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

Кафедра «Техносферная и информационная безопасность**»**

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Б2.В.ДВ.3.1. Математическое моделирование в экологии
(*код и наименование дисциплины в соответствии с РУП*)

Направление подготовки (специальность) *Экология и природопользование*

Профиль образовательной программы *Экология*

Форма обучения *очная*

Оренбург 201_ г.

СОДЕРЖАНИЕ

1. КОНСПЕКТ ЛЕКЦИЙ

1. 1 Лекция №1 (2 часа).

Тема: «Основные статистические понятия»

1.1.1 Вопросы лекции:

1. Математическая статистика как наука.
2. Основные статистические понятия.
3. Этапы статистического исследования.

1.1.2 Краткое содержание вопросов:

Математическая статистика занимается установлением закономерностей, которым подчинены массовые случайные явления, на основе обработки статистических данных, полученных в результате наблюдений. Двумя основными задачами математической статистики являются:
- определение способов сбора и группировки этих статистических данных;
- разработка методов анализа полученных данных в зависимости от целей исследования, к которым относятся:
а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости от других случайных величин и т.д.;
б) проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения.
Для решения этих задач необходимо выбрать из большой совокупности однородных объектов ограниченное количество объектов, по результатам изучения которых можно сделать прогноз относительно исследуемого признака этих объектов.
Определим основные понятия математической статистики.
Генеральная совокупность – все множество имеющихся объектов.
Выборка – набор объектов, случайно отобранных из генеральной совокупности.
Объем генеральной совокупности *N* и объем выборки *n* – число объектов в рассматривае-мой совокупности.
Виды выборки:
Повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность;
Бесповторная – отобранный объект в генеральную совокупность не возвращается.

Замечание. Для того, чтобы по исследованию выборки можно было сделать выводы о поведе-нии интересующего нас признака генеральной совокупности, нужно, чтобы выборка правиль-но представляла пропорции генеральной совокупности, то есть была **репрезентативной** (представительной). Учитывая закон больших чисел, можно утверждать, что это условие выполняется, если каждый объект выбран случайно, причем для любого объекта вероятность попасть в выборку одинакова.

Первичная обработка результатов.
Пусть интересующая нас случайная величина X принимает в выборке значение x_1 n_1 раз, x_2 – n_2 раз, …, $x_к$ – $n_к$ раз, причем $\sum_{i=1}^{k} n_k = n,$ где n – объем выборки. Тогда наблюдаемые значения случайной величины x_1, x_2,…, $x_к$ называют **вариантами**, а n_1, n_2,…, $n_к$ – **частотами**. Если разделить каждую частоту на объем выборки, то получим

относительные частоты $w_i = \frac{n_i}{n}$. Последовательность вариант, записанных в порядке возрастания, называют **вариационным** рядом, а перечень вариант и соответствующих им частот или относительных частот – **стати-стическим рядом**:

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
w_i	w_1	w_2	…	w_k

Пример.
При проведении 20 серий из 10 бросков игральной кости число выпадений шести очков оказалось равным 1,1,4,0,1,2,1,2,2,0,5,3,3,1,0,2,2,3,4,1.Составим вариационный ряд: 0,1,2,3,4,5. Статистический ряд для абсолютных и относительных частот имеет вид:

x_i	0	1	2	3	4	5
n_i	3	6	5	3	2	1
w_i	0,15	0,3	0,25	0,15	0,1	0,05

Если исследуется некоторый непрерывный признак, то вариационный ряд может состоять из очень большого количества чисел. В этом случае удобнее использовать **группированную выборку**. Для ее получения интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько равных частичных интервалов длиной h, а затем находят для каждого частичного интервала n_i – сумму частот вариант, попавших в i-й интервал. Составленная по этим результатам таблица называется **группированным статистическим рядом**:

Номера интервалов	1	2	…	k
Границы интервалов	$(a, a + h)$	$(a + h, a + 2h)$	…	$(b – h, b)$
Сумма частот вариант, попав-ших в интервал	n_1	n_2	…	n_k

Полигон частот. Выборочная функция распределения и гистограмма.

Для наглядного представления о поведении исследуемой случайной величины в выборке можно строить различные графики. Один из них – **полигон частот**: ломаная, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2),…, (x_k, n_k), где x_i откладываются на оси абсцисс, а n_i – на оси ординат. Если на оси ординат откладывать не абсолютные (n_i), а относительные (w_i) частоты, то получим **полигон относительных частот** (рис.1)**.**

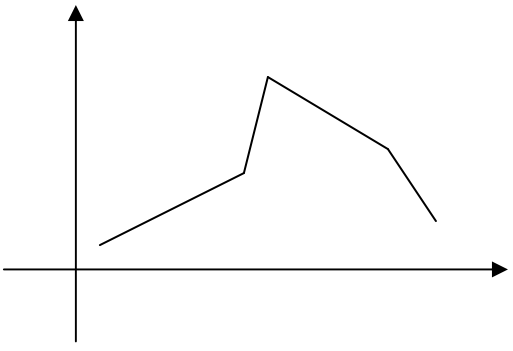


Рис. 1.

По аналогии с функцией распределения случайной величины можно задать некоторую функцию, относительную частоту события $X < x$.

Выборочной (эмпирической) функцией распределения называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$. Таким образом,

$$F^*(x) = \frac{n_x}{n},$$

где n_x – число вариант, меньших x, n – объем выборки.

Замечание. В отличие от эмпирической функции распределения, найденной опытным путем, функцию распределения $F(x)$ генеральной совокупности называют *теоретической функцией распределения*. $F(x)$ определяет вероятность события $X < x$, а $F^*(x)$ – его относительную частоту. При достаточно больших n, как следует из теоремы Бернулли, $F^*(x)$ стремится по вероятности к $F(x)$.

Из определения эмпирической функции распределения видно, что ее свойства совпадают со свойствами $F(x)$, а именно:

1) $0 \le F^*(x) \le 1$.
2) $F^*(x)$ – неубывающая функция.
3) Если x_1 – наименьшая варианта, то $F^*(x) = 0$ при $x \le x_1$; если $x_к$ – наибольшая варианта, то $F^*(x) = 1$ при $x > x_к$.

Для непрерывного признака графической иллюстрацией служит **гистограмма**, то есть ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высотами – отрезки длиной n_i / h (гистограмма частот) или w_i / h (гистограмма относительных частот). В первом случае площадь гистограммы равна объему выборки, во втором – единице (рис.2).

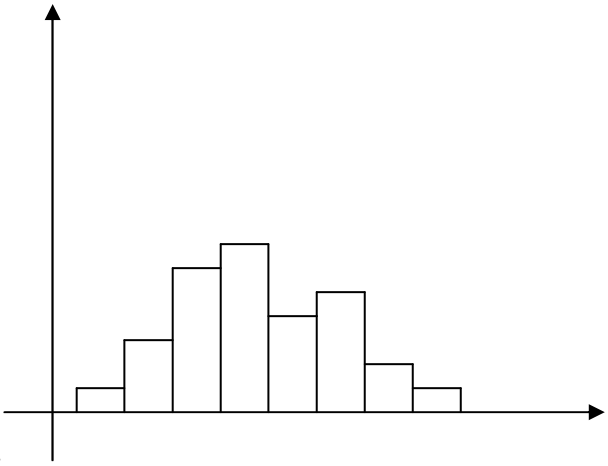


Рис.2.

1. 2 Лекция №2 (1 час).

Тема: «Выборочные характеристики»

1.2.1 Вопросы лекции:

1. Основные выборочные характеристики варьирующих объектов.
2. Средняя арифметическая.
3. Мода.
4. Медиана.

1.2.2 Краткое содержание вопросов:

Одна из задач математической статистики: по имеющейся выборке оценить значения числовых характеристик исследуемой случайной величины.

Определение 16.1. **Выборочным средним** называется среднее арифметическое значений случайной величины, принимаемых в выборке:

$$\bar{x}_B = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n},$$

где x_i – варианты, n_i - частоты.

Замечание. Выборочное среднее служит для оценки математического ожидания исследуемой случайной величины. В дальнейшем будет рассмотрен вопрос, насколько точной является такая оценка.

Определение 16.2. **Выборочной дисперсией** называется

$$D_B = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_B)^2}{n} = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n},$$

а **выборочным средним квадратическим отклонением** –

$$\sigma_B = \sqrt{D_B}.$$

Так же, как в теории случайных величин, можно доказать, что справедлива следующая формула для вычисления выборочной дисперсии:

$$D = \overline{x^2} - (\bar{x})^2.$$

Пример 1. Найдем числовые характеристики выборки, заданной статистическим рядом

x_i	2	5	7	8
n_i	3	8	7	2

$$\bar{x}_B = \frac{2 \cdot 3 + 5 \cdot 8 + 7 \cdot 7 + 8 \cdot 2}{20} = 5{,}55;$$

$$D_B = \frac{4 \cdot 3 + 25 \cdot 8 + 49 \cdot 7 + 64 \cdot 2}{20} - 5{,}55^2 = 3{,}3475; \quad \sigma_B = \sqrt{3{,}3475} = 1{,}83.$$

Другими характеристиками вариационного ряда являются:

- **мода M_0** – варианта, имеющая наибольшую частоту (в предыдущем примере $M_0 = 5$).
- **медиана m_e** - варианта, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно ($n = 2k + 1$), то $m_e = x_{k+1}$, а при четном $n = 2k$ $m_e = \frac{x_k + x_{k+1}}{2}$. В частности, в примере 1 $m_e = \frac{5+7}{2} = 6.$

Оценки начальных и центральных моментов (так называемые эмпирические моменты) определяются аналогично соответствующим теоретическим моментам:

- начальным эмпирическим моментом порядка *k* называется

$$M_k = \frac{\sum n_i x_i^k}{n}.$$

В частности, $M_1 = \frac{\sum n_i x_i}{n} = \bar{x}_B$, то есть начальный эмпирический момент первого порядка равен выборочному среднему.

- **центральным эмпирическим моментом порядка *k*** называется

$$m_k = \frac{\sum n_i (x_i - \bar{x}_B)^k}{n}.$$

В частности, $m_2 = \frac{\sum n_i (x_i - \bar{x}_B)^2}{n} = D_B$, то есть центральный эмпирический момент второго порядка равен выборочной дисперсии.

Статистическое описание и вычисление характеристик двумерного случайного вектора.

При статистическом исследовании двумерных случайных величин основной задачей является обычно выявление связи между составляющими.

Двумерная выборка представляет собой набор значений случайного вектора: (x_1, y_1), (x_2, y_2), …, (x_n, y_n). Для нее можно определить выборочные средние составляющих: $\bar{x}_B = \frac{\sum x_i}{n}$, $\bar{y}_B = \frac{\sum y_i}{n}$ и соответствующие выборочные дисперсии и средние квадратические отклонения. Кроме того, можно вычислить **условные средние**: $\bar{y}_x$ - среднее арифметическое наблюдав-шихся значений Y, соответствующих $X = x$, и $\bar{x}_y$ - среднее значение наблюдавшихся значений X, соответствующих $Y = y$.

Если существует зависимость между составляющими двумерной случайной величины, она может иметь разный вид: функциональная зависимость, если каждому возможному значению X соответствует одно значение Y, и статистическая, при которой изменение одной величины приводит к изменению распределения другой. Если при этом в результате изменения одной величины меняется среднее значение другой, то статистическую зависимость между ними называют корреляционной.

1. 3 Лекция №3 (1 час).

Тема: «Показатели изменчивости»

1.3.1 Вопросы лекции:

1. Понятие показателей изменчивости.
2. Размах.
3. Дисперсия.
4. Среднее квадратическое отклонение.
5. Коэффициент вариации.
6. Нормированное отклонение.
7. Показатели ассиметрии и эксцесса.

1.3.2 Краткое содержание вопросов:

Получив статистические оценки параметров распределения (выборочное среднее, выбороч-ную дисперсию и т.д.), нужно убедиться, что они в достаточной степени служат приближе-нием соответствующих характеристик генеральной совокупности. Определим требования, которые должны при этом выполняться.

Пусть Θ* - статистическая оценка неизвестного параметра Θ теоретического распределения. Извлечем из генеральной совокупности несколько выборок одного и того же объема n и вычислим для каждой из них оценку параметра Θ: $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Тогда оценку Θ* можно рассматривать как случайную величину, принимающую возможные значения $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Если математическое ожидание Θ* не равно оцениваемому параметру, мы будем получать при вычислении оценок систематические ошибки одного знака (с избытком, если $M(\Theta^*) > \Theta$, и с недостатком, если $M(\Theta^*) < \Theta$). Следовательно, необходимым условием отсутствия систе-матических ошибок является требование $M(\Theta^*) = \Theta$.

Статистическая оценка Θ* называется **несмещенной**, если ее математичес-кое ожидание равно оцениваемому параметру Θ при любом объеме выборки:

$$M(\Theta^*) = \Theta.$$

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Однако несмещенность не является достаточным условием хорошего приближения к истин-ному значению оцениваемого параметра. Если при этом возможные значения Θ* могут значительно отклоняться от среднего значения, то есть дисперсия Θ* велика, то значение, найденное по данным одной выборки, может значительно отличаться от оцениваемого параметра. Следовательно, требуется наложить ограничения на дисперсию.

Статистическая оценка называется **эффективной**, если она при заданном объеме выборки n имеет наименьшую возможную дисперсию.

При рассмотрении выборок большого объема к статистическим оценкам предъявляется еще и требование состоятельности.

Состоятельной называется статистическая оценка, которая при $n \to \infty$ стре-мится по вероятности к оцениваемому параметру (если эта оценка несмещенная, то она будет состоятельной, если при $n \to \infty$ ее дисперсия стремится к 0).

Убедимся, что $\bar{x}_B$ представляет собой несмещенную оценку математического ожидания $M(X)$.

Будем рассматривать $\bar{x}_B$ как случайную величину, а x_1, x_2,…, x_n, то есть значения исследуемой случайной величины, составляющие выборку, – как независимые, одинаково распределенные случайные величины X_1, X_2,…, X_n, имеющие математическое ожидание a. Из свойств математического ожидания следует, что

$$M(\bar{X}_B) = M\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = a.$$

Но, поскольку каждая из величин X_1, X_2,…, X_n имеет такое же распределение, что и генеральная совокупность, $a = M(X)$, то есть $M(\bar{X}_B) = M(X)$, что и требовалось доказать.

Выборочное среднее является не только несмещенной, но и состоятельной оценкой математического ожидания. Если предположить, что X_1, X_2,…, X_n имеют ограниченные дисперсии, то из теоремы Чебышева следует, что их среднее арифметическое, то есть $\bar{X}_B$, при увеличении n стремится по вероятности к математическому ожиданию a каждой их величин, то есть к $M(X)$. Следовательно, выборочное среднее есть состоятельная оценка математического ожидания.

В отличие от выборочного среднего, выборочная дисперсия является смещенной оценкой дисперсии генеральной совокупности. Можно доказать, что

$$M(D_B) = \frac{n-1}{n} D_\Gamma,$$

где D_Γ – истинное значение дисперсии генеральной совокупности. Можно предложить другую оценку дисперсии – **исправленную дисперсию s^2**, вычисляемую по формуле

$$s^2 = \frac{n}{n-1} D_B = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n-1}.$$

Такая оценка будет являться несмещенной. Ей соответствует **исправленное среднее квадратическое отклонение**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n-1}}.$$

Оценка некоторого признака называется **асимптотически несмещенной**, если для выборки $x_1, x_2, \ldots, x_n$

$$\lim_{n\to\infty} \frac{x_1 + x_2 + ... + x_n}{n} = X,$$

где X – истинное значение исследуемой величины.

Способы построения оценок.

1. Метод наибольшего правдоподобия.

Пусть X – дискретная случайная величина, которая в результате n испытаний приняла значения x_1, x_2, …, x_n. Предположим, что нам известен закон распределения этой величины, определяемый параметром Θ, но неизвестно численное значение этого параметра. Найдем его точечную оценку.

Пусть $p(x_i, \Theta)$ – вероятность того, что в результате испытания величина X примет значение x_i. Назовем **функцией правдоподобия** дискретной случайной величины X функцию аргумента Θ, определяемую по формуле:

$L(x_1, x_2, \ldots, x_n; \Theta) = p(x_1,\Theta)p(x_2,\Theta)\ldots p(x_n,\Theta)$.

Тогда в качестве точечной оценки параметра Θ принимают такое его значение Θ* = Θ(x_1, x_2, …, x_n), при котором функция правдоподобия достигает максимума. Оценку Θ* называют **оценкой наибольшего правдоподобия**.
Поскольку функции L и ln$L$ достигают максимума при одном и том же значении Θ, удобнее искать максимум ln L – **логарифмической функции правдоподобия**. Для этого нужно:

1) найти производную $\frac{d \ln L}{d\Theta}$;

2) приравнять ее нулю (получим так называемое *уравнение правдоподобия*) и найти критическую точку;

3) найти вторую производную $\frac{d^2 \ln L}{d\Theta^2}$; если она отрицательна в критической точке, то это – точка максимума.

Достоинства метода наибольшего правдоподобия: полученные оценки состоятельны (хотя могут быть смещенными), распределены асимптотически нормально при больших значениях n и имеют наименьшую дисперсию по сравнению с другими асимптотически нормальными оценками; если для оцениваемого параметра Θ существует эффективная оценка Θ*, то уравнение правдоподобия имеет единственное решение Θ*; метод наиболее полно использует данные выборки и поэтому особенно полезен в случае малых выборок.
Недостаток метода наибольшего правдоподобия: сложность вычислений.

Для непрерывной случайной величины с известным видом плотности распределения $f(x)$ и неизвестным параметром Θ функция правдоподобия имеет вид:

$$L(x_1, x_2, \ldots, x_n; \Theta) = f(x_1,\Theta) f(x_2,\Theta) \ldots f(x_n,\Theta).$$

Оценка наибольшего правдоподобия неизвестного параметра проводится так же, как для дискретной случайной величины.

2. Метод моментов.

Метод моментов основан на том, что начальные и центральные эмпирические моменты являются состоятельными оценками соответственно начальных и центральных теоретических моментов, поэтому можно приравнять теоретические моменты соответствующим эмпирическим моментам того же порядка.
Если задан вид плотности распределения $f(x, \Theta)$, определяемой одним неизвестным параметром Θ, то для оценки этого параметра достаточно иметь одно уравнение. Например, можно приравнять начальные моменты первого порядка:

$$\bar{x}_B = M(X) = \int_{-\infty}^{\infty} x f(x;\Theta) dx = \varphi(\Theta),$$

получив тем самым уравнение для определения Θ. Его решение Θ* будет точечной оценкой параметра, которая является функцией от выборочного среднего и, следовательно, и от вариант выборки:

$$\Theta = \psi(x_1, x_2, \ldots, x_n).$$

Если известный вид плотности распределения $f(x, \Theta_1, \Theta_2)$ определяется двумя неизвестными параметрами Θ_1 и Θ_2, то требуется составить два уравнения, например

$$\nu_1 = M_1, \ \mu_2 = m_2.$$

Отсюда $\begin{cases} M(X) = \bar{x}_B \\ D(X) = D_B \end{cases}$ - система двух уравнений с двумя неизвестными Θ_1 и Θ_2. Ее решениями будут точечные оценки Θ_1* и Θ_2* - функции вариант выборки:

$$\Theta_1 = \psi_1(x_1, x_2, \ldots, x_n),$$
$$\Theta_2 = \psi_2(x_1, x_2, \ldots, x_n).$$

3. Метод наименьших квадратов.

Если требуется оценить зависимость величин y и x, причем известен вид связывающей их функции, но неизвестны значения входящих в нее коэффициентов, их величины можно оценить по имеющейся выборке с помощью метода наименьших квадратов. Для этого функция $y = \varphi(x)$ выбирается так, чтобы сумма квадратов отклонений наблюдаемых значений $y_1, y_2, \ldots, y_n$ от $\varphi(x_i)$ была минимальной:

$$\sum_{i=1}^{n}(y_i - \varphi(x_i))^2 = \min.$$

При этом требуется найти стационарную точку функции $\varphi(x; a, b, c\ldots)$, то есть решить систему:

$$\begin{cases} \sum_{i=1}^{n}(y_i - \varphi(x_i;a,b,c...))\left(\frac{\partial\varphi}{\partial a}\right)_i = 0 \\ \sum_{i=1}^{n}(y_i - \varphi(x_i;a,b,c...))\left(\frac{\partial\varphi}{\partial b}\right)_i = 0 \\ \sum_{i=1}^{n}(y_i - \varphi(x_i;a,b,c...))\left(\frac{\partial\varphi}{\partial c}\right)_i = 0 \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

(решение, конечно, возможно только в случае, когда известен конкретный вид функции φ).

Рассмотрим в качестве примера подбор параметров линейной функции методом наименьших квадратов.

Для того, чтобы оценить параметры a и b в функции $y = ax + b$, найдем $\left(\frac{\partial\varphi}{\partial a}\right)_i = x_i; \left(\frac{\partial\varphi}{\partial b}\right)_i = 1.$ Тогда $\begin{cases} \sum_{i=1}^{n}(y_i - (ax_i + b))x_i = 0 \\ \sum_{i=1}^{n}(y_i - (ax_i + b)) = 0 \end{cases}$. Отсюда

$\begin{cases} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - bn = 0 \end{cases}$. Разделив оба полученных уравнения на n и вспомнив определения эмпирических моментов, можно получить выражения для a и b в виде:

$a = \frac{(K_{xy})_B}{(D_x)_B}$, $b = \bar{y}_B - \frac{(K_{xy})_B}{(D_x)_B}\bar{x}_B$. Следовательно, связь между x и y можно задать в виде:

$$y - \bar{y}_B = \frac{(K_{xy})_B}{(D_x)_B}(x - \bar{x}_B).$$

4. Байесовский подход к получению оценок.

Пусть (Y, X) – случайный вектор, для которого известна плотность $p(y|x)$ условного распреде-ления Y при каждом значении $X = x$. Если в результате эксперимента получены лишь значения Y, а соответствующие значения X неизвестны, то для оценки некоторой заданной функции $\varphi(x)$ в качестве ее приближенного значения предлагается искать условное математическое ожидание $M(\varphi(x)|Y)$, вычисляемое по формуле:

$\psi(Y) = \dfrac{\int \varphi(x)\, p(Y \mid x)\, p(x) d\mu(x)}{q(Y)}$, где $q(y) = \int p(y \mid x)\, p(x) d\mu(x)$, $p(x)$ – плотность безусловного распределения X, $q(y)$ – плотность безусловного распределения Y. Задача может быть решена только тогда, когда известна $p(x)$. Иногда, однако, удается построить состоятельную оценку для $q(y)$, зависящую только от полученных в выборке значений Y.

1. 4 Лекция №4 (2 часа).

Тема: «Точечные оценки генеральных параметров»

1.4.1 Вопросы лекции:

1. Статистические ошибки.
2. Точечные оценки.
3. Интервальные оценки.

1.4.2 Краткое содержание вопросов:

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, что приводит к грубым ошибкам. Поэтому в таком случае лучше пользоваться *интервальными оценками*, то есть указывать интервал, в который с заданной вероятностью попадает истинное значение оцениваемого параметра. Разумеется, чем меньше длина этого интервала, тем точнее оценка параметра. Поэтому, если для оценки Θ* некоторого параметра Θ справедливо неравенство | Θ* - Θ | < δ, число δ > 0 характеризует **точность оценки** (чем меньше δ, тем точнее оценка). Но статистические методы позволяют говорить только о том, что это неравенство выполняется с некоторой вероятностью.

Надежностью (доверительной вероятностью) оценки Θ* параметра Θ называется вероятность γ того, что выполняется неравенство | Θ* - Θ | < δ. Если заменить это неравенство двойным неравенством – δ < Θ* - Θ < δ, то получим:

$$p\,(\Theta^* - \delta < \Theta < \Theta^* + \delta) = \gamma.$$

Таким образом, γ есть вероятность того, что Θ попадает в интервал (Θ* - δ, Θ* + δ).

Доверительным называется интервал, в который попадает неизвестный параметр с заданной надежностью γ.

Построение доверительных интервалов.

1. Доверительный интервал для оценки математического ожидания нормального распределения при известной дисперсии.

Пусть исследуемая случайная величина X распределена по нормальному закону с известным средним квадратическим σ, и требуется по значению выборочного среднего $\overline{x}_B$ оценить ее математическое ожидание a. Будем рассматривать выборочное среднее $\overline{x}_B$ как случайную величину $\overline{X}$, а значения вариант выборки x_1, x_2,…, x_n как одинаково распределенные независимые случайные величины X_1, X_2,…, X_n, каждая из которых имеет математическое ожидание a и среднее квадратическое отклонение σ. При этом $M(\overline{X}) = a$,

$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ (используем свойства математического ожидания и дисперсии суммы независимых случайных величин). Оценим вероятность выполнения неравенства $|\overline{X} - a| < \delta$. Применим формулу для вероятности попадания нормально распределенной случайной величины в заданный интервал:

$p\ (|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right)$. Тогда , с учетом того, что $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$, $p\ (|\overline{X} - a| < \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) =$

$=2\Phi(t)$, где $t = \frac{\delta\sqrt{n}}{\sigma}$. Отсюда $\delta = \frac{t\sigma}{\sqrt{n}}$, и предыдущее равенство можно переписать так:

$$p\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}} < a < \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma.$$

Итак, значение математического ожидания a с вероятностью (надежностью) γ попадает в интервал $\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}}; \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right)$, где значение t определяется из таблиц для функции Лапласа так, чтобы выполнялось равенство $2\Phi(t) = \gamma$.

Пример. Найдем доверительный интервал для математического ожидания нормально распреде-ленной случайной величины, если объем выборки $n = 49$, $\bar{x}_B = 2{,}8, \sigma = 1{,}4$, а доверительная вероятность $\gamma = 0{,}9$.

Определим t, при котором $\Phi(t) = 0{,}9:2 = 0{,}45$: $t = 1{,}645$. Тогда

$2{,}8 - \frac{1{,}645 \cdot 1{,}4}{\sqrt{49}} < a < 2{,}8 + \frac{1{,}645 \cdot 1{,}4}{\sqrt{14}}$, или $2{,}471 < a < 3{,}129$. Найден доверительный интервал, в который попадает a с надежностью 0,9.

2. Доверительный интервал для оценки математического ожидания нормального распределения при неизвестной дисперсии.

Если известно, что исследуемая случайная величина X распределена по нормальному закону с неизвестным средним квадратическим отклонением, то для поиска доверительного интервала для ее математического ожидания построим новую случайную величину

$$T = \frac{\bar{x}_B - a}{\frac{s}{\sqrt{n}}},$$

где $\bar{x}_B$ - выборочное среднее, s – исправленная дисперсия, n – объем выборки. Эта случайная величина, возможные значения которой будем обозначать t, имеет распределение Стьюдента (см. лекцию 12) с $k = n - 1$ степенями свободы.

Поскольку плотность распределения Стьюдента $s(t,n) = B_n\left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}$, где

$B_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)}$, явным образом не зависит от a и σ, можно задать вероятность ее

попадания в некоторый интервал (- t_γ , t_γ), учитывая четность плотности распределения, следующим образом: $p\left(\left|\frac{\bar{x}_B - a}{\frac{s}{\sqrt{n}}}\right| < t_\gamma\right) = 2\int_0^{t_\gamma} s(t,n)dt = \gamma$. Отсюда получаем:

$$p\left(\bar{x}_B - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma s}{\sqrt{n}}\right) = \gamma. \qquad)$$

Таким образом, получен доверительный интервал для *a*, где t_γ можно найти по соответствую-щей таблице при заданных *n* и γ.

Пример. Пусть объем выборки *n* = 25, $\bar{x}_B$ = 3, *s* = 1,5. Найдем доверительный интервал для *a* при γ = 0,99. Из таблицы находим, что t_γ (*n* = 25, γ = 0,99) = 2,797. Тогда $3 - \frac{2{,}797 \cdot 1{,}5}{\sqrt{25}} < a < 3 + \frac{2{,}797 \cdot 1{,}5}{\sqrt{25}}$, или 2,161< *a* < 3,839 – доверительный интервал, в который попадает *a* с вероятностью 0,99.

3. Доверительные интервалы для оценки среднего квадратического отклонения нормального распределения.

Будем искать для среднего квадратического отклонения нормально распределенной случайной величины доверительный интервал вида (*s* – *δ, s* +*δ*), где *s* – исправленное выборочное среднее квадратическое отклонение, а для δ выполняется условие: *p* (|σ – *s*| < δ) = γ.

Запишем это неравенство в виде: $s\left(1 - \frac{\delta}{s}\right) < \sigma < s\left(1 + \frac{\delta}{s}\right)$ или, обозначив $q = \frac{\delta}{s}$,

$$s(1-q) < \sigma < s(1+q).$$

Рассмотрим случайную величину χ, определяемую по формуле

$$\chi = \frac{s}{\sigma}\sqrt{n-1},$$

которая распределена по закону «хи-квадрат» с *n*-1 степенями свободы (см. лекцию 12). Плотность ее распределения

$$R(\chi, n) = \frac{\chi^{n-2} e^{-\frac{\chi^2}{2}}}{2^{\frac{n-3}{2}} \Gamma\left(\frac{n-1}{2}\right)}$$

не зависит от оцениваемого параметра σ, а зависит только от объема выборки *n*. Преобразуем неравенство (18.4) так, чтобы оно приняло вид $\chi_1 < \chi < \chi_2$. Вероятность выполнения этого неравенства равна доверительной вероятности γ, следовательно, $\int_{\chi_1}^{\chi_2} R(\chi, n)d\chi = \gamma.$ Предполо-жим, что $q < 1$, тогда неравенство (18.4) можно записать так:

$$\frac{1}{s(1+q)} < \frac{1}{\sigma} < \frac{1}{s(1-q)},$$

или, после умножения на $s\sqrt{n-1}$, $\frac{\sqrt{n-1}}{1+q} < \frac{s\sqrt{n-1}}{\sigma} < \frac{\sqrt{n-1}}{1-q}$. Следовательно, $\frac{\sqrt{n-1}}{1+q} < \chi < \frac{\sqrt{n-1}}{1-q}$. Тогда $\int_{\frac{\sqrt{n-1}}{1+q}}^{\frac{\sqrt{n-1}}{1-q}} R(\chi, n)d\chi = \gamma.$ Существуют таблицы для распределения «хи-

квадрат», из которых можно найти q по заданным n и γ, не решая этого уравнения. Таким образом, вычислив по выборке значение s и определив по таблице значение q, можно найти доверительный интервал (18.4), в который значение σ попадает с заданной вероятностью γ.

Замечание. Если $q > 1$, то с учетом условия $\sigma > 0$ доверительный интервал для σ будет иметь границы

$$0 < \sigma < s(1+q).$$

Пример.
Пусть $n = 20$, $s = 1{,}3$. Найдем доверительный интервал для σ при заданной надежности $\gamma = 0{,}95$. Из соответствующей таблицы находим $q\ (n = 20,\ \gamma = 0{,}95\ ) = 0{,}37$. Следовательно, границы доверительного интервала: 1,3(1-0,37) = 0,819 и 1,3(1+0,37) = 1,781. Итак, $0{,}819 < \sigma < 1{,}781$ с вероятностью 0,95.

1. 5 Лекция №5 (2 часа).

Тема: «Критерии достоверности оценок. Статистические гипотезы и их проверка»

1.5.1 Вопросы лекции:

1. Статистические гипотезы.
2. Параметрические критерии.
3. Непараметрические критерии.

1.5.2 Краткое содержание вопросов:

Статистической гипотезой называют гипотезу о виде неизвестного распределения генеральной совокупности или о параметрах известных распределений.

Нулевой (основной) называют выдвинутую гипотезу H_0. **Конкурирую-щей (альтернативной)** называют гипотезу H_1, которая противоречит нулевой.

Пример. Пусть H_0 заключается в том, что математическое ожидание генеральной совокупности $a = 3$. Тогда возможные варианты H_1: а) $a \neq 3$; б) $a > 3$; в) $a < 3$.

Простой называют гипотезу, содержащую только одно предположение, **сложной** – гипотезу, состоящую из конечного или бесконечного числа простых гипотез.

Пример. Для показательного распределения гипотеза H_0: $\lambda = 2$ – простая, H_0: $\lambda > 2$ – сложная, состоящая из бесконечного числа простых (вида $\lambda = c$, где c – любое число, большее 2).

В результате проверки правильности выдвинутой нулевой гипотезы (такая проверка называется **статистической**, так как производится с применением методов математической статистики) возможны ошибки двух видов: **ошибка первого рода**, состоящая в том, что будет отвергнута правильная нулевая гипотеза, и **ошибка второго рода**, заключающаяся в том, что будет принята неверная гипотеза.

Замечание. Какая из ошибок является на практике более опасной, зависит от конкретной задачи. Например, если проверяется правильность выбора метода лечения больного, то ошибка первого рода означает отказ от правильной методики, что может замедлить лечение, а ошибка второго рода (применение неправильной методики) чревата ухудшением состояния больного и является более опасной.

Вероятность ошибки первого рода называется **уровнем значимости α.**

Основной прием проверки статистических гипотез заключается в том, что по имеющейся выборке вычисляется значение некоторой случайной величины, имеющей известный закон распределения.

Статистическим критерием называется случайная величина K с известным законом распределения, служащая для проверки нулевой гипотезы.

Критической областью называют область значений критерия, при которых нулевую гипотезу отвергают, **областью принятия гипотезы** – область значений критерия, при которых гипотезу принимают.

Итак, процесс проверки гипотезы состоит из следующих этапов:
1) выбирается статистический критерий K;
2) вычисляется его наблюдаемое значение $K_{набл}$ по имеющейся выборке;
3) поскольку закон распределения K известен, определяется (по известному уровню значимости α) **критическое значение** $k_{кр}$, разделяющее критическую область и область принятия гипотезы (например, если $p(K > k_{кр}) = \alpha$, то справа от $k_{кр}$ распо-лагается критическая область, а слева – область принятия гипотезы);
4) если вычисленное значение $K_{набл}$ попадает в область принятия гипотезы, то нулевая гипотеза принимается, если в критическую область – нулевая гипотеза отвергается.

Различают разные виды критических областей:
- **правостороннюю** критическую область, определяемую неравенством $K > k_{кр}$ ($k_{кр} > 0$);
- **левостороннюю** критическую область, определяемую неравенством $K < k_{кр}$ ($k_{кр} < 0$);
- **двустороннюю** критическую область, определяемую неравенствами $K < k_1$, $K > k_2$ ($k_2 > k_1$).

Мощностью критерия называют вероятность попадания критерия в критическую область при условии, что верна конкурирующая гипотеза.
Если обозначить вероятность ошибки второго рода (принятия неправильной нулевой гипотезы) β, то мощность критерия равна 1 – β. Следовательно, чем больше мощность критерия, тем меньше вероятность совершить ошибку второго рода. Поэтому после выбора уровня значимости следует строить критическую область так, чтобы мощность критерия была максимальной.

Критерий для проверки гипотезы о вероятности события.

Пусть проведено n независимых испытаний (n – достаточно большое число), в каждом из которых некоторое событие A появляется с одной и той же, но неизвестной вероятностью p, и найдена относительная частота $\frac{m}{n}$ появлений A в этой серии испытаний. Проверим при заданном уровне значимости α нулевую гипотезу H_0, состоящую в том, что вероятность p равна некоторому значению p_0.

Примем в качестве статистического критерия случайную величину

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}, \qquad)$$

имеющую нормальное распределение с параметрами $M(U) = 0$, $\sigma(U) = 1$ (то есть нормированную). Здесь $q_0 = 1 - p_0$. Вывод о нормальном распределении критерия следует из теоремы Лапласа (при достаточно большом n относительную частоту можно приближенно считать нормально распределенной с математическим ожиданием p и средним квадратическим отклонением $\sqrt{\frac{pq}{n}}$).

Критическая область строится в зависимости от вида конкурирующей гипотезы.

1) Если H_0: $p = p_0$, а H_1: $p \neq p_0$, то критическую область нужно построить так, чтобы вероятность попадания критерия в эту область равнялась заданному уровню значимости α. При этом наибольшая мощность критерия достигается тогда, когда критическая область состоит из двух интервалов, вероятность попадания в каждый из которых равна $\frac{\alpha}{2}$. Поскольку U симметрична относительно оси Оy, вероятность ее попадания в интервалы (-∞; 0) и (0; +∞) равна 0,5, следовательно, критическая область тоже должна быть симметрична относительно Оy. Поэтому $u_{кр}$ определяется по таблице значений функции Лапласа из условия $\Phi(u_{кр}) = \frac{1-\alpha}{2}$, а критическая область имеет вид $(-\infty;-u_{кр}) \cup (u_{кр};+\infty)$.

Замечание. Предполагается, что используется таблица значений функции Лапласа, заданной в виде $\Phi(x) = \int_0^x e^{-\frac{t^2}{2}} dt$, где нижний предел интегрирования равен 0, а не -∞. Функция Лапласа, заданная таким образом, является нечетной, а ее значения на 0,5 меньше, чем значения стандартной функции $\Phi(x)$ (см. лекцию 6).

Далее нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}} \quad .$$

Если $|U_{набл}| < u_{кр}$, то нулевая гипотеза принимается.
Если $|U_{набл}| > u_{кр}$, то нулевая гипотеза отвергается.

2) Если конкурирующая гипотеза H_1: $p > p_0$, то критическая область определяется неравенством $U > u_{кр}$, то есть является правосторонней, причем $p(U > u_{кр}) = \alpha$. Тогда $p(0 < U < u_{кр}) = \frac{1}{2} - \alpha = \frac{1-2\alpha}{2}$. Следовательно, $u_{кр}$ можно найти по таблице значений функции Лапласа из условия, что $\Phi(u_{кр}) = \frac{1-2\alpha}{2}$. Вычислим наблюдаемое значение критерия по формуле (19.2).
Если $U_{набл} < u_{кр}$, то нулевая гипотеза принимается.
Если $U_{набл} > u_{кр}$, то нулевая гипотеза отвергается.

3) Для конкурирующей гипотезы H_1: $p < p_0$ критическая область является левосторонней и задается неравенством $U <- u_{кр}$, где $u_{кр}$ вычисляется так же, как в предыдущем случае.
Если $U_{набл} > - u_{кр}$, то нулевая гипотеза принимается.
Если $U_{набл} < - u_{кр}$, то нулевая гипотеза отвергается.

Пример. Пусть проведено 50 независимых испытаний, и относительная частота появления события A оказалась равной 0,12. Проверим при уровне значимости $\alpha = 0{,}01$ нулевую гипотезу H_0: $p = 0{,}1$ при конкурирующей гипотезе H_1: $p > 0{,}1$. Найдем $U_{набл} = \dfrac{(0{,}12 - 0{,}1)\sqrt{50}}{\sqrt{0{,}1 \cdot 0{,}9}} = 0{,}471.$ Критическая область является правосторонней, а $u_{кр}$ находим из равенства $\Phi(u_{кр}) = \dfrac{1 - 2 \cdot 0{,}01}{2} = 0{,}49.$ Из таблицы значений функции Лапласа определяем $u_{кр} = 2{,}33$. Итак, $U_{набл} < u_{кр}$, и гипотеза о том, что $p = 0{,}1$, принимается.

Критерий для проверки гипотезы о математическом ожидании.

Пусть генеральная совокупность X имеет нормальное распределение, и требуется проверить предположение о том, что ее математическое ожидание равно некоторому числу a_0. Рассмотрим две возможности.

1) Известна дисперсия σ^2 генеральной совокупности. Тогда по выборке объема n найдем выборочное среднее $\bar{x}_B$ и проверим нулевую гипотезу H_0: $M(X) = a_0$.

Учитывая, что выборочное среднее $\overline{X}$ является несмещенной оценкой $M(X)$, то есть $M(\overline{X}) = M(X)$, можно записать нулевую гипотезу так: $M(\overline{X}) = a_0$. Для ее проверки выберем критерий

$$U = \frac{\overline{X} - a_0}{\sigma(\overline{X})} = \frac{(\overline{X} - a_0)\sqrt{n}}{\sigma}. \qquad)$$

Это случайная величина, имеющая нормальное распределение, причем, если нулевая гипотеза справедлива, то $M(U) = 0$, $\sigma(U) = 1$.

Выберем критическую область в зависимости от вида конкурирующей гипотезы:

- если H_1: $M(\overline{X}) \neq a_0$, то $u_{кр}$: $\Phi(u_{кр}) = \dfrac{1-\alpha}{2}$, критическая область двусторонняя, $U_{набл} = \dfrac{(\bar{x} - a_0)\sqrt{n}}{\sigma}$, и, если $|U_{набл}| < u_{кр}$, то нулевая гипотеза принимается; если $|U_{набл}| > u_{кр}$, то нулевая гипотеза отвергается.

- если H_1: $M(\overline{X}) > a_0$, то $u_{кр}$: $\Phi(u_{кр}) = \dfrac{1-2\alpha}{2}$, критическая область правосторонняя, и, если $U_{набл} < u_{кр}$, то нулевая гипотеза принимается; если $U_{набл} > u_{кр}$, то нулевая гипотеза отвергается.

- если H_1: $M(\overline{X}) < a_0$, то $u_{кр}$: $\Phi(u_{кр}) = \dfrac{1-2\alpha}{2}$, критическая область левосторонняя, и, если $U_{набл} > -u_{кр}$, то нулевая гипотеза принимается; если $U_{набл} < -u_{кр}$, то нулевая гипотеза отвергается.

2) Дисперсия генеральной совокупности неизвестна.

В этом случае выберем в качестве критерия случайную величину

$$T = \frac{(\overline{X} - a_0)\sqrt{n}}{S},$$

где S – исправленное среднее квадратическое отклонение. Такая случайная величина имеет распределение Стьюдента с $k = n - 1$ степенями свободы. Рассмотрим те же, что и в предыдущем случае, конкурирующие гипотезы и соответствующие им критические области. Предварительно вычислим наблюдаемое значение критерия:

$$T_{набл} = \frac{(\bar{x}_B - a_0)\sqrt{n}}{S}.$$

- если H_1: $M(\overline{X}) \neq a_0$, то критическая точка $t_{двуст.кр.}$ находится по таблице критических точек распределения Стьюдента по известным α и $k = n - 1$.

Если $|T_{набл}| < t_{двуст.кр.}$, то нулевая гипотеза принимается.

Если $|T_{набл}| > t_{двуст.кр.}$, то нулевая гипотеза отвергается.

- если H_1: $M(\overline{X}) > a_0$, то по соответствующей таблице находят $t_{правост.кр.}(\alpha, k)$ – критическую точку правосторонней критической области. Нулевая гипотеза принимается, если $T_{набл} < t_{правост.кр.}$.

- при конкурирующей гипотезе H_1: $M(\overline{X}) < a_0$ критическая область является левосторонней, и нулевая гипотеза принимается при условии $T_{набл} > - t_{правост.кр.}$. Если $T_{набл} < - t_{правост.кр.}$., нулевую гипотезу отвергают.

Критерий для проверки гипотезы о сравнении двух дисперсий.

Пусть имеются две нормально распределенные генеральные совокупности X и Y. Из них извлечены независимые выборки объемов соответственно n_1 и n_2, по которым вычислены исправленные выборочные дисперсии s_X^2 и s_Y^2. Требуется при заданном уровне значимости α проверить нулевую гипотезу H_0: $D(X) = D(Y)$ о равенстве дисперсий рассматриваемых генеральных совокупностей. Учитывая несмещенность исправленных выборочных дисперсий, можно записать нулевую гипотезу так:

$$H_0: M(s_X^2) = M(s_Y^2).$$

Замечание. Конечно, исправленные дисперсии, вычисленные по выборкам, обычно оказываются различными. При проверке гипотезы выясняется, является ли это различие незначимым и обусловленным случайными причинами (в случае принятия нулевой гипотезы) или оно является следствием того, что сами генеральные дисперсии различны.

В качестве критерия примем случайную величину

$$F = \frac{S_\sigma^2}{S_M^2} -$$

- отношение большей выборочной дисперсии к меньшей. Она имеет распределение Фишера-Снедекора со степенями свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 – объем выборки, по которой вычислена большая исправленная дисперсия, а n_2 – объем второй выборки. Рассмотрим два вида конкурирующих гипотез:

- пусть H_1: $D(X) > D(Y)$. Наблюдаемым значением критерия будет отношение большей из исправленных дисперсий к меньшей: $F_{набл} = \frac{s_\sigma^2}{s_M^2}$. По таблице критических точек распределения Фишера-Снедекора можно найти критическую точку $F_{набл}(\alpha; k_1; k_2)$. При $F_{набл} < F_{кр}$ нулевая гипотеза принимается, при $F_{набл} > F_{кр}$ отвергается.

- если H_1: $D(X) \neq D(Y)$, то критическая область является двусторонней и определяется неравенствами $F < F_1$, $F > F_2$, где $p(F < F_1) = p(F > F_2) = \alpha/2$. При этом достаточно найти правую критическую точку $F_2 = F_{кр}(\frac{\alpha}{2}, k_1, k_2)$. Тогда при $F_{набл} < F_{кр}$ нулевая гипотеза принимается, при $F_{набл} > F_{кр}$ отвергается.

1. 6 Лекция №6 (4 часа).

Тема: «Элементы корреляционного анализа»

1.6.1 Вопросы лекции:

1. Понятие корреляции.
2. Коэффициент корреляции Пирсона.
3. Регрессия.
4. Коэффициент регрессии.

1.6.2 Краткое содержание вопросов:

В предыдущей лекции рассматривались гипотезы, в которых закон распределения генеральной совокупности предполагался известным. Теперь займемся проверкой гипотез о предполагаемом законе неизвестного распределения, то есть будем проверять нулевую гипотезу о том, что генеральная совокупность распределена по некоторому известному закону. Обычно статистические критерии для проверки таких гипотез называются **критериями согласия**.

Критерий Пирсона.

Достоинством критерия Пирсона является его универсальность: с его помощью можно проверять гипотезы о различных законах распределения.

1. Проверка гипотезы о нормальном распределении.

Пусть получена выборка достаточно большого объема n с большим количеством различных значений вариант. Доя удобства ее обработки разделим интервал от наименьшего до наибольшего из значений вариант на s равных частей и будем считать, что значения вари ант, попавших в каждый интервал, приближенно равны числу, задающему середину интервала. Подсчитав число вариант, попавших в каждый интервал, составим так называемую сгруппированную выборку:

варианты……..$x_1 \quad x_2 \quad \dots \quad x_s$

частоты……$n_1 \quad n_2 \quad \dots \quad n_s$,

где x_i – значения середин интервалов, а n_i – число вариант, попавших в i-й интервал (эмпирические частоты).

По полученным данным можно вычислить выборочное среднее $\bar{x}_B$ и выборочное среднее квадратическое отклонение σ_B. Проверим предположение, что генеральная совокупность распределена по нормальному закону с параметрами $M(X) = \bar{x}_B$, $D(X) = \sigma_B^2$. Тогда можно найти количество чисел из выборки объема n, которое должно оказаться в каждом интервале при этом предположении (то есть теоретические частоты). Для этого по таблице значений функции Лапласа найдем вероятность попадания в i-й интервал:

$$p_i = \Phi\left(\frac{b_i - \bar{x}_B}{\sigma_B}\right) - \Phi\left(\frac{a_i - \bar{x}_B}{\sigma_B}\right),$$

где a_i и b_i - границы i-го интервала. Умножив полученные вероятности на объем выборки п, найдем теоретические частоты: $n_i = n \cdot p_i$. Наша цель – сравнить эмпирические и теоретические частоты, которые, конечно, отличаются друг от друга, и выяснить, являются ли эти различия несущественными, не опровергающими гипотезу о нормальном распределении исследуемой случайной величины, или они настолько велики, что противоречат этой гипотезе. Для этого используется критерий в виде случайной величины

$$\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}. \qquad (20.1)$$

Смысл ее очевиден: суммируются части, которые квадраты отклонений эмпирических частот от теоретических составляют от соответствующих теоретических частот. Можно доказать, что вне зависимости от реального закона распределения генеральной совокупности закон распределения случайной величины (20.1) при $n \to \infty$ стремится к закону распределения χ^2 (см. лекцию 12) с числом степеней свободы $k = s - 1 - r$, где r – число параметров предполагаемого распределения, оцененных по данным выборки. Нормальное распределение характеризуется двумя параметрами, поэтому $k = s - 3$. Для выбранного критерия строится правосторонняя критическая область, определяемая условием

$$p(\chi^2 > \chi^2_{kp}(\alpha, k)) = \alpha, \qquad (20.2)$$

где α – уровень значимости. Следовательно, критическая область задается неравенством $\chi^2 > \chi^2_{kp}(\alpha, k)$, а область принятия гипотезы - $\chi^2 < \chi^2_{kp}(\alpha, k)$.

Итак, для проверки нулевой гипотезы H_0: генеральная совокупность распределена нормально – нужно вычислить по выборке наблюдаемое значение критерия:

$$\chi^2_{набл} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}, \qquad (20.1`)$$

а по таблице критических точек распределения χ^2 найти критическую точку $\chi^2_{кр}(\alpha, k)$, используя известные значения α и $k = s - 3$. Если $\chi^2_{набл} < \chi^2_{kp}$ - нулевую гипотезу принимают, при $\chi^2_{набл} > \chi^2_{kp}$ ее отвергают.

2. Проверка гипотезы о равномерном распределении.

При использовании критерия Пирсона для проверки гипотезы о равномерном распределении генеральной совокупности с предполагаемой плотностью вероятности

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a,b) \\ 0, & x \notin (a,b) \end{cases}$$

необходимо, вычислив по имеющейся выборке значение $\bar{x}_B$, оценить параметры a и b по формулам:

$$a^* = \bar{x}_B - \sqrt{3}\sigma_B, \quad b^* = \bar{x}_B + \sqrt{3}\sigma_B, \qquad (20.3)$$

где a^* и b^* - оценки a и b. Действительно, для равномерного распределения $M(X) = \dfrac{a+b}{2}$, $\sigma(x) = \sqrt{D(X)} = \sqrt{\dfrac{(a-b)^2}{12}} = \dfrac{a-b}{2\sqrt{3}}$, откуда можно получить систему для определения a^* и b^*: $\begin{cases} \dfrac{b^*+a^*}{2} = \bar{x}_B \\ \dfrac{b^*-a^*}{2\sqrt{3}} = \sigma_B \end{cases}$, решением которой являются выражения (20.3).

Затем, предполагая, что $f(x) = \dfrac{1}{b^*-a^*}$, можно найти теоретические частоты по формулам

$$n_1' = np_1 = nf(x)(x_1 - a^*) = n \cdot \frac{1}{b^*-a^*}(x_1 - a^*);$$

$$n_2' = n_3' = \ldots = n_{s-1}' = n \cdot \frac{1}{b^*-a^*}(x_i - x_{i-1}), \quad i = 1,2,\ldots,s-1;$$

$$n'_s = n \cdot \frac{1}{b^* - a^*}(b^* - x_{s-1}).$$

Здесь s – число интервалов, на которые разбита выборка.

Наблюдаемое значение критерия Пирсона вычисляется по формуле (20.1`), а критическое – по таблице с учетом того, что число степеней свободы $k = s - 3$. После этого границы критической области определяются так же, как и для проверки гипотезы о нормальном распределении.

3. Проверка гипотезы о показательном распределении.

В этом случае, разбив имеющуюся выборку на равные по длине интервалы, рассмотрим последовательность вариант $x_i^* = \frac{x_i + x_{i+1}}{2}$, равноотстоящих друг от друга (считаем, что все варианты, попавшие в i – й интервал, принимают значение, совпадающее с его серединой), и соответствующих им частот n_i (число вариант выборки, попавших в i – й интервал). Вычислим по этим данным $\bar{x}_B$ и примем в качестве оценки параметра λ величину $\lambda^* = \frac{1}{\bar{x}_B}$. Тогда теоретические частоты вычисляются по формуле

$$n'_i = n_i p_i = n_i p(x_i < X < x_{i+1}) = n_i (e^{-\lambda x_i} - e^{-\lambda x_{i+1}}).$$

Затем сравниваются наблюдаемое и критическое значение критерия Пирсона с учетом того, что число степеней свободы $k = s - 2$.

Критерий Колмогорова.

Этот критерий применяется для проверки простой гипотезы H_0 о том, что независимые одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$ имеют заданную непрерывную функцию распределения $F(x)$.

Найдем функцию эмпирического распределения $F_n(x)$ и будем искать границы двусторонней критической области, определяемой условием

$$D_n = \sup_{|x|<\infty} | F_n(x) - F(x) | > \lambda_n .$$

А.Н.Колмогоров доказал, что в случае справедливости гипотезы H_0 распределение статистики D_n не зависит от функции $F(x)$, и при $n \to \infty$

$$p(\sqrt{n} D_n < \lambda) \to K(\lambda), \quad \lambda > 0,$$

где
$$K(\lambda) = \sum_{m=-\infty}^{\infty} (-1)^m e^{-2m^2\lambda^2} \quad - \qquad (20.4)$$

- критерий Колмогорова, значения которого можно найти в соответствующих таблицах. Критическое значение критерия $\lambda_n(\alpha)$ вычисляется по заданному уровню значимости α как корень уравнения $p(D_n \geq \lambda) = \alpha$.

Можно показать, что приближенное значение вычисляется по формуле

$$\lambda_n(\alpha) \approx \sqrt{\frac{z}{2n}} - \frac{1}{6n},$$

где z – корень уравнения $1 - K\left(\sqrt{\frac{\lambda}{2}}\right) = \alpha.$

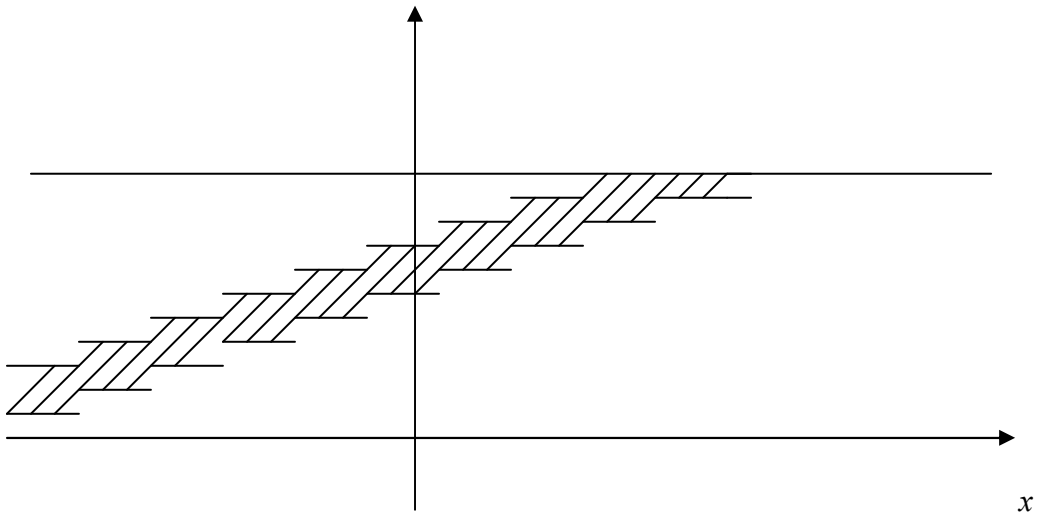
На практике для вычисления значения статистики D_n используется то, что

$D_n = \max(D_n^+, D_n^-)$, где $D_n^+ = \max\limits_{1 \leq m \leq n}\left(\frac{m}{n} - F(X_{(m)})\right)$, $D_n^- = \max\limits_{1 \leq m \leq n}\left(F(X_{(m)}) - \frac{m-1}{n}\right)$,

а $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ - вариационный ряд, построенный по выборке $X_1, X_2, \ldots, X_n$.

Можно дать следующее геометрическое истолкование критерия Колмогорова: если изобразить на плоскости Oxy графики функций $F_n(x)$, $F_n(x) \pm \lambda_n(\alpha)$ (рис. 1), то гипотеза H_0

верна, если график функции $F(x)$ не выходит за пределы области, лежащей между графиками функций $F_n(x) - \lambda_n(\alpha)$ и $F_n(x) + \lambda_n(\alpha)$.



Приближенный метод проверки нормальности распределения, связанный с оценками коэффициентов асимметрии и эксцесса.

Определим по аналогии с соответствующими понятиями для теоретического распределения асимметрию и эксцесс эмпирического распределения.

Определение 20.1. **Асимметрия эмпирического распределения** определяется равенством

$$a_s = \frac{m_3}{\sigma_B^3},$$

где m_3 – центральный эмпирический момент третьего порядка.

Эксцесс эмпирического распределения определяется равенством

$$e_k = \frac{m_4}{\sigma_B^4} - 3,$$

где m_4 – центральный эмпирический момент четвертого порядка.

Как известно, для нормально распределенной случайной величины асимметрия и эксцесс равны 0. Поэтому, если соответствующие эмпирические величины достаточно малы, можно предположить, что генеральная совокупность распределена по нормальному закону.

Рассмотрим выборку двумерной случайной величины (X, Y) . Примем в качестве оценок условных математических ожиданий компонент их условные средние значения, а именно: **условным средним** $\overline{y}_x$ назовем среднее арифметическое наблюдавшихся значений Y, соответствующих $X = x$. Аналогично **условное среднее** $\overline{x}_y$ **-** среднее арифметическое наблюдавшихся значений X, соответствующих $Y = y$. В лекции 11 были выведены уравнения регрессии Y на X и X на Y:

$$M(Y/x) = f(x), \quad M(X/y) = \varphi(y).$$

Условные средние $\overline{y}_x$ и $\overline{x}_y$ являются оценками условных математических ожиданий и, следовательно, тоже функциями от x и y, то есть

$$\overline{y}_x = f^*(x) - \qquad (22.1)$$

- **выборочное уравнение регрессии *Y* на *X*,**

$$\overline{x}_y = \varphi^*(y) - \qquad (22.2)$$

- **выборочное уравнение регрессии *X* на *Y*.**

Соответственно функции $f^*(x)$ и $\varphi^*(y)$ называются **выборочной регрессией *Y* на *X* и *X* на *Y***, а их графики – **выборочными линиями регрессии**. Выясним, как определять параметры выборочных уравнений регрессии, если сам вид этих уравнений известен.

Пусть изучается двумерная случайная величина (*X, Y*), и получена выборка из *n* пар чисел (x_1, y_1), (x_2, y_2),…, (x_n, y_n). Будем искать параметры прямой линии среднеквадратической регрессии *Y* на *X* вида

$$Y = \rho_{yx}x + b\,, \qquad (22.3)$$

Подбирая параметры ρ_{yx} и *b* так, чтобы точки на плоскости с координатами (x_1, y_1), (x_2, y_2), …, (x_n, y_n) лежали как можно ближе к прямой (22.3). Используем для этого метод наименьших квадратов и найдем минимум функции

$$F(\rho,b) = \sum_{i=1}^{n}(Y_i - y_i)^2 = \sum_{i=1}^{n}(\rho x_i + b - y_i)^2\,. \qquad (22.4)$$

Приравняем нулю соответствующие частные производные:

$$\begin{cases} \dfrac{\partial F}{\partial \rho} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i)x_i = 0 \\ \dfrac{\partial F}{\partial b} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i) = 0 \end{cases}.$$

В результате получим систему двух линейных уравнений относительно ρ и *b*:

$$\begin{cases} \left(\sum x^2\right)\rho + \left(\sum x\right)b = \sum xy \\ \left(\sum x\right)\rho + nb = \sum y \end{cases}. \qquad (22.5)$$

Ее решение позволяет найти искомые параметры в виде:

$$\rho_{xy} = \frac{n\sum xy - \sum x\cdot\sum y}{n\sum x^2 - \left(\sum x\right)^2};\quad b = \frac{\sum x^2\cdot\sum y - \sum x\cdot\sum xy}{n\sum x^2 - \left(\sum x\right)^2}\,. \qquad (22.6)$$

При этом предполагалось, что все значения *X* и *Y* наблюдались по одному разу.

Теперь рассмотрим случай, когда имеется достаточно большая выборка (не менее 50 значений), и данные сгруппированы в виде *корреляционной таблицы*:

Y	*X*				
	x_1	x_2	…	x_k	n_y
y_1	n_{11}	n_{21}	…	n_{k1}	$n_{11}+n_{21}+\ldots+n_{k1}$
y_2	n_{12}	n_{22}	…	n_{k2}	$n_{12}+n_{22}+\ldots+n_{k2}$
…	…	…	…	…	……..
y_m	n_{1m}	n_{2m}	…	n_{km}	$n_{1m}+n_{2m}+\ldots+n_{km}$
n_x	$n_{11}+n_{12}+\ldots+n_{1m}$	$n_{21}+n_{22}+\ldots+n_{2m}$	…	$n_{k1}+n_{k2}+\ldots+n_{km}$	$n=\sum n_x = \sum n_y$

Здесь n_{ij} – число появлений в выборке пары чисел (x_i, y_j).

Поскольку $\bar{x} = \dfrac{\sum x}{n}$, $\bar{y} = \dfrac{\sum y}{n}$, $\overline{x^2} = \dfrac{\sum x^2}{n}$, заменим в системе (22.5) $\sum x = n\bar{x}$,

$\sum y = n\bar{y}$, $\sum x^2 = n\overline{x^2}$, $\sum xy = \sum n_{xy}xy$, где n_{xy} – число появлений пары чисел (*x, y*). Тогда система (22.5) примет вид:

$$\begin{cases} (n\overline{x^2})\rho_{yx} + (n\bar{x})b = \sum n_{xy}xy \\ (\bar{x})\rho_{yx} + b = \bar{y} \end{cases}.$$

Можно решить эту систему и найти параметры ρ_{yx} и b, определяющие выборочное уравнение прямой линии регрессии:

$$\overline{y}_x = \rho_{yx}\overline{x} + b\,.$$

Но чаще уравнение регрессии записывают в ином виде, вводя **выборочный коэффициент корреляции**. Выразим b из второго уравнения системы

$$b = \overline{y} - \rho_{yx}\overline{x}\,.$$

Подставим это выражение в уравнение регрессии: $\overline{y}_x - \overline{y} = \rho_{yx}(x - \overline{x})$.

$$\rho_{yx} = \frac{\sum n_{xy}xy - n\overline{xy}}{n(\overline{x^2} - (\overline{x})^2)} = \frac{\sum n_{xy}xy - n\overline{xy}}{n\tilde{\sigma}_x^2},$$

где $\tilde{\sigma}_x^2 = \overline{x^2} - (\overline{x})^2$. Введем понятие **выборочного коэффициента корреляции**

$$r_B = \frac{\sum n_{xy}xy - n\overline{x}\,\overline{y}}{n\tilde{\sigma}_x\tilde{\sigma}_y}$$

и умножим равенство (22.8) на $\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}$: $\rho_{yx}\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} = r_B$, откуда $\rho_{yx} = r_B\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}$. Используя это соотношение, получим выборочное уравнение прямой линии регрессии Y на X вида

$$\overline{y}_x - \overline{y} = r_B\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x - \overline{x})\ .$$

1. 7 Лекция №7 (4 часа).

Тема: «Элементы математического моделирования»

1.7.1 Вопросы лекции:

1. Модели. Виды моделей.
2. Модель численности популяции, межвидовой конкуренции, хищник-жертва, кооперация видов.
3. Закон Харди-Вайнберга.
4. Модель отбора и приспособленности.
5. Уравнение Лотки-Вольтерра.

1.7.2 Краткое содержание вопросов:

Под экологией следует понимать сферу знаний, которая представляет собой взаимодействие всех живых организмов. Еще в первой половине двадцатого века данная наука была одной из биологических дисциплин, однако на данный момент экология учитывает такие важные аспекты как контроль за состоянием окружающей среды. Именно благодаря математической экологии, которая включает в себя различные методы и модели, возможно решение экологических проблем.

К сожалению, невозможно охарактеризовать сложного уровня экосистемы используя простые модели. Для описания необходимо использовать сложные имитационные модели, которые объединяют знания в одну сложную систему или интегрированные модели упрощенного вида.

Имитационные модели, разработанные на компьютерах, содержат представления об элементах системы, их взаимодействии в виде математических объектов: формул, уравнений, матриц, логических процедур, графиков, таблиц, баз данных, оперативной информации экологического мониторинга. С помощью многомерных моделей становится возможно объединить любую информацию относительно экологии и экономики, выработать модели оптимальных стратегий. При имитационном подходе обычно используют высокоразвитую вычислительную, поэтому наибольшее распространение данная наука получила не так давно.

Классы задач и математический аппарат

Сегодня в экологии математические модели делятся на три класса. Первый – модели описательные типа: регрессионные и другие эмпирически установленные количественные зависимости, которые не претендуют на раскрытие системы описываемого процесса. Такие модели принято использовать для описания отдельных процессов и зависимостей и включать в качестве фрагментов в имитационные модели. Второй - модели качественного типа. Данные модели строят для того, чтобы выяснить динамический механизм изучаемого процесса, а также способность воспроизвести наблюдаемые динамические эффекты в поведении систем, такие, например, как колебательный характер изменения биомассы или образование неоднородной в пространстве структуры. Как правило, данные модели не очень большие, поддаются качественному исследованию с применением методов аналитического характера и компьютерного. Третий класс - имитационные модели конкретных экологических и эколого-экономических систем. Такой тип моделей учитывает всю имеющуюся информацию об объекте. Главной целью является подробное и детальное прогнозирование поведения сложных систем или решение оптимизационной задачи их эксплуатации.

По мере того, насколько хорошо изучена сложная экологическая система, зависит обоснование математической модели. В том случае, если наблюдается тесная связь экспериментального исследования и математического моделирования математическая модель может служить необходимым промежуточным звеном между опытными данными и основанной на них теорией изучаемых процессов. Для решения практических задач можно использовать модели всех трех типов.

А. Биологические характеристики компонентов и взаимоотношения между ними не изменяются. Система считается однородной в пространстве. Изучаются изменения во времени численности компонентов системы.

Б. При сохранении гипотезы однородности вводится предположение о закономерном изменении системы отношений между компонентами. Это может соответствовать либо закономерному изменению внешних условий, либо заданному характеру эволюций форм, образующих систему.

Для изучения этих двух классов задач используют системы обыкновенных дифференциальных и дифференциально-разностных уравнений с постоянными (А) и переменными (Б) коэффициентами.

В. Объекты считаются разнородными по своим свойствам и подверженными действию отбора. Предполагается, что эволюция форм определяется условиями существования системы. В этих условиях изучается, с одной стороны, кинетика численности компонентов, с другой - дрейф характеристик популяций. При решении таких задач используют аппарат теории вероятностей.

Г. Отказ от территориальной однородности и учет зависимости усредненных концентраций от координат. Здесь возникают вопросы, связанные с пространственным перераспределением живых и косных компонентов системы. Для описания таких систем необходимо привлечение аппарата дифференциальных уравнений в частных производных. В имитационных моделях часто вместо непрерывного пространственного описания применяют разбиение всей системы на несколько пространственных блоков.

Принципы лимитирования в экологии

По причине того, что процессы в экологической системе довольно сложные, необходимо выделить значимые факторы, взаимодействие которых качественно определяет судьбу системы. Практически все модели, которые характеризуют рост популяций и сообществ, основаны на «принципе лимитирующих факторов» или на «законе совокупного действия факторов». Изначально данные принципы были рассчитаны для популяций одного только вида, но позднее стали применяться для характеристики многовидовых сообществ и систем. Суть лимитирующих факторов принадлежит немецкому агрохимику Юстусу Либиху. Он предложил знаменитый закон минимума, который гласит следующее: "Каждое поле содержит одно или несколько питательных веществ в минимуме и одно или несколько других в максимуме. Урожаи находятся в соответствии с этим минимумом питательных веществ". Либих понимал под этим относительный минимум питательного вещества по сравнению с содержанием других веществ. Позже в экологической литературе фактор, находящийся в минимуме, стали называть лимитирующим фактором. Закон «лимитирующего фактора» для фотосинтетических процессов в 1905 г. предложил Ф.Блэкман, а в 1965 г. Н.Д.Иерусалимский сформулировал этот закон для ферментативных процессов.

Закон толерантности и функции отклика.

В современном мире метод функций отклика в науке используется для исследования зависимости реакции экологической системы от каких-либо факторов. Данный метод получил широкую известность и наиболее часто используется в инженерных науках. Суть метода состоит в использовании информации об отклике системы на известные воздействия для получения оператора перехода по схеме: воздействие реакция. В терминах теории сложных систем, динамика сложной открытой системы характеризуется описанием связи между входными и выходными сигналами.

Американский ученый В.Шелфорд. сформулировал «закон толерантности» в 1913 г. Согласно данному закону, как недостаток, так и избыток любого внешнего фактора может быть вредным для биологического объекта. В доказательство был приведен факт, что функции отклика - зависимости количественных оценок тех или иных характеристик популяций от главных факторов внешней среды (содержания питательных веществ, температуры), которые имеют колоколообразную форму. Под диапазоном толерантности следует понимать пределы, в которых может существовать живой организм. В таком случае, лимитирующий фактор – это фактор, который приближается или выходит за пределы толерантности. Сегодня в экологической литературе закон толерантности, как правило, рассматривают в качестве продолжения и расширения принципа Либиха. Лимитирующим фактором является фактор, по которому для достижения заданного относительного изменения функции отклика необходимо минимальное относительное изменение значения фактора. Такое определение требует подробного изучения зависимости функций отклика от всей совокупности экологических факторов в каждом конкретном случае. Это связано с использованием приемов многофакторного эксперимента и аппарата многомерной математической статистики. Практическое использование такого подхода к исследованию большинства природных экосистем затруднено из-за недостатка экспериментальных данных и отсутствия систематических наблюдений. Из-за ссложности экологических систем функциональную связь между компонентами системы трудно описать традиционными методами.

Модели водных экосистем

Можно с уверенностью сказать, что модели водных экосистем играют важную роль в математической экологии. Водные системы дают людям, животным, сельскому хозяйству и промышленности воду. Океаны, моря и реки обеспечивают в разных странах от 20% до 80% потребности людей в белковой пище. К сожалению, необходимо отметить тот факт, что качественная характеристика воды в водоемах и их продуктивность резко снижается. Безусловно, это связаноЭто связано в первую очередь с тем, что водоемы традиционно использовались людьми как бесплатные системы по переработке отходов, что привело к

их существенному загрязнению, нарушению естественных биологических и химических процессов. Потребности оптимизации использования водных систем и понимания происходящих в них процессов привели к быстрому развитию математического моделирования водных систем. В настоящее время насчитываются тысячи моделей разной степени сложности и подробности. Планирование любого водохозяйственного мероприятия сопровождается и предваряется построением математической модели водной системы. В 70-80 годы особенно активно развивались модели озерных экосистем. Одной из важнейших задач была выработка борьбы с эфтрификацией - "цветением" озер в связи с увеличением количества поступающего в них органического вещества, а также биогенных веществ, в первую очередь азота, вместе со стоками вод из сельскохозяйственных угодий.
Математические модели помогают разработать оптимальную стратегию управления водными ресурсами, в том числе рыбным хозяйством. Дело в том, что наряду с ухудшением состояния воды причиной падения продуктивности водоемов являются систематические переловы. В биологическом смысле они приводят к такому состоянию рыбного стада, когда воспроизводительная способность популяции не может компенсировать убыль в результате вылова. Перелов в экономическом смысле - это сокращение поголовья рыбного стада настолько, что промысел становится нерентабельным. Решение задачи оптимизации систематического лова рыбы восходит к работам Баранова (1918). Представив коэффициенты общей смертности в виде суммы коэффициентов естественной и промысловой гибели в формуле численности рыбного стада, Баранов оценил величину улова и смог подойти к постановке задачи оптимального вылова. Значительный шаг в решении этой проблемы сделали Риккер (1958) и Бивертон и Холт (1957), связавшие модели с конкретным статистическим материалом рыбоводства и ихтиологии и предложившие методики решения задач управления. Особенно большой вклад в моделирование рыбных популяций внес В.В.Меншуткин, ("Математическое моделирование популяций и сообществ водных животных", Л.,1971), который представил схему взаимодействий в водной экосистеме как контур с обратными связями. Такая система может обладать устойчивым стационарным состоянием, в ней могут возникать колебательные или квазистохастические режимы. Подобные схемы, часто весьма детальные, были положены в основу моделей рыбного стада многих озер и морей.

Модели продукционного процесса растений

Сегодня в области математической экологии моделирование продукционного процесса растений является довольно изученной и продвинутой сферой. Это определяется практической значимостью таких моделей для оптимизации агрокультуры и тепличного хозяйства. В таких случаях математические модели обычно применяют для выбора наилучшей стратегии проведения различных мероприятий в области сельского хозяйства. К последним относятся: орошение, полив, внесение удобрений, выбор сроков посева или посадки растений с целью получения максимального урожая. В том случае, если тепличное хозяйство находится под полным контролем, можно построить модель, которая позволит охарактеризовать весь цикл процессов в соответствующих условиях. Выделяются биотический и абиотический блоки. Абиотические блоки состоят из моделей, описывающих формирование теплового, водного режима почвы и приземных слоев воздуха, концентрации и передвижения биогенных и токсических солей, различных остатков, ростовых веществ и метаболитов в почве, концентрации углекислого газа в посеве. Благодаря блочной структуре можно изучать, изменять и детализировать одни блоки, не меняя других. Обычно количество параметров внутри самих блоков намного больше количества параметров, которыми блоки соединяются между собой. На основе блоков синтезируются целостные динамические модели, которые способны предсказывать временные изменения ряда характерных параметров растений.

Оценка загрязнения атмосферы и поверхности земли.

В математической экологии достаточно серьезной проблемой является загрязнение окружающей среды. Именно благодаря данной науке возможно рассчитать

распространение загрязнений от предприятий и спланировать наилучшее место для размещения предприятий, соблюдая санитарные нормы. Распространение выбросов и последующее загрязнение окружающей среды обусловлено турбулентными пульсациями воздуха. Изменения направления ветра в течение года имеют большое значение в теории распространения. За данное время массы воздуха, которые содержат примеси различного рода, несколько раз могут изменять направление и скорость. В статистике многолетние изменения принято описывать с помощью диаграммы, которая имеет название роза ветров. В данном типе диаграммы величина вектора пропорциональна числу повторяющихся событий, связанных с движениями воздушных масс в данном направлении. Наибольшие значения диаграммы розы ветров соответствуют преобладающим в данном районе ветрам. Такая информация используется в качестве исходной при планировании новых индустриальных объектов. Следует помнить, что оценивая уровень допустимых загрязнений предприятий, расположенных среди большого числа экологически значимых зон, необходимо учитывать загрязнения от уже существующих предприятий региона. Благодаря математическим моделям можно оценить загрязнение атмосферы и поверхности различными примесями. Такие модели построены на основе уравнений аэродинамики в частных производных.

В России большой вклад в это направление внесли работы школы академика Г.И.Марчука. На территории Европы и США модели данного типа довольно распространены особенно при разрешении судебных исков, предъявляемых населением или местными властями промышленным предприятиям в связи с нанесением определенного ущерба. Чтобы оценить принесенный ущерб принято проводить экспертизу, после которой можно количественно оценить сумму штрафа. Данный штраф необходимо уплатить государственным или местным органам. Стоит отметить, что данные меры довольно действенны, т.к. во многих развитых странах они привели к внедрению очистительных технологий.

Глобальные модели

Важное место в математической экологии занимают такие модели, в которых рассматриваются глобальные изменения в результате различного характера воздействий, или изменений климата в результате космических и других причин. Классической моделью является модель ядерной зимы. Данная модель позволяет предсказать глобальное изменение климата на срок в несколько десятилетий в сторону понижения температур ниже нуля по Цельсию, а также гибель биосферы в случае широкомасштабной ядерной войны. Моделируя глобальные экологические процессы, не стоит забывать, что нужно учитывать огромное число факторов, пространственную неоднородность Земли, физические и химические процессы, антропогенные воздействия, связанные с развитием промышленности и ростом народонаселения. По причине повышенной сложности такой задачи необходимо применение системного подхода, впервые введенного в практику математического моделирования.

2. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ ЛАБОРАТОРНЫХ РАБОТ

2.1 Лабораторная работа №1 (2 часа).

Тема: «Установка пакета анализа данных. Формирование выборки**»**

2.1.1 Цель работы: овладеть навыками расчета числовых характеристик выборки с помощью Надстройки Пакет Анализа ЭТ MS Excel.

2.1.2 Описание (ход) работы:

Пакет Excel оснащен средствами статистической обработки данных. И хотя Excel существенно уступает специализированным статистическим пакетам обработки данных, тем не менее этот раздел математики представлен в Excel наиболее полно. В него включены основные, наиболее часто используемые статистические процедуры: средства описательной статистики, критерии различия, корреляционные и другие методы, позволяющие проводить необходимый статистический анализ экономических, психологических, педагогических и медико-биологических типов данных.

Каждая единица информации занимает свою собственную ячейку (клетку) в создаваемой рабочей таблице. В каждой рабочей таблице 256 столбцов (из которых в новой рабочей таблице на экране видны, как правило, только первые 10 или 11 (от A до J или K) и 65 536 строк (из которых обычно видны только первые 15-20). Каждая новая рабочая книга содержит три чистых листа рабочих таблиц.

Вся помещаемая в электронную таблицу информация хранится в отдельных клетках рабочей таблицы. Но ввести информацию можно только в текущую клетку. С помощью адреса в строке формул и табличного курсора Excel указывает, какая из клеток рабочей таблицы является текущей. В основе системы адресации клеток рабочей таблицы лежит комбинация буквы (или букв) столбца и номера строки, например A2, B12.

При рассмотрении применения методов обработки статистических данных в данной лабораторной работе ограничимся только простейшими и наиболее часто описательными статистиками, реализованными в мастере функций Excel.

Использование специальных функций

В мастере функций Excel имеется ряд специальных функций, предназначенных для вычисления выборочных характеристик.

Функция **СРЗНАЧ** вычисляет среднее арифметическое из нескольких массивов (аргументов) чисел. Аргументы *число1, число2, ...* — это от 1 до 30 массивов для которых вычисляется среднее.

Функция **МЕДИАНА** позволяет получать медиану заданной выборки. Медиана - это элемент выборки, число элементов выборки со значениями больше которого и меньше которого равно.

Функция **МОДА** вычисляет наиболее часто встречающееся значение в выборке.

Функция **ДИСП** позволяет оценить дисперсию по выборочным данным.

Функция **СТАНДОТКЛОН** вычисляет стандартное отклонение.

Функция **ЭКСЦЕСС** вычисляет оценку эксцесса по выборочным данным.

Функция **СКОС** позволяет оценить асимметрию выборочного распределения.

Функция **КВАРТИЛЬ** вычисляет квартили распределения. Функция имеет формат **КВАРТИЛЬ**(массив, значение), где *массив* – интервал ячеек, содержащих значения СВ; *значение* определяет какая квартиль должна быть найдена (0 – минимальное значение, 1 – нижняя квартиль, 2 – медиана, 3 – верхняя квартиль, 4 – максимальное значение распределения).

Пример 1. Провести статистический анализ методом описательной статистики доходов населения в регионе 1 и регионе 2.

1	49

1	51
1	49
1	51
1	49
1	51
1	49
1	51
1	49
491	51

500	500	**сумма**
50	50	**среднее**
24010	1,11	**дисперсия**
154,95	1,05	**станд. отклонение**
1	49	**квартили**
1	51	**квартили**
1	50	**медиана**
1	49	**мода**
10	-2,57	**эксцесс**
3,16	0	**скос(ассиметрия)**

Задания для самостоятельной работы

1. Наблюдение посещаемости четырех внеклассных мероприятий в экспериментальном (20 человек) и контрольном (30 человек) классах дали значения (соответственно): 18, 20, 20, 18 и 15, 23, 10, 28. Требуется найти среднее значение, стандартное отклонение, медиану и квартили этих данных.

2. Найти среднее значение, медиану, стандартное отклонение и квартили результатов бега на дистанцию 100 м у группы студентов (с): 12,8; 13,2; 13,0; 12,9; 13,5; 13,1.

3. Определите верхнюю и нижнюю квартиль, выборочную асимметрию и эксцесс для данных измерений роста групп студенток: 164, 160, 157, 166, 162, 160, 161, 159, 160, 163, 170, 171.

4. Найти наиболее популярный туристический маршрут из четырех реализуемых фирмой, если за неделю последовательно были реализованы следующие маршруты: 1, 3, 3, 2, 1, 1, 4, 4, 2, 4, 1, 3, 2, 4, 1, 4, 4, 3, 1, 2, 3, 4, 1, 1, 3.

Использование инструмента Пакет анализа

В пакете Excel помимо мастера функций имеется набор более мощных инструментов для работы с несколькими выборками и углубленного анализа данных, называемый Пакет анализа, который может быть использован для решения задач статистической обработки выборочных данных.

Для установки пакета **Анализ данных** в Excel сделайте следующее:

- в меню **Сервис** выберите команду **Надстройки**;
- в появившемся списке установите флажок **Пакет анализа**.

Для использования статистического пакета анализа данных необходимо:

- указать курсором мыши на пункт меню **Сервис** и щелкнуть левой кнопкой мыши;
- в раскрывающемся списке выбрать команду **Анализ данных** (если команда Анализ данных отсутствует в меню Сервис, то необходимо установить в Excel пакет анализа данных);
- выбрать строку **Описательная статистика** и нажать кнопку **Ok**
- в появившемся диалоговом окне указать **входной интервал**, то есть ввести ссылки на ячейки, содержащие анализируемые данные;

- указать **выходной интервал**, то есть ввести ссылку на ячейку, в которую будут выведены результаты анализа;
- в разделе **Группирование** переключатель установить в положение по столбцам или по строкам;
- установить флажок в поле **Итоговая статистика** и нажать **Ок.**

Задание для самостоятельной работы

1. В рабочей зоне производились замеры концентрации вредного вещества. Получен ряд значений (в мг./м^3): 12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 17, 23, 16. Необходимо определить основные выборочные характеристики.

2.2 Лабораторная работа №2 (4 часа).

Тема: «Структурирование и отбор данных в электронных таблицах. Создание сводных таблиц**»**

Описание (ход) работы:

Биномиальное распределение

Представляет собой распределение вероятностей числа наступлений некоторого события («удачи») в *n* повторных независимых испытаниях, если при каждом испытании вероятность наступления этого события равна *p*. При этом распределении разброс вариант (есть или нет события) является следствием влияния ряда независимых и случайных факторов.

Примером практического использования биномиального распределения может являться контроль качества партии фармакологического препарата. Здесь требуется подсчитать число изделий (упаковок), не соответствующих требованиям. Все причины, влияющие на качество препарата, принимаются одинаково вероятными и не зависящими друг от друга. Сплошная проверка качества в этой ситуации не возможна, поскольку изделие, прошедшее испытание, не подлежит дальнейшему использованию. Поэтому для контроля из партии наудачу выбирают определенное количество образцов изделий (*n*). Эти образцы всестороннее проверяют и регистрируют число бракованных изделий (*k*). Теоретически число бракованных изделий может быть любым, от 0 до *n*.

В Excel функция **БИНОМРАСП** применяется для вычисления вероятности в задачах с фиксированным числом тестов или испытаний, когда результатом любого испытания может быть только успех или неудача.

Функция использует следующие параметры:

БИНОМРАСП (число_успехов; число_испытаний; вероятностъ_успеха; интегральная), где

число_успехов — это количество успешных испытаний;

число_испытаний — это число независимых испытаний (число успехов и число испытаний должны быть целыми числами);

вероятность_ успеха — это вероятность успеха каждого испытания;

интегральный — это логическое значение, определяющее форму функции.

Если данный параметр имеет значение **ИСТИНА** (=1), то считается интегральная функция распределения (вероятность того, что число успешных испытаний не менее значения *число_ успехов*);

если этот параметр имеет значение **ЛОЖЬ** (=0), то вычисляется значение функции плотности распределения (вероятность того, что число успешных испытаний в точности равно значению аргумента *число_ успехов*).

Пример 1. Какова вероятность того, что трое из четырех новорожденных будут мальчиками?

Решение:

1. Устанавливаем табличный курсор в свободную ячейку, например в **A1**. Здесь должно оказаться значение искомой вероятности.

2. Для получения значения вероятности воспользуемся специальной функцией: нажимаем на панели инструментов кнопку **Вставка функции (fx)**.

3. В появившемся диалоговом окне **Мастер функций** - шаг 1 из 2 слева в поле **Категория** указаны виды функций. Выбираем **Статистическая**. Справа в поле Функция выбираем функцию **БИНОМРАСП** и нажимаем на кнопку **ОК**.

Появляется диалоговое окно функции. В поле **Число_s** вводим с клавиатуры количество успешных испытаний (3). В поле **Испытания** вводим с клавиатуры общее количество испытаний (4). В рабочее поле **Вероятность_s** вводим с клавиатуры вероятность успеха в отдельном испытании (0,5). В поле **Интегральный** вводим с клавиатуры вид функции распределения — интегральная или весовая (0). Нажимаем на кнопку **ОК**.

В ячейке A1 появляется искомое значение вероятности **p = 0,25**. Ровно 3 мальчика из 4 новорожденных могут появиться с вероятностью 0,25.

Если изменить формулировку условия задачи и выяснить вероятность того, что появится не более трех мальчиков, то в этом случае в рабочее поле **Интегральный** вводим 1 (вид функции распределения интегральный). Вероятность этого события будет равна 0,9375.

Задания для самостоятельной работы

1. Какова вероятность того, что восемь из десяти студентов, сдающих зачет, получат «незачет». (0,04)

Нормальное распределение

Нормальное распределение - это совокупность объектов, в которой крайние значения некоторого признака — наименьшее и наибольшее — появляются редко; чем ближе значение признака к математическому ожиданию, тем чаще оно встречается. Например, распределение студентов по их весу приближается к нормальному распределению. Это распределение имеет очень широкий круг приложений в статистике, включая проверку гипотез.

Диаграмма нормального распределения симметрична относительно точки *a* (математического ожидания). Медиана нормального распределения равна тоже *a*. При этом в точке *a* функция f(x) достигает своего максимума, который равен $\frac{1}{\sigma\sqrt{2\pi}}$.

В Excel для вычисления значений нормального распределения используются функция **НОРМРАСП,** которая вычисляет значения вероятности нормальной функции распределения для указанного среднего и стандартного отклонения.

Функция имеет параметры:

НОРМРАСП ***(x; среднее; стандартное_откл; интегральная)***, где:

x — значения выборки, для которых строится распределение;

среднее — среднее арифметическое выборки;

стандартное_откл — стандартное отклонение распределения;

интегральный — логическое значение, определяющее форму функции. Если интегральная имеет значение ИСТИНА(1), то функция НОРМРАСП возвращает интегральную функцию распределения; если это аргумент имеет значение ЛОЖЬ (0), то вычисляет значение функция плотности распределения.

Если среднее = 0 и стандартное_откл = 1, то функция **НОРМРАСП** возвращает стандартное нормальное распределение.

Пример 2. Построить график нормальной функции распределения f(x) при *x,* меняющемся от 19,8 до 28,8 с шагом 0,5, *a*=24,3 и σ =1,5.

Решение

1. В ячейку A1 вводим символ случайной величины ***x***, а в ячейку B1 — символ функции плотности вероятности — ***f(x)***.

2. Вводим в диапазон A2:A21 значения *x* от 19,8 до 28,8 с шагом 0,5. Для этого воспользуемся маркером автозаполнения: в ячейку A2 вводим левую границу диапазона (19,8), в ячейку A3 левую границу плюс шаг (20,3). Выделяем блок A2:A3. Затем за правый нижний угол протягиваем мышью до ячейки A21 (при нажатой левой кнопке мыши).

3. Устанавливаем табличный курсор в ячейку B2 и для получения значения вероятности воспользуемся специальной функцией — нажимаем на панели инструментов кнопку **Вставка функции (fx)**. В появившемся диалоговом окне Мастер функций - шаг 1 из 2 слева в поле **Категория** указаны виды функций. Выбираем **Статистическая**. Справа в поле **Функция** выбираем функцию **НОРМРАСП**. Нажимаем на кнопку **ОК**.

4. Появляется диалоговое окно **НОРМРАСП**. В рабочее поле **X** вводим адрес ячейки A2 щелчком мыши на этой ячейке. В рабочее поле **Среднее** вводим с клавиатуры значение математического ожидания (24,3). В рабочее поле **Стандартное_откл** вводим с клавиатуры значение среднеквадратического отклонения (1,5). В рабочее поле **Интегральная** вводим с клавиатуры вид функции распределения (0). Нажимаем на кнопку **ОК**.

5. В ячейке B2 появляется вероятность p = 0,002955. Указателем мыши за правый нижний угол табличного курсора протягиванием (при нажатой левой кнопке мыши) из ячейки B2 до B21 копируем функцию **НОРМРАСП** в диапазон B3:B21.

6. По полученным данным строим искомую диаграмму нормальной функции распределения. Щелчком указателя мыши на кнопке на панели инструментов вызываем **Мастер диаграмм**. В появившемся диалоговом окне выбираем тип диаграммы **График**, вид — левый верхний. После нажатия кнопки **Далее** указываем диапазон данных — B1:B21 (с помощью мыши). Проверяем, положение переключателя Ряды в: столбцах. Выбираем закладку **Ряд** и с помощью мыши вводим диапазон подписей оси X: A2:A21. Нажав на кнопку **Далее**, вводим названия осей X и У и нажимаем на кнопку **Готово**.

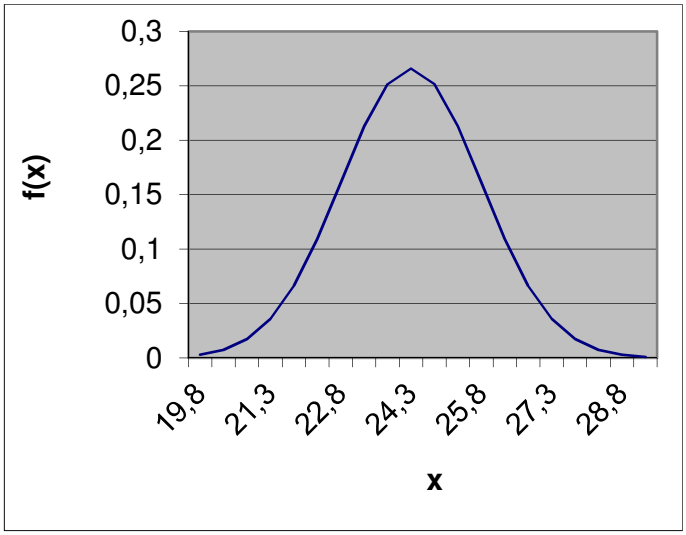


Рис. 1 График нормальной функции распределения

Получен приближенный график нормальной функции плотности распределения (см. рис.1).

Задания для самостоятельной работы

1. Построить график нормальной функции плотности распределения f(x) при *x,* меняющемся от 20 до 40 с шагом 1 при σ =3.

Генерация случайных величин

Еще одним аспектом использования законов распределения вероятностей является генерация случайных величин. Бывают ситуации, когда необходимо получить последовательность случайных чисел. Это, в частности, требуется для моделирования объектов, имеющих случайную природу, по известному распределению вероятностей.

Процедура генерации случайных величин используется для заполнения диапазона ячеек случайными числами, извлеченными из одного или нескольких распределений.

В MS Excel для генерации СВ используются функции из категории **Математические**:

СЛЧИС () – выводит на экран равномерно распределенные случайные числа больше или равные 0 и меньшие 1;

СЛУЧМЕЖДУ *(ниж_граница; верх_граница)* – выводит на экран случайное число, лежащее между произвольными заданными значениями.

В случае использования процедуры **Генерация случайных чисел** из пакета **Анализа** необходимо заполнить следующие поля:

- **число переменных** вводится число столбцов значений, которые необходимо разместить в выходном диапазоне. Если это число не введено, то все столбцы в выходном диапазоне будут заполнены;

- **число случайных чисел** вводится число случайных значений, которое необходимо вывести для каждой переменной, если число случайных чисел не будет введено, то все строки выходного диапазона будут заполнены;

- в поле **распределение** необходимо выбрать тип распределения, которое следует использовать для генерации случайных переменных:

1. ***равномерное*** - характеризуется верхней и нижней границами. Переменные извлекаются с одной и той же вероятностью для всех значений интервала.

2. ***нормальное*** — характеризуется средним значением и стандартным отклонением. Обычно для этого распределения используют среднее значение 0 и стандартное отклонение 1.

3. ***биномиальное*** — характеризуется вероятностью успеха (величина *p)* для некоторого числа попыток. Например, можно сгенерировать случайные двухальтернативные переменные по числу попыток, сумма которых будет биномиальной случайной переменной;

4. ***дискретное*** — характеризуется значением СВ и соответствующим ему интервалом вероятности, диапазон должен состоять из двух столбцов: левого, содержащего значения, и правого, содержащего вероятности, связанные со значением в данной строке. Сумма вероятностей должна быть равна 1;

5. распределения ***Бернулли***, ***Пуассона*** и ***Модельное***.

- в поле **случайное рассеивание** вводится произвольное значение, для которого необходимо генерировать случайные числа. Впоследствии можно снова использовать это значение для получения тех же самых случайных чисел.

- **выходной диапазон** вводится ссылка на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные.

Рассмотрим пример.

Пример 3. Повар столовой может готовить 4 различных первых блюда (уха, щи, борщ, грибной суп). Необходимо составить меню на месяц, так чтобы первые блюда чередовались в случайном порядке.

Решение

1. Пронумеруем первые блюда по порядку: 1 — уха, 2 — щи, 3 — борщ, 4 — грибной суп. Введем числа 1-4 в диапазон A2:A5 рабочей таблицы.
2. Укажем желаемую вероятность появления каждого первого блюда. Пусть все блюда будут равновероятны (p=1/4). Вводим число *0,25* в диапазон B2:B5.
3. В меню **Сервис** выбираем пункт **Анализ данных** и далее указываем строку **Генерация случайных чисел**. В появившемся диалоговом окне указываем **Число переменных** — *1,* **Число случайных чисел** — *30* (количество дней в месяце). В поле **Распределение** указываем **Дискретное** (только натуральные числа). В поле **Входной интервал** значений и вероятностей вводим (мышью) диапазон, содержащий номера супов и их вероятности. – A2:B5.
4. Указываем выходной диапазон и нажимаем **ОК.** В столбце С появляются случайные числа: 1, 2, 3, 4.

Задание для самостоятельной работы

1. Сформировать выборку из 10 случайных чисел, лежащих в диапазоне от 0 до 1.
2. Сформировать выборку из 20 случайных чисел, лежащих в диапазоне от 5 до 20.
3. Пусть спортсмену необходимо составить график тренировок на 10 дней, так чтобы дистанция, пробегаемая каждый день, случайным образом менялась от 5 до 10 км.
4. Составить расписание внеклассных мероприятий на неделю для случайного проведения: семинаров, интеллектуальных игр, КВН и спец. курса.
5. Составить расписание на месяц для случайной демонстрации на телевидении одного из четырех рекламных роликов турфирмы. Причем вероятность появления рекламного ролика №1 должна быть в два раза выше, чем остальных рекламных роликов.

2.3 Лабораторная работа №3 (2 часа).

Тема: «Построение графиков и гистограмм в электронных таблицах**»**

2.3.1 Цель работы: Научиться строить графики и гистограммы в электронных таблицах

2.3.2 Описание (ход) работы:

Рассмотренные в лабораторной работе 2 распределения вероятностей СВ опираются на знание закона распределения СВ. Для практических задач такое знание – редкость. Здесь закон распределения обычно неизвестен, или известен с точностью до некоторых неизвестных параметров. В частности, невозможно рассчитать точное значение соответствующих вероятностей, так как нельзя определить количество общих и благоприятных исходов. Поэтому вводится статистическое определение вероятности. По этому определению вероятность равна отношению числа испытаний, в которых событие произошло, к общему числу произведенных испытаний. Такая вероятность называется статистической частотой.

Связь между **эмпирической функцией распределения** и функцией распределения (теоретической функцией распределения) такая же, как связь между частотой события и его вероятностью.

Для построения выборочной функции распределения весь диапазон изменения случайной величины X (выборки) разбивают на ряд интервалов (карманов) одинаковой ширины. Число интервалов обычно выбирают не менее 3 и не более 15. Затем определяют число значений случайной величины X, попавших в каждый интервал (абсолютная частота, частота интервалов).

Частота интервалов – число, показывающее сколько раз значения, относящиеся к каждому интервалу группировки, встречаются в выборке. Поделив эти числа на общее количество наблюдений *(n)*, находят **относительную частоту (частость)** попадания случайной величины *X* в заданные интервалы.

По найденным относительным частотам строят гистограммы выборочных функций распределения. **Гистограмма распределения частот** – это графическое представление выборки, где по оси абсцисс (OX) отложены величины интервалов, а по оси ординат (OУ) – величины частот, попадающих в данный классовый интервал. При увеличении до бесконечности размера выборки выборочные функции распределения превращаются в теоретические: гистограмма превращается в график плотности распределения.

Накопленная частота интервалов – это число, полученное последовательным суммированием частот в направлении от первого интервала к последнему, до того интервала включительно, для которого определяется накопленная частота.

В Excel для построения выборочных функций распределения используются специальная функция **ЧАСТОТА** и процедура **Гистограмма** из пакета анализа**.**

Функция **ЧАСТОТА** ***(массив_данных, двоичный_массив)*** вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив цифр, где

• *массив_данных* — это массив или ссылка на множество данных, для которых вычисляются частоты;

• *двоичный_массив* — это массив интервалов, по которым группируются значения выборки.

Процедура **Гистограмма** из **Пакета анализа** выводит результаты выборочного распределения в виде таблицы и графика. Параметры диалогового окна **Гистограмма:**

• **Входной диапазон** - диапазон исследуемых данных (выборка);

• **Интервал карманов** - диапазон ячеек или набор граничных значений, определяющих выбранные интервалы (карманы). Эти значения должны быть введены в возрастающем порядке. Если диапазон карманов не был введен, то набор интервалов, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

• **выходной диапазон** предназначен для ввода ссылки на левую верхнюю ячейку выходного диапазона.

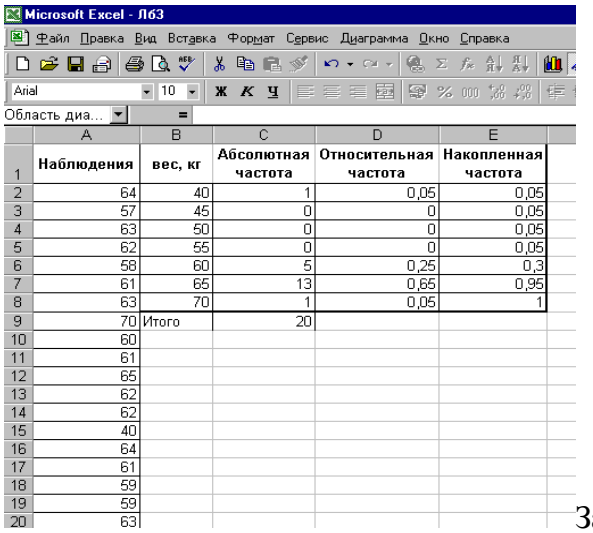
• переключатель **Интегральный процент** позволяет установить режим включения в гистограмму графика интегральных процентов.

• переключатель **Вывод графика** позволяет установить режим автоматического создания встроенной диаграммы на листе, содержащем выходной диапазон.

Пример 1. Построить эмпирическое распределение веса студентов в килограммах для следующей выборки: 64, 57, 63, 62, 58, 61, 63, 70, 60, 61, 65, 62, 62, 40, 64, 61, 59, 59, 63, 61.

Решение

1. В ячейку A1 введите слово ***Наблюдения***, а в диапазон A2:A21 — значения веса студентов (см. рис. 1).
2. В ячейку B1 введите названия интервалов *Вес, кг*. В диапазон B2:B8 введите граничные значения интервалов *(40, 45, 50, 55, 60, 65, 70).*
3. Введите заголовки создаваемой таблицы: в ячейки C1 — ***Абсолютные частоты***, в ячейки D1 — ***Относительные частоты***, в ячейки E1 — ***Накопленные частоты***.(см. рис. 1).
4. С помощью функции **Частота** заполните столбец абсолютных частот, для этого выделите блок ячеек C2:C8. С панели инструментов **Стандартная** вызовите **Мастер функций** (кнопка fx). В появившемся диалоговом окне выберите категорию **Статистические** и функцию **ЧАСТОТА**, после чего нажмите кнопку **ОК**. Указателем мыши в рабочее поле **Массив_данных** введите диапазон данных наблюдений (A2:A8). В рабочее поле **Двоичный_массив** мышью введите диапазон интервалов (B2:B8). Слева на клавиатуре последовательно нажмите комбинацию клавиш **Ctrl+Shift+Enter.** В столбце C должен появиться массив абсолютных частот (см. рис.1).
5. В ячейке C9 найдите общее количество наблюдений. Активизируйте ячейку C9, на панели инструментов **Стандартная** нажмите кнопку **Автосумма**. Убедитесь, что диапазон суммирования указан правильно и нажмите клавишу **Enter**.



	A	B	C	D	E
1	Наблюдения	вес, кг	Абсолютная частота	Относительная частота	Накопленная частота
2	64	40	1	0,05	0,05
3	57	45	0	0	0,05
4	63	50	0	0	0,05
5	62	55	0	0	0,05
6	58	60	5	0,25	0,3
7	61	65	13	0,65	0,95
8	63	70	1	0,05	1
9	70	Итого	20		
10	60				
11	61				
12	65				
13	62				
14	62				
15	40				
16	64				
17	61				
18	59				
19	59				
20	63				

6. Заполните столбец **относительных** частот. В ячейку введите формулу для вычисления относительной частоты: ***=C2/C9***. Нажмите клавишу **Enter**. Протягиванием (за правый нижний угол при нажатой левой кнопке мыши) скопируйте введенную формулу в диапазон и получите массив относительных частот.
7. Заполните столбец **накопленных** частот. В ячейку D2 скопируйте значение относительной частоты из ячейки E2. В ячейку D3 введите формулу: ***=E2+D3***. Нажмите клавишу **Enter**. Протягиванием (за правый нижний угол при нажатой левой кнопке мыши) скопируйте введенную формулу в диапазон D3:D8. Получим массив накопленных частот.

Рис. 1. Результат вычислений из примера 1

8. Постройте диаграмму относительных и накопленных частот. Щелчком указателя мыши по кнопке на панели инструментов вызовите **Мастер диаграмм**. В появившемся диалоговом окне выберите закладку **Нестандартные** и тип диаграммы **График/гистограмма.** После редактирования диаграмма будет иметь такой вид, как на рис. 2.

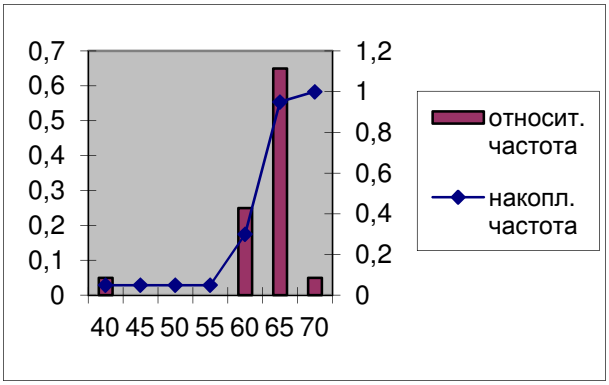


Рис. 2 Диаграмма относительных и накопленных частот из примера 1

Задания для самостоятельной работы

1. Для данных из примера 1 построить выборочные функции распределения, воспользовавшись процедурой **Гистограмма** из пакета **Анализа**.

2. Построить выборочные функции распределения (относительные и накопленные частоты) для роста в см. 20 студентов: 181, 169, 178, 178, 171, 179, 172, 181, 179, 168, 174, 167, 169, 171, 179, 181, 181, 183, 172, 176.

3. Найдите распределение по абсолютным частотам для следующих результатов тестирования в баллах: 79, 85, 78, 85, 83, 81, 95, 88, 97, 85 (используйте границы интервалов 70, 80, 90).

4. Рассмотрим любой из критериев оценки качеств педагога-профессионала, например, «успешное решение задач обучения и воспитания». Ответ на этот вопрос анкеты типа «да», «нет» достаточно груб. Чтобы уменьшить относительную ошибку такого измерения, необходимо увеличить число возможных ответов на конкретный критериальный вопрос. В табл. 1 представлены возможные варианты ответов.

Обозначим этот параметр через x. Тогда в процессе ответа на вопрос величина x примет дискретное значение x, принадлежащее определенному интервалу значений. Поставим в соответствие каждому из ответов определенное числовое значение параметра x (см. табл. 1).

Табл. 1 Критериальный вопрос: успешное решение задач обучения и воспитания

п/п	Варианты ответов	
	Абсолютно неуспешно	,1
	Неуспешно	,2
	Успешно в очень малой степени	,3
	В определенной степени успешно, но еще много недостатков	,4
	В среднем успешно, но недостатки имеются	,5
	Успешно с некоторыми оговорками	,6
	Успешно, но хотелось бы улучшить результат	,7
	Достаточно успешно	,8
	Очень успешно	,9
0	Абсолютно успешно	

При проведении анкетирования в каждой отдельной анкете параметр x принимает случайное значение, но только в пределах числового интервала от 0,1 до 1.

Тогда в результате измерений мы получаем неранжированный ряд случайных значений (см. табл. 2).

Таблица 2. Результаты опроса ста учителей

,6	,7		,6	,2	,8	,3	,5	,9	,3
,5	,1	,4	,5	,5	,4	,4	,6	,5	,4
,6	,9	,7	,9	,8	,5	,5	,6	,8	,4
,4	,4	,8	,7	,6	,6	,7	,8	,5	,6
,7	,6	,7	,3	,2	,7	,5	,3	,4	,5
,9	,7	,6	,5	,7	,6	,2	,8	,8	,3
,7	,5	,7	,6	,2	,5	,8	,3	,7	,8
,7	,6	,6	,8	,4	,6	,6	,6	,9	,7
,7	,5	,7	,6	,9	,4	,8	,7	,5	,8
,8	,9	,4	,3	,4	,6	,4	,5	,3	,5

Сгруппируйте полученную выборку, рассчитайте среднее значение выборки, стандартное отклонение, абсолютную и относительную частоту появления параметра, а также постройте график плотности вероятности f(x)= $\frac{W(x)}{\sigma\sqrt{2\pi}}$, где

W(x) – относительная частота наступления события;

σ - стандартное отклонение;

π =3,14.

Постройте график функции f(x) и сравните его с нормальным распределением Гаусса.

2.4 Лабораторная работа №4 (4 часа).

Тема: «Описательная статистика. Расчет основных выборочных характеристик**»**

Описание (ход) работы:

В процессе обучения постоянно ощущается потребность в хорошо разработанных методах измерения уровня обученности в самых различных областях знаний. Известно, что профессиональное тестирование было начато еще в 2200 году до нашей эры, когда служащие Китайского императора тестировались, чтобы определить их пригодность для императорской службы. По некоторым оценкам в 1986 году по крайней мере 800 профессий лицензировались в Соединенных Штатах на основании тестирования **(А.А. Захаров, А.В. Колпаков Современные математические методы объективных педагогических измерений)**

Почти каждый педагог разрабатывает тестовые задания по своей дисциплине, но не каждый может грамотно обработать и интерпретировать результаты теста. Напротив, грамотное конструирование теста на основе знания теории тестирования позволит педагогу-исследователю создать инструмент, позволяющий провести объективное измерение знаний, умений и навыков по данному курсу с необходимой точностью.

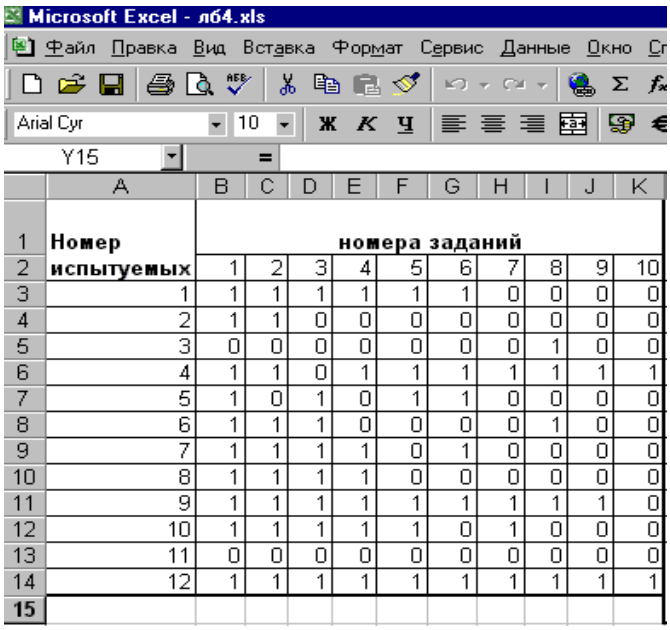
В настоящее время существуют два теоретических подхода к созданию тестов: классическая теория и современная теория IRT (Item Response Theory). Оба подхода базируются на последующей статистической обработке так называемого сырого балла (raw score), то есть балла, набранного в результате тестирования. Только после проведения многократных статистических обработок можно говорить о создании теста с устойчивыми параметрами качества (надежностью и валидностью).

Для обработки данных, полученных на этапе тестирования, воспользуемся пакетом MS Office 2000 и электронными таблицами MS Excel.

После сбора эмпирических данных необходимо провести статистическую обработку, которую будем проводить на ЭВМ. Этап математико–статистической обработки разобьем на ряд шагов.

Шаг 1. Формирование матрицы тестовых результатов.

Результаты ответов учеников на задания тестов оцениваются в дихотомической шкале: за каждый правильный ответ учащийся получает один балл, а за неправильный ответ или за пропуск задания – нуль баллов (см. рис. 1).



	A	B	C	D	E	F	G	H	I	J	K
1	Номер	номера заданий									
2	испытуемых	1	2	3	4	5	6	7	8	9	10
3	1	1	1	1	1	1	1	0	0	0	0
4	2	1	1	0	0	0	0	0	0	0	0
5	3	0	0	0	0	0	0	0	1	0	0
6	4	1	1	0	1	1	1	1	1	1	1
7	5	1	0	1	0	1	1	0	0	0	0
8	6	1	1	1	0	0	0	0	1	0	0
9	7	1	1	1	1	0	1	0	0	0	0
10	8	1	1	1	1	0	0	0	0	0	0
11	9	1	1	1	1	1	1	1	1	1	0
12	10	1	1	1	1	1	0	1	0	0	0
13	11	0	0	0	0	0	0	0	0	0	0
14	12	1	1	1	1	1	1	1	1	1	1
15											

Рис. 1. Матрица результатов тестирования

Шаг 2. Преобразование матрицы тестовых результатов.

На втором шаге из матрицы тестовых результатов устраняются строки и столбцы, состоящие только из нулей или только из единиц. В приведенном выше примере таких столбцов нет, а строк только две. Одна из них, нулевая строка соответствует ответам одиннадцатого испытуемого, который не смог выполнить правильно ни одного задания в тесте.

В этом случае вывод довольно однозначен: тест непригоден для оценки знаний такого ученика. Для выявления его уровня знаний тест необходимо облегчить, добавив несколько более легких заданий, которые, скорее всего, выполнит правильно большинство остальных испытуемых группы.

Столь же непригоден, но уже по другой причине, тест для оценки знаний двенадцатого ученика, который выполнил правильно все без исключения задания теста. Причина непригодности теста заключается в его излишней легкости, не позволяющий выявить истинный уровень подготовки двенадцатого ученика. Возможно, двенадцатый ученик знает много чего другого и в состоянии выполнить по контролируемым разделам содержания гораздо более трудные задания, которые просто не были включены в тест.

Таким образом, на данном шаге необходимо удалить из матрицы данных 11 и 12 строки.

Шаг 3. Подсчет индивидуальных баллов испытуемых и количество правильных ответов на каждое задание теста.

Индивидуальный балл испытуемого получается суммированием всех единиц, полученных им за правильное выполнение задания теста. В Excel для суммирования данных по строке можно воспользоваться кнопкой **Автосумма** Σ на панели инструментов **Стандартная.** Для удобства полученные индивидуальные баллы (X_i) приводятся в последнем столбце матрицы результатов (см. рис. 2).

Число правильных ответов на задания теста (Y_i) также получается суммированием единиц, но уже расположенным по столбцам.(см. рис. 2)

Шаг 4. Упорядочение матрицы результатов.

Значения индивидуальных баллов необходимо отсортировать по возрастанию, для этого в MS Excel:

1. выделим блок ячеек, содержащих номера испытуемых, матрицу результатов и индивидуальные баллы. Начинать выделение необходимо со столбца X (индивидуальные баллы).

M	N	O	P	Q	R	S	T	U	V	W	X
Номер	номера заданий										Индивиду-альные
испытуемых	1	2	3	4	5	6	7	8	9	10	баллы (X)
1	1	1	1	1	1	1	0	0	0	0	6
2	1	1	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	1	0	0	1
4	1	1	0	1	1	1	1	1	1	1	9
5	1	0	1	0	1	1	0	0	0	0	4
6	1	1	1	0	0	0	0	1	0	0	4
7	1	1	1	1	0	1	0	0	0	0	5
8	1	1	1	1	0	0	0	0	0	0	4
9	1	1	1	1	1	1	1	1	1	0	9
10	1	1	1	1	1	0	1	0	0	0	6
число правильных ответов (Y)	9	8	7	6	5	5	3	4	2	1	50

2. на панели инструментов **Стандартная** нажимаем на кнопку **Сортировка по возрастанию**  . Матрица результатов примет вид, изображенный на рис. 3.

	M	N	O	P	Q	R	S	T	U	V	W	X
	Номер	номера заданий										Индивиду-альные
	испытуемых	1	2	3	4	5	6	7	8	9	10	баллы (X)
	3	0	0	0	0	0	0	0	1	0	0	1
	2	1	1	0	0	0	0	0	0	0	0	2
	5	1	0	1	0	1	1	0	0	0	0	4
	6	1	1	1	0	0	0	0	1	0	0	4
	8	1	1	1	1	0	0	0	0	0	0	4
	7	1	1	1	1	0	1	0	0	0	0	5
	1	1	1	1	1	1	1	0	0	0	0	6
	10	1	1	1	1	1	0	1	0	0	0	6
	4	1	1	0	1	1	1	1	1	1	1	9
	9	1	1	1	1	1	1	1	1	1	0	9
	число	9	8	7	6	5	5	3	4	2	1	50
	правильных											
	ответов (Y)											

Рис. 2. Матрица с подсчетом итоговых сумм

	A	B	C
1	Номер	Балл	Ранг
2	3	1	1
3	2	2	2
4	5	4	3
5	6	4	3
6	8	4	3
7	7	5	6
8	1	6	7
9	10	6	7
10	4	9	9
11	9	9	9
12			

Рис. 3 Упорядоченная матрица результатов

Шаг 5. Графическое представление данных.

Эмпирические результаты тестирования можно представить в виде полигона частот, гистограммы, сглаженной кривой или графика.

Для построения кривых упорядочим результаты эксперимента и подсчитаем частоту получения баллов (см. рис. 4-6).

	A	B
	Номер	Балл
2	1	6
3	2	2
4	3	1
5	4	9
6	5	4
7	6	4
8	7	5
9	8	4
10	9	9
11	10	6

Рис. 4. Несгруппированный ряд

Балл	Частота
1	1
2	1
4	3
5	1
6	2
9	2

Рис. 6. Частотное распределение

Рис. 5. Ранжированный ряд

Для расчета рейтинга (ранга) каждого учащегося по индивидуальным балам необходимо применить функцию **РАНГ,** которая возвращает ранг числа в списке чисел. Ранг числа – это его величина относительно других значений в списке.

В MS Excel 2000 для вычисления ранга используется функция

РАНГ (число; ссылка; порядок), где

Число – адрес на ячейку, для которой определяется ранг.

Ссылка - ссылка на массив индивидуальных баллов (выборка).

Порядок – число, определяющее способ упорядочения. Если порядок равен 0 (нулю), или опущен, то Excel определяет ранг числа так, как если бы ссылка была списком, отсортированным в порядке убывания. Если порядок – любое ненулевое число, то Excel определяет ранг числа так, как если бы ссылка была списком, отсортированным в порядке возрастания.

Примечание. Функция РАНГ присваивает повторяющимся числам одинаковый ранг. При этом наличие повторяющихся чисел влияет на ранг последующих чисел. Например, если в списке целых чисел дважды встречается число 10, имеющее ранг 5, число 11 будет иметь ранг 7 (ни одно из чисел не будет иметь ранг 6).

По частотному распределению можно построить гистограмму (см. рис.7).

Гистограмму можно построить и по индивидуальным баллам (см. рис. 8).

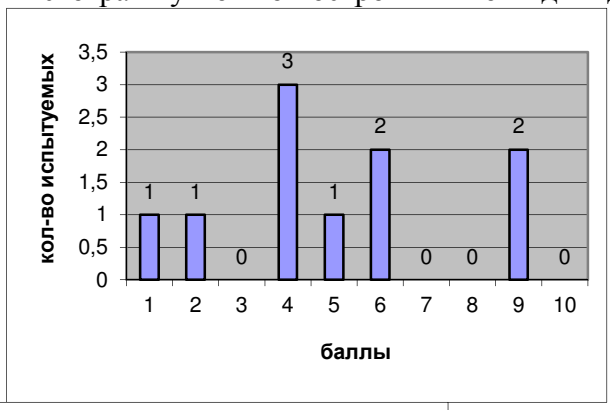


Рис. 7. Столбиковая гистограмма

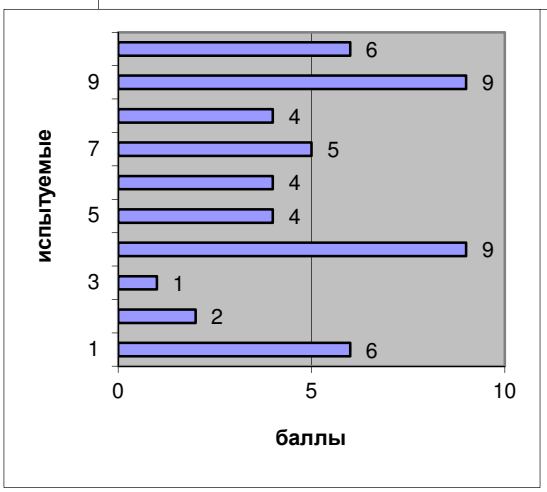


Рис. 8. Гистограмма распределения инд. баллов

При разработке тестов необходимо помнить о том, что кривая распределения индивидуальных баллов, получаемых по репрезентативной выборке, является следствием кривой распределения трудности заданий теста. Этот факт удачно иллюстрируется на рис.9.

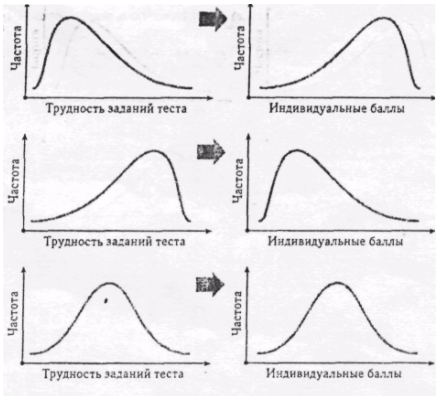


Рис. 9. Связь распределения индивидуальных баллов и трудности заданий теста

Для первого распределения слева характерно явное смещение в тесте в сторону легких заданий, что, несомненно, приведет к появлению большого числа завышенных баллов у репрезентативной выборки учеников. Большая часть учеников выполнит почти все задания теста.

Второй случай (слева) отражает существенное смещение в сторону трудных заданий при разработке теста, что не может не сказаться на снижении результатов учеников, поэтому распределение индивидуальных баллов имеет явно выраженный всплеск вблизи начала горизонтальной оси. Основная часть учеников выполнит незначительное число наиболее легких заданий теста.

В третьем случае задания теста обладают оптимальной трудностью, поскольку распределение имеет вид нормальной кривой. Отсюда автоматически возникает нормальность распределения индивидуальных баллов репрезентативной выборки учеников, что в свою очередь позволяет считать полученное распределение устойчивым по отношению к генеральной совокупности.

В профессионально разработанных нормативно-ориентированных тестах типичным является результат, когда приблизительно 70% учеников выполняют правильно от 30 до 70% заданий теста. а наиболее часто встречается результат в 50%.

Шаг 6. Определение выборочных характеристик результатов.

На данном этапе необходимо вычислить среднее значение, моду, медиану, дисперсию, стандартное отклонение выборки, ассиметрию и эксцесс (см. рис.10).

Степень отклонения распределения наблюдаемых частот выборки от симметричного распределения, характерного для нормальной кривой, оценивается с помощью асимметрии. Наличие асимметрии легко установить визуально, анализируя полигон частот или гистограмму Более тщательный анализ можно провести с помощью обобщенных статистических характеристик, предназначенных для оценки величины асимметрии в распределении.

Функция **СКОС MS Excel** возвращает ассиметрию распределения.

СКОС (число 1; число 2), где *число1* – ссылка на массив данных, содержащих индивидуальные баллы учеников.

При интерпретации полученного значения асимметриии 0,277 необходимо обратить внимание на то, что величина ассиметрии получилась положительной и небольшой (см. рис. 10, 11).

	A	L
1	**Номер**	**Индивиду-**
2	**испытуемых**	**альные баллы**
3	1	6
4	2	2
5	3	1
6	4	9
7	5	4
8	6	4
9	7	5
10	8	4
11	9	9
12	10	6
13	**среднее**	**5**
14	**мода**	**4**
15	**медиана**	**4.5**
16	**дисперсия**	**6.889**
17	**ст. отклонение**	**2.625**
18	**ассиметрия**	**0.277**
19	**эксцесс**	**-0.4117**

Рис. 10. Описательные характеристики выборки

Рис. 11. Кривые распределения с отрицательной, нулевой и положительной ассиметрией (слева направо) соответственно.

Асимметрия распределения положительна, если основная часть значений индивидуальных баллов лежит справа от среднего значения, что обычно характерно для излишне легких тестов.

Асимметрия распределения баллов отрицательна, если большинство учеников получили оценки ниже среднего балла. Эффект отрицательной асимметрии встречается в излишне трудных тестах, не сбалансированных правильно по трудности при отборе заданий

В хорошо сбалансированном по трудности тесте, как уже отмечалось ранее, распределение баллов имеет вид нормальной кривой. Для нормального распределения характерна нулевая асимметрия, что вполне естественно, так как при полной симметрии каждое значение балла, меньшее среднего значения, уравновешивается другим симметричным, большим чем среднее.

С помощью эксцесса можно получить представление о том, является ли функция распределения частот островершинной, средневершинной или плоской.

Для расчета данного параметра применим функцию **ЭКСЦЕСС** (число1; число2; …), где *число1* – ссылка на массив данных, содержащих индивидуальные баллы учеников.

В том случае, когда распределение данных бимодально (имеет две моды), необходимо говорить об эксцессе в окрестности каждой моды. Бимодальная конфигурация указывает на то, что по результатам выполнения теста выборка учеников разделилась на две группы. Одна группа справилась с большинством легких, а другая с большинством трудных заданий теста.

2.5 Лабораторная работа №5 (6 часов).

Тема: «Проверка критериев Стьюдента и Фишера**»**

2.5.1 Цель работы: ознакомиться с основными понятиями статистической проверки гипотез; рассмотреть использование критериев Стьюдента, Фишера, Вилкоксона-Манна-Уитни.

2.5.2Описание (ход) работы:

Почему статистические методы используют в медицине?

Прямое суждение об эффективности того или иного метода лечения ненадёжно из-за многих причин: биологической изменчивости, субъективности оценок, психотерапевтического эффекта и других причин. С примерами ненадёжной, ненаучной медицинской информации и пациенты, и врачи стакиваются, например, в рекламных роликах о высокой эффективности новых лекарств или новых методик лечения. И если пациентам простительно идти на поводу у рекламы, то врачи должны критически подходить к такой информации: необходимо знать, на каком количестве пациентов, в ходе какого типа исследования были получены результаты и т. д. Для современного врача навыки критической оценки столь же важны и необходимы, как, например, умение аускультировать больного.

Развитие идей критической оценки медицинской информации привело к возникновению в 80-х годах прошлого века концепции доказательной медицины (ДМ). Основные положения ДМ: 1) каждое решение врача должно основываться на научных данных; 2) вес каждого факта тем больше, чем строже методика исследования, в ходе которого этот факт получен. “Золотым стандартом” считаются рандомизированные (т. е. полученные в результате случайного отбора) контролируемые исследования. Индивидуальный врачебный опыт и мнение экспертов или “авторитетов” рассматриваются как не имеющие достаточной научной основы.

Одним из важнейших компонентов ДМ является использование научно-обоснованных статистических методов, одним из которых является проверка статистических гипотез.

Понятие статистической гипотезы

Статистическая гипотеза (Н) – это предположение о *виде* или о *параметрах* генеральной совокупности, которое проверяется на основе выборочных данных.

Например: 1) **Н:** вес новорождённых распределён по нормальному закону (гипотеза о виде распределения);

2) **Н:** средние значения артериального давления в двух группах пациентов равны, т.е. обе выборки извлечены из одной генеральной совокупности (гипотеза о параметрах распределения).

Не все научные гипотезы являются статистическими: так, гипотеза де Бройля о волновых свойствах электронов не является статистической, так как в ней не присутствует ни закон распределения, ни параметры.

К проверке статистических гипотез сводятся задачи проверки и оценки различных процессов: сравнение лечебных методик, характеристик препаратов и медицинской техники, эффективности лечения, продолжительности болезни, экономичности и т. п.

Нулевая и альтернативная гипотезы

Проверяемую гипотезу называют нулевой и обозначают H_0. Нулевая гипотеза всегда отвергает эффект вмешательства. Наряду с нулевой рассматривают и одну из альтернативных (конкурирующих) гипотез, которую обозначают H_1.

Например, пусть партию фармпрепарата контролируют по небольшой выборке и сравнивают с нормой; тогда нулевая гипотеза H_0: выпущенная партия фармпрепарата нестандартна (брак), а конкурирующая H_1: партия соответствует норме.

Задача проверки статистических гипотез

Задача проверки гипотез заключается в том, чтобы на основании анализа выборочных данных (неполная информация) принять решение о справедливости одной из гипотез.

Ошибки первого и второго рода

При проверке гипотез из-за наличия неполной информации могут быть допущены ошибки двух видов (см. таблицу):

1. *Принимается H_1, когда на самом деле верна H_o* - ***ошибка первого рода*** (ошибочное заключение о существовании различий, которых на самом деле нет - гипердиагностика).

2. *Принимается H_o, когда верна H_1* -***ошибка второго рода*** (не найти действительно существующих различий - гиподиагностика).

Вероятность совершить ошибку первого рода должна быть достаточно мала, так как нужны *веские основания* для признания, например, того, что один метод лечения лучше другого. Эта вероятность *p* называется *уровнем значимости α.*

Уровнем значимости α *называется вероятность отклонения нулевой гипотезы, когда она на самом деле верна.*

Чем серьёзнее последствия ошибки первого рода, тем меньше надо выбирать уровень значимости. В медицинских исследованиях обычно используют $\alpha = 0{,}05$ или $\alpha = 0{,}01$, а значение $\beta = 0{,}2$ или $0{,}1$. Желательно, чтобы *α и β* были как можно *меньше*, чего можно достичь, только *увеличивая объём выборки*. Этот вывод очень важен, так как напрямую связан с планированием эксперимента.

Результат, полученный при проверке	*Что есть на самом деле*	
	Верна гипотеза H_o	Верна гипотеза H_1
Принята гипотеза Но	Правильное решение с вероятностью 1-α	Неправильное решение с вероятностью β, **ошибка *второго* рода**

Принята гипотеза H_1	Неправильное решение с вероятностью α, **ошибка *первого* рода**	Правильное решение с вероятностью 1- β, **мощность**

Разумное соотношение между α и β находят, исходя из тяжести последствий (ущерба) каждой из ошибок. Пусть, например, проверяется гипотеза H_0 об отсутствии у пациента определённого заболевания, а признак заболевания, это, например, величина артериального давления (АД). Тогда H_0:АД в норме, т. е. пациент здоров, H_1: АД отличается от нормы, т. е. пациент болен. Тогда ошибка 1 рода – отклонение H_0, когда она верна – признание человека больным, когда он на самом деле здоров. Последствия этой ошибки – неудобства для пациента, который, например, должен пройти дополнительное обследование или лечение (хотя опасность «залечивания» тоже есть). Иная ситуация в случае ошибки 2 рода: признать человека здоровым, когда он на самом деле болен. Фактически происходит отказ от лечения больного, теряется время, и последствия ошибки 2 рода могут быть плачевными. Итак, в нашем примере с целью уменьшить вероятность ошибки 2 рода, возможно не рассматривать высокий уровень значимости α (то есть принять α равной, например, 0,05, а не 0, 01).

Возможны и противоположные ситуации. Значения α, меньшие, чем 0,01 используются, например, при статистическом выявлении токсичных медицинских препаратов, когда важнейшее значение приобретает гарантия от ошибочного отклонения проверяемой гипотезы.

Итак, при выборе гипотез *нулевой гипотезой* (по сравнению с альтернативной) должна быть та, *которую более опасно ошибочно отвергнуть*.

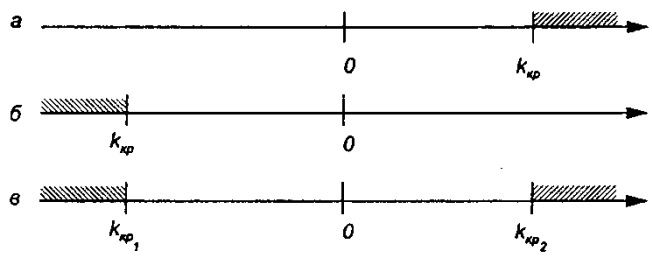
Статистический критерий. Критические области (хвосты)

Для проверки принятой гипотезы используют случайную величину К, являющуюся функцией от выборочных данных и называемую статистическим критерием.

Статистический критерий *– это правило (формула), позволяющее по данным выборки принять либо отвергнуть нулевую гипотезу.*

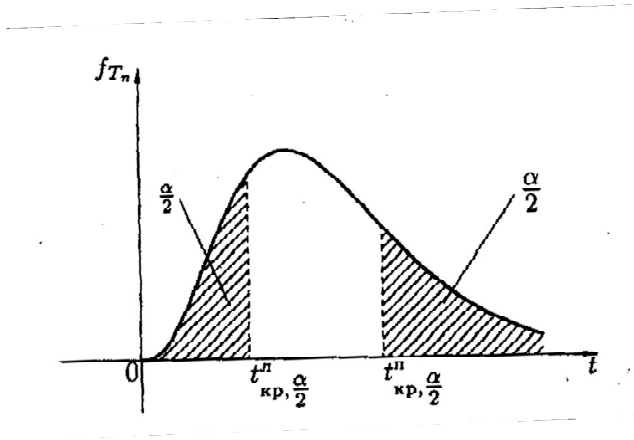
Статистический критерий, являясь случайной величиной, имеет какое-то вероятностное распределение, например, нормальное распределение, распределение Стьюдента, распределение Фишера, χ^2 и др.

Рис. 1



В зависимости от принятого уровня значимости α из всей области допустимых значений критерия К выделяют (см рис. 1) *критическую* область. Это делают с помощью числа $\kappa_{кр.}$, которое находят с помощью таблиц распределения каждого критерия К. Далее работает следующее правило (**основной принцип проверки статистических гипотез**): *если вычисленное по выборке значение* $\kappa_{набл.}$ *критерия К попадает в критическую область, то нулевая гипотеза* H_0 *отвергается в пользу альтернативной* H_1*, а если не попадает то* H_0 *принимается*. Критическая область в зависимости от $\kappa_{кр.}$ может быть *односторонней* (правосторонней или левосторонней) или *двухсторонней*. Для

Рис. 2



правосторонней (левосторонней) критической области значения К удовлетворяют условию: $P(K \geq к_{кр.})=\alpha$ и $P(K \leq к_{кр.})=\alpha$, где $P(\ldots)=\alpha$ - вероятность, того, что критерий К примет значение, большее (или соответственно меньшее) $\mathbf{к_{кр}}$, и. равная площади правого или левого «хвоста» на графике распределения вероятностей (рис. 2). Аналогично, для двухсторонней критической области $P(K \leq к_{кр})+ P(K \geq к_{кр.})=\alpha$, т. е. значение α – это площадь обоих «хвостов» на графике распределения вероятностей.

Одностороннюю критическую область надо использовать тогда, когда интересующий нас процесс должен идти только в одном направлении. Например, есть веские основания утверждать, что определённая диета обязательно снизит вес пациента. Но даже в этом случае необходимо подстраховаться, выбрав двухстороннюю критическую область. В нашем примере это означает, что у некоторых людей предложенная диета может привести к увеличению веса. Поэтому, как показывает практика, *в большинстве исследований применяется двухсторонний критерий*.

Выбор статистического критерия можно сравнить, например, с правилом, по которому рассчитывается проходной балл при поступлении в вуз. Тогда правило расчёта проходного балла – это сам критерий К, $\mathbf{к_{набл.}}$ – набранное количество баллов, а $\mathbf{к_{кр.}}$ – это проходной балл, преодолев который мы принимаем то или иное решение (гипотезу).

Критерии бывают параметрические и непараметрические. *Параметрические критерии* используются, если выборки взяты из генеральной совокупности, которая подчиняется известному, например, нормальному закону распределения. Нормальность распределения выборки должна быть статистически доказана до применения параметрических критериев.

Непараметрические критерии используют, если нет подчинения распределения выборки нормальному закону. Например, если объём выборки настолько мал, что невозможно оценить закон распределения данных в выборке. Параметрические критерии являются более мощными, чем непараметрические в обнаружении реального эффекта.

От исследователя, использующего статистическую проверку гипотез в прикладных задачах, требуется научиться пользоваться существующими критериями.

Рис. 3

Рис. 4

Процедура проверки гипотез

Проверка гипотез обычно проходит следующие этапы.

1. Набирается первичный статистический материал в виде выборок из одной или нескольких генеральных совокупностей.

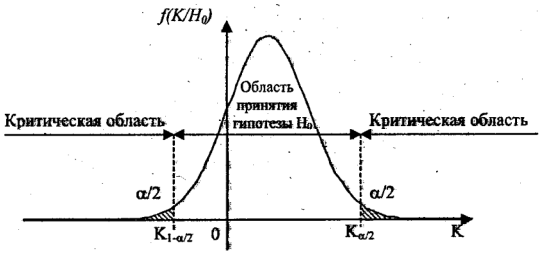
2. Исследователь формулирует основную (H_0) и альтернативную (H_1) гипотезы, а также выбирает уровень значимости α (0,01 или 0,05), соответствующие целям исследования.

3. Выбирается *критерий проверки К,* который подходит в данной ситуации и определяется, какой нужен критерий - односторонний или двухсторонний - и по соответствующим формулам вычисляем *значение статистического критерия* $\boldsymbol{к_{набл.}}$ для имеющихся данных (выборок).

4. По таблицам, соответствующим выбранному методу, находят границу *критической области* $\mathbf{к_{кр}}$ для принятого уровня значимости.

5. Принимается решение о справедливости гипотезы Но или H_1. *Если значение* $\boldsymbol{к_{набл.}}$ *критерия, вычисленного в п.3, принадлежит критической области (п. 4), то основная гипотеза Но отвергается и принимается альтернативная гипотеза* H_1*(различия между наблюдаемыми значениями и теоретическими значимы, т. е. обусловлены ошибочностью*

нулевой гипотезы).



…падают в критическую область, то гипотеза H_o …обусловлены случайными причинами)

Что значит $p < 0,05$?

В ходе проверки статистических гипотез кроме вычисления статистического критерия К в современных статистических пакетах **вычисляется также соответствующее значение *p*, где *p* - это вероятность справедливости H_0.**

Сравнивая полученное значение *p* с принятым уровнем значимости α, делают выводы о гипотезах:

если $p > \alpha$ (α - принятый уровень значимости, **обычно 0,05), то H_0 принимают (различия незначимы);**

если $p < \alpha$, то H_0 отклоняют (различия статистически значимы при $p<0,05$).

Использование круглых чисел 0,05; 0,01 и т. д. в качестве уровня значимости – это следствие ручного проведения статистических расчётов в докомпьютерное время. В настоящее время рекомендуется указывать точное значение *p* (с точностью до трёх знаков), что позволяет читателю самостоятельно оценить статистическую значимость результата, например, значения p=0,049 или p=0,051 следует интерпретировать практически одинаково.

Выборки зависимые и независимые

Примеры *независимых* выборок:

1) параметры двух групп пациентов, к которым применялись различные методики лечения;

2) параметры двух групп пациентов, к одной из которых (опытная группа) применялось воздействие методики, а к другой, контрольной, - нет.

Примеры *зависимых* (связанных или сопряжённых выборок):

1) параметры одной и той же группы пациентов до и после воздействия какого-либо фактора, например, методики лечения;

2) параметры различных частей одного и того же объекта, например, состояние двух конечностей, одна из которых подвергалась лечению, а другая– нет.

В

Перейдем к рассмотрению некоторых наиболее популярных статистических гипотез, используемых в медицинских исследованиях, и примеров их использования.

Проверка гипотез относительно средних
t-критерий Стьюдента

Подобная задача возникает при сравнении двух выборок, например, двух групп больных, подвергшихся определённому воздействию (например, проходящие лечение по различным методикам две группы пациентов, одна из которых принимает определённый лекарственный препарат, а другая, контрольная группа, принимает плацебо (лекарственная форма, содержащая нейтральные вещества – «пустышка»). При этом сравнение средних позволяет судить о степени воздействия, о значимости возможных эффектов или их отсутствии.

1. Постановка задачи. Из двух генеральных совокупностей *X* и *Y,* распределенных по нормальному закону (проверка на нормальность обеих выборок обязательно проводится предварительно), с *равными дисперсиями,* получены выборки, объемы которых n_X и n_Y соответственно. Требуется сравнить между собой *математические ожидания* соответствующих генеральных совокупностей (сравнить генеральные средние).

Рассмотрим последовательность действий при решении этой задачи в соответствии с установленным выше порядком.

2. Проверяемые гипотезы:

H_0 :$M(Y_г)=M(X_г)$ (генеральные средние *одинаковы);*

H_1 : $M(Y_г)\neq M(X_г)$, т. е. выбираем двухстороннюю критическую область.

3. Для проверки гипотез о равенстве генеральных средних при одинаковых генеральных дисперсиях применяется t-критерий Стьюдента, значение которого $t_{набл.}$ вычисляется по формуле:

$$t_{набл.} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_X - 1)\cdot S_X^2 + (n_Y - 1)\cdot S_Y^2}{n_X + n_Y - 2}}} \cdot \sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}}$$

где S_X^2 и S_Y^2 – выборочные дисперсии, $\bar{x}$ и $\bar{y}$ - средние значения выборок.

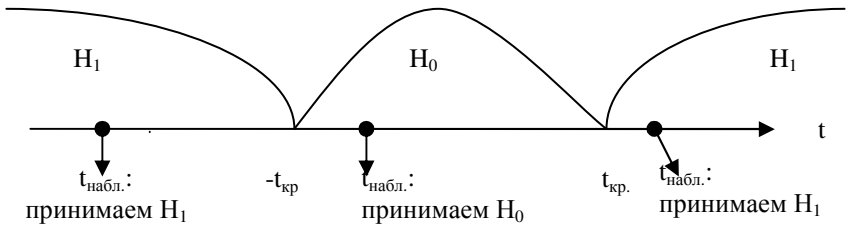
4. По таблице t-распределения находим для уровня значимости α=0,05 двухстороннюю критическую область. Для этого предварительно определяем число степеней свободы по формуле:

$f=n_X + n_Y - 2$

Критическая область для отклонения H_0 (рис. 6):

$|t_{набл}| > t_{кр.}$

Рис. 6



Примечание: Если при проверки гипотез о равенстве генеральных средних при одинаковых генеральных дисперсиях объём выборок n ≥ 30, то используют Z-критерий:

$$Z_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

Найденное значение $|Z_{набл}|$ следует сравнивать с критическим значением $Z_{кр}$ определяемым соотношением: $\Phi(Z_{кр}) = \frac{1-\alpha}{2}$

где *Ф(z)* - функция Лапласа.

Проверка гипотез для дисперсий. F-критерий Фишера

Во многих клинических исследованиях важной является проверка гипотезы о равенстве генеральных дисперсий двух *нормальных* выборок. Эта задача может быть решена с помощью *критерия Фишера.* Подобная задача сравнения дисперсий возникает в

случае сравнения точности измерений, точности приборов, сравнения методик. Поскольку дисперсия характеризует степень рассеяния значений относительно среднего, то наилучшей методикой будет та, у которой дисперсия меньше.

1. *Постановка задачи*. Для случайных величин X и У, распределенных по нормальному закону, получены выборки, объемы которых n_X и n_Y соответственно. Требуется сравнить между собой дисперсии соответствующих генеральных совокупностей.

2. Проверяемые гипотезы: $H_0 : D_г(У) = D_г(X)$ (генеральные дисперсии *одинаковы);*

Конкурирующая гипотеза H_1: $D_г(У) \neq D_г(X)$ (двухсторонняя критическая область).

3. Для проверки гипотез о равенстве дисперсий используем F-критерий Фишера. Вычисляем конкретные значения исправленных выборочных дисперсий S_X^2 и S_Y^2 и находим отношение (наблюдаемое значение критерия):

$$F_{набл.}=\frac{S_Б^2}{S_м^2}$$

где $S_Б^2$ и S_M^2 - большее и меньшее из чисел S_X^2 и S_Y^2.

4. Далее по таблице F- распределения по заданному уровню значимости α и числам степеней свободы $к_1=n_Б - 1$ и $к_2=n_M - 1$ находим критическую точку $к_{кр.}=F_{кр.}(\frac{\alpha}{2}; к_1; к_2)$. Доказано, что в этом случае двухстороннюю критическую область заменить правосторонней, то есть, если $F_{набл.} < к_{кр.}$, то гипотеза H_0 принимается, если $F_{набл.} > к_{кр}$, то различие дисперсий значимо и H_0 отвергается (рис. 7).

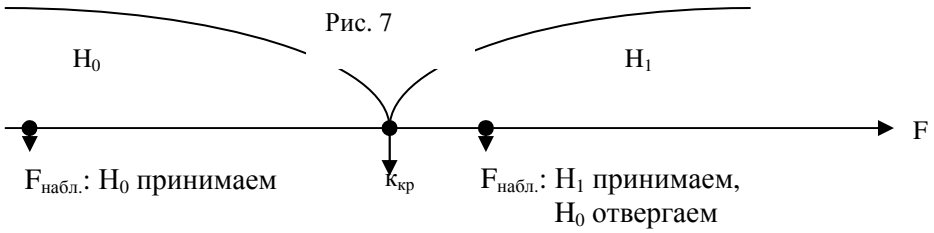


Рис. 7

Критерий Вилкоксона-Манна-Уитни (U-критерий)

Данный критерий является непараметрическим аналогом t-критерия Стьюдента и используется для проверки гипотезы о принадлежности двух *независимых* выборок к одной и той же генеральной совокупности. Здесь нет необходимости, чтобы выборки имели нормальное распределение. Непараметрические критерии, основанные на рангах, используют числа 1,2,3…, описывающие их положение в упорядоченном наборе данных.

1. *Постановка задачи*. Для случайных величин *X* и *Y* с неизвестными законами распределения получены выборки, объемы которых n_X и n_Y, соответственно. Значения элементов представлены в *порядковой шкале.* Требуется проверить гипотезу о принадлежности сравниваемых независимых выборок к одной и той же генеральной совокупности. α=0,05 или 0,01.

2. Проверяемые гипотезы: H_0: M(X)=M(Y). H_1: M(X)≠M(Y).

3. Для расчёта U-критерия необходимо:

- расположить числовые значения выборок в один общий ряд;

- пронумеровать члены общего ряда от 1 до $N=n_1+n_2$, где n_1 и n_2 – объёмы первой и второй выборок. Эти номера и будут рангами членов ряда. Если встречаются одинаковые значения элементов выборки, то им присваиваются одинаковые ранги, равные среднему арифметическому значению рангов одинаковых элементов;

- для каждой выборки найти сумму рангов R_1 и R_2;

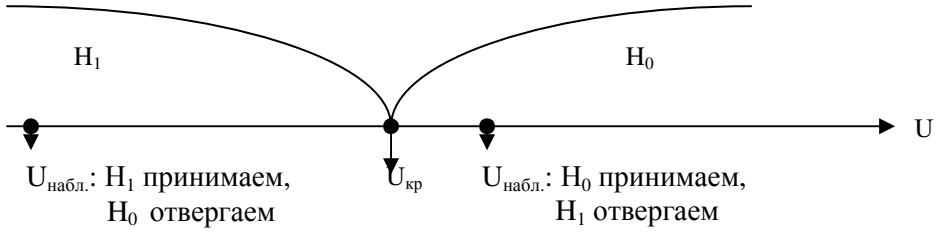
- найти величины U_1 и U_2 (наблюдаемые значения критерия):

$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ и $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$,

и выбрать $U_{набл.}$ как меньшее из U_1 и U_2;

4. По таблице критических значений U-критерия при заданном уровне значимости найти $U_{кр.}$ Если $U_{набл.} > U_{кр.}$, то H_0 принимается (различия статистически незначимы) – рис. 8.

Рис. 8



Примеры использования статистических критериев

t-критерий Стьюдента. F-критерий Фишера

Пример 1. Ш. Хейл и соавторы измеряли диаметр коронарных артерий после приёма нифедипина (препарат, расширяющий сосуды) и после приёма плацебо, и получили две выборки данных диаметра коронарных артерий в мм.

Плацебо: 2,5; 2,2;2,6;2,0;2,1;1,8;2,4;2,3;2,7;2,7;1,9;

Нифедипин:2,5;1,7;1,5;2,5;1,4;1,9;2,3;2,0;2,6;2,3;2,2.

Позволяют ли указанные данные утверждать, что нифедипин влияет на диаметр коронарных артерий?

Другими словами, необходимо проверить, значимо или нет различаются средние, представленные двумя выборками.

Пусть X - генеральная совокупность, из которой извлечена первая выборка (плацебо), Y – вторая (нифедипин). Авторы полагали, что обе генеральные совокупности имеют нормальное распределение (и эту гипотезу надо проверять статистическими методами).

Для корректного использования t-критерия Стьюдента, необходимо вначале проверить равенство дисперсий двух выборок. Сделаем это, воспользовавшись F-критерием Фишера.

1) Выдвигаем гипотезы: H_0: D(X)=D(Y); H_1: D(X)≠D(Y); выбираем уровень значимости α=0,05. Критическая область двухсторонняя.

2) Вычисляем исправленные выборочные дисперсии S_X^2 и S_Y^2 для обеих выборок:

$$\bar{x} = \frac{1}{11}\sum_{i=1}^{11} x_i = \frac{1}{11}(2{,}5 + 2{,}2 + \ldots + 1{,}9) \approx 2{,}29; \qquad \bar{y} = \frac{1}{11}\sum_{j=1}^{11} y_j \approx 2{,}08;$$

$$S_X^2 = \frac{1}{11-1}\sum_{i=1}^{11}(x_i - \bar{x})^2 = \frac{1}{10}((2{,}5-2{,}29)^2 + (2{,}2-2{,}29)^2 + \ldots + (1{,}9-2{,}29)^2) = 0{,}1009;$$

$$S_Y^2 = \frac{1}{11-1}\sum_{j=1}^{11}(y_j - \bar{y})^2 = 0{,}1716$$

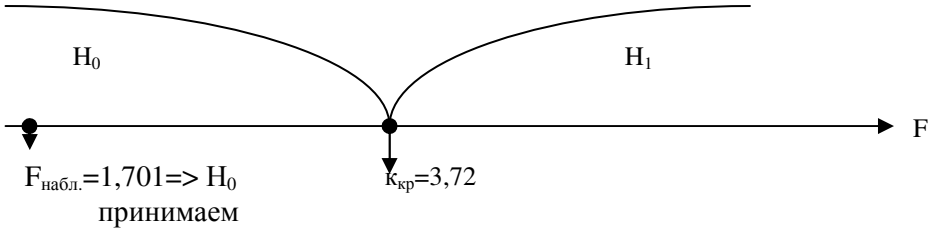
.

Вычисляем $F_{набл}$. Поскольку $S_Y^2 > S_X^2$, то $F_{набл} = \frac{S_Y^2}{S_X^2} \approx 1{,}701$.

3) По таблицам приложения находим $к_{кр.}$:

$к_{кр.} = F_{кр.}(\frac{\alpha}{2}; к_1; к_2) = F_{кр.}(0{,}025; 11\text{-}1; 11\text{-}1) = 3{,}72$.

Рис. 9



Поскольку $F_{набл.} < к_{кр}$, т. е. 1,701<3,72 и $F_{набл.}$ принадлежит области принятия гипотезы (рис. 9), то принимаем нулевую гипотезу о равенстве дисперсий. Это позволяет сравнить средние с помощью критерия Стьюдента.

1). Выдвигаем гипотезы. Нулевая гипотеза H_0: M(Y)=M(X) и альтернативная $\mathbf{H_1}$: M(Y)≠ M(X), т. е. критическая область двухсторонняя. Уровень значимости α=0,05.

2). Вычисляем наблюдаемое значение $t_{набл.}$ критерия:

$$t_{набл.} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_X - 1)\cdot S_X^2 + (n_Y - 1)\cdot S_Y^2}{n_X + n_Y - 2}}} \cdot \sqrt{\frac{n_X \cdot n_Y}{n_X + n_Y}} = \frac{2{,}29 - 2{,}08}{\sqrt{\frac{(11-1)1{,}1009 + (11-1)0{,}1716}{11+11-2}}} \sqrt{\frac{11\cdot 11}{11+11}} \approx 1{,}334$$

3). Число степеней свободы $f = n_X + n_Y - 2 = 11+11-2 = 20$.

По таблице t-распределения находим *для α=0,05* критическое значение $t_{крит.}(\alpha; f) = t_{крит.}(0{,}05; 20) = 2{,}09$. Т. к. критическая область двухсторонняя, то это *интервал* (-∞;-2,09)+(2,09;+∞), а область принятия гипотезы – это интервал (-2,09;+2,09).

Сравнивая $t_{крит.}$ и $t_{набл.}$, видим, что $|t_{набл}| < t_{кр.}$, т. е. 1,334<2,09 и $t_{набл.}$=1.334 попадает в область (-2,09;+2,09) принятия гипотезы H_0 (рис. 10).

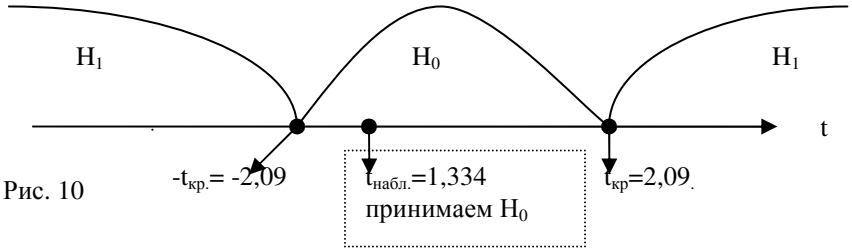


Рис. 10

Вывод: критерий Стьюдента не выявил существенных различий в диаметре коронарных артерий под влиянием нифедипина.

Критерий Вилкоксона-Манна-Уитни (U-критерий).

Пример 2. Пусть имеется 2 группы лабораторных мышей: опытная группа (n_1=9) и группа контрольная (n_2=11). Измерена масса мышей в граммах.

O											

пытная n_1 =9	4	8	0	2	5	6	9	0	3		
Контроль n_2 =11	0	0	2	6	8	9	0	1	3	8	0

Используя U-критерий, оценить значимость различия массы мышей при α=0,01.

1) Выдвигаем гипотезы. H_0: M(X)=M(Y). H_1: M(X)≠M(Y).

2) Располагаем числовые значения выборок в один общий ряд и присваиваем ранги в порядке возрастания с учётом повторяемости:

										0	1	2	3	4	5	6	7	8	9	0
анг	.5	.5				,5	,5		,5	,5	1	2	3	4	5	6	7	8,5	8,5	0
1				4		8			0			2		5	6		9		0	3
2	0	0	2		6		8	9		0	1		3			8		0		

Находим суммы рангов в каждой группе:

R_1= 4+6,5+9,5+12+14+15+17+18,5+20=116,5;

R_2= 1,5+1,5+3+5+6,5+8+9,5+11+13+16+18,5=93,5.

Рассчитываем $U_1 = 116{,}5 - \frac{9(9+1)}{2} = 71{,}5$; $U_2 = 93{,}5 - \frac{11(11+1)}{2} = 27{,}5$

В качестве $U_{набл.}$ выбираем минимальное значение 27,5.

3). В таблице находим $U_{крит.}(\alpha; n_1; n_2) = U_{крит.}(0{,}01; 9; 11) = 19$.

Сравниваем: 27,5>19, т. е. $U_{набл.} > U_{кр.}$, значит принимаем нулевую гипотезу H_0 (рис. 11).

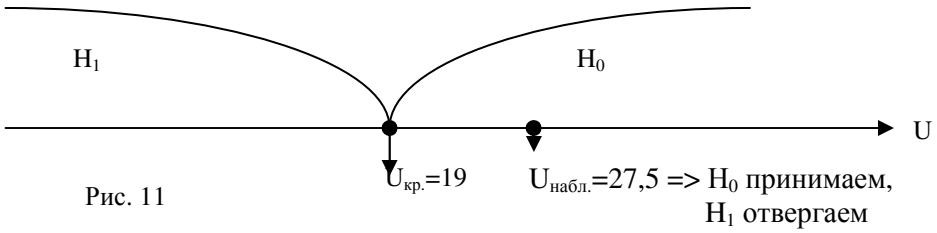


Рис. 11

Вывод: различие между средними значениями массы мышей статистически незначимо.

2.6 Лабораторная работа №6 (4 часа).

Тема: «Проверка критерия согласия Пирсона»

2.6.1 Цель работы: Научиться проверять основные статистические гипотезы: об однородности наблюдений и соответствии результатов измерения закону нормального распределения вероятностей.

2.6.2 Задачи работы:

Основываясь на статистических критериях проверить, не содержат ли результаты измерений x_1, x_2, …, x_n грубых погрешностей. Используя приближенный критерий и критерий согласия Пирсона проверить гипотезу о том, что распределение вероятностей рассматриваемой серии измерений подчиняется нормальному закону.

2.6.3 Описание (ход) работы:

Теоретическая часть

Предварительная обработка результатов измерений преследует в основном две цели: исключение грубых ошибок измерений и проверку гипотезы о соответствии результатов измерений закону нормального распределения.

1. Исключение грубых ошибок измерений.

Трудность обнаружения грубых ошибок обусловлена следующим обстоятельством. Если число измерений n мало, то доверительный интервал широк, и даже значительные отклонения от среднего $\bar{x}$ в него укладываются. Если же n велико, то возрастает вероятность того, что хотя бы одно измерение x_i сильно отклонится от среднего на «законных основаниях», т. е. случайно.

Методы исключения грубых погрешностей измерений для малых выборок изложены в материалах лекционного курса. Для больших выборок на практике используется следующий метод проверки однородности наблюдений.

Пусть произведено n независимых измерений и вычислены значения эмпирического среднего $\bar{x}$ и стандарта s. Сомнительный элемент выборки, резко отличающийся от других, будем обозначать через x_*. Это «крайний» элемент выборки, т. е. $x_* = x_{max}$ или $x_* = x_{min}$.

В основе рассматриваемого метода лежит тот факт, что критические значения максимального относительного отклонения

$$\tau = \frac{|x_* - \bar{x}|}{s} \qquad (1)$$

выражаются через квантили распределения Стьюдента с $n-2$ степенями свободы:

$$\tau_{1-\alpha,n} = \frac{t_{1-\alpha,n-2}\sqrt{n-1}}{\sqrt{n-2+t^2_{1-\alpha,n-2}}}. \qquad (2)$$

На практике обычно вычисляются два значения $\tau_{1-\alpha,n}$ при $\alpha = 0.05$ и $\alpha = 0.001$:

$\tau_1 = \tau_{1-0.05,n}$, $\qquad \tau_2 = \tau_{1-0.001,n}$

Этими значениями вся область изменения τ разбивается на три интервала: 1) $\tau \le \tau_1$; 2) $\tau_1 < \tau < \tau_2$; 3) $\tau_2 \le \tau$. Наблюдения, попавшие в первый интервал, не рекомендуется отбрасывать ни в коем случае. Наблюдения, попавшие во второй интервал можно исключить, если имеются какие-либо дополнительные соображения в пользу их ошибочности. Наконец, наблюдения, попавшие в третий интервал, всегда отбрасываются как грубо ошибочные.

2. Проверка гипотезы о нормальности распределения результатов измерения.

Приближенный метод проверки нормальности распределения основан на вычислении по результатам измерения эмпирических оценок коэффициентов асимметрии, эксцесса и их дисперсий:

$$\hat{A} = \frac{\hat{\mu}_3}{s^3} \approx \frac{1}{s^3(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^3, \qquad \hat{E} = \frac{\hat{\mu}_4}{s^4} \approx \frac{1}{s^4(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3,$$

$$D(A)=\frac{6(n-2)}{(n+1)(n+3)}, \qquad D(E)=\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

Если выборочные асимметрия и эксцесс удовлетворяют неравенствам

$$\left|\widehat{A}\right| \le 3\sqrt{D\left(\widehat{A}\right)} \qquad \left|\widehat{E}\right| \le 5\sqrt{D\left(\widehat{E}\right)}, \qquad (3)$$

то гипотеза о нормальности наблюдаемого распределения принимается, в противном случае гипотеза отклоняется.

Если выборка достаточно велика, применяются иные критерии согласия, наиболее надежным и универсальным из которых является критерий Пирсона χ^2. Применяя данный критерий необходимо выполнить следующие действия.

Область возможных значений случайной величины $(-\infty,+\infty)$ разбивается на конечное число ($m \approx 8 \div 20$) непересекающихся интервалов:

$(-\infty, x_2),(x_2, x_3),(x_3, x_4),\ldots,(x_m,+\infty)$

Для каждого интервала (x_{i-1}, x_i) подсчитывается число n_i элементов выборки, попавших в данный интервал.

Вычисляется теоретическая вероятность p_i попадания в i-й интервал при нормальном законе распределения вероятностей

$$p_i \equiv P(x_{i-1} < X < x_i) = \Phi\left(\frac{x_i-\overline{x}}{s}\right) - \Phi\left(\frac{x_{i-1}-\overline{x}}{s}\right),$$

где $\Phi(x)$ - функция Лапласа.

Проверяется выполнение условия $np_i \ge 5$ для всех интервалов; интервалы, для которых это условие не выполнено, объединяются с соседними интервалами.

Вычисляется сумма

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad (4)$$

имеющая приближенно χ^2 - распределение с $k-3$ степенями свободы.

При заданной доверительной вероятности $p = 1-\alpha$ (α - уровень значимости) и числе степеней свободы $k-3$ вычисляется (или находится по таблицам) *критическое значение* критерия $\chi^2_{p,k-3}$.

Если

$$\chi^2 < \chi^2_{p,k-3}, \qquad (5)$$

то гипотеза принимается, т. е. можно считать, что распределение вероятностей рассматриваемой серии измерений не отличается от нормального.

Необходимо помнить о вероятностном характере выводов, поэтому никакая, даже самая малая величина суммы (4) не может служить *доказательством* нормальности закона распределения.

Порядок выполнения задания

Исключение грубых ошибок измерений

1. Присвойте переменной ORIGIN значение равное единице.

2. Введите вектор выборочных значений (X:=READPRN(“путь к файлу Lab3 1a”)); используя встроенную функцию length(X) вычислите объем выборки.

3. Вычислите выборочные значения среднего, дисперсии и стандартного отклонения: $\overline{x}$, s^2 и s.

4. Изобразите элементы выборки и «трехсигмовый» интервал на графике. Определите грубо-визуально, есть ли среди элементов выборки резко отклоняющиеся значения.

5. Если есть подозрительные элементы, то для удобства дальнейших вычислений, произведите сортировку выборочных значений. Тогда подозрительные элементы будут находится в начале и (или) в конце вариационного ряда.

6. По формулам (1) и (2) вычислите значения τ, τ_1 и τ_2. Если значение τ попадает в третий интервал, исключите его из выборки; по оставшимся элементам выборки заново вычислите параметры $\overline{x}$, s^2, s и переходите к анализу следующего подозрительного элемента и т. д.

7. Сохраните рабочий документ в файле на диске.

Проверка гипотезы о нормальности распределения (1)

1. Присвойте переменной ORIGIN значение равное единице.

2. Введите вектор выборочных значений (Y:=READPRN("путь к файлу Lab3 1b")); используя встроенную функцию length(Y) вычислите объем выборки.

3. Вычислите оценки эмпирических коэффициентов асимметрии, эксцесса и их дисперсий.

4. Сравните вычисленные значения по формуле (3) и сделайте соответствующее заключение.

Проверка гипотезы о нормальности распределения (2)

1. Вычислите оценки эмпирического среднего, дисперсии и стандартного отклонения.

2. Вычислите максимальное и минимальное значения выборки.

3. Присвойте конкретное значение числу интервалов разбиения m и вычислите границы интервалов x_i $i = 1, 2, \ldots, m+1$; крайним границам присвойте значения $x_1 = -\infty$, $x_{m+1} = \infty$.

4. С помощью функции hist(x,X) вычислите частоты попадания выборочных значений в интервалы разбиения, а с помощью функции нормального распределения norm(x,a,s) – теоретические вероятности.

5. Проверьте выполнение условия $np_i \geq 5$ и объедините интервалы так, чтобы это условие было выполнено для всех интервалов.

6. Вычислите сумму (4).

7. Задайте определенный уровень значимости и вычислите критическое значение критерия $\chi^2_{p,m-3}$ - квантиль распределения «хи-квадрат» уровня p с $m-3$ степенями свободы.

8. На основе неравенства (5) сделайте вывод о принятии или отклонении гипотезы о нормальности распределения.

9. Сохраните рабочий документ в файле на диске.

2.7 Лабораторная работа №7 (12 часов).
Тема: «Элементы корреляционного анализа**»**

2.7.1 Цели работы: Познакомиться с методами корреляционного анализа.

2.7.2 Задачи работы:

1. Изучить метод наименьших квадратов для случая линейной зависимости между изучаемыми величинами.

2. Научиться определять наличие и тесноту линейной корреляции двух исследуемых величин; находить уравнение регрессии и находить с его помощью значение зависимой величины.

2.7.3 Перечень приборов, материалов, используемых в лабораторной работе:

1. Методические указания для обучающихся по освоению дисциплины
2. Тетрадь.
3. ПК.

2.7.4 Описание (ход) работы:

Понятие корреляционной зависимости

Переменные величины *X* и *Y* могут быть связаны функциональной и статистической зависимостью. При **функциональной –** "жёсткой" связи между изучаемыми величинами каждому значению *X* соответствует определённое значение *Y*, например, в законе Ома *I=U/R* при *R=const* каждому значению *U* соответствует одно значение *I*. Функциональные связи характерны для законов физики, химии и других естественных наук.

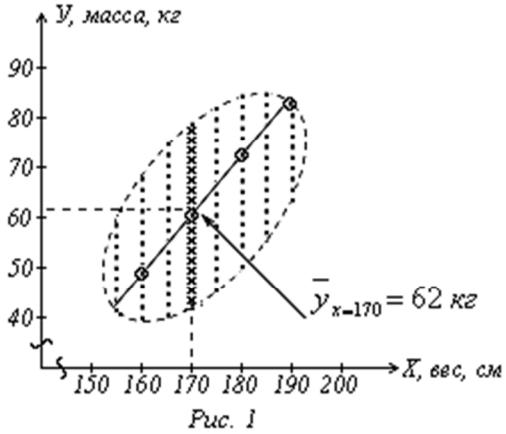


Рис. 1

В медико-биологических исследованиях, однако, чаще встречаются **статистические** зависимости между величинами. Например, при точно определённом изменении возраста пациента не наблюдается строго определённого изменения артериального давления (АД). Объясняется это тем, что изменение АД зависит не только от возраста пациента, но и от ряда других факторов: пола, состояния здоровья и т. п.

Статистическая связь обусловлена несколькими причинами: 1) влиянием на *Y* не только величины *X*, но и других факторов; 2) неизбежностью ошибок при измерении *X* и *Y*.

Частным случаем статистической связи между *X* и *Y* является **корреляционная связь, когда каждому значению *X* ставится в соответствие математическое ожидание (среднее арифметическое значение) распределения другой величины *Y*.** Например, связь между дозой лекарственного препарата *X* и его содержанием в крови *Y*. На *Y* влияет масса пациента, скорость выведения препарата и другие факторы, но у одного и того же пациента с ростом дозы лекарственного препарата содержание его в крови однозначно

увеличивается. Другим примером корреляционной связи является зависимость между ростом человека и его массой, температурой воздуха и количеством заболевших и др.

Человек с ростом, например, 170 см может иметь массу и 50 кг, и 90 кг, но большинство людей имеет вес в интервале 60-80 кг., то есть данному росту соответствует распределение масс, близкое к нормальному, со средним значением *M(Y)*. На рисунке 1 все возможные значения веса человека при *данном* росте 170см отмечены крестиком, а среднее значение (62кг) обведено кружочком. Ясно, что с увеличением роста будет расти и среднее значение веса человека *M(Y)*, то есть мы имеем дело с корреляционной зависимостью между ростом *X* и весом *Y*:

$$M(Y_x) = f(x) \qquad (1)$$

Всё множество значений *X* и *Y* (точки на графике) образует **корреляционное поле**, обведённое пунктиром на рисунке 1.

Уравнение *(1)* называют **уравнением регрессии *Y* на *X***, а его график называют **линией регрессии**. Аналогично можно описать и обратную корреляционную зависимость $M(X_y) = \varphi(y)$, если она существует. Если функции *f(x)* и *φ(y)* – линейные функции, что можно оценить по характеру расположения точек корреляционного поля, то эти функции можно представить в виде:

$$M(Y_x) = ax + b = (наклон)\cdot(x) + (сдвиг),$$

$$M(X_y) = cy + d = (наклон)\cdot(y) + (сдвиг).$$

Для нахождения коэффициентов *a (наклон)* и *b (сдвиг)*, входящих в уравнение прямой, используем метод **наименьших квадратов.**

Метод наименьших квадратов

В 1806 году французский математик Лежандр доказал, что наилучшим образом связь между X и Y будет отражать прямая линия $\bar{y}_x = ax + b$, для которой выполняется условие (см. рис. 2):

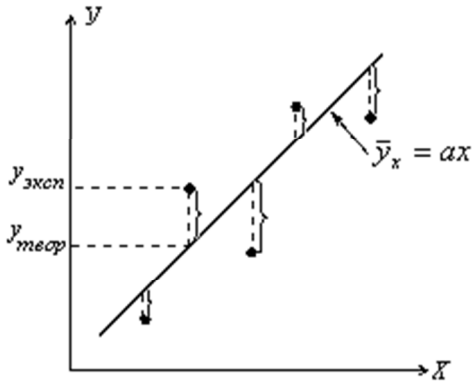


Рис. 2

$$(2)$$

ённое из опыта, $y_{i\,теор}$ – расчётное *y*, лежащее на $x + b$, то условие *(2)* можно записать:

$$\min \qquad (3)$$

значения коэффициентов a и b должны быть ов отклонений ординат экспериментальных точек от ординат точек сглаживающей прямой была бы ***минимальной*** (рис.2).

В соответствии с правилами исследования функции нескольких переменных на минимум должны выполняться следующие условия для частных производных этой функции первого и второго порядков:

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \end{cases} \quad и \quad \begin{cases} \dfrac{\partial^2 S}{\partial a^2} > 0 \\ \dfrac{\partial^2 S}{\partial b^2} > 0 \end{cases} \qquad (4)$$

В соответствии с *(4)* получаем:

$$\begin{cases} \dfrac{\partial S}{\partial a} = 2\sum_{i=1}^{n}(y_i - ax_i - b)(-x_i) = 0 \quad (5) \\ \dfrac{\partial S}{\partial b} = 2\sum_{i=1}^{n}(y_i - ax_i - b)(-1) = 0 \quad (6) \end{cases} \qquad \begin{cases} \dfrac{\partial^2 S}{\partial a^2} = 2\sum_{i=1}^{n} x_i^2 > 0 \\ \dfrac{\partial^2 S}{\partial b^2} = 2\sum_{i=1}^{n} 1 = 2n > 0 \end{cases} \quad (7)$$

Поскольку система неравенств *(7)* удовлетворяется при любых a и b, то решим первую систему уравнений. Делим обе части *(5)* и *(6)* на *2* и умножаем на *(-1):*

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases} \quad (8)$$

Система *(8)* называется системой нормальных уравнений Гаусса. Решая эту систему, найдём *a* и *b*, и получим искомое уравнение прямой $\bar{y}_x = ax + b$:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad (9), \qquad b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad (10).$$

Пример 1. Измерена концентрация *($c=Y_i$)* алкоголя в крови у *n=5* добровольцев с одинаковым весом после нескольких порций алкоголя *(X_i).* Методом наименьших квадратов определите коэффициенты *a* и *b* сглаживающей прямой $\bar{y}_x = ax + b$. Постройте график.

Число порций, X_i					
Концентрация, Y_i	*,05*	*,06*	*,11*	*,13*	*,22*

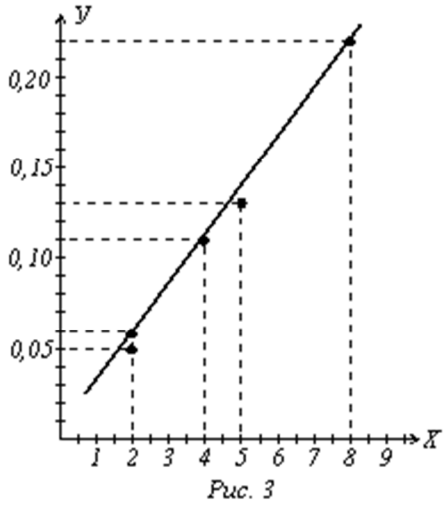


Рис. 3

В соответствии с (9) и (10) найдём предварительно $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} x_i y_i$:

x_i	y_i	x_i^2	$x_i \cdot y_i$
2	*0,05*	*4*	*0,10*

	2	0,06	4	0,12
	4	0,11	16	0,44
	5	0,13	25	0,65
	8	0,22	64	1,76
$\sum$	21	0,57	113	3,07

$$a = \frac{5 \cdot 3{,}07 - 21 \cdot 0{,}57}{5 \cdot 113 - 21^2} = 0{,}027; \qquad b = \frac{113 \cdot 0{,}57 - 21 \cdot 3{,}07}{5 \cdot 113 - 21^2} = -0{,}00048.$$

Искомое уравнение сглаживающей прямой $\bar{y}_x = ax + b$: $\bar{y}_x = 0{,}027x - 0{,}00048$.

График искомой сглаживающей прямой приведён на рис. 3.

Линейная корреляция и её характеристики

Установление **силы и тесноты** корреляционной связи составляет задачу *корреляционного анализа*, а *регрессионный анализ* устанавливает **форму** зависимости между *X* и *У* (линейная, криволинейная).

Исторически теорию корреляции в биологии стали применять раньше, чем в других областях естествознания. Французский биолог Ж. Кювье в 1800-1805 годах в «Лекциях по сравнительной анатомии» сформулировал известный принцип биологической корреляции: любая часть организма непременно согласована с другими частями, следовательно, по одному органу можно судить о целом организме. В 1899 году англичанин К. Пирсон – создатель математической теории корреляции – вывел формулу, связывающую рост современного человека с длиной его бедра. Используя эту формулу, по длине ископаемого бедра Пирсон определил рост доисторического человека.

Для характеристики **формы** уравнения связи *в первую очередь* необходимо учитывать теоретические соображения относительно характера связи между рассматриваемыми величинами. *Во-вторых*, характер расположения точек корреляционного поля также позволяет делать выводы о форме связи. Вытянутая форма корреляционного поля и угол с осями графика, близкий к 45^0, указывает на наличие корреляционной связи (рис. 5г, 5д). Если же скопление точек образует круг или эллипс, длинная ось которой параллельна одной из осей координат, то можно предположить, что связь между величинами отсутствует (рис. 5а).

Силу связи между *X* и *Y* выражает найденный по формуле *(9)* коэффициент «а» (наклон), называемый **коэффициентом регрессии** (его часто обозначают ρ_{yx}). Коэффициент регрессии ρ_{yx} показывает, на сколько единиц изменится в среднем *Y* , если изменение *X* произойдёт ровно на единицу. Чем больше ρ_{yx}, тем связь сильнее. Линейное уравнение регрессии можно записать в общепринятой форме

$$y - \bar{y} = \rho_{yx}(x - \bar{x}), \qquad (11)$$

Тесноту связи (степень разброса точек) оценивают с помощью **коэффициента корреляции *r***. Получим расчётную формулу для *r*.

На первый взгляд, для характеристики разброса точек можно подсчитать произведение $(x_i - \bar{x})(y_i - \bar{y})$ (на графике рис.4 это произведение выражается заштрихованным прямоугольником) и затем найти среднее значение всех произведений (для устранения зависимости от числа пар наблюдений):

$\sum_{i=1}^{n}$

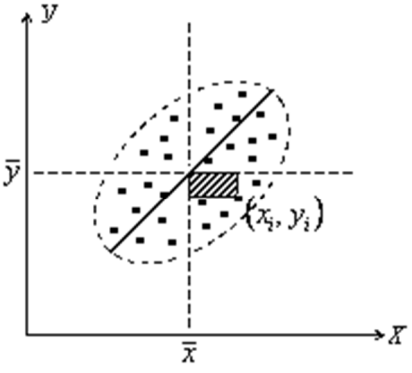


Рис. 4

(12)

ивается **ковариацией**, что означает «сопряжённое
вар от масштаба, выбранного по осям. Этот недостаток
ков поделить *C* на произведение средних квадратических
отк

ю характеристику тесноты связи – **коэффициент**
кор

$$r = \frac{C}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \qquad (13)$$

Раскроем скобки в числителе и учтём, что $\sum_{i=1}^{n} x_i = n\bar{x}$, $\sum_{i=1}^{n} y_i = n\bar{y}$, тогда

$$C = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}}{n} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}n\bar{y} - \bar{y}n\bar{x} + n\bar{x}\bar{y}}{n} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n} = \overline{xy} - \bar{x}\bar{y}$$

, *(14)*

где $\frac{\sum_{i=1}^{n} x_i y_i}{n} = \overline{xy}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. *(15)*

Тогда коэффициент корреляции r с учётом *(14)* равен:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \qquad (16)$$

На практике мы имеем данные не обо всей генеральной совокупности, а только о тех величинах, что получены из эксперимента (выборка). Поэтому определяют *выборочный коэффициент корреляции* r_s, приближённо равный генеральному коэффициенту корреляции *r*. Обозначая средние квадратические отклонения для выборки s_x и s_y, получим:

$$r_s = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}} \qquad (17)$$

где $s_x = \sqrt{\overline{x^2} - (\bar{x})^2}$, $s_y = \sqrt{\overline{y^2} - (\bar{y})^2}$, $\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}$, $\overline{y^2} = \frac{\sum_{i=1}^{n} y_i^2}{n}$. *(18)*

Свойства коэффициента корреляции

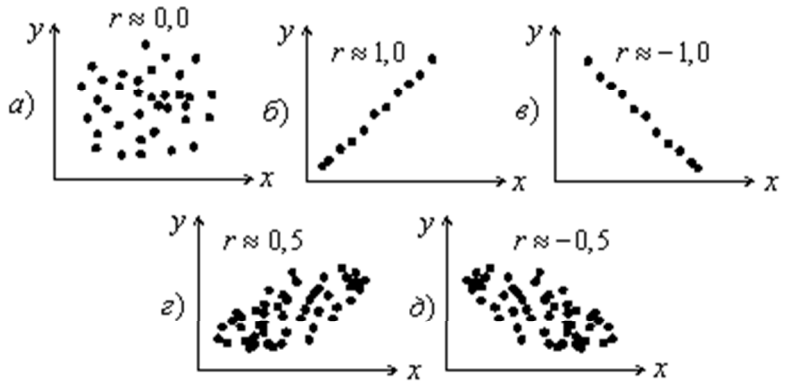


Рис. 5

1) Величина коэффициента корреляции изменяется от -1 до +1, то есть $-1 \le r \le 1$.

2) Чем ближе $|r|$ к единице, тем теснее связь, тем ближе к прямой группируются точки (рис. 5б, 5в). Принята следующая градация тесноты линейной корреляционной связи:

Теснота связи	***Коэффициент корреляции r***
Связь отсутствует	*0*
Связь слабая	*от 0 до 0,3*
Умеренная	*От 0,3 до 0,7*
Сильная	*От 0,7 до 1*
Функциональная	*1*

3) Знак коэффициента корреляции показывает направление связи: прямая (положительная – рис. 5б, 5г) и обратная (отрицательная, рис. 5в, 5д).

Между коэффициентом регрессии ρ_{yx} и коэффициентом корреляции r существует тесная связь: $\rho_{yx} = r\dfrac{s_y}{s_x}$, поэтому $b = \bar{y} - \rho_{yx}\bar{x}$ *(19)*

Тогда **прогнозируемое значение y** при данном значении x равно:

$$y(x) = ax + b = r\frac{s_y}{s_x}x + (\bar{y} - r\frac{s_y}{s_x}\bar{x})$$

Проверка значимости коэффициента корреляции

Так как r_s определяется по данным выборки, то в отличие от коэффициента корреляции *всей* генеральной совокупности, r_s – величина *случайная*. Если $r_s \neq 0$, то возникает вопрос: объясняется ли это действительно существующей линейной связью между X и Y или вызвано случайными факторами. Для ответа на этот вопрос вычисляется величина $t_{ýêñï}$:

$$t_{ýêñï} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \qquad (20)$$

Далее по таблице (Лобоцкая Н.Л. и др. Высшая математика, стр. 302) находим величину $t_{êðèò}$, которая имеет распределение Стьюдента при заданном уровне значимости

p (связанном с доверительной вероятностью соотношением $p=1-\gamma$) и при числе степеней свободы $f=n-2$.

Затем сравнивают $t_{ýêñï}$ и $t_{êðèò}$: **если** $|t_{ýêñï}| > t_{êðèò}$ **, то делают вывод о значимости коэффициента корреляции**, в противном случае линейная связь может быть вызвана случайными факторами. Если коэффициент корреляции оказывается значимым, то можно предсказать значение величины Y при любом значении X.

Замечание: Коэффициент корреляции характеризует связь между величинами, **но** личие корреляции между X и Y может быть вызвано тем, что: Y; величина Y влияет на X; на X и Y влияет третий скрытый фактор, ие связи между X и Y (ложная корреляция). Кроме того, если $r = 0$, ит об отсутствии статистической связи между X и Y – связь может с. 6).

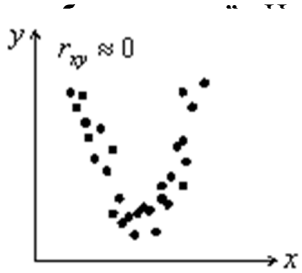


Рис. 6

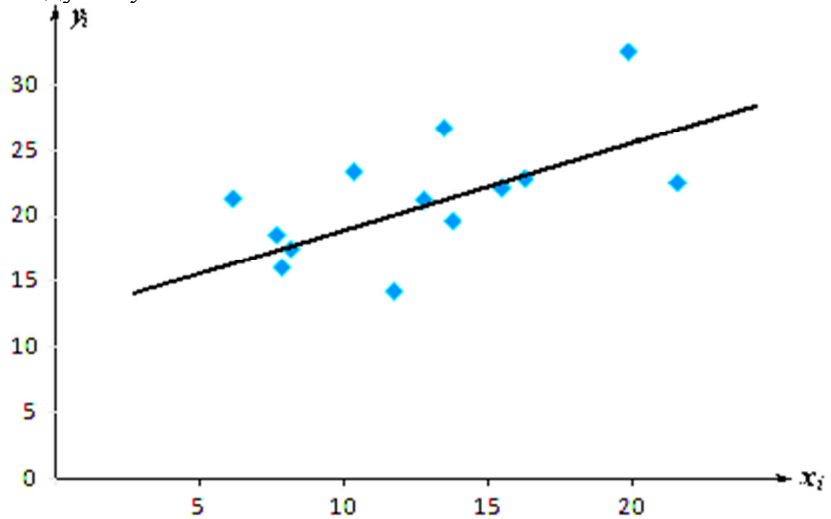
сперименте на 13 кошках получены данные об интрасклеральном влении (y):

$_i$	9,8	,8	2,7	3,4	0,3	3,7	6,2	5,4	1,5	,1	1,7	,6	,1
$_i$	2,5	6,1	1,3	6,8	3,4	9,7	2,9	2,2	2,6	7,6	4,3	8,6	1,4

1. Установите, имеется ли корреляционная связь между x и y; определите коэффициент корреляции r:
2. Определите тесноту корреляционной связи.
3. Проверьте значимость выборочного коэффициента корреляции.
4. Составьте уравнение регрессии и найдите ожидаемое значение для y при x=18.

Решение:

1. Построим график, отложив вдоль оси абсцисс X величину интрасклерального давления x , а вдоль оси ординат Y – величину внутриглазного давления y. Тогда каждой паре значений x и y на графике будет соответствовать определённая точка. По характеру расположения точек можно предположить существование линейной корреляционной связи между x и y.



Вычислим коэффициент линейной корреляции r по формуле

$$r_s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2}\sqrt{\overline{y^2} - (\bar{y})^2}}$$

$$\bar{x} = \frac{1}{13}(19{,}8 + 7{,}8 + ... + 6{,}1) = 12{,}64; \qquad \bar{y} = \frac{1}{13}(32{,}5 + 16{,}1 + ... + 21{,}4) = 21{,}49$$

;

$$\overline{x^2} = \frac{1}{13}(19{,}8^2 + 7{,}8^2 + ... + 6{,}1^2) = 180{,}5;$$

$$\overline{y^2} = \frac{1}{13}(32{,}5^2 + 16{,}1^2 + ... + 21{,}4^2) = 482{,}2;$$

$$\overline{xy} = \frac{1}{13}(19{,}8 \cdot 32{,}5 + 7{,}8 \cdot 16{,}1 + ... + 6{,}1 \cdot 21{,}4) = 283{,}9;$$

$$s_x = \sqrt{180{,}5 - 12{,}64^2} = 4{,}55; \qquad s_y = \sqrt{482{,}2 - 21{,}49^2} = 4{,}51;$$

$$r = \frac{283{,}9 - 12{,}64 \cdot 21{,}49}{4{,}55 \cdot 4{,}51} = 0{,}595.$$

2. Пользуясь таблицей градации оценки тесноты связи, делаем вывод: связь *x* и *y* умеренная, положительная.

3. Для проверки значимости коэффициента корреляции r_s вычислим формуле *(19)* величину $t_{\acute{y}\hat{e}\tilde{n}\ddot{i}}$:

$$t_{эксп} = \frac{0{,}598 \cdot \sqrt{13-2}}{\sqrt{1-0{,}598^2}} = 2{,}47$$

По таблице находим величину $t_{крит}(0{,}05;11) = 2{,}2$. Так как $\left|t_{\acute{y}\hat{e}\tilde{n}\ddot{i}}\right| > t_{\hat{e}\partial\grave{e}\partial}$, то есть 2,47>2.20, то делаем вывод о значимости коэффициента корреляции.

4. По формуле $\rho_{yx} = = r\frac{s_y}{s_x}$ находим коэффициент регрессии:

$$\rho_{yx} = 0{,}595\frac{4{,}51}{4{,}55} = 0{,}589.$$

Далее, подставляя ρ_{yx} в формулу $y - \bar{y} = \rho_{yx}(x - \bar{x})$ и вычисляя *b*, находим уравнение регрессии: $y = 0{,}589x + 14$

И, наконец, вычисляем ожидаемое значение *y* при *x=18*:

$y(18) = 0{,}589 \cdot 18 + 14 = 24{,}6$

Выполнение работы

Используя экспериментальные данные (по указанию преподавателя) выполните следующее задание:

У восьми мужчин были измерены рост (x) и вес (m):

X (см)	65	76	75	68	67	72	75	80
M (кг)	6	5	0	1	2	3	2	0

Что сделать:

1.Установите, имеется ли корреляционная связь между x и y; определите коэффициент корреляции r_s .

2.Определите тесноту корреляционной связи.
3.Проверьте значимость выборочного коэффициента корреляции.
4.Составьте уравнение регрессии $y = \rho_{yx} x + b$ и постройте график.

3. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРОВЕДЕНИЮ ПРАКТИЧЕСКИХ ЗАНЯТИЙ

Не предусмотрено РУП

4. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРОВЕДЕНИЮ СЕМИНАРСКИХ ЗАНЯТИЙ

Не предусмотрено РУП