей, чем  $n$ , например, в границу области, равна нулю.

#### 5. Статистическое определение вероятности

По данным наблюдений рассчитывают отношение  $w_A = \frac{m_A}{n}$ , называемое **частостью** (**относительной частотой**, выборочной долей) события  $A$ , где  $m_A$  называют частотой события  $A$ .

**Статистической вероятностью** события  $A$  называется частость (относительная частота)  $m/n$  появления этого события в  $n$  произведенных испытаниях:

$$P(A) = w = \frac{m}{n},$$

Наблюдаемая частость события  $A$  почти для любой большой серии указанных испытаний мало отличается от некоторой постоянной величины, т. е. проявляется **закон устойчивости** частостей. Поэтому за вероятность случайного события  $A$  принимают на практике либо наблюденную частость, либо число, близкое этому значению, т.е.  $P(A) \approx w_A$ .

## 1.2 Лекция № 2 (2 часа).

Тема: «Повторные независимые испытания» (интерактивная форма)

### 1.2.1 Вопросы лекции:

1. Формула Бернулли
2. Локальная теорема Муавра-Лапласа
3. Интегральная теорема Муавра-Лапласа
4. Теорема Пуассона

### 1.2.2 Краткое содержание вопросов:

#### 1. Формула Бернулли

**Определение:** Если производятся многоократные испытания, в которых вероятность появления события A в каждом испытании не меняется в зависимости от исходов других испытаний, то такие испытания называют *независимыми повторными испытаниями*.

Будем далее рассматривать лишь такие независимые испытания, в которых событие A имеет одну и ту же вероятность.

Пусть производится n независимых испытаний, в каждом из которых событие A может появиться или не появиться. Обозначим вероятность события A – p, т.е.  $P(A) = p$ , тогда вероятность ненаступления события A в каждом испытании также постоянна и равна  $q = 1-p$ .

Под схемой Бернулли понимают проведение серии в n испытаний, в каждом из которых возможны два исхода: либо наступит событие A, либо не наступит, т.е. произойдет противоположное ему событие, и при этом:

- 1) все n испытаний независимы;
- 2) вероятность события A в каждом отдельном испытании постоянна и не меняется от испытания к испытанию.

$$P(A) = p; P(\bar{A}) = 1-p = q.$$

**Теорема.** В случае небольшого числа испытаний n вероятность  $P_n(k)$  того, что в n независимых испытания события A наступит ровно k раз определяется в соответствии с формулой Бернулли:

$$P_n(k) = C_n^k p^k (1-p)^{n-k}$$

или

$$P_n(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

#### 2. Локальная теорема Муавра – Лапласа

**Теорема.** При большом числе испытаний  $n \rightarrow \infty$ , вероятности наступления события A в каждом испытании p, отличной от нуля и единицы, и при выполнении условия  $pqr \geq 20$ , вероятность  $P_n(k)$  того, что в n независимых испытаниях событие A наступит ровно k раз определяется в соответствии с *локальной теоремой Муавра – Лапласа*:

$$P_n(k) \approx \frac{1}{\sqrt{npq}} f(t); \quad t = \frac{k-np}{\sqrt{npq}},$$

где n – число испытаний Бернулли;

k – число испытаний, в которых наступило событие A;

p = P(A) – вероятность наступления события A в каждом испытании;

q = 1-p – вероятность противоположного события ( $\bar{A}$ );

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \text{ - функция Гаусса}$$

Функция Гаусса  $f(t)$  представляет собой плотность стандартного нормального закона распределения (будет рассмотрена далее).

Основные свойства  $f(t)$ , необходимые для применения рассматриваемой теоремы:

- 1)  $f(t)$  – четная функция, т.е.  $f(t) = f(-t)$ ;
- 2)  $f(t)$  – монотонно убывающая функция, т. е.  $f(t) \rightarrow 0$  при  $t \rightarrow \pm\infty$ ; при  $t > 5$  можно считать  $f(t) \approx 0$

### 3. Интегральная теорема Муавра – Лапласа

**Теорема.** При большом числе испытаний  $n \rightarrow \infty$ , вероятности наступления события А в каждом испытании  $p$ , отличной от 0 и 1, и при выполнении условия  $npq \geq 20$ , вероятность  $P_n(k)$  того, что в  $n$  независимых испытаниях событие А наступит от  $a$  до  $b$  раз, определяется в соответствии с интегральной теоремой Муавра – Лапласа:

$$P_n(a \leq k \leq b) \approx \Phi(t_2) - \Phi(t_1);$$

$$t_1 = \frac{a - np}{\sqrt{npq}}, \quad t_2 = \frac{b - np}{\sqrt{npq}},$$

где  $n$  – число испытаний Бернулли;

$k$  – число испытаний, в которых наступило событие А;

$p = P(A)$  - вероятность наступления события А в каждом испытании;

$q = 1-p$  – вероятность противоположного события ( $\bar{A}$ );

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx \text{ – функция Лапласа.}$$

Функция Лапласа  $\Phi(t)$  представляет собой функцию стандартного нормального закона распределения и будет более подробно рассмотрена в теме «Нормальный закон распределения».

Отметим основные свойства  $\Phi(t)$ , необходимые для применения данной теоремы:

- 1)  $\Phi(t)$  – нечетная функция, т. е.  $\Phi(-t) = -\Phi(t)$ .
- 2)  $\Phi(t)$  - монотонно возрастающая функция, т. е.  $\Phi(t) \rightarrow 1$ , при  $t \rightarrow \infty$ ; при  $t > 5$  можно считать  $\Phi(t) \approx 1$  (или 0,5).

#### Задачи, приводящие к интегральной теореме Муавра – Лапласа

Пусть проводится  $n$  – независимых испытаний, в каждом из которых вероятность появления события А постоянна и равна  $p$ ,  $0 < p < 1$ .

1) Поставим задачу: найти вероятность того, что отклонение относительной частоты  $k/n$  от постоянной вероятности  $p$  по абсолютной величине не превышает заданного числа  $\varepsilon \geq 0$ . Другими словами, найдем вероятность осуществления неравенства  $|k/n - p| \leq \varepsilon$ .

$$P\left(\left|\frac{k}{n} - p\right| \leq \varepsilon\right) \approx \Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right);$$

2) Наименьшее число испытаний, которое нужно провести, чтобы с вероятностью, равной  $\beta$ , можно было гарантировать, что частота наступления события А отклониться от вероятности  $p$  не более чем на  $\varepsilon$ :

$$P\left(\left|\frac{k}{n} - p\right| \leq \varepsilon\right) = \beta \rightarrow 2\Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right) = \beta \Rightarrow n = pq \left(\frac{\Phi^{-1}(\beta)}{\varepsilon}\right)^2$$

3) При данной вероятности  $\beta$  и числа испытаний  $n$  границы возможных изменений отклонения частоты наступления события А от вероятности  $p$

$$P\left(\left|\frac{k}{n} - p\right| \leq \varepsilon\right) = \beta \rightarrow \Phi\left(\varepsilon \sqrt{\frac{n}{pq}}\right) = \beta \Rightarrow \varepsilon = \sqrt{\frac{pq}{n}} \Phi^{-1}(\beta);$$

4) Вероятность того, что  $k$  наступлений события А отличается от произведения  $np$  (по модулю) не более чем на величину  $\varepsilon \geq 0$

$$P(|k-np| \leq \varepsilon) \approx \Phi\left(\frac{\varepsilon}{\sqrt{npq}}\right)$$

#### 4. Теорема Пуассона

**Теорема.** При большом числе испытаний  $n \rightarrow \infty$ , постоянной малой вероятности наступления события А в каждом испытании  $p \rightarrow 0$ , и при выполнении условия  $0,1 \leq np \leq 10$ , вероятность  $P_n(k)$  того, что в  $n$  независимых испытаниях событие А наступит ровно  $k$  раз, определяется в соответствии с теоремой Пуассона:

$$P_n(k) \approx \frac{\lambda^k e^{-\lambda}}{k!}; \lambda = np,$$

где  $n$  - число испытаний Бернулли

$k$  - число испытаний, в которых наступило событие А,

$\lambda = np$  - параметр Пуассоновского распределения, называемый еще *средней интенсивностью*.

Значения функции Пуассона  $P_n(k)$  также могут быть определены по таблице «Значение функции Пуассона» при заданных значениях  $k$  и  $\lambda$ .

### 1.3 Лекция № 3 (1 час).

**Тема:** «Дискретная случайная величина (ДСВ)» (интерактивная форма)

#### 1.3.1 Вопросы лекции:

1. Определение дискретной случайной величины. Закон распределения
2. Функция распределения ДСВ
3. Основные числовые характеристики ДСВ

#### 1.3.2 Краткое содержание вопросов:

##### 1. Определение дискретной случайной величины. Закон распределения

*Случайная величина* – это переменная, которая в результате испытания принимает одно из своих возможных значений, причем заранее неизвестно, какое именно, т.к. оно зависит от случая.

*Дискретная случайная величина* – случайная величина, которая принимает конечное или бесконечное, но счетное число отдельных изолированных значений (т.е. их можно перенумеровать натуральными числами).

*Непрерывная случайная величина* – это случайная величина, бесконечное и несчетное множество значений которой есть некоторый интервал (конечный или бесконечный) и она сплошь заполняет этот интервал.

Для определения случайной величины необходимо задать ее закон распределения.

*Законом распределения СВ* называют соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими вероятностями, с которыми случайная величина принимает эти значения. (Др. словами, это соответствие м/у возможными значениями СВ и их вероятностями).

Закон распределения ДСВ можно задать:

1. таблично.
2. графическое представление ряда распределения называется *многоугольником (полиномом) распределения*
3. Третьим способом задания закона распределения ДСВ является интегральная функция распределения или функция накопленных вероятностей.

### 2. Функция распределения ДСВ

*Функция распределения (интегральная функция)  $F(x)$*  определяет для каждого возможного значения  $x$  вероятность того, что случайная величина  $X$  примет значения, меньшее  $x$ :  $F(x) = \sum_{x_i < x} p(X = x_i)$ .

#### Свойства интегральной функции распределения ДСВ

1) Функция распределения может принимать любые значения от 0 до 1, т.к. по определению является вероятностью;

2) Интегральная функция распределения является неубывающей;

3) Функция распределения любой дискретной случайной величины есть разрывная ступенчатая функция, скачки которой происходят в точках, соответствующим возможным значениям случайной величины и равны вероятностям этих значений. Сумма всех скачков равна 1. Эта функция кусочно постоянна на интервалах, на которых нет ее значений.

4) Интегральная функция распределения ДСВ непрерывна слева

Условия непрерывности функции:

$$\lim_{\substack{x \rightarrow x_0 \\ x < x_0}} [F(x) - F(x_0)] = 0$$

$$\lim_{\substack{x \rightarrow x_0 \\ x > x_0}} [F(x) - F(x_0)] = P(X = x_0).$$

5) Вероятность попадания ДСВ в интервал  $[a;b]$  равна приращению функции распределения в этих точках:

$$P(a \leq X \leq b) = F(b) - F(a).$$

6) Если всевозможные значения случайной величины  $X$  принадлежат интервалу  $(x_{min}; x_{max})$ , то

$$F(x) = 0, \text{ при } x \leq x_{min}$$

$$F(x) = 1, \text{ при } x \geq x_{max}$$

7) Если всевозможные значения ДСВ  $X$  расположены на всей числовой оси  $OX$ , то

$$F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0 \quad (\text{как вероятность невозможного события})$$

$$F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = 1 \quad (\text{как вероятность достоверного события})$$

### **3. Основные числовые характеристики ДСВ**

*Математическое ожидание  $M(x)$*  – это число, характеризующее среднее значение случайной величины  $X$ .

*Математическим ожиданием ДСВ* называют сумму произведений всех ее возможных значений на их вероятности, т. е.

$$M(x) = \sum_{i=1}^n x_i p_i = x_1 p_1 + x_2 p_2 + \dots + x_n p_n$$

#### Свойства математического ожидания

1) Математическое ожидание постоянной величины  $C = \text{const}$  равно этой величине:

$$M(C) = C.$$

2) Постоянный множитель можно выносить за знак математического ожидания:

$$M(CX) = C M(X)$$

3) Математическое ожидание алгебраической суммы  $n$  случайных величин  $x_1, x_2, \dots, x_n$  равно сумме математических ожиданий этих СВ:

$$M(X_1 \pm X_2 \pm \dots \pm X_n) = M(X_1) \pm M(X_2) \pm \dots \pm M(X_n).$$

4) Математическое ожидание произведения  $n$  случайных величин  $x_1, x_2, \dots, x_n$  равно произведению математических ожиданий этих СВ:

$$M(X_1 * X_2 * \dots * X_n) = M(X_1) * M(X_2) * \dots * M(X_n).$$

Пусть производится  $n$  независимых испытаний, в каждом из которых вероятность появления события  $A$  постоянна и равна  $p$ .

5) Теорема: Математическое ожидание  $M(x)$  числа появлений события  $A$  в  $n$  независимых испытаниях равно произведению числа испытаний на вероятность появления события в каждом испытании:

$$M(x) = np.$$

Зная математическое ожидание мы не можем судить как значения СВ рассеяны вокруг  $M(X)$ . Другими словами  $M(X)$  полностью не характеризует случайную величину. Рассмотрим показатель дисперсия.

*Дисперсия* характеризует разброс или рассеяние значений СВ около ее математического ожидания.

Дисперсия – это математическое ожидание квадрата отклонения СВ от ее математического ожидания:

$$D(x) = M[X - M(x)]^2.$$

Формула упрощенного вычисления дисперсии имеет вид:

$$D(x) = M(x^2) - (M(x))^2$$

Свойства дисперсии самостоятельно

*Среднее квадратическое отклонение* случайной величины определяется как корень квадратный из дисперсии:

$$\sigma(x) = \sqrt{D(x)}$$

Среднее кв. отклонение было введено как дополнительная характеристика рассеяния значений СВ вокруг ее математического ожидания и, в отличие от дисперсии, она имеет размерность, совпадающую с размерностью СВ.

*Мода*  $M_o(X)$  распределения – это значение СВ, имеющее наиболее вероятное значение.

*Медиана*  $M_e(X)$  – это значение СВ, которое делит таблицу распределения на две части т.о., что вероятность попадания в одну из них равно 0,5

## 1.4 Лекция № 4 (1 час).

**Тема:** «Непрерывная случайная величина (НСВ)» (интерактивная форма)

### 1.4.1 Вопросы лекции:

1. Функция распределения непрерывной случайной величины
2. Функция плотности вероятностей НСВ
3. Основные числовые характеристики НСВ
4. Нормальный закон распределения

### 1.4.2 Краткое содержание вопросов:

#### 1. Функция распределения непрерывной случайной величины

НСВ – это случайная величина, бесконечное и несчетное множество значений, которой есть некоторый интервал (конечный или бесконечный) и она сплошь заполняет этот интервал.

Функция распределения НСВ  $X F(x)$  непрерывна в любой точке и имеет всюду (кроме, возможно, конечного числа точек) непрерывную производную.

**Теорема.** Вероятность любого отдельного взятого значения НСВ равно нулю.

#### Свойства интегральной функции распределения НСВ

1. Функция распределения может принимать значения от 0 до 1, т.к. определяется вероятностью

$$0 \leq F(x) \leq 1$$

2. Интегральная функция распределения является неубывающей  
 $F(x_2) \geq F(x_1)$ , если  $x_2 > x_1$

3. Для НСВ (согласно теореме)

$$P(x_1 \leq X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X \leq x_2) = P(x_1 < X < x_2)$$

4. Вероятность попадания НСВ в интервал  $[x_1; x_2]$  равна приращению функции распределения в этих точках:

$$P(x_1 \leq X < x_2) = F(x_2) - F(x_1)$$

5. Если все возможные значения СВ  $x$  принадлежат интервалу  $(x_1; x_2)$ , то

$$F(x) = 0, \text{при } x \leq x_1$$

$$F(x) = 1, \text{при } x > x_2$$

6. Если все возможные значения НСВ  $x$  расположены на всей числовой оси  $OX$ , то

$$\lim_{x \rightarrow -\infty} F(x) = 0$$

$$\lim_{x \rightarrow +\infty} F(x) = 1.$$

## 2. Функция плотности вероятностей НСВ

Рассмотрим вероятность попадания случайной точки на элементарный участок  $[x; \Delta x]$  длины  $\Delta x$  НСВ  $X$ , имеющей непрерывную и дифференцируемую функцию распространения  $F(x)$  на этом участке.

По 4 свойству функции распределения:

$$P(x \leq X < x + \Delta x) = F(x + \Delta x) - F(x).$$

Определим теперь отношение этой вероятности к длине участка, т.е среднюю вероятность, приходящуюся на единицу длины рассматриваемого участка, и рассмотрим предел при  $\Delta x \rightarrow 0$

$$\lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x).$$

Эта функция, характеризующая плотность, с которой распределяются значения НСВ в точке, и была названа функцией плотности распределения или функцией плотности вероятностей.

*Плотность вероятности (плотностью распределения дифференциальной функцией) СВ  $x$  называется  $f(x)$ , являющаяся первой производной интегральной функции распределения*

$$f(x) = F'(x)$$

Под элементом вероятности для СВ  $x$  понимается величина  $f(x)dx$ , с точностью до бесконечно малых величин высшего порядка отражающая вероятность попадания случайной точки  $x$  в элементарный отрезок  $dx$ , примыкающий к точке  $x$ .

### Свойства функции плотностей вероятностей:

1. Функции плотности вероятностей принимает только неотрицательные значения как производная неубывающей функции распределения  $F(x)$ :

$$f(x) \geq 0$$

2. Вероятность попадания НСВ  $x$  в интервал  $x_1$  до  $x_2$  равна определенному интегралу от функции плотности вероятностей в этих пределах:

Зная плотность распределения, можно вычислить вероятность того, что НСВ примет значение, принадлежащее интервалу  $(x_1, x_2)$ .

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) \quad \stackrel{\substack{\text{по формуле Ньютона-} \\ \text{Лейбница}}}{=} \quad \int_{x_1}^{x_2} F'(x) dx = \int_{x_1}^{x_2} f(x) dx$$

Таким образом, геометрически полученный результат можно истолковать так: вероятность того, что НСВ примет значение принадлежащее интервалу  $(a, b)$ , равна площади криволинейной трапеции, ограниченной осью  $OX$ , кривой распределения  $f(x)$  и прямыми  $x=a$ ,  $x=b$ .

**Определение:** График плотности распределения называют кривой распределения.

Т.о. зная плотность распределения, можно найти функцию распределения.

3. Функция распространения НСВ равна интегралу от функции плотности вероятностей в пределах от  $-\infty$  до  $x$ :

4. Интеграл в бесконечных пределах от функции плотности вероятностей равен 1. (как сумма вероятностей всех возможных значений СВ  $x$ ):

Геометрически: вся площадь криволинейной трапеции ограниченная осью ОХ и кривой распределения равна 1.

### 3.Основные числовые характеристики НСВ

1. Математическое ожидание НСВ:

$$M(x) = \int_{-\infty}^{+\infty} x f(x) dx$$

2. Дисперсия НСВ:

$$D(x) = \int_{-\infty}^{+\infty} ((x - M(x))^2 f(x) dx$$

3. Среднее квадратичное отклонение:

$$\sigma_x = \sqrt{D(x)}$$

### 4. Нормальный закон распределения

Главная особенность нормального закона распределения состоит в том, что он является *пределальным законом*, к которому с ростом числа наблюдений стремятся другие распределения.

**Определение.** НСВ  $X$  имеет нормальный закон распределения (закон Гаусса) с параметрами  $\mu$  и  $\sigma$  (обозначают  $X \sim N(\mu; \sigma)$ ), если ее плотность вероятности имеет вид:

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

$\mu$ - математическое ожидание  $X$ ;

$\sigma^2$  – дисперсия  $X$ ;

$\sigma$  – среднее квадратичное отклонение.

*Свойства функции плотности вероятности нормального закона распределения:*

1.  $f(x) > 0$  существует при любых действительных  $x$ ;

2.  $f(x) \xrightarrow{x \rightarrow \pm\infty} 0$ ;

3. Кривая плотности нормального закона распределения симметрична относительно прямой  $x=\mu$ ;

4. Максимальное значение  $f(x)$  принимает в т.  $x_0 = \mu$ , при этом  $f(x) = \frac{1}{\sigma\sqrt{2\pi}}$ ;

5. Кривая плотности нормального закона распределения имеет 2 точки перегиба с координатами  $(\mu \pm \sigma; \frac{1}{\sigma\sqrt{2\pi}})$

Если  $\sigma = \text{const}$ , и меняется параметр  $\mu$ , т.е. центр симметрии распределения, то нормальная кривая будет смещаться: вдоль оси абсцисс, не меняя формы.

Если  $\mu = \text{const}$  и меняется  $\sigma$ , то меняется ордината максимальная кривой -  

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}.$$

Таким образом, параметр  $\mu$  характеризует положение центра, а параметр  $\sigma$  - форму кривой плотности вероятности.

*Функция распределения случайной величины, распределенной по нормальному закону.*

Определение: Функция распределения СВ  $X$ , имеющее нормальный закон распределения с параметром  $\mu$  и  $\sigma$ , определяется по формуле:  $F(x) = \frac{1}{2} + \Phi(t)$ ,  $t = \frac{x-\mu}{\sigma}$   
 $\Phi(t)$ - функция Лапласа

Геометрически функция распределения представляет собой площадь под нормальной кривой на интервале  $(-\infty; x)$ . Как видим, она состоит из 2-х частей: первой -  $(-\infty; \mu)$  равной  $\frac{1}{2}$  и второй, на интервале  $(\mu; x)$ , равной  $\frac{1}{2}\Phi(t)$ .

Определение: Нормальный закон распределения СВ  $X$  с параметром  $\mu = 0, \sigma = 1$  ( $N(0;1)$ ) называется *стандартным или нормальным*.

*Свойства СВ, распределенной по нормальному закону распределения*

1. Если равна  $X \subset N(\mu; \sigma)$ , то вероятность попадания СВ  $X$  в интервал  $[x_1; x_2]$  равна

$$P(x_1 \leq x \leq x_2) = \Phi(t_2) - \Phi(t_1),$$

$$t_1 = \frac{x_1 - \mu}{\sigma}, \quad t_2 = \frac{x_2 - \mu}{\sigma}$$

2. Вероятность того, что отклонение случайной величины  $X$ , от её математического ожидания  $\mu$  не превысит величину  $\varepsilon > 0$  (по абсолютной величине), равна

$$P(|X - \mu| \leq \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right).$$

Если заменить  $\varepsilon$  на  $\sigma, 2\sigma, 3\sigma$ , то получим «правило трех сигм»

3. Правило трех сигм

Если случайная величина  $X \subset N(\mu; \sigma)$ , то практически достоверно, что её значения заключены в интервале  $(\mu - 3\sigma, \mu + 3\sigma)$  (Вероятность выброса равна 0,0027)

## 1.5 Лекция № 5 (2 часа).

**Тема: «Статистическое оценивание параметров распределения»** (интерактивная форма)

### 1.5.1 Вопросы лекции:

1. Общие сведения о выборочном методе
2. Понятие оценки параметров.
3. Вариационный ряд и его основные характеристики
4. Методы нахождения оценок
5. Интервальные оценки параметров генеральной совокупности

### 1.5.2 Краткое содержание вопросов:

#### 1. Общие сведения о выборочном методе

Объектом изучения математической статистики являются генеральные совокупности, которые исследуются на основе выборки. Задачи математической

статистики практически сводятся к обоснованному суждению об объективных свойствах генеральной совокупности по результатам случайной выборки.

Под *генеральной совокупностью* понимается совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий.

Понятие генеральной совокупности в определенном смысле аналогично понятию случайной величины (закону распределения вероятностей), так как полностью обусловлены определенным комплексом условий.

*Выборочной совокупностью или выборкой* называется та часть объектов, которая отобрана для непосредственного изучения из генеральной совокупности.

Число объектов в выборке или в генеральной совокупности называется их объемом.

Генеральная совокупность может иметь как конечный так и бесконечный объем.

Выборку можно рассмотреть как некий эмпирический аналог генеральной совокупности.

Для эффективного применения математико-статистических методов анализа необходимо, чтобы выборочная совокупность отражала структуру и особенности распределения признаков генеральной совокупности. Таким образом, выборка должна быть *репрезентативной*. Это достигается случайностью отбора, когда все объекты генеральной совокупности имеют одинаковую вероятность быть отобранными.

Ошибки репрезентативности всегда имеют место быть, однако они могут быть заранее оценены и сведены к минимуму посредством правильной организации выборки.

Различают следующие виды выборок: собственно-случайная, типическая, механическая, серийная.

Способы образования выборки: повторный отбор, бесповторный отбор.

Важнейшей задачей выборочного метода является оценка параметров (характеристик) генеральной совокупности по данным выборки.

Теоретическую основу применимости выборочного метода составляет закон больших чисел, согласно которому при неограниченном увеличении объема выборки практически достоверно, что случайные выборочные характеристики как угодно близко приближаются (сходятся по вероятности) к определенным параметрам генеральной совокупности.

## 2. Понятие оценки параметров

Пусть распределение признака  $X$  – генеральной совокупности описывается законом распределения  $F(x, \Theta)$ , который содержит известный параметр  $\Theta$ . (например :  $\lambda$ , в случае распределения Пуассона или экспоненциальном распределении, или  $\bar{x}$  и  $\sigma$  – нормального распределения). Для вычисления параметра  $\Theta$  исследовать все элементы генеральной совокупности не предоставляет возможным, поэтому извлекается из генеральной совокупности выборка  $x_1, x_2, \dots, x_n$ , и по ее результатам судят о параметре  $\Theta$ .

**Определение:** Оценкой параметра  $\Theta$  называют некоторую функцию результата наблюдений  $\tilde{\theta}_n = \tilde{\theta}_n(x_1, x_2, \dots, x_n)$  с помощью которой судят о значении параметра  $\Theta$ .

Поскольку  $x_1, x_2, \dots, x_n$  – случайные величины то и оценка  $\tilde{\theta}_n$  (в от оцениваемого параметра  $\Theta$ - величины неслучайной, детерминированной) является случайной величиной, зависящей от закона распределения случайной величины  $X$  и числа наблюдений  $n$ .

Основная задача теории оценивания состоит в том, чтобы произвести выбор оценки  $\tilde{\theta}_n$  параметра  $\Theta$ , позволяющей получить наилучшее приближение к оцениваемому параметру.

**Например:** если  $\Theta$  - математическое ожидание СВ  $X$  т.е генеральной средней  $\bar{x}_0$  то в качестве оценки  $\tilde{\theta}_n$  можно взять выборочную среднюю,  $M_0, M_e, (x_{\min} + x_{\max}) / 2$  и т.д.

Назвать «наилучшей» оценкой такую, которая наиболее близка к истинному значению оцениваемого параметра, невозможно, т.к.  $\tilde{\theta}_n$  – случайная величина, поэтому невозможно предсказать индивидуальное значение оценки в данном частном случае. Так что о качестве оценки следует судить не по индивидуальным ее значениям, а лишь по распределению ее значений в большом количестве испытаний, т.е по выборочному распределению оценки.

Выбор той или иной функции в качестве оценки параметра  $\Theta$  проводиться с учетом удовлетворения следующих требований.

#### Свойства точечных оценок :

1. *Несмещенност*. Статистическая оценка  $\tilde{\theta}_n$  параметра называется несмещенной, если ее математическое ожидание равно оцениваемому параметру, т.е.  $M(\tilde{\theta}_n) = \Theta$

Если оценка является смещенной, то смещение определяется как  $B_n = M(\tilde{\theta}_n) - \Theta$

Требования несмещенности гарантирует отсутствие систематических ошибок при оценивании ( $M(\tilde{\theta}_n) > \Theta$  или  $M(\tilde{\theta}_n) < \Theta$ )

2. *Состоятельность*. Оценка  $\tilde{\theta}_n$  параметра  $\Theta$  называется состоятельной, если она удовлетворяет закону больших чисел, т.е сходится по вероятности к оцениваемому параметру  $\Theta$  :

$$\lim_{n \rightarrow \infty} P(|\tilde{\theta}_n - \theta| < 1) = 1$$

или

$$\tilde{\theta}_n \xrightarrow{n \rightarrow \infty} \theta$$

В данном случае оправдывается увеличение объема выборки.

3. *Эффективность*. Несмещенная оценка  $\tilde{\theta}_n$  параметра  $\Theta$  называется эффективной, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра  $\Theta$ , вычисленных по выборкам одного и того же объема  $n$ .

В качестве статистических оценок параметров генеральной совокупности желательно использовать оценки, удовлетворяющие одновременно 3 –м требованиям. Однако достичь этого удается не всегда. Может оказаться, что для простоты расчетов целесообразно использовать незначительно смещенные оценки или оценки, обладающие большой дисперсией по сравнению с эффективными оценками.

#### Основные точечные оценки

Наиболее часто использующиеся точечные оценки представлены в таблице 1.

Таблица 1. – Точечные оценки основных параметров распределения

Оцениваемый параметр генеральной совокупности	Оценка его выборочная точечна	
	Простая выборка	Сгруппированные данные
Генеральная средняя или математическое ожидание $\mu$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{\sum n_i}$
Генеральная дисперсия $\sigma^2$	$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ $S^2 = \bar{x}^2 - (\bar{x})^2$	$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 n_i}{\sum n_i}$ $S^2 = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i - \bar{x}$

Исправленная дисперсия	$\hat{S} = \frac{n}{n-1} S^2$	
Генеральное среднее квадратическое отклонение $\sigma$	$S = \sqrt{S^2}$	$S = \sqrt{\hat{S}^2}$
Начальные моменты $l$ -го порядка	$\tilde{V}_k = \frac{1}{n} \sum_{i=1}^n x_i^l$	$\tilde{V}_k = \frac{1}{n} \sum_{i=1}^k x_i^l n_i$
Центральные моменты $l$ -го порядка	$\tilde{\mu}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^l$	$\tilde{\mu}_k = \frac{\sum (x_i - \bar{x})^l n_i}{\sum n_i}$

### 3. Вариационные ряды и их основные характеристики

Результаты наблюдений – это, в общем случае, ряд чисел, расположенных в беспорядке, который для изучения необходимо упорядочить (проранжировать).

Операция, заключенная в расположении значений признака по возрастанию, называется ранжированием данных.

После операции ранжирования данных можно сгруппировать, т.е разбить варианты на отдельные интервалы.

1) Если признак – непрерывная СВ.

Число интервалов  $k$  следует брать не очень большим, чтобы после группирования ряд не был громоздким, и не очень малым, чтобы не потерять особенности распределения признака. Определяется число интервалов (групп).

Согласно функциям Стерджесса (в случае нормально распространенной совместимости) рекомендуемое число интервалов  $k=1+3,32 \lg n$

$h$  - величина интервала (ширина интервала)

$$h = \frac{X_{\max} - X_{\min}}{k}$$

Образуя интервалы группировки, путем последовательного прибавления к нижней границе интервала его длины  $h$ , начиная с  $X_{\min}$ .

Отнесение той или иной единицы наблюдения к образовательным группировкам и подсчет итогов по группе и в целом по совместимости.

Числа, показывающие, сколько раз встречаются варианты из данного интервала называются **частотами** (обозначаем  $m_i$ ), а отношение их к общему числу наблюдений –

$W = \frac{m_i}{n}$ . Частоты и частности называются весами.

**Вариационным рядом называется** ранжированный в порядке возрастания (или убывания) ряд вариантов с соответствующими им весами (частотами или частностями).

2) Если Х-ДСВ, то число групп определяется количеством вариантов.

Различные значения признака (СВХ) называются **вариантами** (обозначаются ч/з х).

**Дискр. вариационным рядом называется** вариационный ряд, варианты которого представлены в виде ДСВ.

**Непрерывным вариационным рядом называется** такой вариационный ряд, варианты которого представлены в виде интервалов.

Графическое изображение вариационных рядов

**Полигон** – ломанная, отрезки которой соединяют точки  $(x_1, m_1), (x_2, m_2), \dots, (x_k, m_k)$ . Для относительных частот аналогично с заменой  $m_i$  на  $\frac{m_i}{n}$ .

**Гистограмма** – ступенчатая фигура, состоящая из прямоугольников с основаниями, равными интервалам значений признака и высотами, равными частотам (частностям) интервалов.

Вариационный ряд является статистическим аналогом (реализацией) распределения признака (СВХ). В этом смысле полигон (гистограмма) аналогичен кривой распределения, а эмпирическая – функции распределения СВХ).

Основными характеристиками вариационного ряда являются средняя величина, медиана, мода, дисперсия, среднее квадратическое отклонение.

**Медианой Me** ВР называется значение признака, приходящийся на середину ранжированного ряда наблюдений.

1. Определяется номер варианта, приходящийся на середину

$$N = \frac{n+1}{2}$$

2. Рассчитываются накопленные частоты варианта, накопленной частоты, который первый раз превысила N- является Me

Для интервального ряда варианта-интервал.

$$Me = \frac{x_{0+h} - 0,5h - 5M_{e-1}}{m_M e}$$

**Модой Mo** вариационного ряда называется вариант, которому соответствует наибольшая частота.

$$M_0 = x_0 + h \frac{[(m]_{M_0} - m_{M_{0-1}})}{[(m]_{M_0} - m_{M_{0-1}}) + (m_{M_0} - m_{M_{0+1}})}$$

Ме и Мо не изменяются от изменений значений крайних членов ряда.

К относительным показателям вариации относят: коэффициент осцилляции, линейный коэффициент вариации, относительное линейное отклонение и др.

**Размах вариации R.** Определяется как разность между самым большим и самым малым значениями признака у единиц данной совокупности:

$$R = X_{\max} - X_{\min}$$

Размах вариации (размах колебаний) - важный показатель колеблемости признака, но он дает возможность увидеть только крайние отклонения, что ограничивает область его применения.

Для более точной характеристики вариации признака на основе учета его колеблемости используются другие показатели.

**Среднее квадратическое отклонение  $\sigma$**  и **среднее квадратическое отклонение в квадрате  $\sigma^2$** , которое называют дисперсией.

Средняя квадратическая простая

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Средняя квадратическая взвешенная

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}}$$

Формулы дисперсии взвешенной  $(\sigma_{\text{зв}}^2)$  и простой  $(\sigma_{\text{пр}}^2)$ :

$$\sigma_{\text{зв}}^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}; \quad \sigma_{\text{пр}}^2 = \frac{\sum (x_i - \bar{x})^2}{n}.$$

Данные показатели рассчитываются как отношение размаха вариации к средней величине признака (коэффициент осцилляции), отношение среднего линейного отклонения к средней величине признака (линейный коэффициент вариации), отношение среднего квадратического отклонения к средней величине признака (коэффициент вариации) и, как правило, выражаются в процентах.

Формулы расчета относительных показателей вариации:

$$V_R = \frac{R}{\bar{x}} \cdot 100\%, \quad V_d = \frac{\bar{d}}{\bar{x}} \cdot 100\%, \quad V_s = \frac{\sigma}{\bar{x}} \cdot 100\%,$$

где  $V_R$  - коэффициент осцилляции;  $V_d$  - линейный коэффициент вариации;  $V_s$  - коэффициент вариации.

#### 4. Методы нахождения оценок

1. *Метод моментов*, предложенный К. Пирсоном: определенное количество выборочных моментов (начальных  $\tilde{V}_k$  или центральных  $\tilde{M}_k$ , или тех и других) приравнивается к соответствующим теоретическим моментам распределения ( $V_k$  или  $M_k$ ) случайной величины  $X$ .

Оценки метода моментов обычно состоятельны, однако по эффективности они не являются «наилучшими», их эффективности  $e(\theta_n)$  часто значительно ниже единицы. Метод моментов часто используются на практике, т.к приводит к сравнительно простым вычислениям.

#### 2. *Метод наибольшего правдоподобия* (Р. Фишер)

Пусть  $x_1, x_2, \dots, x_n$  - случайная выборка из генеральной совокупности  $X$ ,  $f(x; \Theta_1; \Theta_2; \dots; \Theta_k)$  - функция вероятностей (для ДСВ) или плотность (для НСВ).

$\Theta_1; \Theta_2; \dots; \Theta_k$  параметры закона распределения, подлежащие оцениванию по случайной выборке.

МНП состоит в том, что в качестве оценки неизвестного параметра  $\theta$  распределения СВ  $X$  выбирается то его значение, при котором *полученное значение  $x_1, x_2, \dots, x_n$  выборки имеет наибольшую вероятность* (в случае ДСВ) или *плотность* (в случае НСВ).

*Функция правдоподобия наблюдений  $x_1, x_2, \dots, x_n$  представляет собой:*

- В случае ДСВ – вероятность получить в качестве первого элемента выборки значение  $x_1$ , второго – значение  $x_2$ , ...  $n$ -го значение  $x_n$ , т.е

$$L(x_1, x_2, \dots, x_n; \Theta_1; \Theta_2; \dots; \Theta_k) = P(X_1=x_1, X_2=x_2, \dots, X_n=x_n; \Theta_1; \Theta_2; \dots; \Theta_k)$$

т.к. наблюдения независимы, то  $L(x_1, x_2, \dots, x_n; \Theta_1; \Theta_2; \dots; \Theta_k)$

$$= \prod_{i=1}^n P(x_i; \Theta_1; \Theta_2; \dots; \Theta_k).$$

- В случае НСВ – совместная  $n$  мерная плотность вероятности, описывающая закон распределения вероятности  $n$  наблюдений.

Т.к. отдельные наблюдения независимы, то  $L(x_1, x_2, \dots, x_n; \Theta_1; \Theta_2; \dots; \Theta_k) = \prod_{i=1}^n F(x_i; \Theta_1; \Theta_2; \dots; \Theta_k)$

За оценки наибольшего правдоподобия принимают такие значения  $\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k$ , которые максимизируют функцию правдоподобия из условия:

$$L(x_1, x_2, \dots, x_n; \tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k) = \max L(x_1, x_2, \dots, x_n; \Theta_1; \Theta_2; \dots; \Theta_k)$$

Т.е. оценки  $\hat{\theta}_k$  таковы, что имеющиеся наблюдения  $x_1, x_2, \dots, x_n$  являются наиболее правдоподобными.

Для упрощения вычислений максимизируют  $\ln L$ . Для отыскания оценки параметров  $\hat{\theta}_k$  надо решить (уравнение) систему правдоподобия, получаемую приравниваем частных производных нулю по параметрам

$$\frac{d \ln L}{d \theta} = 0$$

Найденную т. максимума  $\hat{\theta}_k$  принимают в качестве оценки наибольшего правдоподобия параметра  $\Theta$ .

Метод дает оценки, которые состоятельны (смещенными) распределены асимптотически нормально (при больших значениях  $n$  приближенно нормальны) и имеют наименьшую дисперсию по сравнению с другими оценками. Особенно полезен метод для малых выборок.

### 3. Метод наименьших квадратов

Оценка определяется из условия минимизации суммы квадратов отклонений выборочных данных от определяемой оценки.

Например, для нахождения оценки генеральной средней  $\theta = \bar{x}_0$  по методу наименьших квадратов  $\hat{\theta}_n$ .

Преимущества: 1) не требует знание закона распределения выборочных данных; 2) достаточно хорошо разработан в плане вычислительной реализации.

### 5. Интервальные оценки параметров генеральной совокупности

Пусть  $\hat{\theta}$  – точечная оценка неизвестного параметра  $\theta$  генеральной совокупности. Однако  $\hat{\theta}$  – это лишь приближенное значение  $\theta$  и для выборки малого объема может существенно отличается от  $\theta$ .

Чтобы получить представление о точности и надежности оценки  $\hat{\theta}_n$  параметра  $\theta$  используют интервальную оценку параметра.

Определение: Доверительным интервалом  $[\hat{\theta} - \Delta; \hat{\theta} + \Delta]$  для параметра  $\theta$  называется такой интервал, относительно которого можно утверждать с определенной, близкой к единице, вероятностью  $\gamma$ , что он содержит неизвестное значение параметра  $\theta$ .

Также доверительный интервал называют интервальной оценкой параметра  $\theta$ .

Верить  $\gamma$  называется доверительной вероятностью, уровнем доверия или надежностью оценки.

Наибольшее отклонение  $\Delta$  оценки  $\hat{\theta}$  от оцениваемого параметра  $\theta$ , которое возможно с заданной доверительной вероятностью  $\gamma$  называется предельной ошибкой выборки.

Ошибка  $\Delta$  является ошибкой репрезентативности (случайной).

Общий подход к построению интервальной оценки параметра  $\theta$  на основе случайной выборки  $x_1, x_2, \dots, x_n \in X$  генеральная совокупность состоит в определении границ интервала удовлетворяющего условию:

$$P\{\theta_{\min} \leq \theta \leq \theta_{\max}\} = \gamma.$$

#### 1. Построение доверительного интервала для генеральной средней.

а) При известной дисперсии  $\sigma^2$  – генеральной совокупности.

Будем рассматривать  $\bar{x}$  как случайную величину  $\bar{X}$  ( $\bar{x}$  изменяется от выборки к выборке) определяется по выборке.

Рассмотрим  $n$  независимых выборок одинаково нормально распределенных т. е.  $x_1, x_2 \dots, x_n$ , с математическим ожиданием  $M(x) = \bar{x}_0$  и средним квадратическим отклонением  $\sigma$ .

Вероятность того, что отклонение выборочной средней от генеральной средней не превзойдет число  $\Delta > 0$  (по абсолютной величине) равна:

$$P(|\bar{X} - \bar{x}_0| < \Delta) = \Phi\left(\frac{\Delta}{\sigma}\right) = \gamma$$

Если определить  $t$  (с заданной надежностью от вероятности), то доверительный интервал можно определить:

$$\bar{x} - \Delta \leq \mu \leq \bar{x} + \Delta$$

Из (\*) видно, что: 1) при возрастании  $n$ ,  $\Delta$  уменьшается, и, следовательно, точность оценки увеличивается;

2) при увеличении надежности  $\gamma$ ,  $t$  увеличивается и увеличивается  $\Delta$ .

Поясним смысл, который имеет заданная надежность (вероятность). Надежность, (вероятность)  $\gamma = 0,95$  указывает, что если проведено достаточно большое число выборок, то 95 % из них определяет такие доверительные интервалы, в которых параметр действительно заключен; лишь в 5% случаев он может выйти за границы доверительного интервала.

б) Доверительный интервал для  $\mu$  при неизвестной дисперсии  $\sigma^2$ .

$$\boxed{\bar{x} - \Delta \leq \mu \leq \bar{x} + \Delta, \Delta = t_{\alpha} \frac{s}{\sqrt{n-1}}}$$

$t_{\alpha}$  - значение обратной функции распределения Стьюдента (t-распределение) соответствующее  $v = n - 1$  степеням свободы и вероятности  $\gamma$

$$t_{\alpha} = St^{-1}(\alpha = 1 - \gamma; V = n - 1)$$

Определение: Число степеней свободы  $v$  определяется как общее число  $n$  наблюдений (вариантов) случайной величины  $X$  минус число уравнений  $l$  связывающих эти наблюдения, т. е.  $V = n - l$ .

$$\text{Например: } t = \frac{\bar{x} - \mu}{s} \cdot \sqrt{n-1} \quad V = n - 1$$

$$\text{Наблюдения связаны одним уравнением } \bar{x} = \frac{\sum x_i}{n}$$

## 2) Интервальные оценки доли или вероятности $p$

Пусть в  $n$  независимых испытаний некоторое событие  $A$ , вероятность появления которого в каждом испытании равна  $p$ , поступило  $m$  раз;  $0 \leq m \leq n$

Точечная оценка генеральной доли используется  $w = \frac{m}{n}$  – частность.

а) Доверительный интервал для  $p$  при достаточно больших  $n (n > 30)$

$$q = 1 - p \quad \Delta_w = t_{\gamma} \sqrt{\frac{p(1-p)}{n}} = t_{\gamma} \sqrt{\frac{\sigma_p}{n}}$$

б) – по малой выборке.

Если доля признака в генеральной совокупности равна  $p$ , то вероятность того, что в повторной выборке объема  $n$   $m$  элементов обладают этим признаком определяется по формуле Бернулли:

$$P_n(m) = C_n^m p^m q^{n-m}, \quad q = 1 - p$$

т. к. при  $p \neq 0,5$  биноминальное распределение несимметрично, то в качестве доверительного интервала для  $p$  берут такой интервал  $(p_1; p_2)$

$$\sum_{i=m}^n C_n^i p_1^i (1-p_1)^{n-i} = \frac{\gamma}{2}; \quad \sum_{i=1}^m C_n^i p_2^i (1-p_2)^{n-i} = \frac{\gamma}{2}$$

который решается приближенно.

**Теорема:** Вероятность того, что отклонение выборочной средней и выборочной средней (или доли) от генеральной средней (доли) не произойдет число  $\Delta > 0$  (по абсолютной величине), равно

$$P(|\bar{x} - \bar{x}_0| \leq \Delta) = \Phi(t) = \gamma$$

$$t = \frac{\Delta}{\sigma_x} \quad t = \frac{\Delta}{\sigma_w}$$

**Определение:** Среднее квадратическое отклонение выборочной средней  $\sigma_{\bar{x}}$  и выборочной доли  $\sigma_w$  собственно-случайной выборки называется средней квадратической (стандартной) ошибкой выборки.

$$\Delta = t \cdot \sigma_x, \quad \Delta = t \cdot \sigma_w$$

Если задать  $t$ , то интервальные оценки (доверительные интервалы) по формулам

$$\bar{x} - \Delta \leq \bar{x}_0 \leq \bar{x} + \Delta$$

$$w - \Delta \leq p \leq w + \Delta$$

Таблица 2 – Формулы средних квадратических ошибок выборки

Оцениваемый параметр	Повторная выборка	Бесповторная выборка
Средняя	$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} \approx \sqrt{\frac{s^2}{n}}$	$\sigma_{\bar{x}} \approx \sqrt{\frac{s^2}{n}(1 - \frac{n}{N})}$
Доля	$\sigma_w = \sqrt{\frac{pq}{n}} \approx \sqrt{\frac{w(1-w)}{n}}$	$\sigma_w = \sqrt{\frac{w(w-1)}{n}(1 - \frac{n}{N})}$

## 1.6 Лекция № 6 (1 час).

Тема: «Корреляционный анализ» (интерактивная форма)

### 1.6.1. Вопросы лекции:

1. Функциональная, статистическая и корреляционная зависимости
2. Сущность и задачи корреляционно-регрессионного анализа
3. Коэффициент корреляции, его расчет
4. Корреляционное отношение и индекс корреляции

### 1.6.2 Краткое содержание вопросов:

1. Функциональная, статистическая и корреляционная зависимости.

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины  $Y$  от одной или нескольких других величин. Рассмотрим зависимость  $Y$  от одной случайной (или неслучайной) величины  $X$ .

Две случайные величины могут быть связаны либо функциональной зависимостью, либо зависимостью другого рода, называемой статистической, либо быть независимыми.

**Статистической (или стохастической, вероятностной)** называют зависимость, при которой изменение одной из величин влечет изменение распределения другой.

Статистическая зависимость между двумя переменными при которой каждому значению одной переменной соответствует определенное (среднее значение) условное математическое ожидание другой называется **корреляционной**.

Возникновение понятия статистической связи объясняется тем, что зависимая переменная подвержена влиянию ряда неконтролируемых или неучтенных факторов, а также тем, что измерение значений переменных неизбежно сопровождается случайными ошибками.

## 2. Сущность и задачи корреляционно-регрессионного анализа

Статистические связи между переменными можно изучать методами корреляционного и регрессионного анализа.

Корреляционный анализ является одним из методов статистического анализа взаимозависимости нескольких признаков. В настоящее время он определяется как метод, применяемый тогда, когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

*Основная задача корреляционного анализа* — выявление связи между случайными переменными и оценка ее тесноты.

*Основная задача регрессионного анализа* — установление формы и изучение зависимости между переменными.

## 3. Коэффициент корреляции, его расчет

Рассмотрим генеральную совокупность с двумя признаками  $X$  и  $Y$ , совместное распределение которых задано плотностью двумерного нормального закона распределения:

$$f(X, Y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[ -\frac{1}{2\sqrt{1-\rho^2}} \left( \left( \frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \left( \frac{x-\mu_x}{\sigma_x} \right) \left( \frac{y-\mu_y}{\sigma_y} \right) + \left( \frac{y-\mu_y}{\sigma_y} \right)^2 \right) \right]$$

определенного пятью параметрами:  $MX = \mu_x$ ,  $MY = \mu_y$ ,  $DX = \sigma_x^2$ ,  $DY = \sigma_y^2$ ,  $\rho$ .

**Парный коэффициент корреляции  $\rho$**  — показатель, который характеризует тесноту линейной связи между случайными величинами  $X$  и  $Y$  и определяется как:

$$\rho = M\left[ \frac{X - \mu_x}{\sigma_x} \cdot \frac{Y - \mu_y}{\sigma_y} \right].$$

*Свойства парного коэффициента корреляции  $\rho$ :*

- 1)  $-1 < \rho < 1$  —  $\rho$  принимает значения на отрезке  $[-1; 1]$ , при этом, чем ближе  $|\rho|$  к единице, тем теснее связь между признаками  $X$  и  $Y$ ;
- 2) при  $|\rho| = 1$  корреляционная связь между признаками  $X$  и  $Y$  представляет собой линейную функциональную зависимость ( $y = f(x)$ );
- 3) при  $\rho = 0$  линейная корреляционная связь между  $X$  и  $Y$  отсутствует (но это не означает невозможность наличия между ними нелинейной связи);
- 4)  $\rho < 0$  указывает на наличие обратной зависимости между переменными  $X$  и  $Y$  (при увеличении одной переменной другая уменьшается);

- 5)  $p > 0$  указывает на наличие прямой зависимости между переменными X и Y (при увеличении (уменьшении) одной переменной другая тоже возрастает (уменьшается));  
 6) если все значения признаков увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

**Коэффициент детерминации  $\rho^2$**  (равный квадрату коэффициента корреляции) указывает долю дисперсии одной случайной величины, обусловленную вариацией другой.

Соответственно,  $(1 - \rho^2)$ , показывает долю дисперсии случайной величины, объясняемой неключенными в рассматриваемую двумерную модель факторами.

Пусть из генеральной совокупности (X, Y) взята случайная выборка объемом n:  $(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$ .

1. Если объем выборки n невелик, то статистические характеристики генеральной совокупности вычисляются непосредственно по ряду наблюдений  $(x_v, y_v)$ .

$x_1$	$x_2$	...	$x_i$	...	$x_n$
$y_1$	$y_2$	...	$y_i$	...	$y_n$

n- объем выборки.

2. Если выборка из генеральной совокупности велика, то ряд наблюдений преобразовывается к двумерному вариационному ряду, представляемому в виде таблицы, называемой корреляционной.

В первой строке в возрастающем порядке расположены варианты  $x_i$ , а в первом столбце — варианты  $y_i$ . На пересечении столбца  $x_i$ - и строки  $y_i$  находится частота  $m_{ij}$ , обозначающая число точек выборки, значения признаков у которых равны  $(x_i, y_j)$ , где  $i = 1, 2, \dots, k, j = 1, 2, \dots, l$ .

	$x_1$	...	$x_i$	...	$x_k$	Итого по y ( $m_y$ )
$y_1$	$m_{11}$	...	$m_{i1}$	...	$m_{kl}$	$m_{*1}$
...	...	...	...	...	...	...
$y_j$	$m_{1j}$	...	$m_{ij}$	...	$m_{kj}$	$m_{*j}$
...	...	...	...	...	...	...
$y_l$	$m_{1l}$	...	$m_{il}$	...	$m_{kl}$	$m_{*l}$
Итого по x ( $m_x$ )	$m_{1*}$	...	$m_{i*}$	...	$m_{k*}$	n

В строке  $m_x$  помещены частоты одномерного вариационного ряда x, полученные путем суммирования значений  $m_{ij}$  в  $x_i$ -м столбце:  $m_{i*} = \sum_{j=1}^l m_{ij}$

В столбце  $m_y$  помещены частоты ряда y:  $m_{*j} = \sum_{i=1}^k m_{ij}$

Наконец, n — объем выборки.

Основные формулы для вычисления оценок статистических характеристик генеральной совокупности для различных объемов выборок приведены в табл. 2.  
 Таблица 2. Точечные оценки параметров двумерной корреляционной модели

Оцениваемый параметр генеральной совокупности	Его выборочная точечная оценка	
	По простым выборкам (n мало, данные не группированы)	По сгруппированным данным (n велико, данные сгруппированы)
Генеральная средняя X, $\mu_x$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i m_{i*}$
Генеральная средняя Y, $\mu_y$	$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$	$\bar{y} = \frac{1}{n} \sum_{j=1}^l y_j m_{*j}$
Генеральная средняя $X^2$ , $M(X^2)$	$\bar{x}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$	$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i m_{i*}$
Генеральная средняя $Y^2$ , $M(Y^2)$	$\bar{y}^2 = \frac{1}{n} \sum_{i=1}^n y_i^2$	$\bar{y}^2 = \frac{1}{n} \sum_{j=1}^l y_j^2 m_{*j}$
Генеральная средняя XY, $M(XY)$	$\bar{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$	$\bar{xy} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^l x_i y_j m_{ij}$
Генеральные дисперсии $\sigma_x^2$ , $\sigma_y^2$	$S_x^2 = \bar{x}^2 - \bar{x}^2$ , $S_y^2 = \bar{y}^2 - \bar{y}^2$	
Генеральный коэффициент корреляции $\rho$	Выборочный коэффициент корреляции $r = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}$	

Выборочный коэффициент детерминации  $r^2$  (равен квадрату коэффициента корреляции) и указывает долю дисперсии одной случайной величины, обусловленную вариацией другой.

Соответственно,  $(1 - r^2)$ , показывает долю остаточной дисперсии случайной величины, объясняемой не включенными в рассматриваемую двумерную модель факторами.

#### 4. Корреляционное отношение и индекс корреляции.

Рассмотренный выше коэффициент корреляции является полноценным показателем тесноты связи лишь в случае линейной зависимости между переменными. Однако часто возникает необходимость в достоверном показателе тесноты связи при любой форме зависимости.

Для получения такого показателя вспомним правило сложения дисперсий:  $Q_{общ} = Q_F + Q_{ост}$

где  $Q_{общ}$  — общая дисперсия переменной;

$Q_{ост}$  — средняя групповых дисперсий или остаточная дисперсия;

$Q_F$  — межгрупповая дисперсия

Остаточной дисперсией измеряют ту часть колеблемости Y, которая возникает из-за изменчивости неучтенных факторов не зависящих от X. Межгрупповая дисперсия выражает ту часть вариации Y, которая обусловлена изменчивостью X. Величина

$$\eta_{yx} = \sqrt{\frac{Q_F}{Q_{общ}}}$$

называется **эмпирическим корреляционным отношением** Y по X.

Чем теснее связь, тем большее влияние на вариацию переменной Y оказывает изменчивость X по сравнению с неучтеными факторами, тем выше  $\eta_{yx}$ .

Величина  $\eta_{yx}^2$ , называется **эмпирическим коэффициентом детерминации**, показывает, какая часть общей вариации Y обусловлена вариацией X.

*Основные свойства корреляционного отношения* (при достаточно большом объеме выборки n):

- 1) Корреляционное отношение есть неотрицательная величина, не превосходящая  $1: 0 < \eta < 1$ .
- 2) Если  $\eta = 0$ , то корреляционная связь отсутствует.
- 3) Если  $\eta = 1$ , то между переменными существует функциональная зависимость.
- 4)  $\eta_{yx} \neq \eta_{xy}$ , т.е. в отличие от коэффициента корреляции  $r$  (для которого  $r_{yx} = r_{xy} = r$ ) при вычислении корреляционного отношения существенно, какую переменную считать независимой, а какую — зависимой.

Эмпирическое корреляционное отношение  $\eta_{yx}$  является показателем рассеяния точек корреляционного поля относительно эмпирической линии регрессии, выражаемой ломаной, соединяющей значения  $\bar{y}_i$ . Однако в связи с тем, что закономерное изменение  $\bar{y}_i$  нарушается случайными зигзагами ломаной, возникающими вследствие остаточного действия неучтенных факторов,  $\eta_{yx}$  преувеличивает тесноту связи. Поэтому наряду с  $\eta_{yx}$  рассматривается показатель тесноты связи  $R_{yx}$ , характеризующий рассеяние точек корреляционного поля относительно линии регрессии ух.

Показатель  $R_{yx}$ , получил название **теоретического корреляционного отношения или**

$$R_{yx} = \sqrt{\frac{Q_F}{Q_{общ}}} = \sqrt{1 - \frac{Q_{ост}}{Q_{общ}}}$$

**индекса корреляции** Y по X.

где дисперсии  $Q_{общ}$  и  $Q_{ост}$  определяются по формулам, в которых групповые средние  $\bar{y}_i$  заменены условными средними  $u_{xi}$ , вычисленными по уравнению регрессии.

Достоинством рассмотренных показателей  $\eta$  и  $R$  является то, что они могут быть вычислены при любой форме связи между переменными. Хотя  $\eta$  и завышает тесноту связи по сравнению с  $R$ , но для его вычисления не нужно знать уравнение регрессии. Корреляционные отношения  $\eta$  и  $R$  связаны с коэффициентом корреляции  $r$  следующим образом:

$$0 \leq |r| \leq R \leq \eta \leq 1$$

В случае линейной модели индекс корреляции  $R_{yx}$  равен коэффициенту корреляции  $r$  (по абсолютной величине).

**Коэффициент детерминации**  $R^2$ , равный квадрату индекса корреляции (для парной линейной модели —  $r^2$ ), показывает долю общей вариации зависимой переменной, обусловленной регрессией или изменчивостью объясняющей переменной.

Чем ближе  $R^2$  к единице, тем лучше регрессия аппроксимирует эмпирические данные, тем теснее наблюдения примыкают к линии регрессии. Если  $R^2 = 1$ , то эмпирические точки (x, y) лежат на линии регрессии и между переменными Y и X существует линейная функциональная зависимость. Если  $R^2 = 0$ , то вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели переменных, и линия регрессии параллельна оси абсцисс.

Расхождение между  $\eta^2$  и  $R^2$  (или  $r^2$ ) может быть использовано для проверки линейности корреляционной зависимости.

Проверка значимости корреляционного отношения  $\eta$  основана на том, что статистика

$$F = \frac{\eta^2(n-m)}{(1-\eta^2)(m-1)}$$

(где  $m$  — число переменных) имеет F-распределение Фишера—Сnedекора с  $k_1 = m - 1$ ,  $k_2 = n - m$  степенями свободы. Поэтому  $\eta$  значимо отличается от нуля, если  $F > F(\alpha, k_1, k_2)$ , где  $F(\alpha, k_1, k_2)$  — табличное значение F-критерия на уровне значимости  $\alpha$  при числе степеней свободы  $k_1 = m - 1$ ,  $k_2 = n - m$ .

## 1.7 Лекция № 7 (1 час).

**Тема:** «Регрессионный анализ» (интерактивная форма)

### 1.10.1. Вопросы лекции:

1. Виды регрессионных моделей. Оценка параметров парной линейной регрессии
2. Проверка статистической значимости уравнения регрессии в целом и его параметров

### 1.10.2 Краткое содержание вопросов:

#### 1. Виды регрессионных моделей. Оценка их параметров

I. Различают однофакторные (с одной независимой переменной) и многофакторные (с несколькими независимыми переменными) модели.

Предположим, что для оценки параметров линейной функции регрессии взята выборка  $n$  пар переменных  $(x_i, y_i)$   $i=1, \dots, n$ .

Оценкой модели (\*) по выборке является уравнение регрессии  $\hat{y} = b_0 + b_1 x$ .

При однофакторных моделях связь часто называют парной, а коэффициент регрессии  $b_1$  — полным. Он показывает на сколько единиц изменяется зависимая переменная  $y$  при изменении независимой переменной  $x$  на единицу своего измерения.

Параметры  $b_0$ ,  $b_1$  определяются на основе МНК согласно которому неизвестные параметры  $b_0$  и  $b_1$ , выбираются таким образом, чтобы сумма квадратов отклонений эмпирических данных результативного признака от значений  $\hat{y}$ , найденные по уравнению регрессии была минимальной:

$$S = \sum (y_i - \hat{y})^2 = \sum (y_i - b_0 - b_1 x_i)^2 \xrightarrow{b_0, b_1} \min$$

Дифференцируя  $S$  по  $b_0$  и  $b_1$  и приравнивая к нулю ее частные производные, получим:

$$\begin{cases} \frac{dS}{db_0} = 2 \sum (y_i - b_0 - b_1 x_i) = 0 \\ \frac{dS}{db_1} = 2 \sum (y_i - b_0 - b_1 x_i) \cdot x_i = 0 \end{cases}$$

После преобразования получим систему нормальных уравнений:

$$\begin{cases} b_0 n + b_1 \sum x_i = \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum y_i x_i \end{cases}$$

решая которую относительно  $b_0$  и  $b_1$  получаем:

$$\begin{cases} b_1 = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} \\ b_0 = \frac{1}{n} \sum y_i - b_1 \frac{1}{n} \sum x_i \end{cases} \quad \text{или} \quad \begin{cases} b_1 = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2} \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases}.$$

Многофакторная модель с линейной связью признаков имеет вид:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k.$$

Коэффициенты регрессии  $b_j$  при факторных признаках называются чистыми.

**Коэффициенты чистой регрессии** показывает, на сколько единиц изменится результативный признак при изменении соответствующего факторного признака на единицу своего измерения, при условии, что другие факторы, включенные в уравнение, зафиксированы на одном уровне.

Регрессионный анализ дополняют расчетом показателя силы корреляционной связи – коэффициента эластичности.

**Коэффициент эластичности** показывает на сколько % изменится в среднем значение результативного признака при изменении факторного признака на 1 % .

$$\bar{\epsilon}_j = b_j \frac{\bar{x}_j}{\bar{y}}.$$

Величина указывает на силу реакции результативного признака на относительный прирост фактора.

Если  $|\bar{\epsilon}_j| \leq 1$  – эластичность слабая;

$|\bar{\epsilon}_j| = 1$  - эластичность средняя;

$|\bar{\epsilon}_j| \geq 1$  - эластичность высокая.

Во множественной корреляции по коэффициентам эластичности проводится сравнительный анализ силы влияния факторов на результативный признак.

## II. Различают линейную и нелинейную форму связи.

Нелинейная форма парной модели может быть выражена любым видом криволинейной функции:

-  $\hat{y} = b_0 + b_1 \ln x$  – полулогарифмическая функция

-  $\hat{y} = b_0 + b_1^x$  – показательной

-  $\hat{y} = b_0 x^{b_1}$  – степенной

-  $\hat{y} = b_0 + b_1 \frac{1}{x}$  - гиперболической и т.д

Многофакторная модель с нелинейной формой связи может выглядеть различным образом, например:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2^2.$$

Для правильного выбора формы связи необходимо провести глубокий теоретический анализ, а затем подкрепить свои логические предположения графическим методом, методом статистических группировок, сопоставлением параллельных рядов.

## 2. Проверка статистической значимости уравнения регрессии в целом и его параметров

Так как выборочные характеристики являются случайными величинами, то проверим значимость уравнения регрессии и его параметров.

Значимость уравнения регрессии в целом оценивается с помощью F критерия Фишера и на основе дисперсии анализа.

1. Для проверки значимости уравнения регрессии, т.е гипотезы  $H_0: \beta_1=0$  (Для множественной регрессии  $H_0: \beta_1=\beta_2=\dots=\beta_k$ ), используют критерий основанный на статистике:

$$F_{\text{набл}} = \frac{Q_R / k}{Q_{\text{ост}} / (n - 2)} = \frac{S_R^2}{S_{\text{ост}}^2},$$

При истинности гипотезы  $H_0$  приведенная статистика имеет F распределение с числом степеней свободы  $v_1=k$ ,  $v_2=n-2$ .

$Q_R$ - сумма квадратов отклонений, обусловленных регрессией,

$$Q_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

$$S_R^2 = \frac{Q_R}{k} \text{ - дисперсия, обусловленная регрессией.}$$

Qост = сумма квадратов отклонений обусловленных остаточными, не включенными в модель факторами.

$$Q_{\text{ост}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

$$S_{\text{ост}}^2 = \frac{Q_{\text{ост}}}{n - k - 1} \text{ - остаточная дисперсия.}$$

Результаты расчетов, как правило, оформляют в виде таблицы

Таблица 1- Дисперсионный анализ

Компоненты дисперсии	Число степеней свободы df	Сумма квадратов	Дисперсии	F критерий	
				фактический	табличный
Регрессия	$k$	$Q_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$S_R^2 = \frac{Q_R}{k}$	$F_{\text{набл}} = \frac{S_R^2}{S_{\text{ост}}^2}$	$F_{\text{кр}}(\alpha; v_1; v_2)$
Остаточная	$n-k-1$	$Q_{\text{ост}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$S_{\text{ост}}^2 = \frac{Q_{\text{ост}}}{n - k - 1}$		
Общая	$n-1$	$Q_{\text{общ}} = \sum_{i=1}^n (y_i - \bar{y})^2$			

Табличное значение находим по таблице « F –распределения Фишера» при уровне значимости  $\alpha$  и числе степеней свободы  $v_1=k$ ,  $v_2=n-k-1$ ,  $F_{\text{кр}}(\alpha, v_1, v_2)$ .

Если  $F_{\text{набл}} > F_{\text{крит}}$ , то уравнение регрессии значимо (достоверно) с вероятностью ошибки  $\alpha$ . Гипотеза  $H_0$  – отвергается. В противном случае гипотеза не отвергается.

2. Для оценки значимости коэффициентов (параметров) регрессии используется критерий t- нормального распределения (большие выборки) или t- Стьюдента (малые выборки).

$$H_0: \beta_1=0$$

$$H_1: \beta_1 \neq 0$$

1) Рассчитывается средняя (стандартная) ошибка выборки для коэффициента регрессии

$$m_{b_1} = \sqrt{\frac{S_{\text{ост}}^2}{\sigma_x^2 \cdot n}} \text{ или } m_{b_1} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot (n - 2)}}$$

$$2) \text{Рассчитывается } t_{\text{набл}} = \frac{b_1}{m_{b_1}}$$

3) Сравнивается  $t_{\text{набл}}$  с  $t_{\text{крит}}$  ( $\alpha, v=n-2$ )

Если  $|t_{\text{набл}}| > t_{\text{крит}}$ , то отвергают  $H_0$  и с вероятностью  $\gamma=1-\alpha$  принимают альтернативную гипотезу о значимости коэффициента регрессии  $b_1$ ;  
 $|t_{\text{набл}}| < t_{\text{крит}}$  – гипотеза  $H_0$  принимается, с вероятностью ошибки  $\alpha$ .

#### Интервальные оценки параметров регрессии

Если гипотеза  $H_0: \beta=0$  отвергается, то представляет интерес определение с надежностью  $\gamma$  интервальных оценок параметров регрессии.

-для коэффициента регрессии  $\beta_1$

$$b_1 - t_\alpha m_{b_1} \leq \beta_1 \leq b_1 + t_\alpha m_{b_1};$$

-для интервала предсказания в точке  $x=x_{n+1}$

$$\tilde{y}_{n+1} \in \left[ \hat{y}_{n+1} \pm t_\alpha \cdot \hat{S} \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

$t_\alpha$  - определяются по таблице распределения Стьюдента.

При  $n > 30$  можно пользоваться - «Нормальный закон распределения».

## 2. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРОВЕДЕНИЮ ПРАКТИЧЕСКИХ ЗАНЯТИЙ

### 2.1 Практическое занятие №1 (1 час).

Тема: «Случайные события. Вероятность события»

#### 1 Задание для работы:

Вопросы к занятию:

1. Теория вероятностей как наука. Краткая историческая справка
2. Случайные события и их виды
3. Операции над событиями
4. Классическое определение вероятности

Типовые задачи:

1. Рассмотрим случайные события – выпадение определенного числа на верхней грани, которые могут произойти при бросании простого 6-гранного игрального кубика.  
Обозначения случайных событий:

A – выпадение числа 2, B – выпадение числа 3, C – выпадение нечетного числа, D – выпадение любого из чисел 1, 3 или 5.

Какие события являются достоверными, невозможными, несовместными, совместными, равносильными и равновозможными?

2. Брошены две игральные кости. Найти вероятность, что а) на гранях кубиков выпадут одинаковые цифры; б) сумма выпавших очков четная, при этом на грани хотя бы одной из костей появится тройка; в) сумма выпавших очков равна восьми, а разность четырем; г) сумма выпавших очков равна шести, если известно, что их разность равна четырем.

3. Куб, все грани которого окрашены, распилен на тысячу кубиков одинакового размера, которые затем тщательно перемешаны. Найти вероятность того, что наудачу извлеченный кубик имеет окрашенных граней а) одну; б) две; в) три; г) хотя бы одну.

**4.** Сколькими способами можно случайным образом из 25 лучших студентов курса выбрать 2-х для поездки в Англию и Америку?

**5.** Из 50 сотрудников фирмы 30 человек владеют английским языком. Для участия в международной конференции случайным образом отбирается 5 человек. Какова вероятность того, что а) все выбранные сотрудники владеют английским языком; б) 3 человека из 5 выбранных знают английский язык; в) хотя бы один владеет английским языком.

**6.** Для изучения качества произведенных деталей техническим отделом было обследовано 200 деталей, из которых 10 оказались бракованными. Найти вероятность изготовления бракованных деталей.

## **2 Краткое описание проводимого занятия:**

1. Решение задач по теме занятия. При решении задач необходимо акцентировать внимание правильном применении формулы расчета вероятности.

## **3 Результаты и выводы:**

Усвоение студентами знаний и закрепление навыков по теме практического занятия.

### **2.2 Практическое занятие №2 (1 час).**

**Тема: «Повторные независимые испытания»**

#### **1 Задание для работы:**

Вопросы к занятию:

1. Формула Бернулли
2. Локальная теорема Муавра-Лапласа
3. Интегральная теорема Муавра-Лапласа
4. Вероятность отклонения относительной частоты от постоянной вероятности в независимых испытаниях
5. Теорема Пуассона

Типовые задачи:

Вероятность того, что посетитель продуктового магазина купит молоко равна 0,7. Определите вероятность того, что а) из 10 покупателей ровно 4 купят молоко; б) из 80 покупателей ровно 50 купят молоко; в) из 100 покупателей от 65 до 80 человек купят молоко.

## **2 Краткое описание проводимого занятия:**

1. Решение задач.

## **3. Результаты и выводы:**

Усвоение студентами знаний и закрепление навыков по теме практического занятия.

### **2.3 Практическое занятие № 3 (2 часа).**

**Тема: «Дискретная случайная величина»**

## **1 Задание для работы:**

Вопросы к занятию:

1. Определение дискретной случайной величины. Закон распределения
2. Функция распределения ДСВ
3. Основные числовые характеристики ДСВ

Типовые задачи:

1. Случайная величина X задана законом распределения:

X	1	2	3	4
p	0,2	0,4	0,3	0,1

Найти функцию распределения случайной величины X и построить ее график.

Рассчитать математическое ожидание, дисперсию, среднее квадратическое отклонение.

2. Вероятность опоздания студента на первую пару равна 0,4. Написать биномиальный закон распределения числа опозданий студента из четырех дней занятий. Построить график распределения вероятностей. Найти функцию распределения числа опоздавших поездов и построить ее график. Рассчитать математическое ожидание, дисперсию и среднеквадратическое отклонение.

## **2 Краткое описание проводимого занятия:**

1. Письменно решение задач
2. Проверка самостоятельного изучения вопросов
3. Проверка подготовки к занятиям.

## **3 Результаты и выводы:**

Усвоение студентами знаний и закрепление навыков по теме практического занятия.

## **2.4 Практическое занятие №4 (2 часа).**

**Тема: «Непрерывная случайная величина»**

## **1 Задание для работы:**

Вопросы к занятию:

1. Определение непрерывной случайной величины.
2. Функция распределения НСВ
3. Основные числовые характеристики НСВ
4. Нормальный закон распределения

Типовые задачи:

1. Задана дифференциальная функция распределения случайной величины X:

$$f(x) = \begin{cases} 0, & \text{при } x \leq 3 \\ \frac{1}{2}, & \text{при } 3 < x \leq 5 \\ 0, & \text{при } x > 5 \end{cases}$$

Рассчитать: 1) математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины X; 2) вероятность того, что в результате испытаний случайная величина примет значение большее 3 и меньшее 5. 3) Построить график функции F(x).

**2.** Случайная величина  $X$  равномерно распределена на отрезке  $[0; 2]$ . Построить функцию распределения случайной величины  $X$ . Рассчитать: 1) математическое ожидание, дисперсию и среднее квадратическое отклонение случайной величины  $X$ ; б) вероятность того, что в результате испытаний случайная величина примет значение большее 0,5 и меньшее 1,5. Построить график функции  $F(X)$ .

**3.** При сортировке случайные значения веса зерна распределены нормально со средним значением 0,15 г. и средним квадратическим отклонением 0,03 г. Нормальные всходы дают зерна, вес которых более 0,10 г.

Определить: а) процент семян, от которых следует ожидать нормальные всходы; б) величину, которую не превзойдет вес отдельного зерна с вероятностью 0,99.

## **2 Краткое описание проводимого занятия:**

1. Проверка подготовки к занятию самостоятельно изученных вопросов
2. Решение задач

## **3 Результаты и выводы:**

Усвоение студентами знаний и закрепление навыков по теме практического занятия.

### **2.5 Практическое занятие № 5 (2 часа).**

**Тема: «Статистическое оценивание параметров распределения»**

#### **1 Задание для работы:**

Вопросы к занятию:

1. Общие сведения о выборочном методе
2. Статистическое распределение выборки и эмпирическая функция распределения
3. Понятие оценки параметров
4. Свойства точечных оценок
5. Некоторые характеристики вариационного ряда
6. Оценка параметров генеральной совокупности по собственно-случайной выборке
7. Интервальные оценки параметров генеральной совокупности

Типовые задачи:

1. На предприятии для анализа производительности труда случайным образом было отобрано 40 человек. Получены следующие данные:

Произведено изделий в час	10-14	14-18	18-22	22-26	26-30
Количество работников	5	8	15	10	2

Найти: а) среднее число изделий, произведенных в час, дисперсию, среднее квадратическое отклонение и коэффициент вариации;  
б) моду и медиану;  
в) коэффициент асимметрии и эксцесса.

**2.** Проведено исследование коммерческих фирм по затратам на рекламу в год. Для этого случайным образом отобрано 50 фирм. Результаты представлены в таблице:

Расходы на рекламу, тыс. руб.	Менее 20	20-40	40-60	60-80	80-100	100 и более
Число фирм	3	5	9	16	13	4

Найти: а) границы, в которых с вероятностью 0,95 заключен средний размер расходов на рекламу;  
 б) вероятность, с которой средний расход на рекламу будет отличаться от выборочной средней не более чем на 5 тыс. руб.

## 2 Краткое описание проводимого занятия:

1. Проверка освоения студентами вопросов для самостоятельного изучения
2. Решение задач

## 3 Результаты и выводы:

Усвоение студентами знаний и закрепление навыков по теме практического занятия.

## 2.6 Практическое занятие № 6 (1 час).

**Тема: «Корреляционный анализ»**

### 1 Задание для работы:

Вопросы к занятию:

1. Расчет коэффициента корреляции и детерминации, их интерпретация
2. Проверка статистической значимости коэффициента корреляции: с использованием таблиц Фишера-Ийтса, распределения Стьюдента

Типовые задачи:

На основе выборочных данных о производительности труда (Х), измеряемой в млн. руб. на человека, и себестоимости продукции (У), измеряемой в тыс. руб. на единицу продукции, полученных с однотипных предприятий за месяц:

X	5	4	3	20	10	15
Y	7	10	12	2	5	4

Найти: а) выборочный коэффициент корреляции между Х и У;  
 б) проверить значимость генерального коэффициента корреляции при  $\alpha=0,05$ ;  
 в) с надежностью  $\gamma=0,95$  найти границы доверительного интервала генерального коэффициента корреляции.

## 2 Краткое описание проводимого занятия:

1. Проверка уровня освоения студентами вопросов для самостоятельного изучения
2. Решение задач

## 3 Результаты и выводы:

Усвоение студентами знаний и закрепление навыков по теме практического занятия.

## 2.7 Практическое занятие № 7 (1 час).

## **Тема: «Регрессионный анализ»**

### **1 Задание для работы:**

Вопросы к занятию:

1. Оценка параметров парной линейной регрессии
2. Проверка статистической значимости уравнения регрессии в целом с помощью критерия Фишера
3. Проверка статистической значимости параметров уравнения регрессии с помощью критерия Стьюдента

Типовые задачи:

На основе выборочных данных о производительности труда ( $X$ ), измеряемой в млн. руб. на человека, и себестоимости продукции ( $Y$ ), измеряемой в тыс. руб. на единицу продукции, полученных с однотипных предприятий за месяц:

X	5	4	3	20	10	15
Y	7	10	12	2	5	4

Найти: а) параметры линейного уравнения регрессии между  $X$  и  $Y$ ;  
б) проверить значимость генерального коэффициента регрессии при  $\alpha=0,05$  и  
уравнения регрессии в целом.

### **2 Краткое описание проводимого занятия:**

3. Проверка уровня освоения студентами вопросов для самостоятельного изучения
4. Решение задач

### **3 Результаты и выводы:**

Усвоение студентами знаний и закрепление навыков по теме практического занятия.