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

Методические рекомендации для самостоятельной работы обучающихся по дисциплине

Б1.Б.08 Теория вероятностей и математическая статистика

Направление подготовки (специальность) 09.03.01 Информатика и вычислительная техника

Профиль образовательной программы "Автоматизированные системы обработки информации и управления**"**

Форма обучения заочная

СОДЕРЖАНИЕ

1. ОРГАНИЗАЦИЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

1.1. Организационно-методические данные дисциплины

п.п.	Наименование темы	Общий объем часов по видам самостоятельной работы (*из табл. 5.1 РПД*)				
		подготовка курсового проекта (работы)	подготовка реферата/эссе	индивидуальные домашние задания (К.Р.)	самостоятельное изучение вопросов (СИВ)	подготовка к занятиям (ПкЗ)
	2	3	4	5	6	7
1	Классическое определение вероятности события. Относительная частота наступления события и статистическая вероятность. Формулы умножения и сложения вероятностей случайных событий. Повторение испытаний: формулы Бернулли, локальные и интегральные теоремы Лапласа, формула Пуассона, простейший поток событий.	0	0	0	4	5
2	Понятие случайной величины примеры. Виды случайных величин. Закон распределения вероятностей. Функция распределения случайных величин. Свойства. Плотность распределения вероятностей. Числовые характеристики: математическое ожидание, свойства; дисперсия, свойства; среднее квадратичное отклонение и его свойства.	0	0	0	8	5
3	Статистический материал. Статистические параметры распределения. Статистические оценки параметров распределения. Понятие статистической гипотезы. Виды гипотез. Статистический критерий. Критическая область. Мощность критерия. Критерии согласия: критерий Пирсона. Выравнивание рядов.	0	0	0	12	6
4	Понятие функциональной, стохастической и корреляционной зависимости. Функция регрессии. Корреляционное отношение. Его свойства, значимость. Линейная функция регрессии. Коэффициент корреляции его.	0	0	0	14	**6**
	Итого в соответствии с РПД	-	-	-	**38**	**22**

2. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО САМОСТОЯТЕЛЬНОМУ ИЗУЧЕНИЮ ВОПРОСОВ

2.1 Условная вероятность, формула полной вероятности, формула Байеса.

Вероятность события А, найденную при условии, что наступило событие В ($P_B(A)$), будем называть **условной вероятностью** события А при условии В.

Например, в урне 3 белых и 2 черных шара. Наудачу вынимают один шар, затем еще один. Событие В: появление белого шара при первом вынимании; событие А: появление белого шара при втором вынимании. Тогда $Pв(A) = \frac{2}{4} = \frac{1}{2}$.

Теорема (формула полной вероятности).

Пусть $B_1, B_2, ..., B_n$ - образуют полную группу несовместных событий, т.е. $\sum_{i=1}^{n} P(B_i) = 1$. Если событие А может осуществляться только при условии совмещении с одним из событий $B_1, B_2, ..., B_n$, то

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + ... + P(B_n) \cdot P_{B_n}(A).$$

Задача. По цели произведено 3 последовательных выстрела. Вероятность попадания при первом выстреле p_1=0,3; вероятность попадания при втором выстреле p_2=0,6; вероятность попадания при третьем выстреле p_3=0,8. При одном попадании вероятность поражения цели λ_1=0,4; при двух попаданиях – λ_2=0,7; при трех попаданиях – λ_3=1,0. Определить вероятность поражения цели при трех выстрелах?

Решение.

Событие А: поражение цели при трех выстрелах. Рассмотрим полную группу несовместных событий:

B_1: было одно попадание при трех выстрелах;

B_2 : было два попадание при трех выстрелах;

B_3 : было три попадание при трех выстрелах;

B_4 : не было ни одного попадания.

Определим вероятность каждого события:

$$P(B_1) = p_1(1-p_2)(1-p_3) + (1-p_1)p_2(1-p_3) + (1-p_1)(1-p_2)p_3 = 0{,}332$$

$$P(B_2) = p_1 p_2(1-p_3) + p_1(1-p_2)p_3 + (1-p_1)p_2 p_3 = 0{,}468$$

$$P(B_3) = p_1 p_2 p_3 = 0{,}144$$

$$P(B_4) = (1-p_1)(1-p_2)(1-p_3) = 0{,}056.$$

Условные вероятности поражения цели при осуществлении каждого из этих событий:

$$P_{B_1}(A) = 0{,}4; \quad P_{B_2}(A) = 0{,}7; \quad P_{B_3}(A) = 1; \quad P_{B_4}(A) = 0.$$

Подставим все данные в формулу из теоремы:

$$P(A) = P(B_1)\cdot P_{B_1}(A) + P(B_2)\cdot P_{B_2}(A) + P(B_3)\cdot P_{B_3}(A) + P(B_4)\cdot P_{B_4}(A) =$$

$$= 0{,}332\cdot 0{,}4 + 0{,}468\cdot 0{,}7 + 0{,}144\cdot 1 + 0{,}0560 = 0{,}6044.$$

Замечание. Если событие А не зависит от события В, то $P(A) = P_B(A)$. Следовательно, $P(A\cdot B) = P(A)\cdot P(B)$.

Пусть $B_1, B_2, ..., B_n$ - полная группа несовместных событий, $P(B_1), P(B_2), ..., P(B_n)$ - соответствующие вероятности. Событие А может наступить только вместе с каким-либо из событий $B_1, B_2, ..., B_n$, которые мы будем называть гипотезами. Тогда справедлива формула полной вероятности: $P(A) = P(B_1)\cdot P_{B_1}(A) + P(B_2)\cdot P_{B_2}(A) + ... + P(B_n)\cdot P_{B_n}(A)$.

Допустим, что событие А уже наступило. Это изменит вероятности гипотез $P(B_1), P(B_2), ..., P(B_n)$. Требуется определить условные вероятности этих гипотез $P_A(B_1), ..., P_A(B_n)$, в предположении, что событие А уже наступило.

Найдем

$$P(A\cdot B_1) = p(B_1)\cdot P_{B_1}(A) = p(A)\cdot P_A(B_1) \Rightarrow P_A(B_1) = \frac{P(A\cdot B_1)}{P(A)} = \frac{p(B_1)\cdot P_{B_1}(A)}{P(A)}$$

Заменим $P(A)$ формулой полной вероятности события:

$$P_A(B_1) = \frac{p(B_1)\cdot P_{B1}(A)}{\sum_{i=1}^{n} P(B_i)\cdot P_{B_i}(A)}$$ Аналогично определяется $P_A(B_2), ..., P_A(B_n)$.

Окончательно получаем формулу Байеса или формулу из теоремы гипотез:

$$P_A(B_k) = \frac{p(B_k)\cdot P_{B_k}(A)}{\sum_{i=1}^{n} P(B_i)\cdot P_{B_i}(A)}.$$

Задача. 30% приборов собирает специалист высокой квалификации и 70% - средней квалификации. Надежность работы прибора, собранного специалистом высокой квалификации – 0,9 и надежность работы прибора, собранного специалистом средней квалификации – 0,8. Взятый наудачу прибор оказался надежным. Определить вероятность того, что он собран специалистом высокой квалификации.

Событие А: безотказная работа прибора.

Для проверки прибора возможны гипотезы:

B_1: прибор собран специалистом высокой квалификации;

B_2: прибор собран специалистом средней квалификации.

По условию задачи:

$P_{B_1}(A) = 0{,}9; \quad P_{B_2}(A) = 0{,}8$.

Определим вероятности гипотез B_1 и B_2 при условии, что событие А наступило:

$$P_A(B_1) = \frac{0{,}3 \cdot 0{,}9}{0{,}3 \cdot 0{,}9 + 0{,}7 \cdot 0{,}8} = 0{,}325; \quad P_A(B_2) = \frac{0{,}7 \cdot 0{,}8}{0{,}3 \cdot 0{,}9 + 0{,}7 \cdot 0{,}8} = 0{,}675$$

2.2 Простейший поток и его свойства. Интенсивность потока. Вероятность события с заданной интенсивностью

Особое внимание следует обратить на простейший поток событий.

Потоком событий называют последовательность событий, которые наступают в случайные моменты времени. Примеры потоков: поступление вызовов на АТС, поступление вызовов на пункт неотложной медицинской помощи, прибытие кораблей в порт, последовательность отказов элементов некоторого устройства.

Простейшим называют поток, обладающий свойствами стационарности, отсутствием последействия и ординарности.

Свойство стационарности характеризуется тем, что вероятность появления k событий за время длительностью t не зависит от начала отсчета промежутка времени, а зависит лишь от его длительности. Так вероятности появления пяти событий на промежутках времени (1; 4), (6; 9), (8; 11) одинаковой длительности $t = 3$ единицы времени равны между собой.

Свойство отсутствия последействия характеризуется тем, что вероятность появления k событий на любом промежутке времени не зависит от того, сколько событий появилось до начала рассматриваемого промежутка.

Свойство ординарности характеризуется тем, что вероятность появления двух и более событий пренебрежимо мала, сравнительно с вероятностью появления одного события.

Интенсивностью потока λ называют среднее число событий, которые появляются в единицу времени. Доказано, что если известна постоянная интенсивность потока λ, то вероятность появления k событий простейшего потока за время длительности t определяется формулой:

$$P_t(k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}.$$

Пример: Среднее число заявок, поступающих на предприятие бытового обслужи-

вания за 1 час, равно трем. Найти вероятность того, что за 2 часа поступит 5 заявок. Предполагается, что поток заявок - простейший.

Решение. По условию $\lambda = 3$, $t = 2$, $k = 5$. Воспользуемся формулой $P_t(k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$. Искомая вероятность того, что за 2 часа поступит 5 заявок, равна $P_2(5) = \frac{(6)^5 \cdot 0{,}00248}{120} \approx 0{,}268$.

Пример: Среднее число заявок, поступающих на АТС в одну минуту, равно двум. Найти вероятности того, что за четыре минуты поступит:

а) три вызова;

б) менее трех вызовов;

в) не менее трех вызовов.

Решение, а) По условию $\lambda = 3$, $t = 2$, $k=5$. Воспользуемся формулой: $P_t(k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$

Подставив данные условия задачи, получим: $P_4(3) = \frac{8^3 \cdot e^{-8}}{3!} = \frac{512 \cdot 0{,}000335}{6} \approx 0{,}03$.

б) Найдем вероятность того, что за четыре минуты поступит менее трех вызовов, т.е. ни одного вызова, или один вызов, или два вызова. Поскольку эти события несовместны, применим теорему суммы несовместных событий:

$$P_4(k<3) = P_4(0) + P_4(1) + P_4(2) = e^{-8} + 8 \cdot e^{-8} \cdot \frac{8^2 \cdot e^{-8}}{2!} = 41 \cdot 0{,}000335 \approx 0{,}01.$$

в) Найдем вероятность того, что за четыре минуты поступит не менее трех вызовов: так как события «поступило менее трех вызовов» и «поступило не менее трех вызовов» - противоположные, то сумма вероятностей этих событий равна единице: $P_4(k<3) + P_4(k\geq 3) = 1$. Поэтому $P_4(k \geq 3) = 1 - P_4(k<3) = 1 - [P_4(0) + P_4(1) + P_4(2)] = 1 - 0{,}01 = 0{,}99$.

2.3 Биномиальное распределение, его свойства, числовые характеристики.

Биномиальное распределение связано с повторными независимыми испытаниями и формулой Бернулли. Оно задается фиксированным числом испытаний N И вероятностью «успеха» в одном испытании P. Отличительные черты биномиального эксперимента:

1. все N испытаний абсолютно одинаковы;

2. результаты разных испытаний не зависят друг от друга;

3. для каждого испытания возможны только два исхода: «успех» и «неудача»; «успех» - когда интересующее нас событие появилось, и «неудача», - когда не появилось;

4. для каждого испытания вероятность появления «успеха» постоянна и равна P.

Число «успехов» в N независимых испытаниях будет случайной величиной X, распределенной по биномиальному закону. Вероятность того, что случайная величина X, распределенная по биномиальному закону примет значение K, вычисляется по известной формуле Бернулли:

$$P(X=k)=P_n(k)=C_n^k \cdot p^k \cdot q^{n-k}$$

X	0	1	2	…	$n-1$	n
P_i	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	$C_n^2 p^2 q^{n-2}$	…	$C_n^{n-1} p^{n-1} q^1$	$C_n^n p^n q^0$

Ряд распределения X принимает вид:

Числовые характеристики биномиального распределения.

1. Математическое ожидание равно произведению числа испытаний N на вероятность «успеха» в одном испытании P: M(X)= Np;

2. Дисперсия равна произведению числа испытаний N на вероятность «успеха» в одном испытании P И на вероятность «неудачи» Q: D(X)= Npq.

2.4 Распределение Пуассона, его свойства, числовые характеристики. Связь распределений ДСВ с нормальным распределением.

Приведем примеры, приводящие к случайным величинам, распределенным по закону Пуассона:

· Автоматическая телефонная станция получает в среднем за минуту A вызовов. Какова вероятность того, что за данную минуту она получит ровно M вызовов? Случайное число вызовов за данную минуту распределено по закону Пуассона.

· Автодорожная инспекция регистрирует количество аврий за неделю на определенном участке дороги. Какова вероятность того, что в течение данной недели произойдет ровно M дорожных аварий? Случайное число аварий За неделю распределено по закону Пуассона.

Аналогичные примеры можно привести не только для временных интервалов (минута, неделя), но и при учете дефектов дорожного покрытия На километр пути или опечаток На страницу текста.

Отличительные черты эксперимента, приводящего к распределению Пуассона (на примере временных интервалов):

1. каждый малый интервал времени может рассматриваться как испытание, результатом которого служит либо «успех» - поступление телефонного вызова, либо «неудача». Интервалы столь малы, что может быть только один «успех» в одном интервале, вероятность которого мала и неизменна.

2. Число «успехов» в одном большом интервале не зависит от их числа в другом. То есть попадание «успехов» в неперекрывающиеся интервалы – события независимые, и «успехи» беспорядочно разбросаны по временным промежуткам;

3. среднее число «успехов» в большом интервале для разных интервалов постоянно на протяжении всего времени.

Число «успехов» на заданном интервале будет случайной величиной, распределенной по закону Пуассона. Случайное число аварий за неделю может принимать значения 0, 1, 2, 3, … (верхнего предела нет). Вероятность того, что случайная величина X, распределенная по закону Пуассона примет значение M, вычисляется по известной формуле Пуассона:

$P_m = \frac{a^m}{m!} \cdot e^{-a}$, M = 0, 1, 2, …

Числовые характеристики распределения Пуассона.

2.5 Равномерное распределение.

Для Равномерного закона плотность вероятности и функция распределения задаются формулами

$$f(x) = \begin{cases} \frac{1}{b-a}, x \in [a,b] \\ 0, x \notin [a,b] \end{cases}, \quad F(x) = \begin{cases} 0, x < a \\ \frac{x-a}{b-a}, a \le x \le b \\ 1, x > b \end{cases},$$

А числовые характеристики M(X)= $\frac{a+b}{2}$, D(X)= $\frac{(b-a)^2}{12}$.

2.6 Показательное распределение.

Для Показательного закона плотность вероятности и функция распределения задаются формулами

$$f(x) = \begin{cases} 0, x < 0 \\ \alpha \cdot e^{-\alpha x}, x \ge 0 \end{cases}, \quad F(x) = \begin{cases} 0, x < 0 \\ 1 - e^{-\alpha x}, x \ge 0 \end{cases},$$

А числовые характеристики M(X)= 1/a, D(X)= 1/a2.

Эти формулы можно использовать при решении задач.

2.7 Нормальное распределение, его свойства.

Нормальный (гауссовский) закон распределения задается плотностью распределения по формуле

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{(x-a)^2}{2\sigma^2}}$$, - ¥ < x < +¥

Числа a Î R и s > 0 называются параметрами нормального закона. Нормальный закон с такими параметрами обозначается N(a, s).

При a = 0 функция f(x) четная (f(-x) = f(x)) , ее график симметричен относительно оси OY, и поэтому среднее значение M(X) = 0. График f(x) для закона N(a, s) получается из графика f(x) для N(0,s) сдвигом на a единиц вправо (это известно из курса средней школы), поэтому в общем случае M(X) = a для нормального закона.

Дисперсия же вычисляется по формуле D(X) =s2.

Пример. Случайная величина X распределена по нормальному закону с плотностью вероятности

$$f(x)-Ae^{-\frac{x^2}{2}+2x-2}$$

Найти A, M (X), D(X), P(-3<X<3).

Т. к. $-\frac{x^2}{2}+2x-2=-\frac{(x-2)^2}{2}$, то $f(x)=A\cdot e^{-\frac{(x-2)^2}{2}}$

Показатель экспоненты $-\frac{(x-2)^2}{2}$ приравняем к $-\frac{(x-a)^2}{2\sigma^2}$, откуда a = 2 , s = 1 .

Числовой коэффициент $\frac{1}{\sigma\sqrt{2\pi}}=\frac{1}{\sqrt{2\pi}}$ Должен быть равен A, следовательно,

$A=\frac{1}{\sqrt{2\pi}}$, M (X) = a = 2, D(X) = s 2 = 1.

P (-3 < X < 3) = F(3) - F(-3) =

$$=\int_{-\infty}^{3}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-2)^2}{2}}dx-\int_{-\infty}^{-3}\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{(x-2)^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-3}^{3}e^{-\frac{(x-2)^2}{2}}dx$$

Этот интеграл не вычисляется в элементарных функциях, его численное значение можно найти по таблицам.

В большинстве учебников имеются таблицы для вычисления функций

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ или $\Phi 1(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \Phi(x)$

Ф(x) - нечетная функция, т. е. Ф(-x) = - Ф(x). В общем случае

$P(x1 < X < x2) = \Phi\left(\frac{x_2 - a}{\sigma}\right) - \Phi\left(\frac{x_1 - a}{\sigma}\right)$,

Где a и s - параметры нормального закона. Следовательно, для данного примера

P(|X| < 3) = Ф1(1) - Ф1(-5) = Ф(1) - Ф(-5) = Ф(1) + Ф(5) =

= 0,3413 + 0,5 = 0,8413.

2.8 Интервальные оценки, их свойства.

Статистические оценки параметров распределения.

Пусть требуется изучить количественный признак генеральной совокупности. Допустим, что из теоретических соображений удалось установить, какое именно распределение имеет признак. Естественно возникает задача оценки параметров, которыми определяется это распределение.

Обычно в распоряжении исследователя имеются лишь данные выборки, например, значения количественного признака $x_1, x_2, ..., x_n$, полученные в результате наблюдений. Через эти данные и выражают оцениваемый параметр. Рассматривая $x_1, x_2, ..., x_n$ как независимые случайные величины , можно сказать, что найти статистическую оценку неизвестного параметра распределения – это значит найти функцию от наблюдаемых случайных величин, которая и дает приближенное значение оцениваемого параметра.

Несмещенные, эффективные и состоятельные оценки.

Для того чтобы статистические оценки давали «хорошие» приближения оцениваемых параметров, они должны удовлетворять определенным требованиям. Укажем эти требования.

Пусть $x_1, x_2, ..., x_n$ – наблюдаемые значения СВ X. Обозначим через θ^* статистическую оценку неизвестного параметра θ, вычисленного на основе данного статистического материала.

Несмещенной называют статистическую оценку θ^*, математическое ожидание которой равно оцениваемому параметру θ при любом объеме выборки, т.е.

$M[\theta^*] = \theta$.

Эффективной называют статистическую оценку, которая при заданном объеме выборки имеет наименьшую дисперсию.

Состоятельной называют статистическую оценку, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру, т.е. для любого $\varepsilon > 0$ при $n \to \infty$ $P(|\Theta^* - \Theta| < \varepsilon) \to 1$.

Отметим, что смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Числовые характеристики вариационных рядов.

Выборочная средняя.

Пусть для изучения генеральной совокупности относительно количественного признака X извлечена выборка объема .

Выборочной средней $\overline{x_в}$ называют среднее арифметическое значение признака выборочной совокупности.

Если все значения $x_1, x_2, ..., x_n$ признака выборки объема различны, то

$$\overline{x_в} = \frac{x_1 + x_2 + ... + x_n}{n}. \quad (1)$$

Если же значения признака $x_1, x_2, ..., x_k$ имеют соответственно частоты $m_1, m_2, ..., m_k$, причем $m_1 + m_2 + ... + m_k = n$, то

$$\overline{x_в} = \frac{m_1 x_1 + m_2 x_2 + ... + m_k x_k}{n} \text{ или } \overline{x_в} = \frac{\sum_{i=1}^{k} m_i x_i}{n}. \quad (2)$$

Выборочная дисперсия и выборочное среднее квадратическое отклонение.

Для того чтобы охарактеризовать рассеяние наблюдаемых значений количественного признака выборки вокруг своего среднего значения $\overline{x_в}$, вводят такую характеристику как выборочная дисперсия.

Выборочной дисперсией называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\overline{x_в}$.

Если все значения $x_1, x_2, ..., x_n$ признака выборки объема различны, то

$$D_в = \frac{\sum_{i=1}^{n} (x_i - \overline{x_в})^2}{n}. \quad (3)$$

Если же значения признака $x_1, x_2, ..., x_k$ имеют соответственно частоты $m_1, m_2, ..., m_k$, причем $m_1 + m_2 + ... + m_k = n$, то

$$D_в = \frac{\sum_{i=1}^{k} m_i \left(x_i - \overline{x_в}\right)^2}{n} \quad . (4)$$

Пример. Выборочная совокупность задана таблицей распределения:

		2	3	4
m_i	0	15	10	5

Найти выборочную дисперсию.

Решение.

Найдем выборочную среднюю по формуле (2):

$$\overline{x_в} = \frac{20 \cdot 1 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{20 + 15 + 10 + 5} = \frac{100}{50} = 2 \quad .$$

Найдем выборочную дисперсию:

$$D_в = \frac{20 \cdot (1-2)^2 + 15 \cdot (2-2)^2 + 10 \cdot (3-2)^2 + 5 \cdot (4-2)^2}{50} = \frac{50}{50} = 1 \quad .$$

Кроме дисперсии, для характеристики рассеяния значений признака выборочной совокупности вокруг своего среднего значения пользуются средним квадратическим отклонением.

Выборочным средним квадратическим отклонением (стандартом) называют квадратный корень из выборочной дисперсии: $\sigma_в = \sqrt{D_в}$.

Исправленная выборочная дисперсия.

Выборочная дисперсия является смещенной оценкой генеральной дисперсии, поэтому в статистике применяют также исправленную выборочную дисперсию, которая является несмещенной оценкой генеральной дисперсии и обозначается s^2 .

Исправленная выборочная дисперсия находится по формуле:

$$s^2 = \frac{n}{n-1} D_в = \frac{\sum_{i=1}^{k} m_i \left(x_i - \overline{x_в}\right)^2}{n-1} \quad . (5)$$

Для оценки среднего квадратического отклонения генеральной совокупности используют «исправленное» среднее квадратическое отклонение, которое равно квадратному корню из исправленной дисперсии:

$$s = \sqrt{\frac{\sum_{i=1}^{k} m_i \left(x_i - \overline{x_в}\right)^2}{n-1}}$$. (6)

Отметим, что s не является несмещенной оценкой.

Замечание. Сравнивая формулы (4) и (6), видим, что они отличаются только знаменателями. Очевидно, что при больших значениях объема выборки выборочная и исправленная дисперсии отличаются мало. На практике пользуются исправленной дисперсией, если примерно $n < 30$.

2.9 Метод доверительных интервалов при заданных условиях.

Для вычисления характеристик выборки удобно пользоваться эмпирическими моментами, определения которых аналогичны определениям соответствующих теоретических моментов. В отличие от теоретических эмпирических моментов вычисляют по данным наблюдений.

Обычным эмпирическим моментом порядка k называют среднее значение k-х степеней разностей $x_i - C$: $M_k' = \frac{\sum m_i (x_i - C)^k}{n}$,

где – наблюдаемая варианта, m_i – частота варианты, $n = \sum m_i$ – объем выборки, C – произвольное постоянное число (ложный нуль).

Начальным эмпирическим моментом порядка k называют обычный момент порядка k при $C = 0$: $M_k = \frac{\sum m_i x_i^k}{n}$. В частности, $M_1 = \frac{\sum m_i x_i}{n} = \overline{x_в}$.

Центральным эмпирическим моментом порядка k называют обычный момент порядка k при $C = \overline{x_в}$: $m_k = \frac{\sum m_i \left(x_i - \overline{x_в}\right)^k}{n}$.

2.10 Метод моментов.

Точечной называют оценку, которая определяется одним числом. Все оценки, рассмотренные выше, – точечные.

Можно доказать, что начальные и центральные эмпирические моменты являются состоятельными оценками соответственно начальных и центральных теоретических моментов того же порядка. На этом основан метод моментов, предложенный К. Пирсоном.

Метод моментов точечной оценки неизвестных параметров заданного распределения состоит в приравнивании теоретических моментов рассматриваемого распределения соответствующим эмпирическим моментам того же порядка.

Если распределение определяется одним параметром, то для его отыскания приравнивают один теоретический момент одному эмпирическому моменту того же порядка. Например, можно приравнять начальный теоретический момент первого порядка начальному эмпирическому моменту первого порядка: $v_1 = M_1$. Учитывая, что и $M_1 = \overline{x_в}$, получим: $M(X) = \overline{x_в}$. (*)

Математическое ожидание является функцией от неизвестного параметра заданного распределения, поэтому, решив уравнение (*) относительно неизвестного параметра, тем самым получим его точечную оценку.

Если распределение определяется двумя параметрами, то приравнивают два теоретических момента двум соответствующим эмпирическим моментам того же порядка. Например, можно приравнять начальный теоретический момент первого порядка начальному эмпирическому моменту первого порядка и центральный теоретический момент второго порядка центральному эмпирическому моменту второго порядка: $v_1 = M_1$, $\mu_2 = m_2$.

Разумеется, для вычисления выборочной средней $\overline{x_в}$ и выборочной дисперсии надо располагать выборкой $x_1, x_2, ..., x_n$.

Интервальные оценки. Доверительная вероятность (надежность). Доверительный интервал.

При выборке малого объема точечная оценка может приводить к грубым ошибкам. По этой причине при небольшом объеме выборки следует пользоваться интервальными оценками.

Интервальной называют оценку, которая определяется двумя числами – концами интервала. Интервальные оценки позволяют установить точность и надежность оценок.

Пусть найденная по данным выборки статистическая характеристика Θ^* служит оценкой неизвестного параметра Θ. Будем считать Θ постоянным числом (Θ может быть и случайной величиной). Понятно, что Θ^* тем точнее определяет параметр Θ, чем меньше абсолютная величина разности . Другими словами, если $\delta > 0$ и $|\Theta - \Theta^*| < \delta$, то чем меньше δ, тем оценка точнее. Таким образом, положительное число δ характеризует точность оценки.

Однако статистические методы не позволяют категорически утверждать, что оценка Θ^* удовлетворяет неравенству $|\Theta - \Theta^*| < \delta$; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется.

Надежностью (доверительной вероятностью) оценки θ по θ^* называют вероятность γ, с которой осуществляется неравенство $|\theta-\theta^*|<\delta$. Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Пусть вероятность того, $|\theta-\theta^*|<\delta$, равна γ.

Далее имеем: $-\delta<\theta-\theta^*<\delta$ или $\theta^*-\delta<\theta<\theta^*+\delta$.

Тогда вероятность того, что интервал $(\theta^*-\delta;\theta^*+\delta)$ заключает в себе (покрывает) неизвестный параметр θ, равна γ, т.е. $P(\theta^*-\delta<\theta<\theta^*+\delta)=\gamma$.

Доверительным называют интервал $(\theta^*-\delta;\theta^*+\delta)$, который покрывает неизвестный параметр с заданной надежностью γ.

Замечание. Интервал $(\theta^*-\delta;\theta^*+\delta)$ имеет случайные концы (их называют доверительными границами). Поэтому доверительные границы сами являются случайными величинами – функциями от $x_1, x_2, \dots, x_n$.

Интервальные оценки для математического ожидания и среднего квадратического отклонения СВ, имеющей нормальное распределение.

Пусть количественный признак генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения может быть известно или неизвестно. Требуется оценить неизвестное математическое ожидание a по выборочной средней $\overline{x_в}$.

Интервальной оценкой (с надежностью γ) математического ожидания a нормально распределенного количественного признака X по выборочной средней $\overline{x_в}$ при известном среднем квадратическом отклонении σ генеральной совокупности служит доверительный интервал

$$\overline{x_в}-t\cdot\frac{\sigma}{\sqrt{n}}<a<\overline{x_в}+t\cdot\frac{\sigma}{\sqrt{n}}, \quad (7)$$

где $t\cdot\frac{\sigma}{\sqrt{n}}=\delta$ – точность оценки, – объем выборки, – значение аргумента функции Лапласа $\Phi(t)$, при котором $\Phi(t)=\frac{\gamma}{2}$; при неизвестном σ (и объеме выборки $n<30$)

$$\overline{x_в} - t_\gamma \cdot \frac{s}{\sqrt{n}} < a < \overline{x_в} + t_\gamma \cdot \frac{s}{\sqrt{n}}$$, (8)

где s – «исправленное» выборочное среднее квадратическое отклонение, t_γ находят по таблице приложений по заданным n и γ.

Пусть количественный признак X генеральной совокупности распределен нормально. Требуется оценить неизвестное генеральное среднее квадратическое отклонение σ по «исправленному» выборочному среднему квадратическому отклонению s.

Интервальной оценкой (с надежностью γ) среднего квадратического отклонения σ нормально распределенного количественного признака X по «исправленному» выборочному среднему квадратическому отклонению s служит доверительный интервал (для $n \geq 30$)

$s \cdot (1-q) < \sigma < s \cdot (1+q)$ (при $q < 1$);

$0 < \sigma < s \cdot (1+q)$ (при $q > 1$), (9)

где q находят по таблице приложений по заданным n и γ.

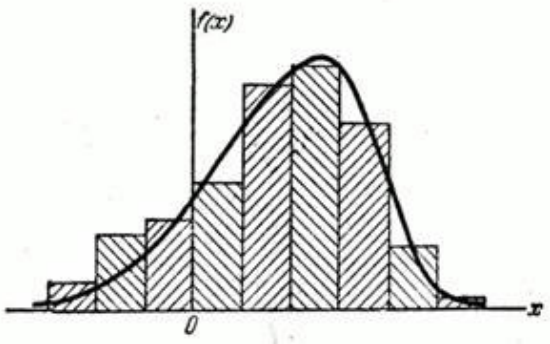
Пример. Найти доверительный интервал для оценки с надежностью 0,95 неизвестного математического ожидания a нормально распределенного признака X генеральной совокупности, если генеральное среднее квадратическое отклонение $\sigma = 5$, выборочная средняя $\overline{x_в} = 14$, объем выборки $n = 25$.

Решение. Найдем доверительный интервал по формуле (7). Все величины, кроме , известны. Найдем из соотношения $\Phi(t) = \frac{0,95}{2} = 0,475$. По таблице приложений находим $t = 1,96$. Следовательно, искомый доверительный интервал:

$12,04 < a < 15,96$.

2.11 Выравнивание статистических рядов.

Во всяком статистическом распределении неизбежно присутствуют элементы случайности, связанные с тем, что число наблюдений ограничено, что произведены именно те, а не другие опыты, давшие именно те, а не другие результаты. Только при очень большом числе наблюдений эти элементы случайности сглаживаются, и случайное явление обнаруживает в полной мере присущую ему закономерность. На практике мы почти никогда не имеем дела с таким большим числом



наблюдений и вынуждены считаться с тем, что любому статистическому распределению свойственны в большей или меньшей мере черты случайности. Поэтому при обработке статистического материала часто приходится решать вопрос о том, как подобрать для данного статистического ряда теоретическую кривую распределения, выражающую лишь существенные черты статистического материала, но не случайности, связанные с недостаточным объемом экспериментальных данных. Такая задача называется задачей выравнивания (сглаживания) статистических рядов.

Задача выравнивания заключается в том, чтобы подобрать теоретическую плавную кривую распределения, с той или иной точки зрения наилучшим образом описывающую данное статистическое распределение (рис. 1).

Задача о наилучшем выравнивании статистических рядов, как и вообще задача о наилучшем аналитическом представлении эмпирических функций, есть задача в

Рис. 1

значительной мере неопределенная, и решение ее зависит от того, что условиться считать «наилучшим».

Например, при сглаживании эмпирических зависимостей очень часто исходят из так называемого принципа или метода наименьших квадратов, считая, что наилучшим приближением к эмпирической зависимости в данном классе функций является такое, при котором сумма квадратов отклонений обращается в минимум. При этом вопрос о том, в каком именно классе функций следует искать наилучшее приближение, решается уже не из математических соображений, а из соображения, связанных с физикой решаемой задачи, с учетом характера полученной эмпирической кривой и степени точности произведенных наблюдений. Часто принципиальный характер функции, выражающей исследуемую зависимость, известен заранее из теоретических соображении, из опыта же требуется получить лишь некоторые численные параметры, входящие в выражение функции; именно эти параметры подбираются с помощью метода наименьших квадратов.

Аналогично обстоит дело и с задачей выравнивания статистических рядов. Как правило, принципиальный вид теоретической кривой выбирается заранее из соображений, связанных с существом задачи, а в некоторых случаях просто с внешним видом статистического распределения. Аналитическое выражение выбранной кривой распределения зависит от некоторых параметров; задача выравнивания статистического ряда переходит в задачу рационального выбора тех значений параметров, при которых соответствие между статистическим и теоретическим распределениями оказывается наилучшим.

Предположим, например, что исследуемая величина X есть ошибка измерения, возникающая в результате суммирования воздействий множества независимых элемен-

тарных ошибок; тогда из теоретических соображений можно считать, что величина X подчиняется нормальному закону:

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

и задача выравнивания переходит в задачу о рациональном выборе параметров m и σ в выражении (1).

Бывают случаи, когда заранее известно, что величина X распределяется статистически приблизительно равномерно на некотором интервале; тогда можно поставить задачу о рациональном выборе параметров того закона равномерной плотности

$$f(x)=\begin{cases}\dfrac{1}{\beta-\alpha} & \text{при } \alpha<x<\beta,\\ 0 & \text{при } x<\alpha \text{ или } x>\beta\end{cases}$$

которым можно наилучшим образом заменить (выровнять) заданное статистическое распределение.

Следует при этом иметь в виду, что любая аналитическая функция $f(x)$, с помощью которой выравнивается статистическое распределение, должна обладать основными свойствами плотности распределения:

$$\left.\begin{array}{l} f(x)\geq 0;\\ \int\limits_{-\infty}^{\infty} f(x)dx=1\end{array}\right\} \qquad (2)$$

Предположим, что, исходя из тех или иных соображений, нами выбрана функция $f(x)$, удовлетворяющая условиям (2), с помощью корой мы хотим выровнять данное статистическое распределение; в выражение этой функции входит несколько параметров $a,b,\cdots$; требуется подобрать эти параметры так, чтобы функция $f(x)$ наилучшим образом описывала данный статистический материал. Один из методов, применяемых для решения этой задачи, - это так называемый метод моментов.

Согласно методу моментов, параметры $a,b,\cdots$ выбираются с таким расчетом, чтобы несколько важнейших числовых характеристик (моментов) теоретического распределения были равны соответствующим статистическим характеристикам. Например, если теоретическая кривая $f(x)$ зависит только от двух параметров a и b, эти параметры выбираются так, чтобы математическое ожидание m_x и дисперсия D_x^* теоретического рас-

пределения совпадали с соответствующими статистическими характеристиками m_x^* и D_x^*. Если кривая $f(x)$ зависит от трех параметров, можно подобрать их так, чтобы совпали первые три момента и т.д. При выравнивании статистических рядов может оказаться полезной специально разработанная система кривых Пирсона, каждая из которых зависит в общем случае от четырех параметров. При выравнивании эти параметры выбираются с тем расчетом, чтобы сохранить первые четыре момента статистического распределения (математическое ожидание, дисперсию, третий и четвертый моменты). Оригинальный набор кривых распределения, построенных по иному принципу, дал Н.А. Бородачев. Принцип, на котором строится система кривых Н.А. Бородачева, заключается в том, что выбор типа теоретической кривой основывается не на внешних формальных признаках, а на анализе физической сущности случайного явления или процесса, приводящего к тому или иному закону распределения.

Следует заметить, что при выравнивании статистических рядов нерационально пользоваться моментами порядка выше четвертого, так как точность вычисления моментов резко падает с увеличением их порядка.

Пример. Произведено 500 измерений боковой ошибки наводки при стрельбе с самолета по наземной цели. Результаты измерений (в тысячных долях радиана) сведены в статистический ряд:

l_i	-4; -3	-3; -2	-2; -1	-1; 0	0; 1	1; 2	2; 3	3; 4
m_i	6	25	72	133	120	88	46	10
p_i^*	0,012	0,050	0,144	0,266	0,240	0,176	0,092	0,020

Требуется выровнять это распределение с помощью нормального закона:

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

Нормальный закон зависит от двух параметров: m и σ. Подберем эти параметры так, чтобы сохранить первые два момента – математическое ожидание и дисперсию – статистического распределения.

Вычислим приближенно статистическое среднее ошибки наводки, причем за представителя каждого разряда примем его середину:

$$m_x^*=-3,5\cdot 0,012-2,5\cdot 0,050-1,5\cdot 0,144-0,5\cdot 0,266+$$
$$+0,5\cdot 0,240+1,5\cdot 0,176+2,5\cdot 0,092+3,5\cdot 0,020=0,168$$

Для определения дисперсии вычислим сначала второй начальный момент), полагая $s=2, k=8$

$$\alpha_2^* = \sum_{i=1}^{\infty} \tilde{x}_i^2 p_i^* = 2,216$$

Пользуясь выражением дисперсии через второй начальный момент, получим:

$$D_x^* = \alpha_2^* - \left(m_x^*\right)^2 = 2,126 - 0,028 = 2,098$$

Выберем параметры m и σ нормального закона так, чтобы выполнялись условия:

$$m = m_x^*, \quad \sigma^2 = D_x^*,$$

то есть примем:

$$m = 0,168; \quad \sigma = 1,448$$

Напишем выражение нормального закона:

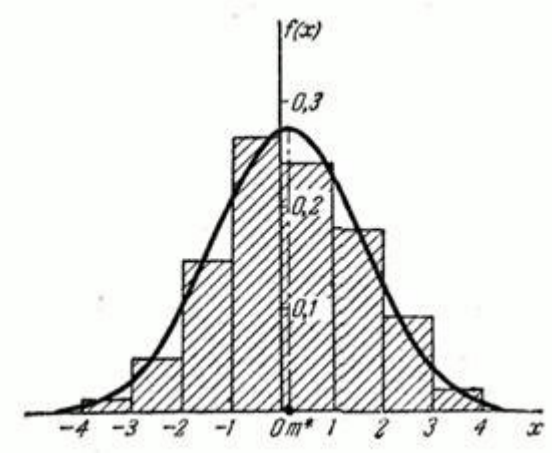
$$f(x) = \frac{1}{1,448\sqrt{2\pi}} e^{-\frac{(x-0,168)^2}{2 \cdot 1,448^2}}$$

Пользуясь таблицами, вычислим значения $f(x)$ на границах разрядов

x	-4	-3	-2	-1	0	1	2	3	4
$f(x)$	0,004	0,025	0,090	0,199	0,274	0,234	0,124	0,041	0,008

Построим на одном графике (рис. 2) гистограмму и выравнивающую ее кривую распределения.

Из графика видно, что теоретическая кривая распределения $f(x)$, сохраняя, в основном существенные особенности статистического распределения, свободна от случайных неправильностей хода гистограммы, которые, по-видимому, могут быть отнесены за счет случайных причин; более серьезное обоснование последнему суждению будет дано в следующем параграфе.



Примечание. В данном примере при определении D_x^*, мы воспользовались выражением статистической дисперсии через второй начальный момент. Этот прием можно рекомендовать только в случае, когда математическое ожидание m_x^* иссле-

дуемой случайной величины X сравнительно невелико; в противном случае выражают дисперсию D_x^* как разность близких чисел и получают весьма малую точность.

Пример. С целью исследования закона рас

пределения ошибки измерения дальности с помощью радиодальномера произведено 400 измерений дальности. Результаты опытов представлены в виде статистического ряда:

$l_i\,(м)$	20;30	30;40	40;50	50;60	60;70	70;80	80;90	90;100
m_i	21	72	66	38	51	56	64	32
p_i^*	0,052	0,180	0,165	0,095	0,128	0,140	0,160	0,080

Выровнять статистический ряд с помощью закона равномерной плотности.

Решение. Закон равномерной плотности выражается формулой

$$f(x)=\begin{cases}\dfrac{1}{\beta-\alpha} & \text{при } \alpha<x<\beta,\\ 0 & \text{при } x<\alpha \text{ или } x>\beta\end{cases}$$

и зависит от двух параметров α и β. Эти параметры следует выбрать так, чтобы сохранить первые два момента статистического распределения – математическое ожидание m_x^* и дисперсию D_x^*. Из примера n^o 5.8 имеем выражения математического ожидания и дисперсии для закона равномерной плотности:

$$m_x=\frac{\alpha+\beta}{2};$$

$$D_x=\frac{(\beta-\alpha)^2}{12}.$$

Для того, чтобы упростить вычисления, связанные с определением статистических моментов, перенесем начало отсчета в точку $x_0=60$ и примем за представителя его разряда его середину. Ряд распределения имеет вид:

z_i'	−35	−25	−15	−5	5	15	25	35
p_i^*	0,052	0,180	0,165	0,095	0,128	0,140	0,160	0,080

где z_i' - среднее для разряда значение ошибки радиодальномера X' при новом начале отсчета.

Приближенное значение статистического среднего ошибки X' равно:

$$m_{x'}^*=\sum_{i=1}^{k} z_i' p_i^*=0,26$$

Второй статистический момент величины X' равен:

$$\alpha_2^* = \sum_{i=1}^{k} (\tilde{x}_i')^2 p_i^* = 447,8$$

откуда статистическая дисперсия:

$$D_{x'}^* = \alpha_2^* - \left(m_{x'}^*\right)^2 = 447,7$$

Переходя к прежнему началу отсчета, получим новое статистическое среднее:

$$m_x^* = m_{x'}^* + 60 = 60,26$$

в ту же статистическую дисперсию:

$$D_x^* = D_{x'}^* = 447,7$$

Параметры закона равномерной плотности определяются уравнениями:

$$\frac{\alpha+\beta}{2} = 60,26; \quad \frac{(\beta-\alpha)^2}{12} = 447,7$$

Решая эти уравнения относительно α и β, имеем:

$$\alpha \approx 23,6; \quad \beta \approx 96,9$$

откуда

$$f(x) = \frac{1}{\beta-\alpha} = \frac{1}{73,3} \approx 0,0136$$

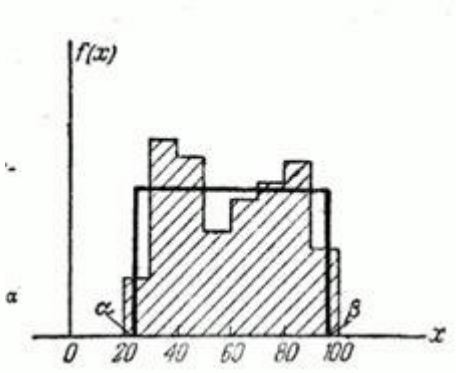


Рис. 3

На рис. 3. показаны гистограмма и выравнивающий ее закон равномерной плотности $f(x)$.

2.12 Виды зависимостей между величинами. Функция регрессии.

Условимся обозначать через X независимую переменную, а через Y – зависимую переменную.

Зависимость величины Y от X называется **функциональной**, если каждому значению величины X соответствует единственное значение величины Y. С функциональной зависимостью мы встречаемся, например, в математике, при изучении физических законов. Обратим внимание на то, что если X – детерминированная величина (т.е. принимающая вполне определённые значения), то и функционально зависящая от неё величина Y тоже является детерминированной; если же X – случайная величина, то и Y также случайная величина.

Однако гораздо чаще в окружающем нас мире имеет место не функциональная, а **стохастическая**, или **вероятностная**, **зависимость**, когда каждому фиксированному значению независимой переменной X соответствует не одно, а множество значений переменной Y, причём сказать заранее, какое именно значение примет величина Y, нельзя. Более частое появление такой зависимости объясняется действием на результирующую переменную не только контролируемого или контролируемых факторов (в данном случае таким контролируемым фактором является переменная X), а и многочисленных неконтролируемых случайный факторов. В этой ситуации переменная Y является случайной величиной. Переменная же X может быть, как детерминированной, так и случайной величиной. Следует заметить, что со стохастической зависимостью мы уже сталкивались в дисперсионном анализе.

Допустим, что существует стохастическая зависимость случайной переменной Y от X. Зафиксируем некоторое значение x переменной X. При $X = x$ переменная Y в силу её стохастической зависимости от X может принять любое значение из некоторого множества, причём какое именно – заранее неизвестно. Среднее этого множества называют **групповым генеральным средним** переменной Y при $X = x$ или **математическим ожиданием случайной величины Y, вычисленным при условии, что X = x; это условное математическое ожидание обозначают так: M(Y/X = x)**. Если существует стохастическая зависимость Y от X, то прежде всего стараются выяснить, изменяются или нет при изменении x условные математические ожидания $M(Y/X=x)$. Если при изменении x условные математические ожидания $M(Y/X=x)$ изменяются, то говорят, что имеет место **корреляционная зависимость** величины Y от X; если же условные математические ожидания остаются неизменными, то говорят, что корреляционная зависимость величины Y от X отсутствует.

Функция $\varphi(x)=M(Y/X=x)$, описывающая изменение условного математического ожидания случайной переменной Y при изменении значений x переменной X, называется **функцией регрессии**.

Выясним, почему именно при наличии стохастической зависимости интересуются поведением условного математического ожидания.

Рассмотрим пример. Пусть X – уровень квалификации рабочего, Y – его выработка за смену. Ясно, что зависимость Y от X не функциональная, а стохастическая: на выработку помимо квалификации влияет множество других факторов. Зафиксируем значение x уровня квалификации: ему соответствует некоторое множество значений выработки Y. Тогда $M(Y/X = x)$ – средняя выработка рабочего при условии, что его уровень квалификации равен x, или, иначе говоря, $M(Y/X = x)$ – это норматив выработки при уровне квали-

фикации, равном x. Зная зависимость этого норматива от уровня квалификации, можно для любого уровня квалификации рассчитать норматив выработки и, сравнив его с реальной выработкой, оценить работу рабочего.

Обратим внимание на то, что введённые понятия стохастической и корреляционной зависимости относились к генеральной совокупности. Поясним эти понятия числовым примером.

Пример. Допустим, что одновременно изучаются две случайные величины X и Y, или, иначе говоря, двумерная случайная величина (X, Y), которая задана табл. 1.

Таблица 1.

	i			
	x_i / y_i	$_1 = 2$	$_2 = 5$	$_3 = 8$
	$y_1 = 0,4$	,15	,12	,03
	$y_2 = 0,8$	,05	,30	,35

Табл. 1 называют **таблицей распределения двумерной величины (X, Y)**;её следует понимать так. Случайная величина X может принять одно из следующих значений: 2, 5 и 8. Случайная величина Y – значения 0,4 и 0,8. Число 0,15 – это вероятность того, что X = 2 и одновременно Y = 0,4, или, иначе говоря, вероятность произведения двух событий; события, состоящего в том, что X = 2, и события, состоящего в том, что Y = 0,4, т.е. $P((X=2)(Y=0,4)) = 0,15$. Аналогично, вероятность $P((X=2)(Y=0,8)) = 0,05$ и т.д. Обратим внимание на следующее: поскольку в табл. 1 указаны все возможные значения величин X и Y, сумма вероятностей, стоящих в таблице, должна быть равна единице: 0,15 + 0,05 + 0,12 + 0,30 + 0,03 + 0,35 = 1.

Прежде чем выяснить тип зависимости величины Y от X, найдём:

а) Закон распределения величины X. Он представлен табл. 2.

Таблица 2.

x	$x_1 = 2$	$x_2 = 5$	$x_3 = 8$	
P $(X = x)$	0,15 + 0,05 = 0,2	0,12 + 0,30 = 0,42	0,35 + 0,03 = 0,38	= 1

M(X) = 5,54, D(X) = 4,9284

Действительно, например, величина X примет значение, равное 2, только в том случае, когда одновременно с этим величина Y примет значение 0,4 или 0,8, т.е.

P(X = 2) = P((X = 2)(Y = 0,4)) + P((X = 2)(Y = 0,08)) = 0,15 + 0,05 = 0,2.

Справа от ряда распределения величины X находятся её математическое ожидание и дисперсия.

б) Закон распределения величины Y. Он имеет вид табл. 3.

Таблица 3.

y	$y_1 = 0,4$	$y_2 = 0,8$	
P $(Y = y)$	0,15 + 0,12 + 0,03 = 0,30	0,05 + 0,30 + 0,35 = 0,7	= 1

M(Y) = 0,68, D(Y) = 0,0336

в) Условные законы распределения величины Y, а именно закон распределения величины Y сначала при условии, что X = 2, затем при условии, что X=5, и наконец, при условии, что X = 8.

Итак, пусть X = 2. Тогда условная вероятность

$$P(\mathrm{Y} = 0{,}4/\mathrm{X} = 2) = \frac{P(Y = 0{,}4)(X = 2))}{P(X = 2)} = \frac{0{,}15}{0{,}2} = 0{,}75,$$

а условная вероятность

$$P(\mathrm{Y} = 0{,}8/\mathrm{X} = 2) = \frac{P((Y = 0{,}8)(X = 2))}{P(X = 2)} = \frac{0{,}05}{0{,}2} = 0{,}25.$$

Таким образом, закон распределения величины Y при условии, что X = 2, задан табл. 4.

Таблица 4.

y	$y_1 = 0,4$	$y_2 = 0,8$	
$P(Y = y/X = 2)$	0,75	0,25	= 1

M(Y/X = 2) = 0,4*0,75 + 0,8*0,25 = 0,5, D(Y/X = 2) = 0,03

Справа помещено условное математическое ожидание и значение условной дисперсии. Покажем, как вычисляется условная дисперсия. Общая формула условной дисперсии имеет вид

$$D(Y/X = x) = M[(Y/X = x) - M(Y/X = x)]^2. \qquad (23)$$

Для табл. 4 получаем

$$D(Y/X = 2) = M\,[(Y/X = 2) - M(Y/X = 2)]^2 = M\,[(Y/X = 2) - 0{,}5]^2 = \sum_{i=1}^{2} (y_i - 0{,}5)^2 \cdot P(Y = y_i/X = 2) = (0{,}4 - 0{,}5)^2 \cdot 0{,}75 + (0{,}8 - 0{,}5)^2 \cdot 0{,}25 = 0{,}03.$$

Пусть X = 5. Тогда $P(Y = 0{,}4/X = 5) = \frac{P((Y = 0{,}4)(X = 5))}{P(X = 5)} = \frac{0{,}12}{0{,}42} = \frac{2}{7}$; $P(Y=0{,}8/X=5) = \frac{P((Y = 0{,}8)(X = 5))}{P(X = 5)} = \frac{0{,}30}{0{,}42} = \frac{5}{7}$.

Таким образом, закон распределения величины Y при условии, что X = 5, имеет вид табл. 5.

Таблица 5.

y	,4	,8	
$P(Y = y/X = 5)$	/7	/7	= 1

$M(Y/X = 5) = \frac{24}{35} \approx 0{,}686$, $D(Y/X = 5) = 0{,}03265$.

И наконец, при X = 8 ряд распределения задан таблицей 14.

Таблица 6.

y	,4	,8	
$P(Y = y/X = 8)$			= 1

$M(Y/X = 8) = \frac{73}{95} \approx 0{,}768$, $D(Y/X = 8) = 0{,}01163$

Из табл. 4–6 видно, что зависимость Y от X стохастическая, поскольку при каждом фиксированном значении величины X величина Y может быть равной либо 0,4, либо 0,8, причём какому именно из этих чисел она будет равна – сказать заранее нельзя. Ясно прослеживается и корреляционная зависимость величины Y от X, поскольку с изменением значений x величины X меняются и условные математические ожидания $M(Y/X = x)$.

Функция регрессии, т.е. зависимость условного математического ожидания M(Y/X = x) от x, задаётся в виде табл. 7.

Таблица 7.

x		5	8
M(Y/X = *x*)	,5	24/35 ≈ 0,686	73/95 ≈ 0,768

Определение. *Начальным моментом порядка* $k+s$ системы двух случайных величин $(X;Y)$ называется действительное число $\alpha_{k,s}$, определяемое по формуле:

$\alpha_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{ij}$, если $(X;Y)$ – система двух дискретных случайных величин;

$\alpha_{k,s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^k y^s f_{X,Y}(x,y)\,dx\,dy$, если $(X;Y)$ – система двух непрерывных случайных величин.

Определение. *Центральным моментом порядка* $k+s$ системы двух случайных величин $(X;Y)$ называется действительное число $\mu_{k,s}$, определяемое по формуле:

$\mu_{k,s} = \sum_i \sum_j (x_i - m_X)^k (y_j - m_Y)^s p_{ij}$, если $(X;Y)$ – система двух дискретных случайных величин;

$\mu_{k,s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - m_X)^k (y - m_Y)^s f_{X,Y}(x,y)\,dx\,dy$, если $(X;Y)$ – система двух непрерывных случайных величин.

На практике чаще всего встречаются моменты первого и второго порядков. Очевидно, что начальные моменты первого порядка есть не что иное, как математические ожидания компонент *X* и *Y*:

$\alpha_{1,0} = m_X$, $\alpha_{0,1} = m_Y$.

Точка с координатами $(m_X; m_Y)$ на плоскости *xOy* представляет собой характеристику положения случайной точки $(X;Y)$, а ее рассеивание (разброс) происходит вокруг $(m_X; m_Y)$.

Центральные моменты первого порядка, очевидно, равны нулю, т.е.

$\mu_{1,0} = \mu_{0,1} = 0$.

Имеются три начальных момента второго порядка – $\alpha_{2,0}$, $\alpha_{0,2}$ и $\alpha_{1,1}$. Причем первые два из них есть не что иное, как начальные моменты второго порядка компонент *X* и *Y*:

$$\alpha_{2,0} = \alpha_2[X], \quad \alpha_{0,2} = \alpha_2[Y].$$

Имеются три центральных момента второго порядка $\mu_{2,0}$, $\mu_{0,2}$ и $\mu_{1,1}$. Первые два из них представляют собой дисперсии компонент *X* и *Y* соответственно:

$$\mu_{2,0} = \mathrm{D}[X], \quad \mu_{0,2} = \mathrm{D}[Y].$$

2.13 Корреляционное отношение, коэффициент детерминации. Корреляционная зависимость.

Определение. Центральный момент второго порядка $\mu_{1,1}$ называется *ковариацией* случайной величины $(X; Y)$.

Для момента $\mu_{1,1}$ используется обозначение $K_{X,Y} = \mathrm{cov}(X; Y)$.

Замечание. По определению ковариации: $K_{X,Y} = K_{Y,X}$.

В механической интерпретации, когда распределение вероятностей на плоскости *xOy* трактуется как распределение единичной массы на этой плоскости, точка $(m_X; m_Y)$ есть не что иное, как *центр масс* распределения; дисперсии $\mathrm{D}[X]$ и $\mathrm{D}[Y]$ – *моменты инерции* распределения относительно точки $(m_X; m_Y)$ в направлении осей *Ox* и *Oy* соответственно, а ковариация – это *центробежный момент инерции* распределения масс.

Теорема. Если случайные величины *X* и *Y* независимы, то $K_{X,Y} = 0$.

Замечание. Как правило, $K_{X,Y}$ удобнее вычислять по формуле

$$K_{X,Y} = \alpha_{1,1} - \alpha_{1,0} \cdot \alpha_{0,1}.$$

Ковариация $K_{X,Y}$ характеризует не только степень зависимости двух случайных величин $(X; Y)$, но также их рассеивание вокруг точки $(m_X; m_Y)$. Однако размерность ковариации $K_{X,Y}$ равна произведению размерностей случайных величин *X* и *Y*. Чтобы получить безразмерную величину, характеризующую только зависимость, а не разброс, ковариацию $K_{X,Y}$ делят на произведение $\sigma_X \sigma_Y$: $\rho_{X,Y} = \dfrac{K_{X,Y}}{\sigma_X \sigma_Y}$.

Определение. Величина $\rho_{X,Y}$ называется *коэффициентом корреляции* случайных величин *X* и *Y*.

Коэффициент корреляции $\rho_{X,Y}$ характеризует степень зависимости случайных величин *X* и *Y*, причем не любой зависимости, а только *линейной*, проявляющейся в том, что при возрастании одной случайной величины другая проявляет тенденцию также возрастать (или убывать). В первом случае $\rho_{X,Y} > 0$ и говорят, что случайные величины *X* и *Y* *связаны положительной корреляцией*, во втором случае $\rho_{X,Y} < 0$ и говорят, что случайные величины *X* и *Y* *связаны отрицательной корреляцией*. Модуль коэффициента корреляции случайных величин *X* и *Y* характеризует *степень тесноты линейной зависимости* между ними. Если линейной зависимости нет, то $\rho_{X,Y} = 0$.

2.14 Коэффициент детерминации.

Исследование начинается с теории, устанавливающей связь между явлениями. Из всего круга факторов, влияющих на результативный признак, выделяются наиболее существенные факторы. После того, как было выявлено наличие взаимосвязи между изучаемыми признаками, определяется точный вид этой зависимости с помощью регрессионного анализа.

Регрессионный анализ заключается в определении аналитического выражения (в определении функции), в котором изменение одной величины (результативного признака) обусловлено влиянием независимой величины (факторного признака). Количественно оценить данную взаимосвязь можно с помощью построения уравнения регрессии или регрессионной функции.

Базисной регрессионной моделью является модель парной (однофакторной) регрессии. Парная регрессия– уравнение связи двух переменных y и x:

$$y = f(x)$$

где y – зависимая переменная (результативный признак);

x –независимая, объясняющая переменная (факторный признак).

В зависимости от характера изменения ус изменением x различают линейные и нелинейные регрессии.

Линейная регрессия $Y = \beta_0 + \beta_1 X + \varepsilon$

Данная регрессионная функция называется полиномом первой степени и используется для описания равномерно развивающихся во времени процессов.

Наличие случайного члена ε (ошибки регрессии) связано с воздействием на зависимую переменную других неучтенных в уравнении факторов, с возможной нелинейностью модели, ошибками измерения, следовательно, появление случайной ошибки уравнения регрессии может быть обусловлено следующими объективными причинами:

1) нерепрезентативность выборки. В модель парной регрессии включается фактор, не способный полностью объяснить вариацию результативного признака, который может быть подвержен влиянию многих других факторов (пропущенных переменных) в гораздо большей степени. Например, заработная плата может зависеть, кроме квалификации, от уровня образования, стажа работы, пола и пр.;

2) существует вероятность того, что переменные, участвующие в модели, могут быть измерены с ошибкой. Например, данные по расходам семьи на питание составляются на основании записей участников опросов, которые, как предполагается, тщательно фиксируют свои ежедневные расходы. Разумеется, при этом возможны ошибки.

На основе выборочного наблюдения оценивается выборочное уравнение регрессии (линия регрессии):

$$y_x = a + bx,$$

где a, b – оценки параметров уравнения регрессии (α, β).

Аналитическая форма зависимости между изучаемой парой признаков (регрессионная функция) определяется с помощью следующих методов:

На основе теоретического и логического анализа природы изучаемых явлений, их социально-экономической сущности. Например, если изучается зависимость между доходами населения и размером вкладов населения в банки, то очевидно, что связь прямая.

Графический метод, когда характер связи оценивается визуально.

Эту зависимость можно наглядно увидеть, если построить график, отложив на оси абсцисс значения признака x, а на оси ординат – значения признака y. Нанеся на график точки, соответствующие значениям x и y, получим корреляционное поле:

а) если точки беспорядочно разбросаны по всему полю – это говорит об отсутствии зависимости между этими признаками;

б) если точки концентрируются вокруг оси, идущей от нижнего левого угла в верхний правый – то имеется прямая зависимость между признаками;

в) если точки концентрируются вокруг оси, идущей от верхнего левого угла в нижний правый – то обратная зависимость между признаками.

Если на корреляционном поле соединим точки отрезками прямой, то получим ломаную линию с некоторой тенденцией к росту. Это будет эмпирическая линия связи

или эмпирическая линия регрессии. По ее виду можно судить не только о наличии, но и о форме зависимости между изучаемыми признаками.

Построение уравнения парной регрессии

Построение уравнения регрессии сводится к оценке ее параметров. Эти оценки параметров могут быть найдены различными способами. Одним их них является метод наименьших квадратов (МНК). Суть метода состоит в следующем. Каждому значению соответствует эмпирическое (наблюдаемое) значение y. Построив уравнение регрессии, например уравнение прямой линии, каждому значению будет соответствовать теоретическое (расчетное) значение y_x. Наблюдаемые значения не лежат в точности на линии регрессии, т.е. не совпадают с . Разность между фактическим и расчетным значениями зависимой переменной называется остатком:

$$e = y - y_x$$

МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака уот теоретических y_x, т.е. сумма квадратов остатков, минимальна:

$$\Sigma(y - y_x)^2 \to min$$

Для линейных уравнений и нелинейных, приводимых к линейным, решается следующая система относительно а и b:

$$a\,n + b\Sigma x = \Sigma y$$
$$a\Sigma x + b\Sigma x^2 = \Sigma yx$$

где n– численность выборки.

Решив систему уравнений, получим значения а и b, что позволяет записать уравнение регрессии (регрессионное уравнение):

$y_x = a + bx$ где х – объясняющая (независимая) переменная;

y_x –объясняемая (зависимая) переменная;

Линия регрессии проходит через точку (x , y) и выполняются равенства:

$$e = 0, \quad y = y_x$$

Можно воспользоваться готовыми формулами, которые вытекают из этой системы уравнений:

$$b = \frac{cov(x,y)}{\sigma_x^2} = \frac{y \cdot x - y \cdot x}{x^2 - x^2};$$
$$a = \bar{y} - b \cdot \bar{x}$$

где – среднее значение зависимого признака;

– среднее значение независимого признака;

$y \cdot x$ –среднее арифметическое значение произведения, зависимого и независимого признаков;

σ_x^2 –дисперсия независимого признака;

$cov(x, y)$ –ковариация между зависимым и независимым признаками.

2.15 Значимость выборочных коэффициентов. Линейная парная регрессия.

Выборочной ковариацией двух переменных x, y называется средняя величина произведения отклонений этих переменных от своих средних

Параметр b при x имеет большое практическое значение и носит название коэффициента регрессии. Коэффициент регрессии показывает, на сколько единиц в среднем изменяется величина y при изменении факторного признака x на 1 единицу своего измерения.

Знак параметра b в уравнении парной регрессии указывает на направление связи:

если $b > 0$, то связь между изучаемыми показателями прямая, т.е. с увеличением факторного признаках увеличивается и результативный признаку, и наоборот;

если $b < 0$, то связь между изучаемыми показателями обратная, т.е. с увеличением факторного признаках результативный признак y уменьшается, и наоборот.

Значение параметра ав уравнении парной регрессии в ряде случаев можно трактовать как начальное значение результативного признака y. Такая трактовка параметра а возможна только в том случае, если значение $x = 0$ имеет смысл.

После построения уравнения регрессии, наблюдаемые значения y можно представить как: $y = y_x + e$

Остатки e, как и ошибки , являются случайными величинами, однако они, в отличие от ошибок , наблюдаемы. Остаток есть та часть зависимой переменной y, которую невозможно объяснить с помощью уравнения регрессии.

На основании уравнения регрессии могут быть вычислены теоретические значения y x для любых значений x.

В экономическом анализе часто используется понятие эластичности функции. Эластичность функции рассчитывается как относительное изменение y к относительному изменению x. Эластичность показывает, на сколько процентов изменяется функция при изменении независимой переменной на 1%.

Поскольку эластичность линейной функции не является постоянной величиной, а зависит от x, то обычно рассчитывается коэффициент эластичности как средний показатель эластичности.

Коэффициент эластичности показывает, на сколько процентов в среднем по совокупности изменится величина результативного признака у при изменении факторного признаках на 1% от своего среднего значения:

$$Э_x = b\frac{x}{y}$$

где x, y – средние значения переменных x и y в выборке.

Оценка качества построенной модели регрессии

Качество модели регрессии– адекватность построенной модели исходным (наблюдаемым) данным.

Чтобы измерить тесноту связи, т.е. измерить, насколько она близка к функциональной, нужно определить дисперсию, измеряющую отклонения y от yx и характеризующую остаточную вариацию, обусловленную прочими факторами. Они лежат в основе показателей, характеризующих качество модели регрессии.

При исследовании регрессии устанавливается однофакторная или многофакторная будет строиться модель и вид модели (линейный или нелинейный).

Обоснование вида модели состоит в выборе вида функции (некоторого аналитического выражения), с помощью которого можно будет описать изменение исследуемого показателя под воздействием факторов.

К обоснованию вида функции идут двумя путями: Теоретическим (анализируя экономическую природу x_0и x_j, выдвигается гипотеза о характере изменения показателя под действием фактора) И эмпирическим(закон изменения результативного показателя под действием фактора устанавливается путем анализа совокупности фактических данных по полям корреляции).

Наиболее употребительными выражениями при описании связи одного фактора и исследуемого показателя являются:

- Уравнение прямой - $x_0 = a_0 + a_1x_1$, - Уравнение параболы - $x_0 = a_0 + a_1x_1 + a_2x_1^2$,

-Уравнение гиперболы - $x_0 = a_0 + \frac{a_1}{x_1}$.

После обоснования парных взаимосвязей переходят к записи многофакторных моделей. В экономических исследованиях чаще всего применяется линейная многофакторная модель - $x_0 = a_0 + a_1x_1 + \cdots + a_nx_n$.

В качестве нелинейных моделей применяются

- Мультипликативная модель - $x_0 = a_0x_1^{a_1}x_2^{a_2}x_3^{a_3}\ldots$ или $x_0 = a_0a_1^{x_1}a_2^{x_2}a_3^{x_3}\ldots$

Для оценки значений параметров регрессионной модели чаще всего используется Метод наименьших квадратов (МНК). Этот метод можно применить как для линейных

моделей, так и для нелинейных, допускающих преобразование их к линейному виду путем замены переменных или дифференцированием.

При использовании МНК делаются определенные предпосылки относительно случайной составляющей ε. В модели $y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \varepsilon$ случайная составляющая ε представляет собой ненаблюдаемую величину. Поэтому в задачу регрессионного анализа входит не только построение самой модели, но и исследование случайных отклонений ε_i, т. е. остаточных величин.

Остатки представляют собой независимые случайные величины, и их среднее значение равно 0; они имеют одинаковую (постоянную) дисперсию и подчиняются нормальному распределению.

Статистические проверки параметров регрессии, показателей корреляции основаны на непроверяемых предпосылках распределения случайной составляющей ε_i. Связано это с тем, что оценки параметров регрессии должны отвечать определенным критериям: быть Несмещенными, состоятельными и эффективными. Эти свойства оценок, полученных по МНК, имеют чрезвычайно важное практическое значение в использовании результатов регрессии и корреляции.

Коэффициенты регрессии, найденные из системы нормальных уравнений, представляют собой выборочные оценки характеристики силы связи. Их несмещенность является желательным свойством, т. к. только в этом случае они могут иметь практическую значимость.

Несмещенность оценки означает, что математическое ожидание остатков равно нулю. Оценки считаются Эффективными, если они характеризуются наименьшей дисперсией. Поэтому несмещенность оценки должна дополняться минимальной дисперсией. Состоятельность оценок характеризует увеличение их точности с увеличением объема выработки.

Указанные критерии оценок (несмещенность, состоятельность, эффективность) обязательно учитываются при разных способах оценивания. Метод наименьших квадратов строит оценки регрессии на основе минимизации суммы квадратов остатков ($y - \hat{y}_x$).

Исследование остатков ε_i предполагают проверку наличия следующих пяти предпосылок МНК:

- случайный характер остатков; - нулевая средняя величина остатков, не зависящая от x_i; - гомоскедастичность – дисперсия каждого отклонение ε_i одинакова для всех значений x; - отсутствие автокорреляции остатков, т. е. значения остат-

ков ε_i распределены независимо друг от друга; - остатки подчиняются нормальному распределению.

С цель проверки случайного характера остатков ε_i строится график зависимости остатков ε_i от теоретических значений результативного признака $\hat{y}$.

.Если на графике нет направленности в расположении точек ε_i, то остатки ε_i представляют собой случайные величины и МНК оправдан. Также возможны следующие случаи: если ε_i зависит от теоретического значения, то:

Вторая предпосылка МНК относительно нулевой средней величины остатков означает, что $\sum(y - \hat{y}_x) = 0$. Это выполнимо для линейных моделей и моделей, нелинейных относительно включаемых переменных. Для обеспечения несмещенности оценок коэффициентов регрессии, полученных МНК, необходимо выполнение условий независимости случайных остатков ε_i и переменных x, что исследуется в рамках соблюдения второй предпосылки МНК. С целью проверки выполнение этой предпосылки строится график зависимости случайных остатков ε от факторов, включенных в регрессию x_i. Если расположение остатков на графике не имеет направленности, то они независимы от значений x_i . Если же график показывает наличие зависимости ε_i и x_i, то модель неадекватна.

Предпосылка о нормальном распределении остатков позволяет проводить проверку параметров регрессии и корреляции с помощью критериев t и F. Вместе с тем оценки регрессии, найденные с применением МНК, обладают хорошими свойствами даже при отсутствии нормального распределения остатков, т. е. при нарушении пятой предпосылки метода наименьших квадратов.

В соответствии с третьей предпосылкой МНК требуется, чтобы дисперсия остатков была гомоскедастичной. Это означает, что для каждого значения фактора x_i остатки ε_i имеют одинаковую дисперсию. Если это условие применения МНК не соблюдается, то имеет место гетероскедастичность. Используя трехмерной изображение, рассмотрим отличие гомо - и гетероскедастичности.

Наличие гетероскедастичности будет сказываться на уменьшении эффективности оценок b_i, в частности, становится затруднительным использование формулы стандартной ошибки коэффициента регрессии, предполагающей единую дисперсию остатков для любых значений фактора.

2.16 Коэффициент корреляции, его свойства, значимость.

Наличие гетероскедастичности в остатках регрессии можно проверить с помощью ранговой корреляции Спирмэна. Суть проверки заключается в том, что в случае гетероскедастичности абсолютные остатки ε_iкоррелированы со значениями фактора x_i. Эту корреляцию можно измерять с помощью коэффициента ранговой корреляции Спирмэна:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)},$$

где ρ – абсолютная разность между рангами значений x_i и $|\varepsilon_i|$.

Статистическую значимость ρ можно определить с помощью t-критерия:

$$t_\rho = \frac{\rho}{\sqrt{(1-\rho^2)}}\sqrt{(n-2)}.$$

Принято считать, что если $t_{\text{расч}} > t_{\text{табл}}$, то корреляция между ε_i и x_i статистически значима, т. е. имеет место гетероскедастичность остатков. В противном случае принимается гипотеза об отсутствии гетероскедастичности остатков.

При построении регрессионных моделей чрезвычайно важно соблюдение четвертой предпосылки МНК – отсутствие автокорреляции остатков, т. е. распределения остатков ε_i и ε_{i-1} независимы. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений. Находится коэффициент корреляции между ε_i и ε_{i-1}, и если он окажется существенно отличным от нуля, то остатки автокоррелированы и функция плотности вероятности F(ε) зависит от j-ой точки наблюдения и от распределения значений остатков в других точках наблюдения.

Отсутствие автокорреляции остатков обеспечивает состоятельность и эффективность оценок коэффициентов регрессии.

До сих пор в качестве факторов рассматривались экономические переменные, принимающие количественные значения в некотором интервале. Вместе с тем может оказаться необходимым включить в модель фактор, имеющий два или более качественных уровней. Это могут быть разного рода атрибутивные признаки, такие, например, как профессия, пол, образование, климатические условия, принадлежность к определенному региону. Для того, чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные цифровые метки, т. е. качественные переменные необходимо преобразовать в количественные. Такого вида сконструированные переменные в эконометрике принято называть фиктивными переменными.

Качественные признаки могут приводить к неоднородности исследуемой совокупности, что может быть учтено при моделировании двумя путями:

- регрессия строится для каждой качественно отличной группы единиц совокупности, т. е. для каждой группы в отдельности, чтобы преодолеть неоднородность единиц общей совокупности; - общая регрессионная модель строится для совокупности в целом, учитывающей неоднородность данных. В этом случае в регрессионную модель вводятся фиктивные переменные, т. е. строится регрессионная модель с переменной структурой, отражающей неоднородность данных.

Качественный фактор может иметь только два состояния, которым будут соответствовать 1 и 0. Если же число градаций качественного признака-фактора превышает два, то в модель вводится несколько фиктивных переменных, число которых должно быть меньше числа качественных градаций. Только при соблюдении этого положения матрица исходных фиктивных переменных не будет линейно зависима и возможна оценка параметров модели.

Коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение зависимой переменной при переходе от одной категории к другой при неизменных значениях остальных параметров. На основе t-критерия Стьюдента делается вывод о значимости влияния фиктивной переменной, существенности расхождения между категориями.

Такая проверка производится с помощью статистических критериев и на их основе делается вывод о статистической надежности построенного уравнения регрессии, о пригодности модели для анализа и прогнозирования исследуемого показателя.

Для проверки надежности модели в целом используется отношение факторной дисперсии к остаточной $\frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}$ Известно, что отношение этих дисперсий подчиняется распределению Фишера (F-распределение). Расчетное значение F-отношения сравнивается с табличным значением, которое определяется для конкретного уровня значимости α. В экономических исследованиях α принимается равным 0,05 (реже 0,01), число степеней свободы $k_1 = \mathrm{p}, k_2 = n - p - 1$. Если $F_{\text{расч}} > F_{\text{табл}}$, то построенная модель считается статистически надежной, а следовательно, отражает закон изменения исследуемого показателя под действием факторов.

Для проверки полноту модели используется $R^2 \cdot 100\%$. Этот показатель показывает, на сколько процентов изменится вариация результативного показателя под влиянием факторов, включенных в модель.

Проверку надежности параметров уравнения регрессии проводят с использовани-

ем Т - критерия. Расчетное значение вычисляется по формуле $t = \frac{|a_j|}{\sigma_{a_j}}$

, $\sigma_{a_j} = \frac{S_{ост}}{\sigma_{x_j}\sqrt{n}\sqrt{1-R^2_{j,1,2,\dots,j-1,j+1,p}}}$. Фактическое значение Т- критерия сравнивается с таб-

личным и если $t_{\text{факт}} > t_{\text{табл}}\left(t_{\alpha,k}, \alpha = 0{,}05(0{,}01), k = n-p-1\right)$, то тогда соответствующий коэффициент регрессии значим, т. е. отличен от нуля, а влияние J-го фактора следует считать сильным. Факторы, оказывающие несущественное влияние на исследуемый показатель, из модели исключают.

На этом этапе разрабатываются рекомендации об использовании результатов моделирования. Анализируется уравнение регрессии в натуральном масштабе: коэффициент регрессии a_j показывает, на сколько своих единиц измерения в среднем изменится исследуемый показатель, при увеличении J-Го фактора на единицу своего измерения, при условии, что все остальные факторы находятся на постоянном уровне. Свободный член уравнения характеризует изменение результативного показателя за счет изменения факторов, неучтенных в модели.

В связи с тем, что факторы имеют различный физический смысл и различные единицы измерения, коэффициенты регрессии нельзя сравнивать между собой и, следовательно, трудно определить, какой из факторов оказывает наибольшее влияние. Для устранения различий в единицах измерения применяют Частные коэффициенты эластично-

сти $Э_j = a_j \frac{\bar{x}_j}{\bar{x}_0}$ характеризующие, на сколько % в среднем изменится x_0 при увеличении J-го Фактора на 1% при фиксированном положении других факторов.

При определении степени влияния отдельных факторов необходим показатель, который бы учитывал влияние анализируемых факторов с учетом различий в уровне их вариации. Таким показателем является Коэффициент регрессии в стандартизированном

масштабе $\beta_j = a_j \frac{\sigma_{x_j}}{\sigma_{x_0}}$. Коэффициент β_j показывает, на какую часть своего среднеквадратического отклонения изменится x_0 при изменении J-го фактора на одно свое среднеквадратическое отклонение при фиксированном значении остальных факторов. Уравнение

регрессии в стандартизированном масштабе: $t_{0,1,2,\dots,p} = \sum_{j=1}^{p} \beta_j t_j$, где $t_j = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}$.

Т. к. в стандартизованном уравнении все факторы и функция измеряются в одних и тех же единицах измерения – стандартных отклонениях, то по стандартизованным коэффициентам можно судить о влиянии каждого фактора по сравнению с другими.

3. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ПОДГОТОВКЕ К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ

3.1 Практическое занятие № 1 (ПЗ-1). Классическое определение вероятности события. Относительная частота наступления события и статистическая вероятность. Формулы умножения и сложения вероятностей случайных событий. Повторение испытаний: формулы Бернулли, локальные и интегральные теоремы Лапласа, формула Пуассона, простейший поток событий

При подготовке к занятию необходимо обратить внимание на следующие моменты:

- классификацию случайных событий; различные подходы к определению вероятности случайного события;

- комбинаторные формулы, методы непосредственного вычисления вероятности случайного события, основные теоремы; понятие условной вероятности, схемы повторных испытаний; алгоритм применения формул Бернулли, Лапласа, Пуассона.

- классификацию СВ, определение закона распределения вероятностей;

- построение функции распределения ДСВ, вычисление плотности распределения НСВ, вероятности попадания СВ в заданный интервал;

- вычисление числовых характеристик СВ.

3.2 Практическое занятие № 2 (ПЗ-2). Понятие случайной величины примеры. Виды случайных величин. Закон распределения вероятностей. Функция распределения случайных величин. Свойства. Плотность распределения вероятностей. Числовые характеристики: математическое ожидание, свойства; дисперсия, свойства; среднее квадратичное отклонение и его свойства.

При подготовке к занятию необходимо обратить внимание на следующие моменты:

- закон распределения случайной величины;

- ряд распределения;

- функция распределения;

- плотность распределения;

- числовые характеристики ДСВ;

- числовые характеристики НСВ.

3.3 Практическое занятие № 3 (ПЗ-3). Статистический материал. Статистические параметры распределения. Статистические оценки параметров распределения. Понятие статистической гипотезы. Виды гипотез. Статистический критерий. Критическая область. Мощность критерия. Критерии согласия: критерий Пирсона. Выравнивание рядов.

При подготовке к занятию необходимо обратить внимание на следующие моменты:

- определение основных понятий статистики;
- первичную обработку статистических данных;
- точечные и интервальные оценки параметров распределения;
- определение статистического критерия, ошибок первого и второго рода;
- классификацию статистических критериев, их мощность;
- применение критериев согласия.