Лекция 1.

Основные понятия математической статистики. Генеральная совокупность и выборка. Вариационный ряд, статистический ряд. Группированная выборка. Группированный статистический ряд. Полигон частот. Выборочная функция распределения и гистограмма.

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 рассматриваемой 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 поведении интересующего нас признака генеральной совокупности, нужно, чтобы выборка правильно 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 $x_1, x_2, \ldots, x_к$ называют **вариантами**, а $n_1, n_2, \ldots, n_к$ – **частотами**. Если разделить каждую частоту на объем выборки, то получим **относительные частоты.** Последовательность вариант, записанных в порядке возрастания, называют **вариационным** рядом, а перечень вариант и соответствующих им частот или относительных частот – **статистическим 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)**.**

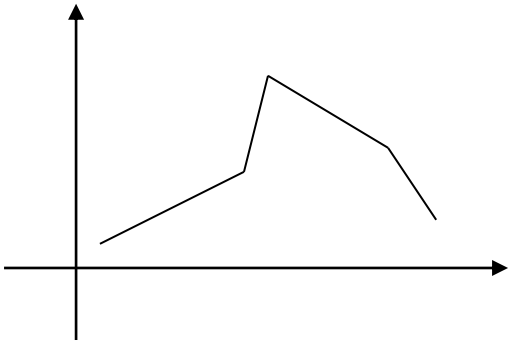


Рис. 1.

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

Определение 1.1. **Выборочной (эмпирической) функцией распределения** называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$. Таким образом,

$$F^*(x) = \frac{n_x}{n}, \qquad (1.1)$$

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).

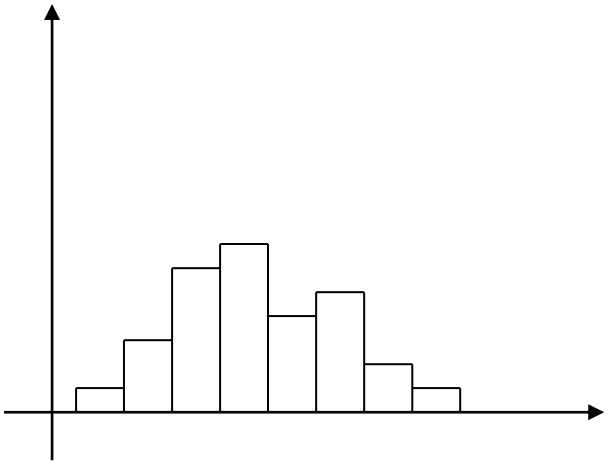


Рис.2.

Лекция 2.

Числовые характеристики статистического распределения: выборочное среднее, оценки дисперсии, оценки моды и медианы, оценки начальных и центральных моментов. Статистическое описание и вычисление оценок параметров двумерного случайного вектора.

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

Определение 2.1. **Выборочным средним** называется среднее арифметическое значений случайной величины, принимаемых в выборке:

$$\bar{x}_B = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n}, \qquad (2.1)$$

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 2.2. **Выборочной дисперсией** называется

$$D_B = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_B)^2}{n} = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n}, \qquad (2.2)$$

а **выборочным средним квадратическим отклонением** –

$$\sigma_B = \sqrt{D_B}. \qquad (2.3)$$

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

$$D = \overline{x^2} - (\bar{x})^2. \qquad (2.4)$$

Пример 1. Найдем числовые характеристики выборки, заданной статистическим рядом

x_i	2	5	7	8
n_i	3	8	7	2

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

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 = 2*k* + 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

$$M_k = \frac{\sum n_i x_i^k}{n}. \qquad (2.5)$$

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

$$m_k = \frac{\sum n_i (x_i - \bar{x}_B)^k}{n}. \qquad (2.6)$$

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 наблюдавшихся 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.

Лекция 3.

Основные свойства статистических характеристик параметров распределения: несмещенность, состоятельность, эффективность. Несмещенность и состоятельность выборочного среднего как оценки математического ожидания. Смещенность выборочной дисперсии. Пример несмещенной оценки дисперсии. Асимптотически несмещенные оценки. Способы построения оценок: метод наибольшего правдоподобия, метод моментов, метод квантили, метод наименьших квадратов, байесовский подход к получению оценок.

Получив статистические оценки параметров распределения (выборочное среднее, выборочную дисперсию и т.д.), нужно убедиться, что они в достаточной степени служат приближением 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 систематических ошибок является требование $M(\Theta^*) = \Theta$.

Определение 3.1. Статистическая оценка Θ* называется **несмещенной**, если ее математическое ожидание равно оцениваемому параметру Θ при любом объеме выборки:

$$M(\Theta^*) = \Theta. \qquad (3.1)$$

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

Однако несмещенность не является достаточным условием хорошего приближения к истинному 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.

Определение 3.2. 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.

Определение 17.3. **Состоятельной** называется статистическая оценка, которая при $n \to \infty$ стремится 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 $x_1, x_2, \ldots, x_n$, то есть значения исследуемой случайной величины, составляющие выборку, – как независимые, одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$, имеющие математическое ожидание a. Из свойств математического ожидания следует, что

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

Но, поскольку каждая из величин $X_1, X_2, \ldots, X_n$ 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 $X_1, X_2, \ldots, X_n$ 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

$$M(D_B) = \frac{n-1}{n} D_\Gamma, \qquad (3.2)$$

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

$$s^2 = \frac{n}{n-1} D_B = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n-1}. \qquad (3.3)$$

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

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n-1}}. \qquad (3.4)$$

Определение 3.4. Оценка некоторого признака называется **асимптотически несмещенной**, если для выборки $x_1, x_2, \ldots, x_n$

$$\lim_{n\to\infty}\frac{x_1 + x_2 + \ldots + x_n}{n} = X\,, \qquad (3.5)$$

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 $x_1, x_2, \ldots, x_n$. 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 $\Theta^* = \Theta(x_1, x_2, \ldots, x_n)$, 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:

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

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:

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

(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:

$\psi(Y) = \frac{\int \varphi(x)\,p(Y \mid x)\,p(x)d\mu(x)}{q(Y)}$, 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.

Лекция 4.
Интервальное оценивание неизвестных параметров. Точность оценки, доверительная вероятность (надежность), доверительный интервал. Построение доверительных интервалов для оценки математического ожидания нормального распределения при известной и при неизвестной дисперсии. Доверительные интервалы для оценки среднего квадратического отклонения нормального распределения.

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 $|\Theta^* - \Theta| < \delta$, число $\delta > 0$ характеризует **точность оценки** (чем меньше δ, тем точнее оценка). Но статистические методы позволяют говорить только о том, что это неравенство выполняется с некоторой вероятностью.

Определение 4.1. **Надежностью (доверительной вероятностью)** оценки Θ^* параметра Θ называется вероятность γ того, что выполняется неравенство $|\Theta^* - \Theta| < \delta$. Если заменить это неравенство двойным неравенством $-\delta < \Theta^* - \Theta < \delta$, то получим:

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

Таким образом, γ есть вероятность того, что Θ попадает в интервал $(\Theta^* - \delta, \Theta^* + \delta)$.

Определение 4.2. **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 $\bar{x}_B$ оценить ее математическое ожидание a. Будем рассматривать выборочное среднее $\bar{x}_B$ как случайную величину $\bar{X}$, а значения вариант выборки $x_1, x_2, \ldots, x_n$ как одинаково распределенные независимые случайные величины $X_1, X_2, \ldots, X_n$, каждая из которых имеет математическое ожидание a и среднее квадратическое отклонение σ. При этом $M(\bar{X}) = a$, $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ (используем свойства математического ожидания и дисперсии суммы независимых случайных

величин). Оценим вероятность выполнения неравенства $|\bar{X} - a| < \delta$. 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 $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, $p\,(|\bar{X} - a| < \delta) =$

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

$$p\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}} < a < \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma\,. \qquad (4.1)$$

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 распределенной случайной величины, если объем выборки n = 49, $\bar{x}_B = 2{,}8, \sigma = 1{,}4$, а доверительная вероятность γ = 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

$$T = \frac{\bar{x}_B - a}{\frac{s}{\sqrt{n}}}, \qquad (4.2)$$

где $\bar{x}_B$ - выборочное среднее, s – исправленная дисперсия, n – объем выборки. Эта случайная величина, возможные значения которой будем обозначать t, имеет распределение Стьюдента 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:

$$p\left(\bar{x}_B - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma s}{\sqrt{n}}\right) = \gamma. \qquad (4.3)$$

Таким образом, получен доверительный интервал для a, где t_γ можно найти по соответствующей 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 $\gamma = 0{,}99$. Из таблицы находим, что t_γ ($n = 25$, $\gamma = 0{,}99$) = 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 $(s - \delta,\ s + \delta)$, где s – исправленное выборочное среднее квадратическое отклонение, а для δ выполняется условие: $p\ (\ |\sigma - s| < \delta\ ) = \gamma$.

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

$$s(1-q) < \sigma < s(1+q). \qquad (4.4)$$

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. 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, $\int_{\chi_1}^{\chi_2} R(\chi, n)d\chi = \gamma$. Предположим, что $q < 1$, тогда неравенство (4.4) 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 (4.4), 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

$$0 < \sigma < s(1+q). \qquad (3.5)$$

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 < σ < 1,781 с вероятностью 0,95.

Лекция 5.

Статистическая проверка статистических гипотез. Общие принципы проверки гипотез. Понятия статистической гипотезы (простой и сложной), нулевой и конкурирующей гипотезы, ошибок первого и второго рода, уровня значимости, статистического критерия, критической области, области принятия гипотезы. Наблюдаемое значение критерия. Критические точки. Мощность критерия. Критерии для проверки гипотез о вероятности события, о математическом ожидании, о сравнении двух дисперсий.

Определение 5.1. **Статистической гипотезой** называют гипотезу о виде неизвестного распределения генеральной совокупности или о параметрах известных распределений.

Определение 5.2. **Нулевой (основной)** называют выдвинутую гипотезу H_0. **Конкурирующей (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) $a < 3$.

Определение 5.3. **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.

Определение 5.4. 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.

Определение 5.5. **Статистическим критерием** называется случайная величина K с известным законом распределения, служащая для проверки нулевой гипотезы.

Определение 5.6. **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 $k_{кр}$ располагается 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 $K < k_1$, $K > k_2$ $(k_2 > k_1)$.

Определение 5.7. **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 $1 - \beta$.
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

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}, \qquad (5.1)$$

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 $\Phi(x)$.

Далее нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}} \quad . \qquad (5.2)$$

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 (5.2).

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 $U_{набл} = \frac{(0{,}12 - 0{,}1)\sqrt{50}}{\sqrt{0{,}1 \cdot 0{,}9}} = 0{,}471$. Критическая область является правосторонней, а $u_{кр}$ находим из равенства

$\Phi(u_{кр}) = \frac{1 - 2 \cdot 0{,}01}{2} = 0{,}49.$

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

$$U = \frac{\overline{X} - a_0}{\sigma(\overline{X})} = \frac{(\overline{X} - a_0)\sqrt{n}}{\sigma}. \qquad (5.3)$$

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 $u_{кр}$: $\Phi(u_{кр}) = \frac{1 - \alpha}{2}$, критическая область двусторонняя, $U_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma}$, 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 $u_{кр}$: $\Phi(u_{кр}) = \frac{1 - 2\alpha}{2}$, 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: $M(\bar{X}) < a_0$, то $u_{кр}$: $\Phi(u_{кр}) = \frac{1-2\alpha}{2}$, 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

$$T = \frac{(\bar{X} - a_0)\sqrt{n}}{S}, \qquad (5.4)$$

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:

$$T_{набл} = \frac{(\bar{x}_B - a_0)\sqrt{n}}{S}. \qquad (5.5)$$

- если H_1: $M(\bar{X}) \neq a_0$, 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: $M(\bar{X}) > a_0$, 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: $M(\bar{X}) < a_0$ 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 рассматривае-мых генеральных совокупностей. Учитывая несмещенность исправленных выборочных дисперсий, можно записать нулевую гипотезу так:

$$H_0: M(s_X^2) = M(s_Y^2). \qquad (5.6)$$

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

$$F = \frac{S_\sigma^2}{S_M^2} \text{ -} \qquad (5.7)$$

- 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.

Лекция 6.

Критерий Пирсона для проверки гипотезы о виде закона распределения случайной величины. Проверка гипотез о нормальном, показательном и равномерном распределениях по критерию Пирсона. Критерий Колмогорова. Приближенный метод проверки нормальности распределения, связанный с оценками коэффициентов асимметрии и эксцесса.

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 x_2 … x_s

частоты……n_1 n_2 … 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

$$\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}. \qquad (6.1)$$

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 (6.1) при $n \to \infty$ стремится к закону распределения χ^2 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

$$p(\chi^2 > \chi^2_{kp}(\alpha, k)) = \alpha, \qquad (6.2)$$

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

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

$$\chi^2_{набл} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'} \ , \qquad (6.1`)$$

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 $\chi^2_{набл} < \chi^2_{кр}$ - нулевую гипотезу принимают, при $\chi^2_{набл} > \chi^2_{кр}$ 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:

$$a^* = \bar{x}_B - \sqrt{3}\sigma_B, \ \ b^* = \bar{x}_B + \sqrt{3}\sigma_B \ , \qquad (6.3)$$

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 (6.3).

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

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

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

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

Здесь s – число интервалов, на которые разбита выборка.

Наблюдаемое значение критерия Пирсона вычисляется по формуле (6.1`), 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 $$K(\lambda) = \sum_{m=-\infty}^{\infty} (-1)^m e^{-2m^2\lambda^2}$$ -

- 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

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

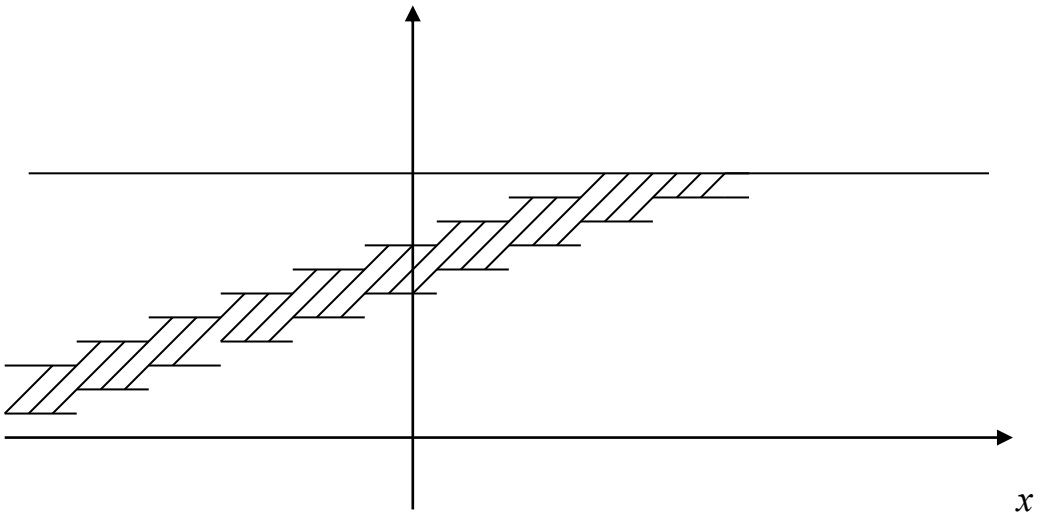
где z – корень уравнения $1 - K\left(\sqrt{\frac{\lambda}{2}}\right) = \alpha.$

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 Оxy 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 6.1. **Асимметрия эмпирического распределения** определяется равенством

$$a_s = \frac{m_3}{\sigma_B^3}, \qquad (6.5)$$

где m_3 – центральный эмпирический момент третьего порядка.

Эксцесс эмпирического распределения определяется равенством

$$e_k = \frac{m_4}{\sigma_B^4} - 3, \qquad (6.6)$$

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.

Лекция 7.
Корреляционный анализ.

Проверка гипотезы о значимости выборочного коэффициента корреляции.

Рассмотрим выборку объема n, извлеченную из нормально распределенной двумерной генеральной совокупности (X, Y). Вычислим выборочный коэффициент корреляции r_B. Пусть он оказался не равным нулю. Это еще не означает, что и коэффициент корреляции генеральной совокупности не равен нулю. Поэтому при заданном уровне значимости α возникает необходимость проверки нулевой гипотезы H_0: $r_г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r_г \neq 0$. Таким образом, при принятии нулевой гипотезы X и Y некоррелированы, то есть не связаны линейной зависимостью, а при отклонении H_0 они коррелированы.
В качестве критерия примем случайную величину

$$T = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}}, \qquad (7.1)$$

которая при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n - 2$ степенями свободы. Из вида конкурирующей гипотезы следует, что критическая область двусторонняя с границами $\pm t_{кр}$, где значение $t_{кр}(\alpha, k)$ находится из таблиц для двусторонней критической области.
Вычислив наблюдаемое значение критерия

$$T_{набл} = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}}$$

и сравнив его с $t_{кр}$, делаем вывод:
- если $|T_{набл}| < t_{кр}$ – нулевая гипотеза принимается (корреляции нет);
- если $|T_{набл}| > t_{кр}$ – нулевая гипотеза отвергается (корреляция есть).

Ранговая корреляция.

Пусть объекты генеральной совокупности обладают двумя качественными признаками (то есть признаками, которые невозможно измерить точно, но которые позволяют сравнивать объекты между собой и располагать их в порядке убывания или возрастания качества). Договоримся для определенности располагать объекты в порядке ухудшения качества.
Пусть выборка объема n содержит независимые объекты, обладающие двумя качественными признаками: A и B. Требуется выяснить степень их связи между собой, то есть установить наличие или отсутствие **ранговой корреляции.**

Расположим объекты выборки в порядке ухудшения качества по признаку A, предполагая, что все они имеют различное качество по обоим признакам. Назовем место, занимаемое в этом ряду некоторым объектом, его **рангом x_i:** $x_1 = 1, x_2 = 2,\ldots, x_n = n$.

Теперь расположим объекты в порядке ухудшения качества по признаку B**,** присвоив им ранги y_i , где номер i равен порядковому номеру объекта по признаку A, а само значение ранга равно порядковому номеру объекта по признаку B. Таким образом, получены две последовательности рангов:

по признаку A … $x_1, x_2,\ldots, x_n$

по признаку B … $y_1, y_2,\ldots, y_n$.

При этом, если, например, $y_3 = 6$, то это означает, что данный объект занимает в ряду по признаку A третье место, а в ряду по признаку B – шестое. Сравним полученные последовательности рангов.

1. Если $x_i = y_i$ при всех значениях i, то ухудшение качества по признаку A влечет за собой ухудшение качества по признаку B, то есть имеется «полная ранговая зависимость».
2. Если ранги противоположны, то есть $x_1 = 1, y_1 = n$; $x_2 = 2, y_2 = n - 1;\ldots,$ $x_n = n, y_n = 1$, то признаки тоже связаны: ухудшение качества по одному из них приводит к улучшению качества по другому («противоположная зависимость»).
3. На практике чаще всего встречается промежуточный случай, когда ряд y_i не монотонен. Для оценки связи между признаками будем считать ранги $x_1, x_2,\ldots, x_n$ возможными значениями случайной величины X, а $y_1, y_2,\ldots, y_n$ – возможными значениями случайной величины Y. Теперь можно исследовать связь между X и Y, вычислив для них выборочный коэффициент корреляции

$$r_B = \frac{\sum n_{uv} uv - n\bar{u}\bar{v}}{n\sigma_u\sigma_v}, \qquad (7.2)$$

где $u_i = x_i - \bar{x},\ v_i = y_i - \bar{y}$ (условные варианты). Поскольку каждому рангу x_i соответствует только одно значение y_i, то частота любой пары условных вариант с одинаковыми индексами равна 1, а с разными индексами – нулю. Кроме того, из выбора условных вариант следует, что $\bar{u} = \bar{v} = 0$, поэтому формула (7.2) приобретает более простой вид:

$$r_B = \frac{\sum u_i v_i}{n\sigma_u\sigma_v}. \qquad (7.3)$$

Итак, требуется найти $\sum u_i v_i$, σ_u и σ_v.

Можно показать, что $\sum u_i^2 = \sum v_i^2 = \frac{n^3 - n}{12}$. Учитывая, что $\bar{x} = \bar{y}$, можно выразить $\sum u_i v_i$ через разности рангов $d_i = x_i - y_i = u_i - v_i$. После преобразований получим: $\sum u_i v_i = \frac{n^3 - n}{12} - \sum \frac{d_i^2}{2}$, $\sigma_u = \sigma_v = \sqrt{\frac{n^2 - 1}{12}}$, откуда $n\sigma_u\sigma_v = \frac{n^3 - n}{12}$. Подставив эти результаты в (7.3), получим **выборочный коэффициент ранговой корреляции Спирмена**:

$$\rho_B = 1 - \frac{6\sum d_i^2}{n^3 - n}. \qquad (7.4)$$

Свойства выборочного коэффициента корреляции Спирмена.

1. Если между A и B имеется «полная прямая зависимость», то есть ранги совпадают при всех i, то $\rho_B = 1$. Действительно, при этом $d_i = 0$, и из формулы (7.4) следует справедливость свойства 1.
2. Если между A и B имеется «противоположная зависимость», то $\rho_B = -1$. В этом случае, преобразуя $d_i = (2i - 1) - n$, найдем, что $\sum d_i^2 = \frac{n^3 - n}{3}$, тогда из (7.4) $\rho_B = 1 - \frac{6(n^3 - n)}{3(n^3 - n)} = 1 - 2 = -1.$
3. В остальных случаях $-1 < \rho_B < 1$, причем зависимость между A и B тем меньше, чем ближе $|\rho_B|$ к нулю.

Итак, требуется при заданном уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции Спирмена $\rho_г$ при конкурирующей гипотезе H_1: $\rho_г \neq 0$. Для этого найдем критическую точку:

$$T_{кр} = t_{кр}(\alpha, k)\sqrt{\frac{1 - \rho_B^2}{n - 2}}, \qquad (7.5)$$

где n – объем выборки, ρ_B – выборочный коэффициент ранговой корреляции Спирмена, $t_{кр}$ (α, k) – критическая точка двусторонней критической области, найденная по таблице критических точек распределения Стьюдента, число степеней свободы $k = n - 2$.
Тогда, если $|\rho_B| < T_{кр}$, то нулевая гипотеза принимается, то есть ранговая корреляционная связь между признаками незначима.
Если $|\rho_B| > T_{кр}$, то нулевая гипотеза отвергается, и между признаками существует значимая ранговая корреляционная связь.

Можно использовать и другой коэффициент – коэффициент ранговой корреляции Кендалла. Рассмотрим ряд рангов $y_1, y_2, \ldots, y_n$, введенный так же, как и ранее, и зададим величины R_i следующим образом: пусть правее y_1 имеется R_1 рангов, больших y_1; правее y_2 – R_2 рангов, больших y_2 и т.д. Тогда, если обозначить $R = R_1 + R_2 + \ldots + R_{n-1}$, то **выборочный коэффициент ранговой корреляции Кендалла** определяется формулой

$$\tau_B = \frac{4R}{n(n-1)} - 1, \qquad (7.6)$$

где n – объем выборки.
Замечание. Легко убедиться, что коэффициент Кендалла обладает теми же свойствами, что и коэффициент Спирмена.

Для проверки нулевой гипотезы H_0: $\tau_г = 0$ (генеральный коэффициент ранговой корреляции Кендалла равен нулю) при альтернативной гипотезе H_1: $\tau_г \neq 0$ необходимо найти критическую точку:

$$T_{кр} = z_{kp}\sqrt{\frac{2(2n+5)}{9n(n-1)}}, \qquad (7.7)$$

где n – объем выборки, а $z_{кр}$ – критическая точка двусторонней критической области, определяемая из условия $\Phi(z_{kp}) = \frac{1-\alpha}{2}$ по таблицам для функции Лапласа.

Если $|\tau_В| < T_{кр}$, то нулевая гипотеза принимается (ранговая корреляционная связь между признаками незначима).

Если $|\tau_В| > T_{кр}$, то нулевая гипотеза отвергается (между признаками существует значимая ранговая корреляционная связь).

Лекция 8.
Регрессионный анализ.

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

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

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

$$\bar{y}_x = f^*(x) - \qquad (8.1)$$

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

$$\bar{x}_y = \varphi^*(y) - \qquad (8.2)$$

- 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

$$Y = \rho_{yx}x + b, \qquad (8.3)$$

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 (8.3).

Используем для этого метод наименьших квадратов и найдем минимум функции

$$F(\rho,b)=\sum_{i=1}^{n}(Y_i-y_i)^2=\sum_{i=1}^{n}(\rho x_i+b-y_i)^2\ . \qquad (8.4)$$

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:

$$\begin{cases}\left(\sum x^2\right)\rho+\left(\sum x\right)b=\sum xy\\ \left(\sum x\right)\rho+nb=\sum y\end{cases}. \qquad (8.5)$$

Ее решение позволяет найти искомые параметры в виде:

$$\rho_{xy}=\frac{n\sum xy-\sum x\cdot\sum y}{n\sum x^2-\left(\sum x\right)^2};\ \ b=\frac{\sum x^2\cdot\sum y-\sum x\cdot\sum xy}{n\sum x^2-\left(\sum x\right)^2}\ . \qquad (8.6)$$

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 (8.5) $\sum x=n\bar{x}$, $\sum y=n\bar{y}$, $\sum x^2=n\overline{x^2}$, $\sum xy=\sum n_{xy}xy$, где n_{xy} – число появлений пары чисел (x, y). Тогда система (8.5) примет вид:

$$\begin{cases}(n\overline{x^2})\rho_{yx}+(n\bar{x})b=\sum n_{xy}xy\\ (\bar{x})\rho_{yx}+b=\bar{y}\end{cases}. \qquad (8.7)$$

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

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

Но чаще уравнение регрессии записывают в ином виде, вводя **выборочный коэффициент корреляции**. Выразим b из второго уравнения системы (8.7):

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

Подставим это выражение в уравнение регрессии: $\bar{y}_x-\bar{y}=\rho_{yx}(x-\bar{x})$. Из (8.7)

$$\rho_{yx} = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n(\overline{x^2} - (\bar{x})^2)} = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\tilde{\sigma}_x^2}, \qquad (8.8)$$

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

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

и умножим равенство (8.8) 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

$$\bar{y}_x - \bar{y} = r_B\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x - \bar{x}) \ . \qquad (8.9)$$

Лекция 9.

Однофакторный дисперсионный анализ.

Пусть генеральные совокупности $X_1, X_2, \ldots, X_p$ распределены нормально и имеют одинаковую дисперсию, значение которой неизвестно. Найдем выборочные средние по выборкам из этих генеральных совокупностей и проверим при заданном уровне значимости нулевую гипотезу H_0: $M(X_1) = M(X_2) = \ldots = M(X_p)$ о равенстве всех математических ожиданий. Для решения этой задачи применяется метод, основанный на сравнении дисперсий и названный поэтому **дисперсионным анализом**.

Будем считать, что на случайную величину X воздействует некоторый качественный фактор F, имеющий p уровней: $F_1, F_2, \ldots, F_p$. Требуется сравнить «факторную дисперсию», то есть рассеяние, порождаемое изменением уровня фактора, и «остаточную дисперсию», обусловленную случайными причинами. Если их различие значимо, то фактор существенно влияет на X и при изменении его уровня групповые средние различаются значимо.

Будем считать, что количество наблюдений на каждом уровне фактора одинаково и равно q. Оформим результаты наблюдений в виде таблицы:

Номер испытания	Уровни фактора F_j			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{qp}
Групповое среднее	$\bar{x}_{гр1}$	$\bar{x}_{гр2}$	…	$\bar{x}_{грp}$

Определим общую, факторную и остаточную суммы квадратов отклонений от среднего:

$$S_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \bar{x})^2 \text{ -} \qquad (9.1)$$

- общая сумма квадратов отклонений наблюдаемых значений от общего среднего $\bar{x}$;

$$S_{факт} = q\sum_{j=1}^{p}(\bar{x}_{грj} - \bar{x})^2 \text{ -} \qquad (9.2)$$

- факторная сумма отклонений групповых средних от общей средней, характеризующая рассеяние между группами;

$$S_{ост} = \sum_{i=1}^{q}(x_{i1} - \bar{x}_{гр1})^2 + \sum_{i=1}^{q}(x_{i2} - \bar{x}_{гр2})^2 + ... + \sum_{i=1}^{q}(x_{ip} - \bar{x}_{грp})^2 \text{ -} \qquad (9.3)$$

- остаточная сумма квадратов отклонений наблюдаемых значений группы от своего группового среднего, характеризующая рассеяние внутри групп.

Замечание. Остаточную сумму можно найти из равенства

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}}\ .$$

Вводя обозначения $R_j = \sum_{i=1}^{q} x_{ij}$, $P_j = \sum_{i=1}^{q} x_{ij}^2$, получим формулы, более удобные для расчетов:

$$S_{общ} = \sum_{j=1}^{p} P_j - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq}, \qquad (9.1`)$$

$$S_{факт} = \frac{\sum_{j=1}^{p} R_j^2}{q} - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq}\ . \qquad (9.2`)$$

Разделив суммы квадратов на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$s_{общ}^2 = \frac{S_{общ}}{pq-1},\ s_{факт}^2 = \frac{S_{факт}}{p-1},\ s_{ост}^2 = \frac{S_{ост}}{p(q-1)}\ . \qquad (9.4)$$

Если справедлива гипотеза H_0, то все эти дисперсии являются несмещенными оценками генеральной дисперсии. Покажем, что проверка нулевой гипотезы сводится к сравнению факторной и остаточной дисперсии по критерию Фишера-Снедекора.

1. Пусть гипотеза H_0 правильна. Тогда факторная и остаточная дисперсии являются несмещенными оценками неизвестной генеральной дисперсии и, следовательно, различаются незначимо. Поэтому результат оценки по критерию Фишера-Снедекора F покажет, что нулевая гипотеза принимается. Таким образом, если верна гипотеза о равенстве математических ожиданий генеральных совокупностей, то верна и гипотеза о равенстве факторной и остаточной дисперсий.

2. Если нулевая гипотеза неверна, то с возрастанием расхождения между математическими ожиданиями увеличивается и факторная дисперсия, а вместе с ней и отношение $F_{набл} = \frac{s_{факт}^2}{s_{ост}^2}$. Поэтому в результате $F_{\text{набл}}$ окажется

больше $F_{кр}$, и гипотеза о равенстве дисперсий будет отвергнута. Следовательно, если гипотеза о равенстве математических ожиданий генеральных совокупностей ложна, то ложна и гипотеза о равенстве факторной и остаточной дисперсий.

Итак, метод дисперсионного анализа состоит в *проверке по критерию F нулевой гипотезы о равенстве факторной и остаточной дисперсий.*

Замечание. Если факторная дисперсия окажется меньше остаточной, то гипотеза о равенстве математических ожиданий генеральных совокупностей верна. При этом нет необходимости использовать критерий F.

Если число испытаний на разных уровнях различно (q_1 испытаний на уровне F_1, q_2 – на уровне F_2, …, q_p - на уровне F_p), то

$$S_{общ} = (P_1 + P_2 + ... + P_p) - (R_1 + R_2 + ... + R_p),$$

где $P_j = \sum_{i=1}^{q_j} x_{ij}^2$ – сумма квадратов наблюдавшихся значений признака на уровне F_j,

$R_j = \sum_{i=1}^{q_j} x_{ij}$ – сумма наблюдавшихся значений признака на уровне F_j.

При этом объем выборки, или общее число испытаний, равен $n = q_1 + q_2 + ... + q_p$.

Факторная сумма квадратов отклонений вычисляется по формуле

$$S_{факт} = \left(\frac{R_1^2}{q_1} + \frac{R_2^2}{q_2} + ... + \frac{R_p^2}{q_p} \right) - \frac{(R_1 + R_2 + ... + R_p)^2}{n}.$$

Остальные вычисления проводятся так же, как в случае одинакового числа испытаний:

$$S_{ост} = S_{общ} - S_{факт}, \quad s_{факт}^2 = \frac{S_{факт}}{p-1}, \quad s_{ост}^2 = \frac{S_{ост}}{n-p}.$$

Лекция 10.

Моделирование случайных величин методом Монте-Карло (статистических испытаний).

Задачу, для решения которой применяется метод Монте-Карло, можно сформулировать так: требуется найти значение a изучаемой случайной величины. Для его определения выбирается случайная величина X, математическое ожидание которой равно a, и для выборки из n значений X, полученных в n испытаниях, вычисляется выборочное среднее:

$$\bar{x} = \frac{\sum x_i}{n},$$

которое принимается в качестве оценки искомого числа a:

$$a \approx a^* = \bar{x}.$$

Этот метод требует проведения большого числа испытаний, поэтому его иначе называют **методом статистических испытаний**. Теория метода Монте-Карло исследует, как наиболее целесообразно выбрать случайную величину X, как найти ее возможные значения, как уменьшить дисперсию

используемых случайных величин, чтобы погрешность при замене a на a^* была возможно меньшей.

Поиск возможных значений X называют **разыгрыванием** случайной **величины**. Рассмотрим некоторые способы разыгрывания случайных величин и выясним, как оценить допускаемую при этом ошибку.

Оценка погрешности метода Монте-Карло.

Если поставить задачу определения верхней границы допускаемой ошибки с заданной доверительной вероятностью γ, то есть поиска числа δ, для которого

$$p(|\overline{X} - a| \le \delta) = \gamma,$$

то получим известную задачу определения доверительного интервала для математического ожидания генеральной совокупности . Воспользуемся результатами решения этой задачи для следующих случаев:

1) случайная величины X распределена нормально и известно ее среднее квадратическое отклонение. Тогда из формулы (4.1) получаем: $\delta = \frac{t\delta}{\sqrt{n}}$, где n – число испытаний, σ - известное среднее квадратическое отклонение, а t – аргумент функции Лапласа, при котором $\Phi(t) = \gamma/2$.
2) Случайная величина X распределена нормально с неизвестным σ. Воспользуемся формулой (4.3), из которой следует, что $\delta = \frac{t_\gamma s}{\sqrt{n}}$, где s – исправленное выборочное среднее квадратическое отклонение, а t_γ определяется по соответствующей таблице.
3) Если случайная величина распределена по иному закону, то при достаточно большом количестве испытаний ($n > 30$) можно использовать для оценки δ предыдущие формулы, так как при $n \to \infty$ распределение Стьюдента стремится к нормальному, и границы интервалов, полученные по формулам (4.1) и (4.3), различаются незначительно.

Разыгрывание случайных величин.

Определение 10.1. **Случайными числами** называют возможные значения r непрерывной случайной величины R, распределенной равномерно в интервале (0; 1).

1. Разыгрывание дискретной случайной величины.

Пусть требуется разыграть дискретную случайную величину X, то есть получить последовательность ее возможных значений, зная закон распределения X:

$$\begin{array}{cccccc} X & x_1 & x_2 & \dots & x_n \\ p & p_1 & p_2 & \dots & p_n \end{array}.$$

Рассмотрим равномерно распределенную в (0, 1) случайную величину R и разобьем интервал (0, 1) точками с координатами $p_1,\ p_1 + p_2, \ldots, p_1 + p_2 + \ldots + p_{n-1}$ на n частичных интервалов $\Delta_1, \Delta_2, ..., \Delta_n$, длины которых равны вероятностям с теми же индексами.

Теорема 10.1. Если каждому случайному числу $r_j (0 \le r_j < 1)$, которое попало в интервал Δ_i, ставить в соответствие возможное значение x_i, то разыгрываемая величина будет иметь заданный закон распределения:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n.

Доказательство.

Возможные значения полученной случайной величины совпадают с множеством x_1, x_2, … x_n, так как число интервалов равно n, а при попадании r_j в интервал Δ_i случайная величина может принимать только одно из значений x_1, x_2, … x_n.

Так как R распределена равномерно, то вероятность ее попадания в каждый интервал равна его длине, откуда следует, что каждому значению соответствует вероятность p_i. Таким образом, разыгрываемая случайная величина имеет заданный закон распределения.

Пример. Разыграть 10 значений дискретной случайной величины X, закон распределения которой имеет вид:

X	2	3	6	8
p	0,1	0,3	0,5	0,1

Решение. Разобьем интервал (0, 1) на частичные интервалы: Δ_1 (0; 0,1), Δ_2 (0,1; 0,4), Δ_3 (0,4; 0,9), Δ_4 (0,9; 1).

Выпишем из таблицы случайных чисел 10 чисел:

0,09; 0,73; 0,25; 0,33; 0,76; 0,52; 0,01; 0,35; 0,86; 0,34.

Первое и седьмое числа лежат на интервале Δ_1, следовательно, в этих случаях разыгрываемая случайная величина приняла значение $x_1 = 2$; третье, четвертое, восьмое и десятое числа попали в интервал Δ_2, что соответствует $x_2 = 3$; второе, пятое, шестое и девятое числа оказались в интервале Δ_3 – при этом $X = x_3 = 6$; на последний интервал не попало ни одного числа. Итак, разыгранные возможные значения X таковы: 2, 6, 3, 3, 6, 6, 2, 3, 6, 3.

2. Разыгрывание противоположных событий.

Пусть требуется разыграть испытания, в каждом из которых событие A появляется с известной вероятностью p. Рассмотрим дискретную случайную величину X, принимающую значения 1 (в случае, если событие A произошло) с вероятностью p и 0 (если A не произошло) с вероятностью $q = 1 – p$. Затем разыграем эту случайную величину так, как было предложено в предыдущем пункте.

Пример. Разыграть 10 испытаний, в каждом из которых событие A появляется с вероятностью 0,3.
Решение. Для случайной величины X с законом распределения

X	1	0
p	0,3	0,7

получим интервалы Δ_1 (0; 0,3) и Δ_2 (0,3; 1). Используем ту же выборку случайных чисел, что и в предыдущем примере, для которой в интервал Δ_1 попадают числа №№1,3 и 7, а остальные – в интервал Δ_2. Следовательно, можно считать, что событие A произошло в первом, третьем и седьмом испытаниях, а в остальных – не произошло.

3. Разыгрывание полной группы событий.

Если события $A_1, A_2, \ldots, A_n$, вероятности которых равны p_1 , p_2 ,… p_n, образуют полную группу, то для из разыгрывания (то есть моделирования последовательности их появлений в серии испытаний) можно разыграть дискретную случайную величину X с законом распределения X 1 2 … n, сделав это так же, как в пункте 1. При этом считаем, что

p p_1 p_2 … p_n

если X принимает значение $x_i = i$, то в данном испытании произошло событие A_i.

4. Разыгрывание непрерывной случайной величины.

а) Метод обратных функций.

Пусть требуется разыграть непрерывную случайную величину X, то есть получить последовательность ее возможных значений x_i ($i = 1, 2, \ldots, n$), зная функцию распределения $F(x)$.

Теорема 10.2. Если r_i – случайное число, то возможное значение x_i разыгрываемой непрерывной случайной величины X с заданной функцией распределения $F(x)$, соответствующее r_i , является корнем уравнения

$$F(x_i) = r_i. \qquad (10.1)$$

Доказательство.

Так как $F(x)$ монотонно возрастает в интервале от 0 до 1, то найдется (причем единственное) значение аргумента x_i , при котором функция распределения примет значение r_i . Значит, уравнение (10.1) имеет единственное решение: $x_i = F^{-1}(r_i)$, где F^{-1}- функция, обратная к F.

Докажем, что корень уравнения (10.1) является возможным значением рассматриваемой случайной величины X. Предположим вначале, что x_i – возможное значение некоторой случайной величины ξ, и докажем, что вероятность попадания ξ в интервал (c, d) равна $F(d) – F(c)$. Действительно, $c < x_i < d \Leftrightarrow F(c) < r_i < F(d)$ в силу монотонности $F(x)$ и того, что $F(x_i) = r_i$.

Тогда $c < \xi < d \Leftrightarrow F(c) < R < F(d)$, следовательно,

$p(c < \xi < d) = p(F(c) < R < F(d)) = F(d) - F(c)$.

Значит, вероятность попадания ξ в интервал (*c*, *d*) равна приращению функции распределения *F*(*x*) на этом интервале, следовательно, ξ = *X*.

Пример.

Разыграть 3 возможных значения непрерывной случайной величины *X*, распределенной равномерно в интервале (5; 8).

Решение.

F(*x*) = $\frac{x-5}{3}$, то есть требуется решить уравнение $\frac{x_i-5}{3}=r_i,\ x_i=3r_i+5.$

Выберем 3 случайных числа: 0,23; 0,09 и 0,56 и подставим их в это уравнение. Получим соответствующие возможные значения *X*:

$x_1=5{,}69;\ x_2=5{,}27;\ x_3=6{,}68.$

б) Метод суперпозиции.

Если функция распределения разыгрываемой случайной величины может быть представлена в виде линейной комбинации двух функций распределения:

$$F(x)=C_1F_1(x)+C_2F_2(x)\ \ (C_{1,2}>0)\,, \qquad (10.2)$$

то $C_1+C_2=1$, так как при $x\to\infty$ $F(x)\to 1$.

Введем вспомогательную дискретную случайную величину *Z* с законом распределения

Z	1	2
p	C_1	C_2

Выберем 2 независимых случайных числа r_1 и r_2 и разыграем возможное значение *Z* по числу r_1 (см. пункт 1). Если *Z* = 1, то ищем искомое возможное значение *X* из уравнения $F_1(x)=r_2$, а если *Z* = 2, то решаем уравнение $F_2(x)=r_2$.

Можно доказать, что при этом функция распределения разыгрываемой случайной величины равна заданной функции распределения.

в) Приближенное разыгрывание нормальной случайной величины.

Так как для *R*, равномерно распределенной в (0, 1), $M(R)=\frac{1}{2},\ D(R)=\frac{1}{12}$, то для суммы *n* независимых, равномерно распределенных в интервале (0,1) случайных величин $\sum_{j=1}^{n}R_j$ $\quad M\left(\sum_{j=1}^{n}R_j\right)=\frac{n}{2},\ D\left(\sum_{j=1}^{n}R_j\right)=\frac{n}{12},\ \sigma=\sqrt{\frac{n}{12}}$.

Тогда в силу центральной предельной теоремы нормированная случайная величина $\dfrac{\sum_{j=1}^{n}R_j-\frac{n}{2}}{\sqrt{\frac{n}{12}}}$ при $n\to\infty$ будет иметь распределение, близкое к

нормальному, с параметрами $a = 0$ и $\sigma = 1$. В частности, достаточно хорошее приближение получается при $n = 12$: $\sum_{j=1}^{12} R_j - 6$.

Итак, чтобы разыграть возможное значение нормированной нормальной случайной величины x, надо сложить 12 независимых случайных чисел и из суммы вычесть 6.