Лекция 1

ЗАКОН БОЛЬШИХ ЧИСЕЛ
ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

Свойство устойчивости массовых, случайных явлений известно человечеству еще с глубоких времен. В какой бы области оно не проявлялось, суть его сводится к следующему: конкретные особенности каждого отдельного случайного явления почти не сказывается на среднем результате массы таких явлений. Именно эта *устойчивость средних* и представляет собой физическое содержание "*закона больших чисел*", понимаемого в широком смысле слова: *при очень большом числе случайных явлений средний их результат практически перестает быть случайным и может быть предсказан с большой степенью определенности*.

Различные формы закона больших чисел вместе с различными формами центральной предельной теоремы образуют совокупность так называемых *предельных теорем* теории вероятности.

Предельные теоремы дают возможность не только осуществить научные прогнозы в области случайных явлений, но и оценивать точность этих прогнозов.

Неравенство Чебышева

Имеется случайная величина (СВ) $X \Rightarrow m_x, D_x$. Неравенство Чебышева утверждает, что каково бы ни было положительное число ε, вероятность того, что СВ X отклонится от своего математического ожидания не меньше, чем на ε, ограничена сверху величиной $\frac{D_x}{\varepsilon^2}$, т.е.

$$P(|X - m_x| > \varepsilon) \le \frac{D_x}{\varepsilon^2}. \qquad (1.1)$$

Доказательство: рассмотрим доказательство для непрерывной СВ X. По определению известно, что

$$P(|X - m_x| > \varepsilon) = \int\limits_{|X - m_x| > \varepsilon} \rho(x)dx. \qquad (1.2)$$

Запишем чему равна дисперсия СВ X:

$$D_x = \int\limits_{-\infty}^{\infty} (x - m_x)^2 \rho(x)dx \ge \int\limits_{|X - m_x| > \varepsilon} (x - m_x)^2 \rho(x)dx. \qquad (1.3)$$

Проиллюстрируем выражение (1.3) рисунком:

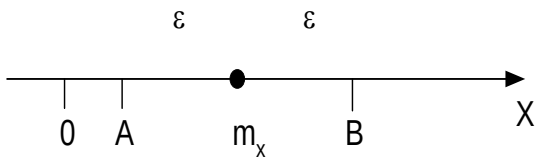


Рис. 1.1. Отрезок интегрирования

В (1.3) $|x - m_x| > \varepsilon$ означает, что интегрирование ведется на внешней части отрезка *AB*. Если в (1.3) принять, что $(x - m_x) = \varepsilon$ (отметим, что неравенство при этом только усиливается), получим:

$$D_x \geq \varepsilon^2 \int\limits_{|X - m_x| > \varepsilon} \rho(x)dx = \varepsilon^2 P(|X - m_x| > \varepsilon). \qquad (1.4)$$

Итак, неравенство Чебышева доказано. Неравенство Чебышева дает грубую оценку сверху и утверждает, что для любой случайной величины *X* вероятность того, что она отклонится на ε от $\overline{x}$ меньше, чем дисперсия деленная на ε^2.

Пример 1: Дана СВ $X \Rightarrow m_x, D_x = \sigma_x^2$. Оценить сверху вероятность того, что *X* отклонится от m_x не меньше, чем на 3σ.

Решение: Согласно неравенству Чебышева запишем

$$P(|X - m_x| > 3\sigma_x) \leq \frac{D_x}{9\sigma^2} = \frac{1}{9}.$$

Для случайной величины, распределенной по нормальному закону $P(|X - m_x| > 3\sigma_x) \cong 0.003.$

Теорема Чебышева

Пусть имеется СВ $X \Rightarrow m_x, D_x$. Над этой величиной производится n независимых опытов и вычисляется среднее арифметическое всех наблюденных значений случайной величины *X*. Необходимо найти характеристики среднего арифметического – *математическое ожидание и дисперсию*. В результате первого опыта СВ X приняла значение x_1, во втором опыте – x_2, ..., в n – ом опыте – x_n.

Рассмотрим среднее арифметическое этих значений:

$$Y = \frac{\sum\limits_{i=1}^{n} x_i}{n}. \qquad (1.5)$$

СВ $Y -$ линейная функция независимых случайных величин $x_1, \ldots, x_n$. Определим:

$$m_Y = M[Y] = \frac{1}{n}\sum_{i=1}^{n} M[x_i] = \frac{1}{n} n \cdot m_x = m_x, \qquad (1.6)$$

$$D_Y = \frac{1}{n^2}\sum_{i=1}^{n} D[x_i] = \frac{D_x}{n} = \frac{\sigma^2}{n}. \qquad (1.7)$$

Т. о. $m_Y -$ не зависит от числа опытов (n), а дисперсия при больших n может стать сколь угодно малой, т.е. СВ Y ведет себя почти не как случайная. Это свойство и устанавливает теорема Чебышева.

При достаточно большом числе независимых опытов среднее арифметическое наблюденных значений случайной величины сходится по вероятности к ее математическому ожиданию.

Говорят что СВ X_n сходится по вероятности к величине a, если при увеличении n вероятность того, что X_n и a будут сколь угодно близки, неограниченно приближается к единице, а это значит, что при достаточно большом n

$$P(|X_n - a| < \varepsilon) > 1 - \delta, \tag{1.8}$$

где ε и δ – произвольно малые положительные числа. Запишем аналогично теорему Чебышева:

$$P(|\frac{\sum_{i=1}^{n} x_i}{n} - m_x| < \varepsilon) > 1 - \delta. \tag{1.9}$$

Докажем справедливость (1.9).

Ранее было показано, что для $Y = \frac{\sum_{i=1}^{n} x_i}{n}$, $m_Y = m_x$, $D_Y = \frac{D_x}{n}$.

Применим к СВ Y неравенство Чебышева:

$$P(|Y - m_Y| \geq \varepsilon) \leq \frac{D_Y}{\varepsilon^2} = \frac{D_x}{n \cdot \varepsilon^2}. \tag{1.10}$$

Как бы ни было мало ε, всегда можно взять такое большое n, чтобы выполнялось неравенство: $\frac{D_x}{n \cdot \varepsilon^2} < \delta$, где $\delta -$ сколь угодно малое число. Тогда получим:

$$P(|\frac{\sum_{i=1}^{n} x_i}{n} - m_x| \geq \varepsilon) < \delta. \tag{1.11}$$

Запишем вероятность события противоположного (1.11)

$$P(|\frac{\sum_{i=1}^{n} x_i}{n} - m_x| < \varepsilon) > 1 - \delta. \tag{1.12}$$

Что и требовалось доказать.

Обобщенная теорема Чебышева

Пусть $X_1, X_2, \ldots, X_n -$ независимые случайные величины с соответствующими математическими ожиданиями и дисперсиями, т.е. $X_1 \Rightarrow m_{x_1}, D_{x_1}; X_2 \Rightarrow m_{x_2}, D_{x_2}; \ldots; X_n \Rightarrow m_{x_n}, D_{x_n}$. И если все дисперсии ограничены сверху одним и тем же числом L таким, что $D_{x_i} < L \quad (i = 1, 2, \ldots, n)$, то при возрастании n среднее арифметическое наблюденных значений величин $X_1, X_2, \ldots, X_n$ сходится по вероятности к среднему арифметическому их математических ожиданий:

$$P(|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{x_i}}{n}| < \varepsilon) > 1 - \delta. \qquad (1.13)$$

Доказательство: Рассмотрим СВ $Y = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$, с соответствующими характеристиками: $m_Y = \frac{1}{n} \sum_{i=1}^{n} m_{x_i}$, $D_Y = \frac{1}{n^2} \sum_{i=1}^{n} D_{x_i}$. Применим с СВ Y неравенство Чебышева

$$P(|Y - m_Y| \geq \varepsilon) \leq \frac{D_Y}{\varepsilon^2}, \text{ или}$$

$$P(|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{x_i}}{n}| \geq \varepsilon) < \frac{L}{n \cdot \varepsilon^2}. \qquad (1.14)$$

Как бы ни было мало ε, можно так выбрать n, что будет выполняться неравенство $\frac{L}{n \cdot \varepsilon^2} < \delta$. Тогда получим:

$$P(|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{x_i}}{n}| \geq \varepsilon) < \delta. \qquad (1.15)$$

Перейдем к противоположному событию:

$$P(|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{x_i}}{n}| < \varepsilon) > 1 - \delta. \qquad (1.16)$$

Что и требовалось доказать.

Теорема Маркова

Если имеются зависимые СВ $X_1, X_2, \ldots, X_n$ и, если при $n \to \infty$ справедливо соотношение $\frac{D[\sum_{i=1}^{n} X_i]}{n^2} \to 0,$ то среднее арифметическое наблюденных значений СВ $X_1, X_2, \ldots, X_n$ сходится по вероятности к среднему арифметическому их математических ожиданий.

Доказательство: Рассмотрим величину $Y = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow D_Y = \frac{1}{n^2} D\sum_{i=1}^{n} X_i$.

Применим к величине Y неравенство Чебышева:

$$P(|Y - m_Y| \geq \varepsilon) \leq \frac{D_Y}{\varepsilon^2}$$

Т.к. по условию теоремы при $n \to \infty$ $D_Y \to 0$, то при достаточно большом n справедливо $P(|Y - m_Y| \geq \varepsilon) < \delta$, или переходе к противоположному событию:

$$P(|Y - m_Y| < \varepsilon) = P(|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} m_{x_i}}{n}| < \varepsilon) > 1 - \delta. \qquad (1.17)$$

Что и требовалось доказать.

Теорема Бернулли

Пусть производится n независимых опытов, в каждом из которых может произойти или не произойти событие A, вероятность которого в каждом опыте равна p. Теорема Бернулли утверждает, что: *при неограниченном увеличении числа опытов n частота события A сходится по вероятности к его вероятности p.*

Обозначим частоту события A в n опытах через p^* и запишем теорему Бернулли в виде формулы:

$$P(|p^* - p| < \varepsilon) > 1 - \delta, \qquad (1.18)$$

где ε и δ – сколько угодные малые положительные числа. Требуется доказать неравенство (1.18) при достаточно большом n.

Доказательство: Рассмотрим независимые случайные величины:

X_1 – число появления события A в первом опыте;

X_2 – число появления события A во втором опыте;

.....

Все эти величины дискретные и имеют один и тот же закон распределения; выраженный рядом распределения:

x_i	0	1
p_i	q	p

Здесь $q = 1 - p$, нетрудно показать, что $m_x = p, \quad D_x = pq$. Частота p^* не что иное, как среднее арифметическое величин $X_1, X_2, \ldots, X_n$, т.е. $p^* = \frac{1}{n}\sum_{i=1}^{n} X_i$. Тогда согласно закону больших чисел p^* сходится по вероятности к общему математическому ожиданию m_x этих величин.

Центральная предельная теорема

Закон больших чисел устанавливает факт приближения средних большого числа случайных величин к *определенным постоянным*. Но этим не ограничиваются закономерности, возникающие в результате суммарного действия случайных величин. Оказывается, что при некоторых условиях совокупное действие случайных величин приводит к *определенному*, а именно – к *нормальному закону распределения*.

Центральная предельная теорема представляет собой группу теорем, посвященных установлению условий, при которых возникает нормальный закон распределения. Среди этих теорем важнейшее место принадлежит теореме Ляпунова.

Теорема Ляпунова *Если $X_1, X_2, \ldots, X_n$ – независимые случайные величины, у каждой из которых существует математическое ожидание $M[X_i] = a_i$, дисперсия $D[X_i] = \sigma_i^2$, абсолютный центральный момент третьего порядка $M\left[|X_i - a_i|^3\right] = m_i$ и*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} m_i}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{3/2}} = 0, \qquad (1.19)$$

то закон распределения суммы $Y_n = X_1 + X_2 + \ldots X_n$ при $n \to \infty$ неограниченно приближается к нормальному закону с математическим ожиданием $\sum_{i=1}^{n} a_i$ и дисперсией $\sum_{i=1}^{n} \sigma_i^2$.

<u>Теорему примем без доказательства.</u>

Смысл условия (1.19) состоит в том, чтобы в сумме $Y_n = \sum_{i=1}^{n} X_i$ не было слагаемых, влияние которых на рассеяние Y_n подавляюще велико по сравнению с влиянием всех остальных, а также не должно быть большого числа случайных слагаемых, влияние которых очень мало по сравнению с суммарным влиянием остальных. Таким образом, *удельный вес каждого*

отдельного слагаемого должен стремиться к нулю при увеличении числа слагаемых.

Следствие *Если* $X_1, X_2, \ldots, X_n -$ *независимые случайные величины, у которых существуют равные математические ожидания* $M\left[X_i\right] = a,$ *дисперсии* $D\left[X_i\right] = \sigma^2$ *и абсолютные центральные моменты третьего порядка* $M\left[\left|X_i - a_i\right|^3\right] = m_i \ \left(i = 1, 2, \ldots, n\right),$ *то закон распределения суммы* $Y_n = X_1 + X_2 + \ldots + X_n$ *при* $n \to \infty$ *неограниченно приближается к нормальному закону.*

Замечание Опираясь на центральную предельную теорему, можно утверждать, что рассмотренные нами ранее случайные величины, имеющие законы распределения – биномиальный, Пуассона, равномерный, χ^2 ("хи-квадрат"), t (Стьюдента), при $n \to \infty$ распределены асимптотически нормально.

Пример 2: Определить вероятность того, что средняя продолжительность 100 производственных операций окажется в пределах от 46 до 49 с, если математическое ожидание одной операции равно 47.4 с, а среднее квадратичное отклонение – 4.9 с.

РЕШЕНИЕ: В этой задаче $X -$ продолжительность наугад взятой производственной операции, $a = M\left[X\right] = 47.4c,$ $\sigma = \sqrt{D[X]} = 4.9c,$ $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$ $n = 100.$ Здесь $\bar{X} -$ средняя продолжительность 100 наугад взятых производственных операций, причем $M\left[\bar{X}\right] = 47.4c,$ $D\left[\bar{X}\right] = \sigma^2 / n = 4.9^2 / 100c^2.$ Теперь вычислим соответствующую вероятность

$$P\left(46 \le \bar{X} \le 49\right) = P\left(-\frac{1.4}{0.49} \le \frac{\bar{X} - 47.4}{0.49} \le \frac{1.6}{0.49}\right) = P\left(-2.857 \le Z \le 3.265\right) =$$

$$= \Phi^*\left(3.265\right) + \Phi^*\left(2.857\right) - 1 = 0.9973.$$

При решении задачи на основании следствия из теоремы Ляпунова распределение центрированного и нормированного среднего $(\bar{X} - 47.4) / 0.49$ *приближенно заменено на распределение стандартной нормальной случайной величины* $Z \sim N(0, 1)$.

ЛЕКЦИЯ №2

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика – это наука, изучающая методы сбора, систематизации и интерпретации числовых (случайных) данных.

В этом определении интерпретация и систематизация данных рассматривается как существенный аспект.

Главная цель статистики – получение осмысленных заключений из несогласованных (подверженных разбросу) данных.

Действительно, исключая тривиальные ситуации, реальные данные всегда являются несогласованными, что требует применения статистических методов. Рассогласованность (разброс) между индивидуальными наблюдениями может быть, например, обусловлена ошибкой при считывании позиции стрелки прибора, когда она расположена между двумя делениями шкалы стрелочного прибора. Изменчивость может быть также следствием нестабильности работы электронного оборудования при передаче сообщений по радио или телеграфу. (В последнем случае для характеристики ситуации используется термин "шум").

Чем же, конкретно занимается математическая статистика?

Какие задачи решает?

1. *Выборочные распределения*

Статистика должна получить свои выводы, используя наличную выборку. Каждое наблюдение является реализацией некоторой случайной величины. Известно множество значений, которые может принимать случайная величина; некоторые из них имеют большую возможность появления, чем другие.

Значение, которое наблюдалось, представляет собой **реализацию**. Вероятности возможных реализаций характеризуются **распределением вероятностей случайных величин** (СВ). Обычно функции распределения вероятностей бывают заданы с точностью до одного двух параметров, значений некоторых неизвестных. Это приводит к проблеме поиска таких комбинаций выборочных значений, которые бы давали наилучшее приближение для неизвестных параметров. Каждая такая комбинация и есть *статистика*. Выборочное распределение статистики позволяет судить, может ли предложенная статистика служить оценкой интересующего нас параметра.

2.*Оценки, тесты* (*критерии, значимости*), *решения*

Проблема оценивания была схематично рассмотрена выше. Ясно, что разумная процедура оценивания не должна ограничиваться лишь выбором приближенного численного значения для неизвестного параметра; она

должна что-то говорить и о надежности этого приближения. Обычно говорят о *точечном оценивании* и *об интервальном оценивании*.

Существуют различные методы конструирования "точечных" оценок и определения их надежности. Наиболее полезным из них является **метод максимального правдоподобия**. Другой известный метод, который можно рассматривать либо как специальный случай ММП (метод максимального правдоподобия), либо как независимая процедура подгонки – **метод наименьших квадратов**.

Интервальное оценивание – связано с определением "доверительных интервалов", правдоподобных интервалов, байесовских интервалов.

Поскольку статистика в целом основана на случайной изменчивости, каждая оценка подвержена ошибке. Так, если получены две различные оценки параметра – одна при одном наборе условий, а другая – при другом, непосредственно неясно, соответствует ли имеющееся между ними различие различию между параметрами.

Вопрос об их различии решается с помощью статистического критерия (теста) или критерия значимости.

Один из подходов к статистическим критериям (проверки гипотез) связан с именем Р.А Фишера, который рассматривает проверку гипотезы как пробный шаг в проведение научного исследования, позволяющий получить ученому объектный критерий, с помощью которого можно судить об истинности гипотезы.

Другой подход связан, в основном, с именами Дж. Неймана и Э. Пирсона, которые рассматривают процедуру проверки гипотезы, как правило, с помощью которого должен быть сделан выбор, либо принято решение об истинности одной гипотезы в противоречии другой.

Одна из частных проблем теории проверки статистических гипотез – оценка пригодности вероятной модели, предложенной для объяснения (интерпретации) данных. При этом необходимо решить: насколько предложенная модель соответствует выборке? И является ли выборочные значения действительно близкими к тем, которые можно ожидать, используя подогнанную модель? Наиболее широко для решения подобных вопросов применяется процедура, предложенная Карлом Пирсоном, и использующая критерий, основанный на ее выборочном распределении. Это пирсоновский критерий согласия *хи-квадрат*.

Выборочные распределения

Статистическая устойчивость случайных явлений проявляется лишь при большом (в пределе – бесконечно большом) числе наблюдений. Однако на практике реальное число наблюдений ограничено. Поэтому характеристики случайных величин (СВ), определенные по малому числу наблюдений, в принципе не должны совпадать с величинами тех же характеристик, определенными по большому числу наблюдений (условия опыта остаются

неизменными). Чтобы провести различие между характеристиками СВ, найденными по достаточно большому и малому числу наблюдений, в математической статистике введено понятие абстрактной *генеральной совокупности* и *выборки*.

Генеральной совокупностью случайной величины ξ называется множество всех значений, которые может принимать случайная величина ξ.

Выборка представляет собой совокупность ограниченного числа наблюдений.

В соответствии с этим различают выборочные характеристики СВ, найденные по ограниченному числу наблюдений (выборке) и зависящие от числа наблюдений, и соответствующие им характеристики в генеральной совокупности, не зависящие от числа наблюдений. При этом выборочные характеристики рассматриваются как оценки соответствующих характеристик в генеральной совокупности.

На практике во многих случаях функция распределения рассматриваемой случайной величины ξ неизвестна; ее определяют по результатам наблюдений, или как говорят, по *выборке*.

Выборкой объемом n для данной случайной величины ξ называется последовательность $X_1, X_2, \ldots, X_n - n$ независимых наблюдений этой величины.

Пусть из генеральной совокупности извлечена выборка, причем

$X_1 -$ наблюдалось ν_1 раз;

$X_2 -$ наблюдалось ν_2 раз;

$X_k -$ наблюдалось ν_k раз;

Объём выборки – $n = \sum_{i=1}^{k} \nu_i$. Наблюдаемые значения X_i называют **вариантами**, а последовательность вариант, записанных в возрастающем порядке – **вариационным рядом.**

Число наблюдений называют **частотами**, а их отношение к объему выборки: $\frac{\nu_i}{n} = w_i -$ **относительными частотами (частостями).**

В статистике различают *малые* и *большие выборки*.

Малой выборкой считают такую выборку, при обработке которой методами, основанными на группировании наблюдений, нельзя достичь заданных точности и достоверности.

Большой считают такую выборку, при обработке которой можно перейти к группированию наблюдений без ощутимой потери информации и достижением заданных значений точности и достоверности.

Если выборка достаточно велика, то построенный на ее основе вариационный ряд неудобен для дальнейшего статистического анализа. В

этом случае строится так называемый ***группированный статистический ряд***.

Группирование данных, гистограмма, полигон

При группировании данных необходимо соблюдать определенные правила. Рассмотрим наиболее важные из них:

1. Объем выборки должен быть достаточно велик $(n \geq 50)$.
2. Число интервалов группирования m (число групп) должно находиться в интервале $5 \leq m \leq 20$. При выборе m в каждом конкретном случае следует помнить, что при малом числе групп определение вида теоретической кривой распределения по эмпирическим данным может быть затруднено из-за маскировки (утраты) резких изменений кривой распределения, если они фактически имели место. При большом числе групп и незначительном объеме выборки будет наблюдаться большое количество пропусков (ноль попаданий в группу), что будет обусловлено не столько видом распределения, сколько недостатком статистики, кроме того, в этом случае даже небольшие случайные колебания приводят к искажению кривой распределения.
3. Необходимо, по возможности, охватывать всю область данных, т.к. при неизвестных предельных значениях невозможно вычислить некоторые числовые характеристики выборки.
4. Интервалы не должны перекрываться. Не должно возникать никаких сомнений относительно того, в какой интервал попадает любое значение.
5. Если заведомо известно, что теоретическая кривая может быть двумодальной, число групп может быть увеличено в 1,5 – 2 раза по сравнению с оптимальным числом m.

Оптимальное число групп m выборки объемом n рассчитывается по формулам:

- при известном значении $\varepsilon = \dfrac{\mu_4}{\sigma^4} \Rightarrow m = \dfrac{\varepsilon^{0.8} \cdot n^{0.4}}{3}$; (2.1)
- при неизвестном значении ε, но известно, что $1.8 \leq \varepsilon \leq 6$:

$$0.55n^4 \leq m \leq 1.25n^{0.4}; \qquad (2.2)$$

- согласно формуле "Стерджесса":

$$m = 1 + \log_2 n. \qquad (2.3)$$

Отсюда (2.3) видно, что для увеличения оптимального количества интервалов на единицу необходимо увеличить объем выборки вдвое.

Шаг группирования (ширина интервала) h определяется по формуле:

$$h = \frac{X_{\max} - X_{\min}}{m}. \qquad (2.4)$$

Для графического изображения вариационных рядов наиболее часто используются *полигон, гистограмма и кумулятивная кривая*.

Гистограммой распределения, или просто ***гистограммой*** называется чертеж в прямоугольной системе координат, горизонтальная ось которого разбивается на m равных интервалов (групп) шириной h. На каждом отрезке, как на основании, строится прямоугольник с высотой, равной частоте (частости) ν_i (ω_i) соответствующего интервала.

Полигоном распределения, или просто ***полигоном***, называется ломаная линия, соединяющая середины верхних оснований каждого столбца гистограммы. За пределами гистограммы, как слева, так и справа, размещают пустые интервалы, в которых точки, соответствующие их серединам, лежат на оси абсцисс.

Кумулятивная кривая (***кумулята***) – кривая накопления частот (частостей). Для дискретного ряда кумулята представляет ломанную соединяющую точки $(x_i, \sum_{k=1}^{i} \nu_k)$ или $(x_i, \sum_{k=1}^{i} \omega_k)$. Для интервального вариационного ряда ломанная начинается с точки, абсцисса которой равна началу первого интервала, а ордината – накопленной частоте (частости), равной нулю. Остальные точки этой ломанной соответствуют концам интервалов.

Пример: Построить полигон, гистограмму и кумуляту по выборке объема $n = 100$. Сгруппированные данные приведены в таблице.

Интервалы	[0, 1]	[1, 2]	[2, 3]	[3, 4]	[4, 5]	[5, 6]	[6, 7]	[7, 8]
Частоты	2	7	14	28	22	20	6	1

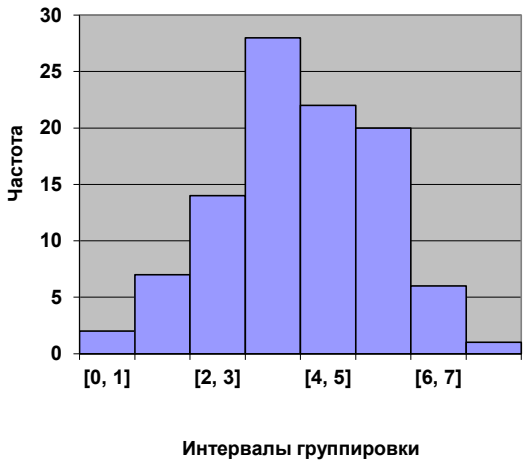


Рис. 2.1. Гистограмма частот

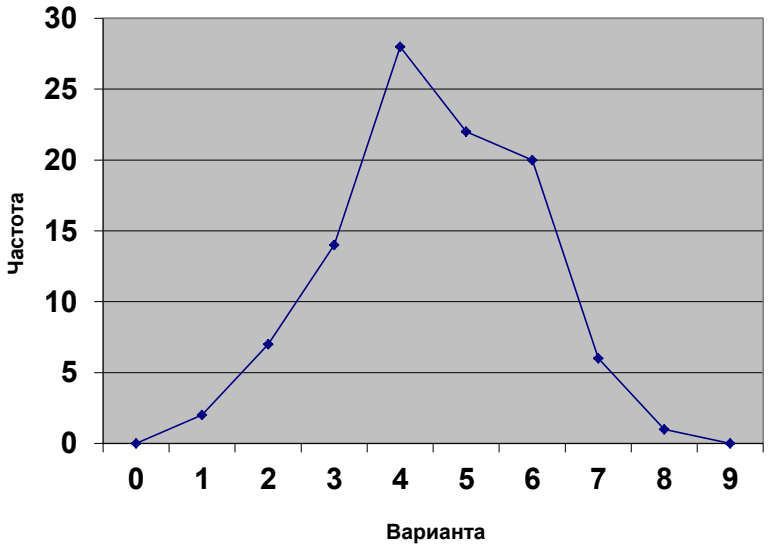


Рис. 2.2. Полигон

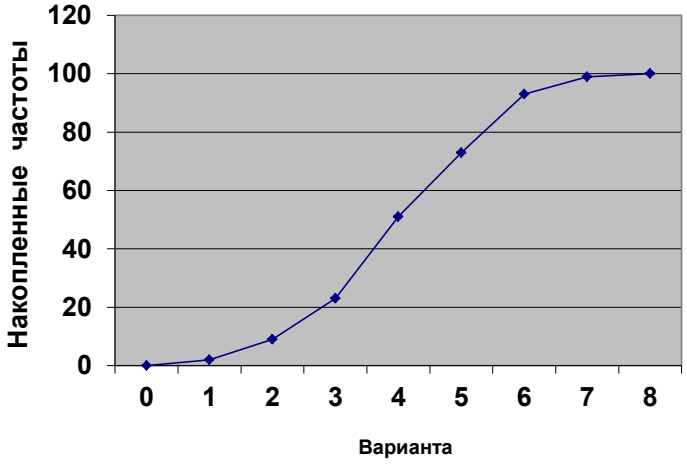


Рис. 2.3. Кумулята

ЛЕКЦИЯ №3

СТАТИСТИЧЕСКАЯ (ЭМПИРИЧЕСКАЯ) ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот (*частостей*).

В теории вероятностей под распределением понимают соответствие между возможными значениями случайной величины и их вероятностями, а в математической статистике – соответствие между наблюдаемыми значениями и их частотами, или относительными частотами.

Пример 1. Задана выборка объемом $n = 20$ с соответствующими частотами. Необходимо найти частости (относительные частоты).

x_i	2	6	12
ν_i	3	10	7
ω_i	3/20	10/20	7/20

Контроль: $$\sum_{i=1}^{3} \omega_i = \frac{3}{20} + \frac{10}{20} + \frac{7}{20} = 1.$$

Пусть, исследуется статистическое распределение, частот количественного признака (случайной величины) X. Введем обозначение:

$\nu_x -$ число наблюдений, при которых наблюдалось значение признака меньшее x;

$n -$ общее число наблюдений (объем выборки).

Очевидно, что относительная частота (частость) события $X < x$ равна $\omega_x = \nu_x / n$.

Статистической функцией распределения случайной величины X называется функция, определяющая для каждого значения x относительную частоту события $X < x$

$$P(X < x) = F^*(x) = \frac{\nu_x}{n} = \omega_x. \qquad (3.1)$$

Сравним статистическую и интегральную функции распределения. Вспомним (*теорема Бернулли*), *что относительная частота события* $X < x$, *т.е.* $F^*(x)$ *стремится по вероятности к вероятности* $F(x)$ *этого события.*

Функция $F^*(x)$ обладает теми же свойствами, что и $F(x)$:

1. Значения $F^*(x) \in [0;1]$.
2. Эмпирическая функция распределения $F^*(x) -$ неубывающая.

3. Если x_1 – наименьшая варианта, то $F^*(x)=0$ при $x \le x_1$.
4. Если x_k – наибольшая варианта, то $F^*(x)=1$ при $x > x_k$.

Пример 2. Построить эмпирическую функцию по данной выборке:

x_i	2	6	10
ν_i	12	18	30

Решение: Найдем объем выборки $n=12+18+30=60$. Теперь найдем статистическую функцию распределения:

x_i	2	6	10	>10
$F^*(x_i)$	0	12 / 60	30 / 60	1

Представим $F^*(x)$ в аналитическом и графическом виде:

$$F^*(x)=\begin{cases} 0, & при \quad x \le 2; \\ 0.2, & при \quad 2 < x \le 6; \\ 0.5, & при \quad 6 < x \le 10; \\ 1, & при \quad x > 10. \end{cases}$$

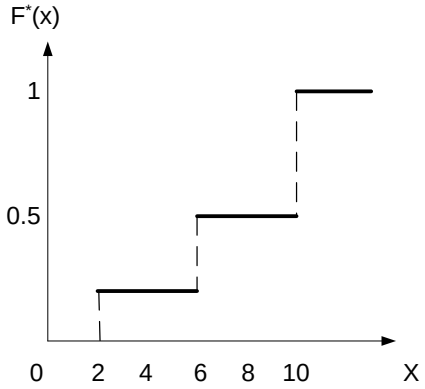


Рис. 3.1. Статистическая функция распределения

Выборочные значения и оценка параметров.

Рассмотрим один из возможных методов (есть и другие) оценивания среднего значения и дисперсии случайной величины x по n независимым наблюдением:

$$\bar{x}=\frac{\sum_{i=1}^{n} x_i}{n}; \tag{3.2}$$

$$S_b^2=\sigma_x^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=\frac{\sum_{i=1}^{n} x_i^2-\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}. \tag{3.3}$$

Здесь $\bar{x}$ и S_b^2 – выборочное среднее и выборочная дисперсия соответственно. Индекс в формуле S_b^2 (см. 3.3) указывает на *смещённость* оценки дисперсии. Наряду с вышеприведенными характеристиками, при обработке результатов наблюдений обычно находят следующие оценки:

- выборочная дисперсия (несмещённая)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}; \qquad (3.4)$$

- среднее квадратическое отклонение

$$S = \sqrt{S^2}; \qquad (3.5)$$

- выборочный коэффициент асимметрии

$$Sk = \frac{\mu_3}{S^3}; \quad \mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3; \qquad (3.6)$$

- выборочный коэффициент эксцесса

$$Ex = \frac{\mu_4}{S^4} - 3; \quad \mu_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4. \qquad (3.7)$$

Для установления качества или "правильности" любой оценки используются *свойства* (требования) "*хороших оценок*".

Требования "хороших оценок"

1. **Несмещённость.**

Во-первых, желательно, чтобы математическое ожидание оценки равнялось оцениваемому параметру:

$$M[\hat{\varphi}] = \varphi. \qquad (3.8)$$

Здесь $\hat{\varphi}$ – оценка параметра φ. Если свойство (3.8) имеет место, то оценка называется *несмещённой*.

2. **Эффективность.**

Во-вторых, желательно, чтобы среднеквадратическая ошибка данной оценки была наименьшей среди всех возможных оценок, т.е.:

$$M[(\hat{\varphi}_1 - \varphi)^2] \le M[(\hat{\varphi}_i - \varphi)^2]. \qquad (3.9)$$

Где $\hat{\varphi}_1$ – исследуемая оценка, а $\hat{\varphi}_i$ – любая другая оценка. Если это свойство имеет место, то оценка $\hat{\varphi}_1$ называется *эффективной*.

3. **Состоятельность.**

В-третьих, желательно, чтобы оценка сходилась к оцениваемому параметру с вероятностью, стремящейся к единице по мере увеличения размера выборки, т.е. для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P[|\hat{\varphi} - \varphi| \ge \varepsilon] = 0. \qquad (3.10)$$

Если выполнено условие (3.10), то оценка называется *состоятельной*. Из неравенства Чебышева следует, что достаточным для выполнения (3.10) является условие:

$$\lim_{n\to\infty} M[(\hat{\varphi}-\varphi)^2]=0 \qquad (3.11)$$

В качестве примера "хорошей оценки" рассмотрим оценку среднего значения (3.2). Математическое ожидание выборочного среднего $\bar{x}$ равно:

$$M[\bar{x}]=M\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]=\frac{1}{n}(n\cdot m_x)=m_x. \qquad (3.12)$$

Следовательно, согласно (3.8), оценка $m_x=\bar{x}-$ ***несмещённая***.

Среднеквадратическая ошибка выборочного среднего $\bar{x}$ равна:

$$M\left[(\bar{x}-m_x)^2\right]=M\left[(\frac{1}{n}\cdot\sum_{i=1}^{n}x_i-m_x)^2\right]=\frac{1}{n^2}M\left[\left(\sum_{i=1}^{n}(x_i-m_x)\right)^2\right]. \qquad (3.13)$$

Поскольку наблюдения x_i- независимы, то математическое ожидание членов, содержащих смешанные произведения, равны нулю. Поэтому из (3.13) получим:

$$M\left[(\bar{x}-m_x)^2\right]=\frac{1}{n^2}M\left[\sum_{i=1}^{n}(x_i-m_x)^2\right]=\frac{1}{n^2}\left(n\sigma_x^2\right)=\frac{\sigma_x^2}{n}. \qquad (3.14)$$

Т.о., согласно (3.11) оценка $m_x=\bar{x}-$ ***состоятельная***. Можно показать, что эта оценка эффективна.

Рассмотрим оценку дисперсии по формуле (3.3).

$$M\left[S_b^2\right]=M\left[\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{1}{n}M\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]. \qquad (3.15)$$

Однако

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}\left((x_i-m_x)+(m_x-\bar{x})\right)^2=\\
&=\sum_{i=1}^{n}(x_i-m_x)^2-2(\bar{x}-m_x)\sum_{i=1}^{n}(x_i-m_x)+\sum_{i=1}^{n}(\bar{x}-m_x)^2=\\
&=\sum_{i=1}^{n}(x_i-m_x)^2-2(\bar{x}-m_x)\cdot n\cdot(\bar{x}-m_x)+n\cdot(\bar{x}-m_x)^2=\\
&=\sum_{i=1}^{n}(x_i-m_x)^2-n\cdot(\bar{x}-m_x)^2.
\end{aligned} \qquad (3.16)$$

Поскольку $M[(x_i-m_x)^2]=\sigma^2$ и $M[(\bar{x}-m_x)^2]=\sigma_x^2/n$, то, подставив в (3.16), получим:

$$M\left[S_b^2\right]=\frac{1}{n}\left[n\cdot\sigma_x^2-\sigma_x^2\right]=\frac{(n-1)\cdot\sigma_x^2}{n}. \qquad (3.17)$$

Следовательно, оценка $\sigma_x^2 = S_b^2$ – ***смещённая***.

Хотя оценка (выборочная дисперсия) S_b^2 и является смещённой оценкой, эта оценка *состоятельна* и *эффективна*. Из (3.17) понятно, что для получения несмещённой оценки σ_x^2 следует взять несколько видоизмененную выборочную дисперсию (3.4).

ЛЕКЦИЯ №4

ИНТЕРВАЛНОЕ ОЦЕНИВАНИЕ

Ранее мы обсудили использование выборочных значений в качестве оценок параметров случайных величин. Однако такие процедуры дают только точечные оценки интересующих нас параметров и не позволяют судить о степени близости выборочных значений к оцениваемому параметру. Более предпочтительная процедура – построения интервала, который накрывает оцениваемый параметр с известной степенью достоверности. Такой подход называется "***интервальным оцениванием***". Сразу отметим следующее: чем больше уверенность в том, что оцениваемый параметр лежит в интервале, тем шире интервал. Так что искать интервал, накрывающий параметр с вероятностью равной единице, бессмысленно. Это вся область $\Theta \subseteq R$, т.е. $I=(-\infty,+\infty)$.

Пусть для параметра a получена *несмещённая* оценка $\tilde{a}$. Мы хотим оценить возможную при этом ошибку. Назначим некоторую достаточно большую вероятность β (например: $\beta = 0.9, 0.95, \ldots$), такую, что событие с вероятностью β можно считать практически достоверным и найдем такое значение ε, для которого выполняется соотношение

$$P(|\tilde{a}-a|<\varepsilon)=\beta. \qquad (4.1)$$

Тогда диапазон практически возможных значений ошибки, возникающий при замене a на $\tilde{a}$, будет равен $\pm\varepsilon$. Ошибки большие по абсолютной величине ε будут появляться с малой вероятностью $\alpha = 1-\beta$. Запишем (4.1) в другом виде:

$$P(\tilde{a}-\varepsilon<a<\tilde{a}+\varepsilon)=\beta. \qquad (4.2)$$

Т.е. с вероятностью β неизвестное значение параметра a попадает в интервал

$$I_\beta=(\tilde{a}-\varepsilon;\tilde{a}+\varepsilon). \qquad (4.3)$$

Ранее (в теории вероятностей) мы рассматривали вероятность попадания случайной величины на некоторый интервал. У нас же a не случайная величина, а интервал – случаен, т.е. здесь корректно говорить о вероятности I_β накрыть точку a.

Вероятность β принято называть ***доверительной вероятностью,*** а интервал I_β – ***доверительным интервалом.***

Перейдем к вопросу о нахождении доверительных границ a_1 и a_2 параметра a, имеющего несмещенную оценку $\tilde{a}$. Если бы нам был известен закон распределения величины $\tilde{a}$, то из выражения (4.1) нахождение ε при заданной β не представляло бы затруднений. Однако, как правило, мы не знаем закон распределения случайной величины X.

Пусть теперь распределение случайной величины X отлично от нормального. Применяя центральную предельную теорему, получаем следующий результат.

С увеличением объема выборки n выборочное распределение выборочного среднего $\bar{x}$ стремится к нормальному распределению независимо от вида распределения исходной случайной величины.

Практически во многих случаях выборочное $\bar{x}$ можно считать нормальным уже при $n > 4$, а при $n > 10$ приближение будет очень хорошим.

В качестве примера рассмотрим задачу о доверительном интервале математического ожидания: Пусть произведено n независимых опытов над случайной величины X с неизвестными m_x, D_x. Для этих параметров выберем оценки:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i; \qquad S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}). \tag{4.4}$$

Необходимо построить доверительный интервал I_β, соответствующий доверительной вероятности β:

$$P(|\bar{x} - m_x| < \varepsilon_\beta) = \beta. \tag{4 .5}$$

Интервальная оценка математического ожидания при известной дисперсии

Пусть СВ X имеет гауссово распределение с параметрами $N(m_x, \sigma_x)$, причём m_x – неизвестно, а значение σ_x^2 – известно. Тогда *эффективной оценкой* параметра m_x будет $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$.

При этом $\bar{x}-$ имеет нормальное распределение $N\left(m_x, \frac{\sigma_x}{\sqrt{n}}\right)$.

Статистика (оценка) СВ $z = \frac{\bar{x} - m_x}{\sigma_x / \sqrt{n}}$ имеет распределение $N(0,1)$, независимо от параметра m_x и как функция $\alpha-$ непрерывна и монотонна. Вспомним, что $\alpha = 1 - \beta$. Тогда, с учетом (4.2) запишем:

$$P\left(U_{\alpha/2} < z < U_{1-\alpha/2}\right) = 1 - \alpha = \beta. \tag{4.6}$$

Здесь $U_{\alpha/2}$ и $U_{1-\alpha/2}-$ квантили стандартного нормального распределения $N(0,1)$, обладающие свойством: $U_{\alpha/2} = -U_{1-\alpha/2}$. Поставив z в явном виде в (4.6) получим:

$$U_{\alpha/2} < \frac{\bar{x} - m_x}{\sigma_x} \cdot \sqrt{n} < U_{1-\alpha/2}. \tag{4.7}$$

Запишем это неравенство относительно m_x:

$$\bar{x} - U_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}} < m_x < \bar{x} + U_{1-\alpha/2} \cdot \frac{\sigma_x}{\sqrt{n}}. \tag{4.8}$$

Квантили стандартного нормального распределения определяются по таблицам, тогда окончательно получим:

$$\varepsilon = U_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}. \tag{4.9}$$

Искомый *доверительный интервал* математического ожидания нормально распределенной СВ с известной дисперсией равен:

$$I_\beta = (\bar{x} - U_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}, \bar{x} + U_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}). \tag{4.10}$$

На рис.4.1. представлена плотность распределения стандартного нормального распределения с отмеченными квантилями $\alpha = 0.05$.

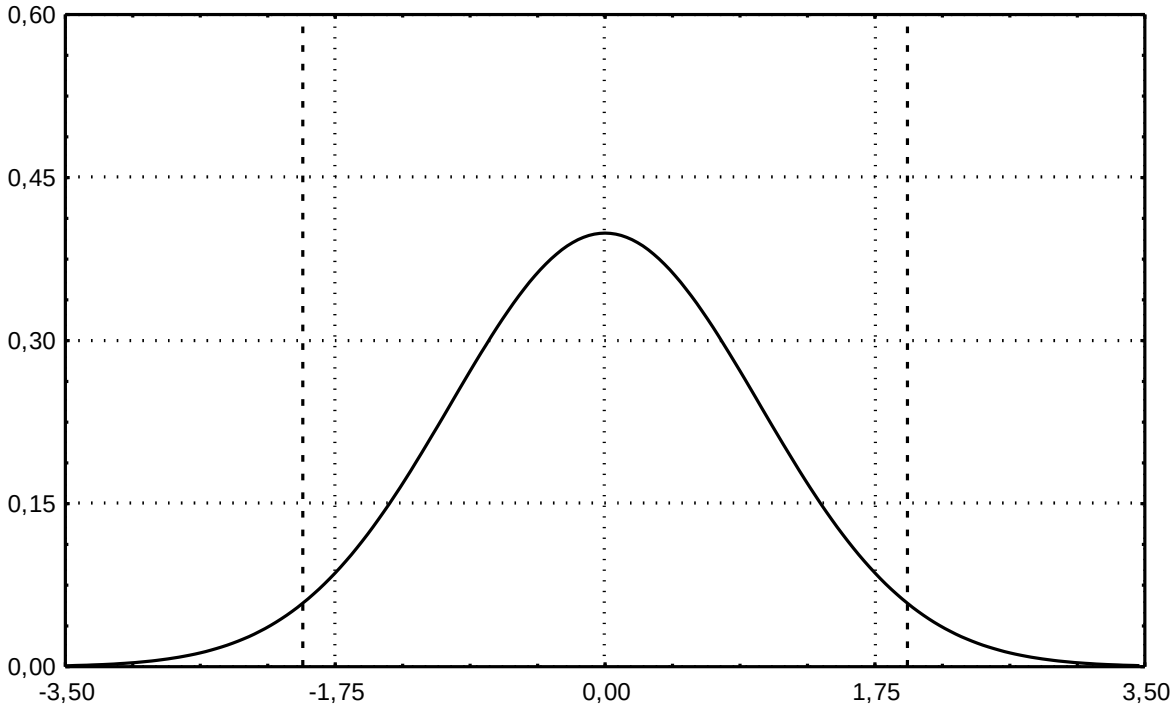


Рис. 4.1. Гауссово (нормальное) распределение
$(N(0,1), \beta = 0.95, U_{\alpha/2} = \pm 1.96)$

Интервальная оценка математического ожидания при неизвестной дисперсии

На практике почти всегда генеральная дисперсия σ_x^2, (как и оцениваемое математическое ожидание m_x) неизвестна. Итак, имеется нормально распределенная СВ $X \Rightarrow N(m_x, \sigma_x)$, с неизвестными параметрами m_x и σ_x^2. По случайной выборке найдем *несмещённые, эффективные* оценки: $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$; $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. Построение интервальной оценки основано на статистике:

$$t_{n-1} = \frac{\bar{x} - m_x}{S/\sqrt{n}}. \tag{4.11}$$

Вспомним, что $S = \sqrt{\frac{n}{n-1}} \cdot S_b$, и подставим в (4.11):

$$t_{n-1} = \frac{\bar{x} - m_x}{S_b} \sqrt{n-1} = \frac{(\bar{x} - m_x) / \frac{\sigma_x}{\sqrt{n}}}{\sqrt{\frac{1}{n-1} \cdot \frac{n \cdot S_b^2}{\sigma_x^2}}}. \tag{4.12}$$

Числитель выражения (4.12), как было показано выше, имеет стандартное нормальное распределение $N(0,1)$. Показано, что величина $n \cdot S_b^2 / \sigma_x^2$ имеет χ^2- распределение с $k = n-1$ степенями свободы. А статистика $t_{n-1}-$ имеет распределение Стьюдента с $k = n-1$ степенями свободы. Распределение Стьюдента не зависит от неизвестных параметров распределения случайной величины X, а зависит лишь от числа k, называемого *числом степеней свободы*.

Следует отметить, что $t-$ распределение Стьюдента напоминает нормальное распределение, и при $k \to \infty$ сколь угодно близко приближается к нему.

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

Так, например, для распределения $t-$ статистики число степеней свободы $k = n-1$, т.к. одна степень свободы "теряется" при определении выборочного среднего $\bar{x}$ (n наблюдений связаны одним уравнением).

Таким образом, по аналогии с (4.6) запишем:

$$P\left(t_{\alpha/2} < t_{n-1} < t_{1-\alpha/2}\right) = 1 - \alpha = \beta. \tag{4.13}$$

$$t_{\alpha/2} < \frac{\bar{x} - m_x}{S} \cdot \sqrt{n} < t_{1-\alpha/2}. \tag{4.14}$$

$$\bar{x} - t_{1-\alpha/2} \frac{S}{\sqrt{n}} < m_x < \bar{x} + t_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}}. \tag{4.15}$$

На рис.4.2. представлена плотность распределения Стьюдента с пятнадцатью степенями свободы.

Доверительный интервал математического ожидания нормально распределенной СВ с неизвестной дисперсией равен:

$$I_\beta = (\bar{x} - t_{1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2} \frac{S}{\sqrt{n}}). \tag{4.16}$$

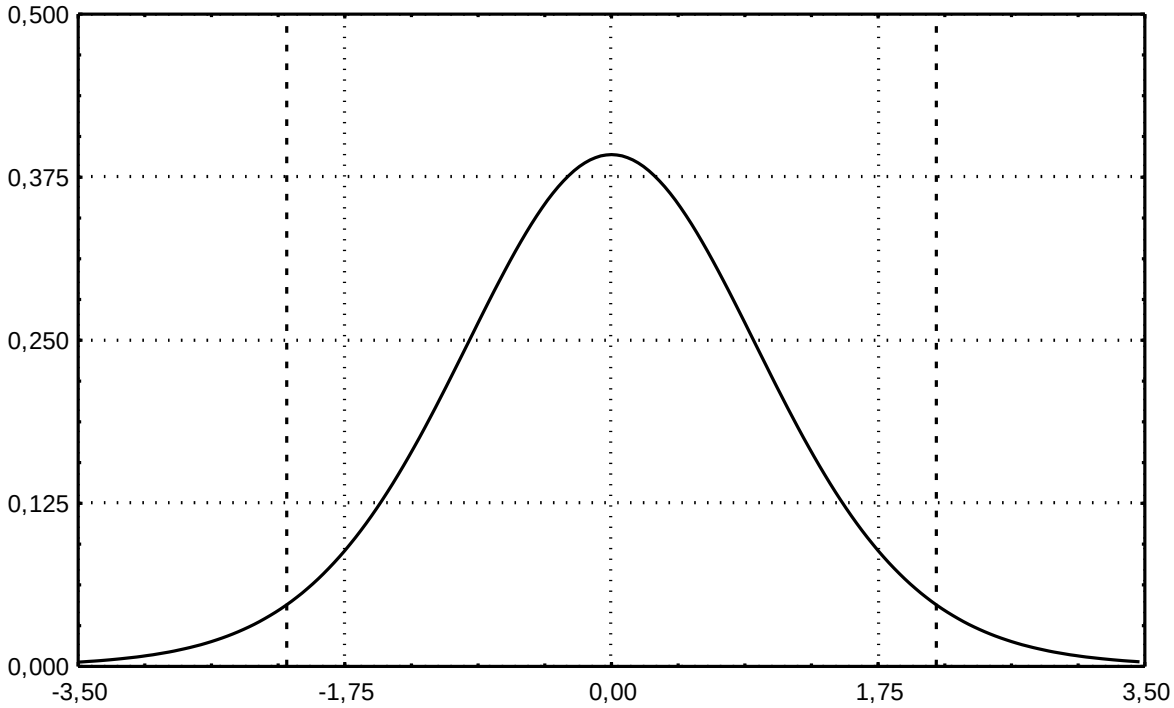


Рис. 4.2. Распределение Стьюдента ($k = 15, \beta = 0.95, t_{15,\alpha/2} = \pm 2.131$)

ЛЕКЦИЯ №5

Интервальная оценка выборочной дисперсии

Доверительный интервал для оценки дисперсии по выборочной дисперсии S^2 для СВ $X \Rightarrow N(m_x, \sigma_x)$ строится аналогичным образом. Естественно, что в качестве математического ожидания и дисперсии гауссовой СВ мы возьмем их *несмещённые* и *эффективные* оценки: $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i; \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. Исходя из вышесказанного, запишем:

$$P\left(\sigma_1^2 < \sigma^2 < \sigma_2^2\right) = \beta, \tag{5.1}$$

$$I_\beta\left(\sigma_1^2; \sigma_2^2\right). \tag{5.2}$$

Это интервал, который с вероятностью β накрывает неизвестную дисперсию.

Из статистики известно, что если СВ X имеет гауссово распределение $N(m_x, \sigma_x)$, а СВ $\bar{x} \Rightarrow N(m_x, \sigma_x / \sqrt{n})$, то справедливо соотношение:

$$(n-1)S^2 = \sigma^2 \chi_{n-1}^2. \tag{5.3}$$

Здесь $\chi_{n-1}^2 -$ хи-квадрат распределения с $n-1$ степенями свободы. Теперь задавая β или, что равносильно α, можно найти квантили

(соответствующие) $\chi^2_{n-1}(\beta)$. При этом следует учесть, что распределение χ^2_{n-1} – не симметрично (см. рис. 5.1.).

Как же решить эту задачу, однозначно? Ведь сдвигая интервал влево или вправо соответствующим образом, можно для заданной доверительной вероятности найти бесконечное множество решений (интервалов). Для обеспечения единообразия выбираются такие квантили (интервал), чтобы площадь под кривой, лежащая левее левого квантиля, равнялась площади под кривой, расположенной правее правого квантиля, т.е.:

$$P(\chi^2 < \chi_1^2) = P(\chi^2 > \chi_2^2) = \frac{1-\beta}{2} = \frac{\alpha}{2}. \qquad (5.4)$$

Тогда из (5.3), учитывая (5.4), получим соответствующие границы интервала:

$$\sigma_1^2 = \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}; \qquad \sigma_2^2 = \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}. \qquad (5.5)$$

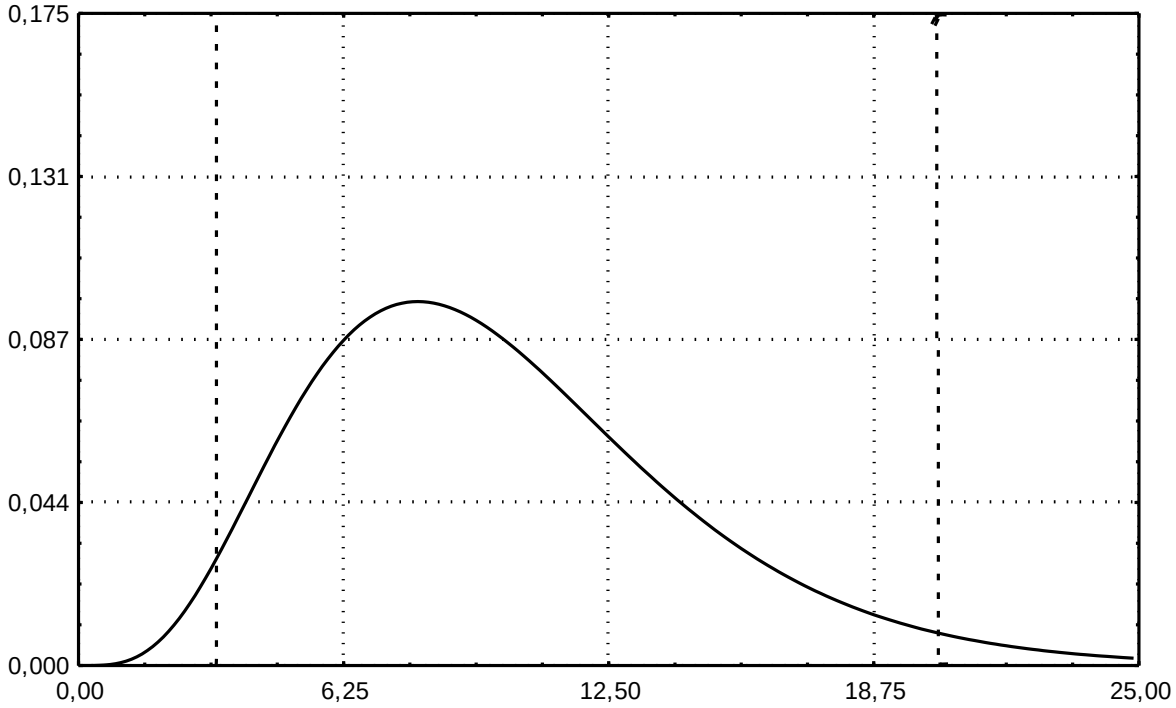


Рис. 5.1. Распределение χ^2

$$(k = 10, \beta = 0.95, \chi^2_{10;0.025} = 3.247, \chi^2_{10;0.975} = 20.486)$$

Пример Дана выборка СВ Y объемом $n = 10$. Предполагается, что СВ Y распределена нормально с неизвестными параметрами (m_y, σ_y).

i	1	2	3	4	5	6	7	8	9	10
y_i	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

Необходимо найти доверительные интервалы для математического ожидания и дисперсии при доверительной вероятности равной 0.97.

Решение: В качестве несмещенных и эффективных оценок вычислим:

$$n = 10; \quad \bar{y} = \sum y_i / 10 = 1.58; \quad S^2 = \sum (y_i - \bar{y})^2 / (n-1);$$

$$9 \cdot S^2 = 13.616; \quad S^2 = 1.513; \quad S = 1.23.$$

а) Вычислим доверительный интервал для математического ожидания, если дисперсия известна (полагаем, что $\sigma^2 \equiv S^2$). Тогда из таблицы нормального распределения получим $U_{0.985} = -U_{0.015} = 2.17$. Подставим полученные значения в (4.9 и 4.10):

$$\varepsilon = U_{1-\alpha/2} \cdot \frac{\sigma_y}{\sqrt{n}} = 2.17 \cdot \frac{1.23}{\sqrt{10}} = 0.844; \quad I_{0.97} = (0.736;\ 2.424).$$

б) Вычислим доверительный интервал для математического ожидания, при неизвестной дисперсии. Воспользуемся таблицей распределения Стьюдента с числом степеней свободы $k = n - 1 = 9$. Соответствующие квантили равны $t_{9;0.985} = -t_{9;0.015} = 2.527$. Подставим полученные значения в (4.15 и 4.16):

$$\varepsilon = t_{1-\alpha/2} \cdot \frac{S_y}{\sqrt{n}} = 2.527 \cdot \frac{1.23}{\sqrt{10}} = 0.983; \quad I_{0.97} = (0.597;\ 2.563).$$

в) Вычислим доверительный интервал для дисперсии. Воспользуемся таблицей распределения χ^2. Симметричный 97% вероятностный интервал с $k = n - 1 = 9$ числом степеней свободы $-(2.33;\ 20.5)$. Подставив полученные значения в (16.21), получим:

$$\sigma_1^2 = \frac{13.616}{20.5} = 0.664; \quad \sigma_1^2 = \frac{13.616}{2.33} = 5.844; \quad I_{0.97} = (0.664;\ 5.844).$$

ЛЕКЦИЯ №6

СТАТИСТИЧЕСКИЕ КРИТЕРИИ

Что такое критерий значимости?

Прежде, чем перейти к рассмотрению понятия *статистической гипотезы*, сформулируем так называемый принцип ***практической уверенности***, лежащий в основе применения выводов и рекомендаций, полученных с помощью теории вероятностей и математической статистики:

Если вероятность события A в данном испытании очень мала, то при однократном испытании можно быть уверенным в том, что событие A не произойдет, и в практической деятельности вести себя так, как будто, событие A вообще невозможно.

Вопрос о том, насколько малой должна быть вероятность α события A, чтобы его можно было считать практически невозможным, выходит за рамки математической теории и решается в каждом отдельном случае с учетом важности последствий, вытекающих из наступления события A.

В ряде случаев можно пренебречь событиями, вероятность которых меньше 0.05, а в других, когда речь идет, например, о разрушении сооружений, гибели судна и т.п. нельзя пренебрегать событиями, которые могут появиться с вероятностью, равной 0.001.

Статистическим критерием (или просто критерием) называют случайную величину K, которая служит для проверки гипотезы.

Критерии значимости (критерии проверки гипотез, иногда – просто *тесты*) – это простейшие, но, наиболее широко используемые статистические средства.

Критерий значимости дает возможность статистику найти разумный ответ на вопрос, подобный следующим:

- Сталь, произведенная разными методами, имеет неодинаковые пределы прочности. "Указывает ли это на то, что производимая разными методами сталь имеет различную прочность, или же выявленное различие можно объяснить выборочными флуктуациями?"
- "Превосходит ли по эффективности одно противогриппозное средство другое?"
- "Способствует ли отказ от курения снижению вероятности раковых заболеваний?"
- "Превосходит ли по воздействию одно удобрение другое при выращивании овощей?"

Проверка гипотез

Статистической – называют гипотезу о виде неизвестного распределения или о параметрах известных распределений.

Рассмотрим простейший вид статистической процедуры, называемой проверкой гипотез.

Пусть дана некоторая оценка $\tilde{\phi}$, построенная по выборке из n независимых наблюдений СВ X. Предположим, что есть основания считать истинное значение оцениваемого параметра равным ϕ_0. Однако даже если истинное значение параметра ϕ равно ϕ_0, выборочное значение $\tilde{\phi}$, вероятно, не будет в точности равняться ϕ_0 из-за выборочной изменчивости,

присущей $\tilde{\phi}$. Поэтому возникает следующий вопрос. Если предположить, что $\phi = \phi_0$, то, при каком отклонении $\tilde{\phi}$ от ϕ_0, эта гипотеза должна быть отвергнута как несостоятельная? На этот вопрос ответ можно дать в статистических терминах, вычислив вероятность любого значимого отклонения $\tilde{\phi}$ от ϕ_0 по выборочному распределению $\tilde{\phi}$. Если вероятность такого отличия мала, то отличие следует считать значимым, и гипотеза $\phi = \phi_0$ должна быть отвергнута. Если же вероятность такого отличия велика, то отклонение следует приписать естественной статистической изменчивости, и гипотеза $\phi = \phi_0$ может быть принята.

Проиллюстрируем общий подход, предположив, что выборочное значение $\tilde{\phi}$, являющееся оценкой параметра ϕ, имеет плотность вероятности нормального распределения $\rho(\tilde{\phi})$. Теперь, если гипотеза $\phi = \phi_0$ верна, то $\rho(\tilde{\phi})$ должна иметь среднее значение ϕ_0 (рис. 6.1).

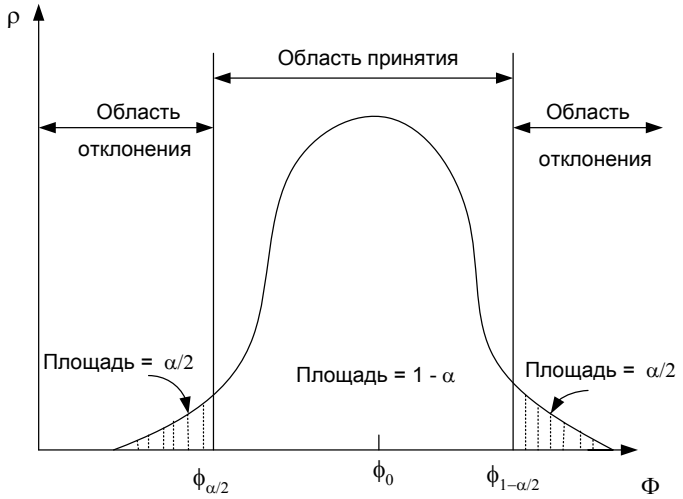


Рис.6.1. Область принятия и отклонения гипотезы (двусторонний критерий)

Вероятность α, использованная при испытании гипотез, называется *уровнем значимости критерия.*

Вероятность того, что $\tilde{\phi}$ окажется меньше нижней границы $\phi_{\alpha/2}$, равна вероятности того, что $\tilde{\phi}$ превзойдет верхнюю границу $\phi_{1-\alpha/2}$ и каждая из них равна $\alpha/2$. Следовательно, вероятность того, что $\tilde{\phi}$ окажется вне интервала, заключенного между этими границами равна α. Область значений $\tilde{\phi}$, при которых гипотеза принимается, называется *областью принятия* гипотезы.

Нулевой (основной) называют выдвинутую гипотезу H_0. В данном примере $H_0 : \phi = \phi_0$.

Область значений $\tilde{\phi}$, при которых гипотеза должна быть отвергнута, называется *областью отклонения* гипотеза или *критической областью*.

Конкурирующей (альтернативной) называют гипотезу, которая противоречит нулевой. В данном примере $H_1 : \phi \neq \phi_0$.

Рассмотренный нами простой критерий испытания гипотез называется *двусторонним критерием*, т.к., когда гипотеза неверна, значение ϕ может быть либо больше, либо меньше ϕ_0.

В других случаях достаточно бывает *односторонних* критериев. Например, пусть основная гипотеза $H_0 : \phi \geq \phi_0$. Тогда альтернативная гипотеза состоит в том: $H_1 : \phi < \phi_0$. Следовательно, в критерии должна использоваться только нижняя (левая) граница ϕ_α, определяемая по плотности вероятности $\rho(\tilde{\phi})$.

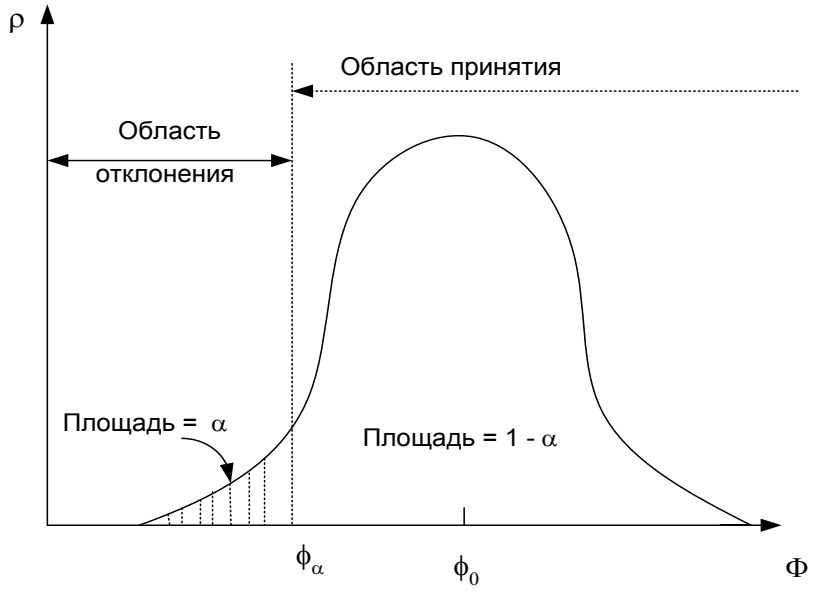


Рис.6.2. Область принятия и отклонения гипотезы (односторонний критерий)

При проверке гипотезы возможны два типа ошибок.

- Во-первых, гипотеза может быть отклонена, хотя фактически она верна. Такая ошибка называется ***ошибкой первого рода***.
- Во-вторых, гипотеза может быть принята, хотя фактически она неверна. Такая ошибка называется ***ошибкой второго рода***.

Проиллюстрируем эти понятия графически (рис. 6.3).

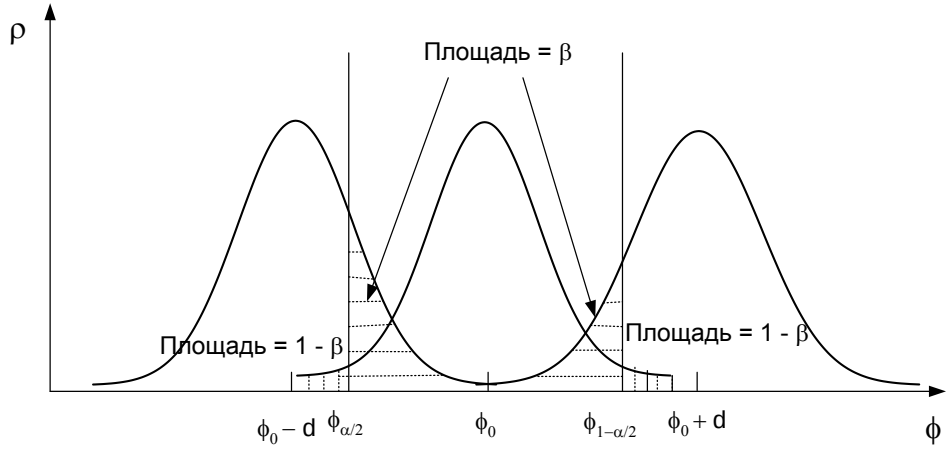


Рис. 6.3. Определение ошибки первого и второго рода при проверке гипотез

Из рисунка видно, что ошибка первого рода происходит в том случае, когда при справедливости гипотезы H_0 значение $\tilde{\phi}$ попадает в область ее отклонения (критическую область). Следовательно, вероятность ошибки первого рода равна α, т.е. уровню значимости критерия.

Для определения вероятности ошибки второго рода предположим, к примеру, что истинный параметр равен либо $\phi_0 + d$, либо $\phi_0 - d$ (см. рис. 6.3). Если гипотеза состоит в том, что $H_0 : \phi = \phi_0$, тогда как на самом деле $\phi = \phi_0 \pm d$, то вероятность того, что $\tilde{\phi}$ попадает в область принятия гипотезы, заключенную между $\phi_{\alpha/2}$ и $\phi_{1-\alpha/2}$ равна β. Следовательно, вероятность ошибки второго рода равна β при выявлении отклонения величиной $\pm d$ от гипотетического значения ϕ_0.

Вероятность $1-\beta$ называется ***мощностью критерия***.

Следует отметить, что вероятности ошибок первого и второго рода вычисляются при ***разных*** предположениях о распределении (если верна гипотеза H_0 и если верна гипотеза H_1), так что никаких раз и навсегда фиксированных соотношений (например $\alpha = 1-\beta$, независимо от вида гипотезы и вида критерия) между ними нет. Таким образом, при фиксированном объеме выборки n, мы можем сколь угодно уменьшать ошибку первого рода, уменьшая уровень значимости α. При этом, естественно, возрастает вероятность β ошибки второго рода (уменьшается мощность критерия). Единственный способ одновременно уменьшить ошибки первого и второго рода (α и β) – увеличить размер выборки n. Именно такие соображения лежат в основе выбора нужного размера выборки в статистических экспериментах.

Пример 1. *ПОСТРОЕНИЕ КРИТЕРИЯ ПРОВЕРКИ ГИПОТЕЗ*

Предположим, что среднее значение СВ X равно $m_x = 10$, также предположим, что дисперсия известна и равна $\sigma_x^2 = 4$. Необходимо найти объем выборки n, позволяющий построить критерий проверки гипотезы $H_0 : m_x = 10$ с 5% – уровнем значимости и 5% – ошибкой второго рода для выявления 10% – отклонений от гипотетического значения. Построим также область принятия гипотезы H_0.

Решение. Выборочное среднее $\bar{x}$, определяемое формулой (3.2), является несмещенной оценкой m_x. Соответствующее выборочное распределение $\bar{x}$ определяется из соотношения (4.6):

$$\bar{x} = \frac{\sigma_x}{\sqrt{n}} z + m_x, \tag{6.1}$$

где z имеет распределение $N(0,1)$. Верхняя и нижняя границы области принятия гипотезы соответственно равны:

$$m_x + z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}, \quad m_x + z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}}. \tag{6.2}$$

Если теперь истинное среднее значение равно $m_x^{'} = m_x \pm d$, то с вероятностью β произойдет ошибка второго рода, если выборочное среднее $\bar{x}$ окажется меньше (левее) верхней границы и больше (правее) нижней. В терминах выборочного распределения $\bar{x}$ со средним $m_x^{'} = m_x + d$ или $m_x^{'} = m_x - d$ для верхней и нижней границ (см. рис. 6.3):

$$m_x + d + \frac{\sigma_x}{\sqrt{n}} z_\beta, \quad m_x - d + \frac{\sigma_x}{\sqrt{n}} z_{1-\beta}. \tag{6.3}$$

Итак, справедливы следующие равенства:

$$\begin{aligned} m_x + z_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}} &= m_x + d + \frac{\sigma_x}{\sqrt{n}} z_\beta, \quad z_{1-\alpha/2} = z_\beta + \frac{\sqrt{n}}{\sigma_x} d; \\ m_x + z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} &= m_x - d + \frac{\sigma_x}{\sqrt{n}} z_{1-\beta}, \quad z_{\alpha/2} = z_{1-\beta} - \frac{\sqrt{n}}{\sigma_x} d. \end{aligned} \tag{6.4}$$

Вспомним, что благодаря симметричности распределения $N(0,1)$ справедливы равенства:

$$z_{\alpha/2} = -z_{1-\alpha/2}; \quad z_\beta = -z_{1-\beta}. \tag{6.5}$$

Теперь из (6.4) с учетом (6.5) найдем требуемый объем выборки:

$$n = \left(\frac{\sigma_x (z_{1-\alpha/2} + z_{1-\beta})}{d} \right)^2. \tag{6.6}$$

Для конкретных значений данного примера:

$(\sigma_x = 2,\ z_{1-\alpha/2} = z_{0.975} = 1.96,\ z_{1-\beta} = z_{0.95} = 1.645,\ d = 0.1 \times 10 = 1).$

Подставим эти значения в (6.6) и получим значение необходимого объема выборки $n = 51.9841$. Таким образом, объем выборки должен быть равен или больше пятидесяти двух. Область принятия гипотезы H_0 определяется соответствующими границами (верхней и нижней) (6.2):

$$m_x + z_{1-\alpha/2}\frac{\sigma_x}{\sqrt{n}} = 10.54, \qquad m_x + z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}} = 9.46.$$

ЛЕКЦИЯ №7

Различие между двумя выборочными средними

Пусть дана выборка из n_1 значений нормально распределённой СВ X и n_2 значений нормально распределенной СВ Y, причем

$$X \Rightarrow N(m_x, \sigma_x^2); \quad Y \Rightarrow N(m_y, \sigma_y^2).$$

Необходимо проверить гипотезу $H_0: \ m_x = m_y$, против гипотезы $H_1: \ m_x \neq m_y$.

Заметим, что дисперсии σ_x^2 и σ_y^2 нам известны. Кроме того, предположение относительно нормальности распределения указанных величин не входит в проверяемую гипотезу.

Как было показано раньше, выборочные средние $\bar{x}$ и $\bar{y}$ являются эффективными и несмещенными оценками соответствующих математических ожиданий с соответствующими дисперсиями: $\frac{\sigma_x^2}{n_1}, \frac{\sigma_y^2}{n_2}$.

В качестве статистики (оценки) возьмем:

$$W = \frac{\bar{x} - \bar{y}}{\sigma_{(\bar{x}-\bar{y})}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}}. \tag{7.1}$$

В (7.1) мы воспользовались свойством, дисперсия разности равна сумме дисперсий. Статистика $W \Rightarrow N(0,1)$, и теперь легко проверить, сравнив с табличными значениями, значимо ли W отличается от нуля.

Если дисперсии σ_x^2 и σ_y^2 неизвестны, то можно воспользоваться объединённой оценкой $\sigma^2 \Rightarrow S^2$, полученной из обеих выборок:

$$S^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2}{n_1 + n_2 - 2}. \qquad (7.2)$$

В этом случае мы используем $t-$статистику с $k = n_1 + n_2 - 2$ степенями свободы:

$$t_k = \frac{\bar{x} - \bar{y}}{S(x-y)} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}} = \frac{\bar{x} - \bar{y}}{S \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}}. \qquad (7.3)$$

Вспомним, что $t-$ статистика имеет распределение Стьюдента с k степенями свободы.

Отметим, что для статистики W и t_k можно рассматривать, как двусторонние, так и односторонние критерии.

Пример 7.1. *СРАВНЕНИЕ ДВУХ ВЫБОРОЧНЫХ СРЕДНИХ*

Имеется 13 мотков пряжи, которые надо исследовать, не изменяются ли при намокании её способность к вытягиванию. Шесть произвольных мотков подвергли испытанию на растяжение, для чего подвешивается груз фиксированной величины. Относительное удлинение мотков (СВ Y) приведено в таблице:

i	1	2	3	4	5	6
y_i	12.3	13.7	10.4	11.4	14.9	12.6

Оставшиеся семь мотков были тщательно намочены и подвергнуты испытаниям с теми же грузами:

i	1	2	3	4	5	6	7
x_i	15.7	10.3	12.6	14.5	12.6	13.8	11.9

Подсчитаем: $\bar{x} = 13.06, \quad \bar{y} = 12.55, \quad \bar{x} - \bar{y} = 0.51.$

Требуется проверить, случайно ли полученное различие?

Для этого выдвинем гипотезы H_0: $\bar{x} = \bar{y}$, H_1: $\bar{x} \neq \bar{y}$. В связи с тем, что дисперсия неизвестна, вычислим объединенную выборочную дисперсию (7.2):

$$S^2 = \frac{12.86 + 19.98}{11} = 2.89; \quad S = 1.7.$$

Подставим полученные значения в (7.3):

$$t_{11} = \frac{0.51}{1.7 \cdot \sqrt{\frac{13}{42}}} = 0.5392 \approx 0.54.$$

По таблицам распределения Стьюдента для одиннадцати степеней свободы определяем, что вероятность P того, что t отличается от нулевого среднего в любую сторону более чем на t_{11} равна 0.6. Поэтому нет оснований считать полученный результат необычным. Т.е. обнаруженная разница ***незначима,*** и верна гипотеза H_0.

Критерий Фишера

Критерий Фишера применяется при проверке гипотезы о равенстве дисперсий двух генеральных совокупностей, распределенных по нормальному закону. Гипотезы о дисперсиях возникают довольно часто, так как дисперсия характеризует такие исключительно важные показатели, как точность машин, приборов, технологических процессов, степень однородности совокупностей, риск, связанный с отклонением доходности активов от ожидаемого уровня, и т.д.

Сформулируем задачу. Пусть имеются две нормально распределенные совокупности, дисперсии которых равны σ_1^2 и σ_2^2. Необходимо проверить нулевую гипотезу о равенстве дисперсий, т.е. H_0: $\sigma_1^2 = \sigma_2^2$ относительно конкурирующей H_1: $\sigma_1^2 > \sigma_2^2$ или H_1': $\sigma_1^2 \neq \sigma_2^2$.

Для проверки гипотезы H_0 из этих совокупностей взяты две независимые выборки объемом n_1 и n_2. Так как оценки дисперсий σ_1^2 и σ_2^2 нам неизвестны, воспользуемся несмещенными выборочными оценками дисперсий S_1^2 и S_2^2.

Тогда при справедливости гипотезы H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ в качестве оценки σ^2 можно взять те же дисперсии S_1^2 и S_2^2, рассчитанные по элементам первой и второй выборок.

Известно, что выборочные характеристики $\frac{(n_1-1)\cdot S_1^2}{\sigma^2}$ и $\frac{(n_2-1)\cdot S_2^2}{\sigma^2}$ имеют распределение χ^2 соответственно с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы, а их отношение $\frac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}$ имеет $F-$ распределение Фишера – Снедекора с k_1 и k_2 степенями свободы. Следовательно, случайная величина F, определяемая отношением:

$$F = \frac{\frac{1}{n_1-1}\left[(n_1-1)\cdot S_1^2/\sigma^2\right]}{\frac{1}{n_2-1}\left[(n_2-1)\cdot S_2^2/\sigma^2\right]} = \frac{S_1^2}{S_2^2}, \tag{7.4}$$

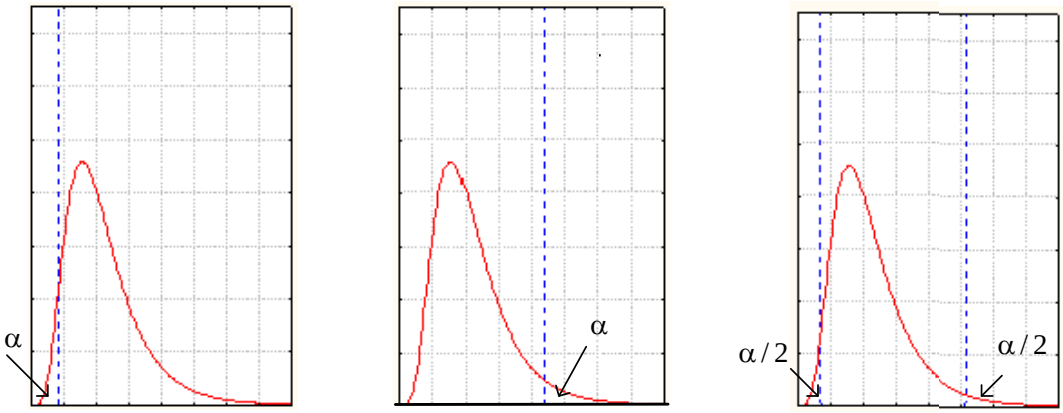
т.е. отношение несмещенных выборочных дисперсий имеет $F-$ *распределение Фишера – Снедекора с* $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы.

Очевидно, что при равенстве дисперсий величина критерия будет равна единице. В остальных случаях она будет больше (меньше) единицы.

При формировании критерия отклонения (принятия) гипотезы H_0 следует учесть, что распределение статистики F (в отличие от нормального или распределения Стьюдента является несимметричным.)

Критерий Фишера $F(\alpha, k_1, k_2)$ – двусторонний критерий, и нулевая гипотеза $H_0 : S_1^2 = S_2^2$ принимается (отвергается альтернативная гипотеза $H_1 : S_1^2 \neq S_2^2$) если $F(\alpha/2, k_1, k_2) < F < F(1-\alpha/2, k_1, k_2)$. Здесь $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1, n_2 – объем первой и второй выборки соответственно.

На рис. 7.1 приведено распределение $F(x, 15{,}20)$. При проверке одностороннего критерия гипотеза H_0 отвергается в пользу альтернативной гипотезы H_1 если $F < F(0.05{,}15{,}20)$ – левосторонняя критическая область (рис. 7.1а), либо если $F > F(0.95{,}15{,}20)$ – правосторонняя критическая область (рис. 7.1б), либо если $F < F(0.025{,}15{,}20)$ или $F > F(0.975{,}15{,}20)$ – в случае двусторонней критической области (рис. 7.1в).



а) б) в)

Рис. 7.1 Критические области распределения Фишера – Снедекора.

Замечание. Для F – критерия доказана справедливость соотношения:

$$F(1-\alpha/2, k_1, k_2) = {}^{1}/_{F(\alpha/2,\ k_1, k_2)}. \qquad (7.5)$$

Пример 7.2. *СРАВНЕНИЕ ДИСПЕРСИЙ ДВУХ ВЫБОРОК*

Пусть поставлена задача, сравнить эффективности обучения двух групп студентов по разным методикам. Успеваемость студентов – случайные величины X_1 и X_2 соответственно, подчинена нормальному закону распределению. В первой группе обучалось $n_1 = 25$ студентов, а во второй – $n_2 = 30$ студентов. Качество обучения (эффективность) характеризуется дисперсией. По данным двух выборок (групп) рассчитаны выборочные несмещенные дисперсии $S_1^2 = 0.50$ и $S_2^2 = 0.08$. Задавая уровень значимости $\alpha = 0.05$, выясним, можно ли считать эффективности обучения двух методик одинаковыми.

Выдвинем нулевую гипотезу H_0: $S_1^2 = S_2^2$, т.е. эффективности обучения двух различных методик – одинаковы. В качестве альтернативной гипотезы рассмотрим H_1: $S_1^2 > S_2^2$.

Вычислим $F = S_1^2 \,/\, S_2^2$ (в числителе должна быть большая дисперсия), $F = {}^{0.5}/_{0.08} = 6.25$. По таблицам (**STATISTICA – Probability Distribution Calculator)** находим критическое значение (правосторонняя область см. рис. 7.1б) $F(0.95, 24, 29) = 1.9005$, которое меньше вычисленного. Следовательно, нулевая гипотеза должна быть отвергнута в пользу альтернативы H_1: $S_1^2 > S_2^2$. Таким образом, эффективность второй методики значительно выше (дисперсия меньше), чем первой.

ЛЕКЦИЯ №8

НЕПАРАМЕТРИЧЕСКИЕ ГИПОТЕЗЫ

Критерий согласия хи-квадрат

Одной из важнейших задач математической статистики является ***установление теоретического закона распределения случайной величины***, характеризующего изучаемый признак по опытному (эмпирическому распределению), представляющему вариационный ряд.

Для решения этой задачи необходимо определить вид и параметры закона распределения.

Предположения о ***виде закона распределения*** может быть выдвинуто исходя из теоретических предпосылок, опыта предшествующих аналогичных исследований, и, наконец, на основании графического изображения эмпирического распределения.

Параметры распределения, как правило, неизвестны, поэтому их заменяют выборочными оценками (несмещенными, эффективными и состоятельными).

Как бы хорошо ни был подобран теоретический закон распределения, между эмпирическим и теоретическим распределениями неизбежны расхождения. Естественно возникает вопрос, обусловлены ли эти расхождения только случайными факторами, связанными с ограниченным числом наблюдений (объемом выборки), или они являются существенными и обусловлены неудачным выбором теоретического закона распределения. Для ответа на этот вопрос и служат ***критерии согласия***.

Для проверки эквивалентности плотности вероятности выборочных данных некоторой гипотетической плотности часто используется особый критерий, называемый ***критерием согласия хи-квадрат***. Общая идея критерия заключается в использовании в качестве меры расхождения наблюдаемой плотности вероятности и гипотетической плотности некоторой статистики (оценки), приближенно подчиняющейся распределению хи-квадрат (χ^2). Затем гипотеза относительно их эквивалентности проверяется путем изучения выборочного распределения этой статистики.

Пусть дана выборка из N независимых наблюдений случайной величины X с плотностью $\rho(x)$. Сгруппируем N наблюдений по $k-$ интервалам, называемым *интервалами группировки*, которые в совокупности образуют *гистограмму частот*. Число наблюдений, попавших в $i-$й интервал, называется *наблюденной частотой* $i-$го интервала; обозначим её f_i. Число наблюдений, которые могли бы попасть в $i-$й интервал, если бы истинной плотностью СВ X была $\rho(x)$, называется *ожидаемой частотой* $i-$го интервала F_i.

Расхождение между наблюденной и ожидаемой частотами в каждом интервале равно $f_i - F_i$.

Для того чтобы измерить общее расхождение по всем интервалам, нормируем квадраты расхождений соответствующими ожидаемыми частотами и просуммируем их. В результате получим выборочную статистику (оценку):

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - F_i)^2}{F_i} \tag{8.1}$$

Показано, что распределение X^2 приближенно совпадает с χ_n^2- распределением. Число степеней свободы n в этом случае равно k (число интервалов группировки) минус число различных независимых линейных ограничений, наложенных на наблюдения. Рассмотрим эти ограничения подробнее:

1. Ограничение связано с тем, что частота в последнем интервале группировки полностью определяется частотами всех остальных интервалов, т.е. не является независимой величиной.

2. Если гипотетическая (предполагаемая) плотность – нормальная, с неизвестным математическим ожиданием и дисперсией, то появятся два дополнительных ограничения, поскольку для подбора нормальной плотности необходимо оценить два параметра $(\bar{x}, S^2)$.

Естественно, при проверке нормальности распределения $n = k - 3$. Для показательного распределения $m_x = \sigma_x = \lambda$. Тогда $n = k - 2$.

После выбора числа степеней свободы величины χ^2 проверка гипотезы производится следующим образом:

- выдвигается гипотеза $H_0 -$ случайная величина X имеет плотность распределения $\rho(x)$;
- группируем выборочные значения по k интервалам и вычисляем ожидаемую частоту для каждого интервала в предположении, что $\rho(x) = \rho_0(x)$;
- по формуле (8.1) находим значение X^2;
- поскольку любое отклонение $\rho(x)$ от $\rho_0(x)$ вызовет увеличение X^2, то используем односторонний критерий (см. рис. 8.1). Тогда область принятия гипотезы H_0 имеет вид: $X^2 \leq \chi^2_{n.1-\alpha}$
- если выборочная оценка X^2 превышает $\chi^2_{n;1-\alpha}$, то гипотеза о том, что $\rho(x) = \rho_0(x)$, отвергается с уровнем значимости α;

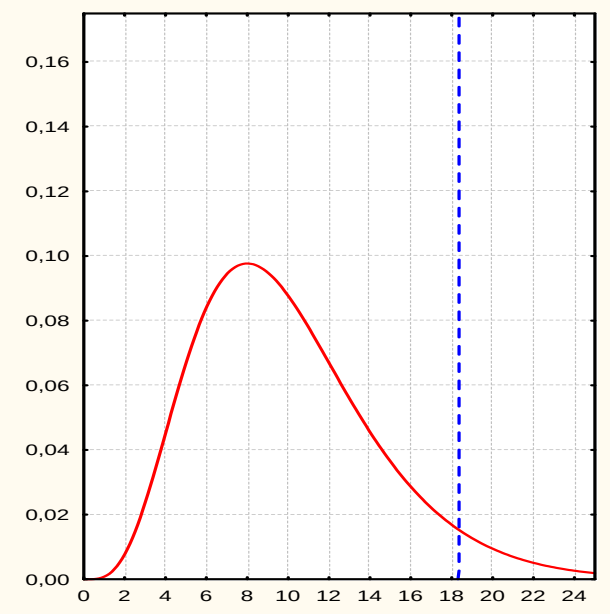


Рис. 8.1 Критическая область критерия Пирсона ($\chi^2_{10,0.95}$).

ПРИМЕЧАНИЯ:

1). Обычно ширину интервала группировки выбирают: $\Delta x = 0.4S$.

2). В качестве первого и последнего интервалов выбирают интервалы, простирающиеся соответственно до $-\infty$ и $+\infty$. При этом должно выполняться условие $F_1, F_k > 3$. Если это условие не выполняется, то объединяем соседние интервалы.

Пример 1. *ПРОВЕРКА НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ*

Имеется выборка из $N = 200$ независимых наблюдений. Необходимо проверить гипотезу о нормальности выборочного распределения с помощью критерия Пирсона с уровнем значимости $\alpha = 0.05$. Для удобства выборочные значения упорядочены по возрастанию и сведены в таблицу:

Таблица 1 Исходная упорядоченная по возрастанию выборка

-7.6	-3.8	-2.5	-1.6	-0.7	0.2	1.1	2.0	3.4	4.6
-6.9	-3.8	-2.5	-1.6	-0.7	0.2	1.1	2.1	3.5	4.8
-6.6	-3.7	-2.4	-1.6	-0.6	0.2	1.2	2.3	3.5	4.8
-6.4	-3.6	-2.3	-1.5	-0.6	0.3	1.2	2.3	3.6	4.9
-6.2	-3.5	-2.3	-1.5	-0.6	0.3	1.3	2.3	3.6	5.0
-6.1	-3.4	-2.3	-1.4	-0.5	0.3	1.3	2.4	3.6	5.2
-6.0	-3.4	-2.2	-1.4	-0.4	0.4	1.3	2.4	3.7	5.3
-5.7	-3.4	-2.2	-1.2	-0.4	0.4	1.4	2.5	3.7	5.4
-5.6	-3.3	-2.1	-1.2	-0.4	0.5	1.5	2.5	3.7	5.6
-5.5	-3.2	-2.1	-1.2	-0.3	0.5	1.5	2.6	3.7	5.9
-5.4	-3.2	-2.0	-1.1	-0.3	0.6	1.6	2.6	3.8	6.1
-5.2	-3.1	-2.0	-1.1	-0.2	0.6	1.6	2.6	3.8	6.3
-4.8	-3.0	-1.9	-1.0	-0.2	0.7	1.6	2.7	3.9	6.3
-4.6	-3.0	-1.9	-1.0	-0.2	0.8	1.7	2.8	4.0	6.5
-4.4	-2.9	-1.8	-1.0	-0.1	0.9	1.8	2.8	4.2	6.9
-4.4	-2.9	-1.8	-0.9	0.0	0.9	1.8	2.9	4.2	7.1
-4.3	-2.9	-1.8	-0.9	0.0	1.0	1.8	3.1	4.3	7.2
-4.1	-2.7	-1.7	-0.8	0.1	1.0	1.9	3.2	4.3	7.4
-4.0	-2.6	-1.7	-0.8	0.1	1.1	1.9	3.2	4..4	7.9

-3.8	-2.6	-1.6	-0.7	0.2	1.1	2.0	3.3	4.4	9.0

Решение:

1. Вначале вычислим выборочные оценки $\bar{x} = 0.24; \quad S = 3.30$.
2. Сгруппируем имеющиеся данные, принимая ширину интервала равной $\Delta x = 0.4S = 1.32$.
3. Квантили нормального (гауссова) распределения возьмем из таблицы, не забывая, что первый интервал начинается на $-\infty$, а последний – заканчивается на $+\infty$. Из таблицы определим вероятности попасть в соответствующий интервал. Полученные результаты сведем в таблицу:

Таблица 2 Результаты расчетов

№	z_α	$x = Sz_\alpha + \bar{x}$	P	$F = NP$	f	$\lvert F - f \rvert$	$(F-f)^2 / F$
1	-2.0	-6.36	0.0228	4.5	4	0.5	0.06
2	-1.6	-5.04	0.0320	6.4	8	1.6	0.40
3	-1.2	-3.72	0.0603	12.1	10	2.1	0.36
4	-0.8	-2.40	0.0968	19.4	21	1.6	0.13
5	-0.4	-1.08	0.1327	26.5	29	2.5	0.24
6	0	0.24	0.1554	31.1	31	0.1	0.00
7	0.4	1.56	0.1554	31.1	27	4.1	0.54
8	0.8	2.88	0.1327	26.5	25	1.5	0.08
9	1.2	4.20	0.0968	19.4	20	0.6	0.02
10	1.6	5.52	0.0603	12.1	13	0.9	0.07
11	2.0	6.84	0.0320	6.4	6	0.4	0.03
12	∞	∞	0.0228	4.5	6	1.5	0.50

4. Как видно из таблицы, F_1 и F_{12} больше трех. В данном случае число степеней свободы равно $n = k - 3 = 9$, а величина $X^2 = 2.43$. По таблице распределения хи-квадрат найдем уровень значимости $\chi^2_{9;0.95} = 16.92$.т.к. $X^2 \le \chi^2_{9;0.95}$, следовательно, гипотеза о нормальности принимается с уровнем значимости $\alpha = 0.05$.

Критерий Колмогорова.

При анализе выборок малого объема невозможно применить критерий χ^2 (группирование данных некорректно). В этом случае часто используется критерий Колмогорова, в котором в качестве меры расхождения между теоретическим и эмпирическим распределениями рассматривают максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения:

$$D = \max_{-\infty \le x \le \infty} \lvert F_n(x) - F(x) \rvert. \tag{8.2}$$

Оценка D называется ***статистикой критерия Колмогорова***.

Доказано, что какова бы ни была функция распределения $F(x)$ непрерывной случайной величины X, при неограниченном увеличении числа наблюдений $(n \to \infty)$ вероятность неравенства $P(D\sqrt{n} \geq \lambda)$ стремится к пределу

$$P(\lambda) = 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2\lambda^2}. \qquad (8.3)$$

Задавая уровень значимости α, из соотношения

$$P(\lambda) = \alpha \qquad (8.4)$$

можно найти соответствующее критическое значение λ_α.

Схема применения критерия Колмогорова

1. Строятся эмпирическая функция распределения $F_n(x)$ и предполагаемая теоретическая функция распределения $F(x)$.
2. Определяется мера расхождения между теоретическим и эмпирическим распределением D по формуле (8.2) и вычисляется величина

$$\lambda = D\sqrt{n}. \qquad (8.4)$$

3. Если вычисленное значение λ окажется больше критического λ_α, определенного на уровне значимости α, то нулевая гипотеза H_0 о том, что случайная величина X имеет заданный закон распределения, отвергается (односторонний критерий). Если $\lambda \leq \lambda_\alpha$, то считают, что гипотеза H_0 не противоречит опытным данным.

Замечание

Можно отметить, что решение подобных задач можно было бы найти с помощью критерия χ^2. Потенциальное преимущества критерия Колмогорова в том, что он не требует группирования данных (с неизбежной потерей информации), а дает возможность рассматривать индивидуальные наблюдаемые значения. Этот критерий можно успешно применять для малых выборок. Считается, что его мощность, вообще говоря, выше, чем у критерия χ^2.

Пример 2. *ПРОВЕРКА НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ*

Получена случайная выборка объема $n = 6$. Построим вариационный ряд и эмпирическую функцию распределения:

x_i	-1.0	-0.6	0.2	1.3	2.1	3.0	> 3
ω_i	1 \ 6	1 \ 6	1 \ 6	1 \ 6	1 \ 6	1 \ 6	
$F_6(x)$	0	1 \ 6	2 \ 6	3 \ 6	4 \ 6	5 \ 6	1

Проверим гипотезу, что эти наблюдения образуют случайную выборку из распределения $N(1,1)$ с уровнем значимости $\alpha = 0.05$. Затем мы можем определить $D_6(x)$ графически либо аналитически, причем эти значения должны появиться в точке x, соответствующей одной из наблюдаемых величин. С этой целью необходимо вычислить пары величин d_1 и d_2 (рис. 8.2) для каждого значения выборки.

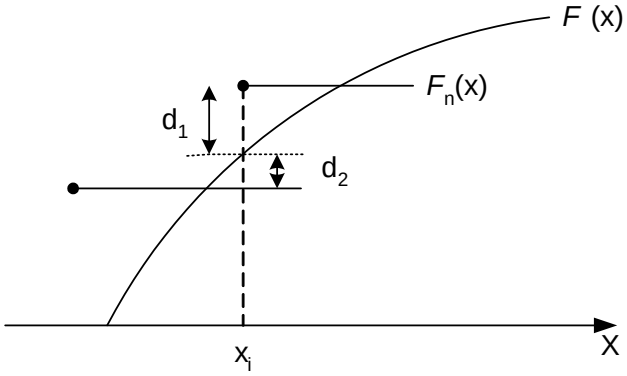


Рис. 8.2 Мера расхождения в точке наблюдения x_i.

Для вычисления $F(x)$ вспомним: $F(x) = P(X < x) = \Phi_0(x-1)$, где $\Phi_0(x-1)$ - функция стандартного нормального распределения. Результаты всех вычислений представим в виде таблицы:

x	$F_6(x)$	$F(x)$	d_1	d_2
-1.0	0.1667	0.0228	0.1439	0.0228
-0.6	0.3333	0.0548	0.2785	0.1119
0.2	0.5	0.2119	0.2881	0.1214
1.3	0.6667	0.6179	0.0488	0.1179
2.1	0.8333	0.8643	0.0310	0.1976
3.0	1.0000	0.9772	0.0228	0.1439

Из таблицы результатов следует: $D_6(x) = 0.2881$; $\lambda = 0.7057$. Из статистических таблиц получим $\lambda_\alpha = 1.36$. Поскольку $\lambda \le \lambda_\alpha$, то принимается гипотеза H_0, т.е. можно считать, что данные подчиняются распределению $N(1,1)$.

ЛЕКЦИЯ №9

Критерий Колмогорова-Смирнова. Проверка гипотезы об однородности выборок

Гипотезы об однородности выборок – это гипотезы о том, что рассматриваемые выборки извлечены из одной и той же генеральной совокупности.

Пусть имеются две независимые выборки, произведенные из генеральных совокупностей с неизвестными теоретическими функциями распределения $F_1(x)$ и $F_2(x)$.

Проверяемая нулевая гипотеза имеет вид H_0: $F_1(x) = F_2(x)$ против конкурирующей $H_1 : F_1(x) \neq F_2(x)$. Будем предполагать, что функции $F_1(x)$ и $F_2(x)$ непрерывны и для оценки используем статистику *Колмогорова – Смирнова*.

Критерий Колмогорова-Смирнова использует ту же самую идею, что и критерий Колмогорова. Однако различие заключается в том, что в критерии Колмогорова сравнивается эмпирическая функция распределения с теоретической, а в критерии Колмогорова-Смирнова сравниваются две эмпирические функции распределения.

Статистика критерия Колмогорова-Смирнова имеет вид:

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max\left|F_{n_1}(x) - F_{n_2}(x)\right|, \qquad (9.1)$$

где $F_{n_1}(x)$ и $F_{n_2}(x)$ – эмпирические функции распределения, построенные по двум выборкам с объемами n_1 и n_2.

Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше критического $\lambda'_{\text{кр}}$, т.е. $\lambda' > \lambda'_{\text{кр}}$, и принимается в противном случае.

При малых объемах выборок $(n_1, n_2 \leq 20)$ критические значения $\lambda'_{\text{кр}}$ для заданных уровней значимости критерия можно найти в специальных таблицах. При $n_1, n_2 \to \infty$ (а практически при $n_1, n_2 \geq 50$) распределение статистики λ' сводится к распределению Колмогорова для статистики λ. В этом случае гипотеза H_0 отвергается на уровне значимости α, если фактически наблюдаемое значение λ' больше критического λ_α, т.е. $\lambda' > \lambda_\alpha$, и принимается в противном случае.

Пример 1. *ПРОВЕРКА ОДНОРОДНОСТИ ДВУХ ВЫБОРОК*

Были осуществлены две проверки торговых точек с целью выявления недовесов. Полученные результаты сведены в таблицу:

Номер интервала	Интервалы недовесов, г	Частоты	
		Выборка 1	Выборка 2
1	0 – 10	3	5
2	10 – 20	10	12
3	20 – 30	15	8
4	30 – 40	20	25
5	40 – 50	12	10
6	50 – 60	5	8
7	60 – 70	25	20
8	70 – 80	15	7

9	80 – 90	5	5

Объем первой выборки был равен $n_1 = 110$, а второй – $n_2 = 100$.

Можно ли считать, что на уровне значимости $\alpha = 0.05$ по результатам двух проверок (случайных выборок) недовесы овощей описываются одной и той же функцией распределения?

Решение:

Обозначим $n_{i_1}^{\text{нак}}$ и $n_{i_2}^{\text{нак}}$ – накопленные частоты выборок 1 и 2; $F_{n_1}(x_i) = n_{i_1}^{\text{нак}}/n_1$, $F_{n_2}(x_i) = n_{i_2}^{\text{нак}}/n_2$ – значения их эмпирических функций распределения соответственно. Обработанные результаты сведем в таблицу:

x_i	$n_{i_1}^{\text{нак}}$	$n_{i_2}^{\text{нак}}$	$F_{n_1}(x_i)$	$F_{n_2}(x_i)$	$\lvert F_{n_1}(x_i) - F_{n_2}(x_i)\rvert$
10	3	5	0.027	0.050	0.023
20	13	17	0.118	0.170	0.052
30	28	25	0.254	0.250	0.004
40	48	50	0.436	0.500	0.064
50	60	60	0.545	0.600	0.055
60	65	68	0.591	0.680	**0.089**
70	90	88	0.818	0.880	0.072
80	105	95	0.955	0.950	0.005
90	110	100	1.000	1.000	0.000

Из последнего столбца таблицы видно, что $\max\lvert F_{n_1}(x_i) - F_{n_2}(x_i)\rvert = 0.089$. По формуле (9.1) получим $\lambda' = \sqrt{\frac{110 \cdot 100}{110+100}} \cdot 0.089 = 0.644$. Из статистических таблиц известно, что $\lambda_{0.05} = 1.36$. Так как $\lambda' < \lambda_{0.05}$, то принимается нулевая гипотеза H_0, т.е. недовесы покупателям описываются одной и той же функцией распределения.

СТАТИСТИЧЕСКАЯ НЕЗАВИСИМОСТЬ И ВЫЯВЛЕНИЕ ТРЕНДА

При анализе случайных данных часто возникает ситуации, когда требуется выяснить, являются ли наблюдения или оценки параметров статистически независимыми или же они подвержены тренду. Это особенно важно при анализе ***нестационарных данных.***

Такие исследования, обычно, проводят на основе *свободных от распределений* или *непараметрических методов*, в которых относительно функции распределения исследуемых данных не делается никаких предположений.

Критерий серий

Рассмотрим последовательность N наблюдённых значений случайной величины X, причём каждое наблюдение отнесено к одному из двух

взаимно исключаемых классов, которые можно обозначить просто (+) или (–). Рассмотрим ряд примеров:

- Бросание монеты: герб (+), цифра (–);
- Пусть имеется выборка x_i $(i = 1, 2, ..., N)$, со средним значением $\bar{x}$.

Тогда, если $x_i \geq \bar{x}$, то (+), если же $x_i < \bar{x}$, то (–);

- Имеется последовательность одновременных измерений двух случайных величин x_i и y_i $(i = 1, 2, ..., N)$. Здесь каждое наблюдение обозначим (+), если $x_i \geq y_i$ и (–), если $x_i < y_i$.

В каждом из этих примерах образуется последовательность вида:

$$\underbrace{++}_{1}\underbrace{-}_{2}\underbrace{++}_{3}\underbrace{-}_{4}\underbrace{+++}_{5}\underbrace{-}_{6}\underbrace{+}_{7}\underbrace{-}_{8}\underbrace{+}_{9}\underbrace{-}_{10}\underbrace{+}_{11}\underbrace{---}_{12}$$

Серией называется последовательность однотипных наблюдений, перед и после которой следуют наблюдения противоположного типа или же вообще нет никаких наблюдений.

В приведенной последовательности число наблюдений равно $N = 20$; а количество серий равно $r = 12$.

Если последовательность N наблюдений состоит из независимых исходов одной и той же случайной величины, т.е. если вероятность отдельных исходов [(+) или (−)] не меняется от наблюдения к наблюдению, то выборочное распределение числа серий в последовательности является случайной величиной r со средним значением и дисперсией:

$$m_r = \left(\frac{2N_1 \cdot N_2}{N}\right) + 1, \tag{9.2}$$

$$\sigma_r^2 = \frac{2N_1 \cdot N_2 (2N_1 \cdot N_2 - N)}{N^2 (N-1)}. \tag{9.3}$$

Здесь N_1 — число исходов (+), а N_2 — число исходов (−), естественно $N = N_1 + N_2$. В частном случае если $N_1 = N_2 = N/2$, то:

$$m_r = \frac{N}{2} + 1, \qquad \sigma_r^2 = \frac{N^2 - 2N}{4(N-1)}. \tag{9.4}$$

Предположим, что есть основание подозревать наличие тренда в последовательности наблюдений, т.е. есть основание считать, что вероятность появления (+) или (−) меняются от наблюдения к наблюдению. Существование тренда можно проверить следующим образом. Примем в качестве нулевой гипотезы H_0 – тренда нет, т.е. предположим, что N наблюдений являются независимыми исходами одной и той же случайной величины. Тогда для проверки гипотезы с любым требуемым уровнем значимости α необходимо сравнить наблюденное число серий с границами области принятия гипотезы равными $r_{n;\alpha/2}$ и $r_{n;1-\alpha/2}$, где $n = N/2$.

Если наблюденное число серий окажется вне области принятия гипотезы, то нулевая гипотеза должна быть отвергнута с уровнем значимости α. В противном случае нулевую гипотезу можно принять.

Пример 2. *ПРИМЕНЕНИЕ КРИТЕРИЯ СЕРИЙ*

Имеется последовательность независимых наблюдений $N = 20$:

5.5	5.1	5.7	5.2	4.8	5.7	5.0	6.5	5.4	5.8
6.8	6.6	4.9	5.4	5.9	5.4	6.8	5.8	6.9	5.5

Проверим независимость наблюдений, подсчитав число серий в последовательности, полученной путем сравнения наблюдений с медианой. Применим критерий с уровнем значимости $\alpha = 0.05$.

Из анализа данных получим, что значение $x = 5.6$ является медианой. Тогда введем обозначения (+) при $x \geq \mu_e$, (–) при $x < \mu_e$. Итак, получим:

$$\underbrace{--}_{1}\underbrace{+}_{2}\underbrace{-}_{3}\underbrace{-}_{4}\underbrace{+}_{5}\underbrace{-}_{6}\underbrace{+}_{7}\underbrace{-}_{8}\underbrace{+++}_{8}\underbrace{-}_{9}\underbrace{-}_{10}\underbrace{+}_{11}\underbrace{-}_{12}\underbrace{+++}_{12}\underbrace{-}_{13}$$

В нашем примере $r = 13$, а область принятия гипотезы имеет вид:

$$[r_{10;\alpha/2} \leq r \leq r_{10;1-\alpha/2}].$$

По статистическим таблицам находим $r_{10;0.025} = 6$, $r_{10;0.975} = 15$. Т.к. $6 \leq r \leq 15$, то нет оснований сомневаться в независимости наблюдений, т.е. верна гипотеза H_0 – тренд отсутствует.

Критерий инверсий

Пусть имеется последовательность из N наблюдений случайной величины X, обозначенных x_i $(i = 1, 2, ..., N)$. Подсчитаем теперь, сколько раз в последовательности имеют место неравенства $x_i > x_j$ при $i < j$. Каждое такое неравенство называется ***инверсией***. Пусть A – общее число инверсий. Формально A вычисляется следующим образом. Определим для множества наблюдений $x_1, x_2, \ldots x_N$ величины

$$\eta_{ij} = \begin{cases} 1, & x_i > x_j, \\ 0, & x_i \leq x_j. \end{cases} \tag{9.5}$$

Тогда:

$$A = \sum_{i=1}^{N-1} A_i; \quad A_i = \sum_{j=i+1}^{N} h_{ij}; \quad (i = \overline{1, N-1}). \tag{9.6}$$

Рассмотрим данный метод на примере последовательности из 8 наблюдений:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
5	3	8	9	4	1	7	5

шаг 1: Т.к. $x_1 > x_2$, $x_1 > x_5$, $x_1 > x_6$, то $A_1 = 3$;

шаг 2: $x_2 > x_6$, $A_2 = 1$;

шаг 3: $x_3 > x_5$, $x_3 > x_6$, $x_3 > x_7$, $x_3 > x_8$, $A_3 = 4$;

шаг 4: $A_4 = 4$;

шаг 5: $A_5 = 1$;

шаг 6: $A_6 = 0$;

шаг 7: $A_7 = 1$.

$$A = \sum_{i=1}^{7} A_i = 14.$$

Если последовательность из N наблюдений состоит из независимых исходов одной и той же случайной величины, то число инверсий является случайной величиной A со средним значением и дисперсией:

$$m_A = \frac{N(N-1)}{4}; \tag{9.7}$$

$$\sigma_A^2 = \frac{2N^3 + 3N^2 - 5N}{72}. \tag{9.8}$$

Критерий инверсий применяется примерно так же, как и критерий серий.

ПРИМЕЧАНИЕ:

Критерий инверсий – более мощный по сравнению с критерием серий при обнаружении монотонного тренда в последовательности наблюдений. Однако этот критерий не столь эффективен при выявлении тренда типа флуктуации.

Пример 3. *ПРИМЕНЕНИЕ КРИТЕРИЯ ИНВЕРСИЙ*

Рассмотрим тот же пример, что и в критерии серий (пример 2). Общее число инверсий в 20 наблюдениях равно $A = 62$. Из соответствующей таблицы при уровне значимости $\alpha = 0.05$ определим область принятия гипотезы: $[A_{20;0.025} = 64, A_{20;0.975} = 125]$.

Следовательно, гипотеза об отсутствии тренда должна быть отвергнута, т.к. $A = 62$ не попадает в область принятия гипотезы H_0.

Этот пример иллюстрирует различную чувствительность двух методов.

ЛЕКЦИЯ №10

ЛИНЕЙНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Исключительный интерес для широкого класса задач представляет обнаружение взаимных связей между двумя и более случайными величинами. Например, существует ли связь между курением и ожидаемой продолжительностью жизни, между умственными способностями и успеваемостью и т.п. В инженерных исследованиях такие задачи, обычно, сводятся к установлению связи между некоторым предполагаемым возбуждением и наблюдаемым откликом изучаемой физической системы.

Существование таких взаимосвязей и их относительную силу можно измерить коэффициентом корреляции.

Основная задача корреляционного анализа *состоит в выявлении связи между случайными переменными путем точечной и интервальной оценки различных (парных, множественных, частных) коэффициентов корреляции*.

Коэффициент корреляции определяется через корреляционный момент (ковариацию) K_{xy} по формуле:

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y} = \frac{M[(X - m_x)(Y - m_y)]}{\sigma_x \sigma_y}. \qquad (10.1)$$

Величина ρ_{xy} характеризует тесноту связи между случайными переменными X и Y в генеральной совокупности. Из свойств коэффициента корреляции известно, что ρ_{xy} является показателем тесноты связи лишь в случае линейной зависимости между двумя переменными. Для линейно независимых случайных величин $\rho_{xy} \equiv 0$. Но даже и для зависимых СВ ρ_{xy} может быть равен 0. В этом случае СВ X и Y называют некоррелированными.

Пусть получена выборка N пар СВ X и Y. Тогда коэффициент корреляции можно оценить по выборочным данным следующим образом:

$$r = \rho_{xy} = \frac{S_{xy}}{S_x S_y}. \qquad (10.2)$$

Вспомним несмещённые, состоятельные и эффективные оценки:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i; \qquad (10.3)$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - N \cdot \bar{x}^2\right]; \qquad (10.4)$$

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N \cdot \bar{y}^2\right]; \qquad (10.5)$$

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i y_i - N \cdot \bar{x}\bar{y}\right]. \qquad (10.6)$$

Тогда ***эмпирический коэффициент корреляции*** определяется по формуле:

$$r_{xy} = \frac{\frac{1}{N-1}\left[\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\bar{y}\right]}{S_x \cdot S_y} = \frac{\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)\left(\sum_{i=1}^{N} y_i^2 - N\bar{y}^2\right)}}. \qquad (10.7)$$

Как и ρ_{xy} выборочный коэффициент корреляции принимает значения в интервале $[-1;+1]$, причем одно из граничных значений принимается только при наличии идеальной линейной связи между наблюдениями. Нелинейная связь и (или) разброс данных, вызванных ошибками измерений или же неполной коррелированностью СВ, приводит к уменьшению абсолютного значения r_{xy} (рис. 10.1).

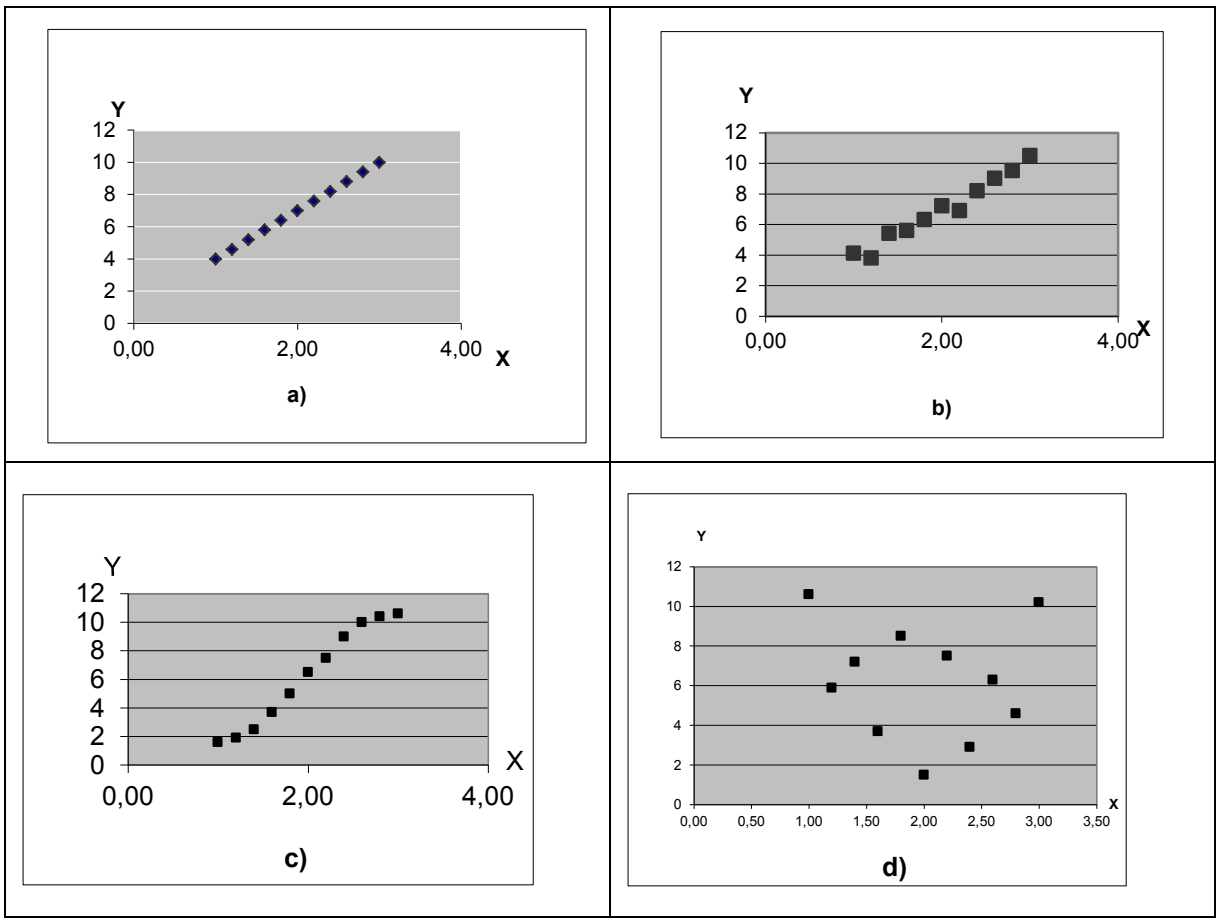


Рис. 10.1. Различные степени корреляции: **a)** – точная линейная корреляция; **b)** – умеренная линейная корреляция; **c)** – нелинейная корреляция; **d)** – отсутствие корреляции.

Для оценки связи случайных величин, основанной на коэффициенте корреляции , используется шкала Чеддока:

Показания тесноты связи	0.1 – 0.3	0.3 – 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 0.99
Сила связи	слабая	умеренная	заметная	высокая	весьма высокая

Эмпирический коэффициент корреляции r_{xy} дает состоятельную, но смещённую оценку. Смещение равно $-\frac{\rho\left(1-\rho^2\right)}{2N}$. Однако при $N>50$ величина смещения составляет менее 1%. Для оценки точности выборочного значения r_{xy} удобно использовать некоторую функцию от r_{xy}:

$$w=\frac{1}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right]. \qquad (10.8)$$

Показано, что распределение случайной величины w можно аппроксимировать нормальным распределением со средним значением и дисперсией:

$$m_w=\frac{1}{2}\ln\left[\frac{1+\rho_{xy}}{1-\rho_{xy}}\right]; \quad \sigma_w^2=\frac{1}{N-3}. \qquad (10.9)$$

Даже для независимых случайных величин эмпирический коэффициент корреляции может быть отличен от "0" вследствие случайного рассеивания результатов измерения. Т.е. из-за выборочной изменчивости необходимо проверять, свидетельствует ли не нулевые значения выборочного коэффициента корреляции о существовании статистически значимой корреляции между исследуемыми случайными величинами X и Y. Сделать это можно, проверив гипотезу H_0: $\rho_{xy}=0$, причем отклонение гипотезы будет свидетельствовать о принятии альтернативной гипотезы H_1 —значимости корреляции.

Из формулы (10.9) следует, что при $\rho_{xy}=0$ выборочное распределение w будет нормальным со средним $m_w=0$ и дисперсией $\sigma_w^2=\frac{1}{(N-3)}$. Поэтому область принятия гипотезы о нулевой корреляции будет иметь вид:

$$z_{\alpha/2}\le\frac{\sqrt{N-3}}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right]<z_{1-\alpha/2}. \qquad (10.10)$$

Здесь $\alpha-$ уровень значимости, $z-$ стандартное нормальное распределение $N(0,1)$.

Пример *ЛИНЕЙНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ*

Имеются данные о росте и массе $N=25$, выбранных наугад студентов. Есть ли основание считать, что рост и масса студентов

коррелированны при уровне значимости $\alpha = 0.05$? Пусть $x -$рост, а $y -$ масса студента. Все данные приведены в таблице:

i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	178	188	178	165	175	185	183	175	183	193	188	183	173
y_i	63	95	67	66	83	75	70	77	79	70	84	84	75
i	14	15	16	17	18	19	20	21	22	23	24	25	
x_i	178	180	173	185	165	185	188	163	183	183	170	185	
y_i	100	84	82	77	61	79	82	68	77	75	66	77	

По данным таблицы получим: $r_{xy} = 0.43$, по формуле (10.8) получим $w = 0.460$, а $w\sqrt{N-3} = 2.16$. По таблицам стандартного нормального распределения получим $z_{1-\alpha/2} = -z_{\alpha/2} = 1.96$. Таким образом, гипотеза H_0 должна быть отвергнута, следовательно, имеются основания считать, что между ростом и массой студентов существует значимая корреляция.

Ранговая корреляция

До сих пор нами рассматривались и анализировались зависимости между *количественными* переменными, измеренными в так называемых *количественных* шкалах. Эти шкалы с непрерывным множеством значений позволяют выявить на сколько (или во сколько раз) проявление признака у одного объекта больше (меньше), чем у другого (например, производительность труда, заработная плата, накладные расходы и т.п.).

Наряду с этим на практике часто возникает необходимость изучения связи между *ординальными* (*порядковыми*) переменными, измеренными в так называемой *порядковой* шкале. В этой шкале можно установить лишь порядок, в котором объекты выстраиваются по степени проявления признака (например, уровень благоустроенности жилья, класс гостиницы, тестовые баллы, экзаменационные оценки и т.п.). Если, скажем, по некоторой дисциплине два студента имеют оценки "отлично" и "удовлетворительно", то можно утверждать, уровень подготовки по этой дисциплине первого студенты выше (больше), чем второго, но нельзя сказать, на сколько или во сколько раз больше.

Оказывается, что в таких случаях проблема оценки тесноты связи разрешима, если упорядочить, или ранжировать, объекты анализа по степени выраженности измеряемых признаков. При этом каждому объекту присваивается определенный номер, называемый *рангом*. Например, объекту с наименьшим проявлением (значением) присваивается ранг 1, следующему за ним – ранг 2 и т.д. Объекты можно располагать и в порядке убывания значений (проявлений) признака. Если объекты ранжированы по

двум признакам, то появляется возможность оценить тесноту связи между признаками, основываясь на рангах, т.е. *ранговые корреляции*.

Коэффициент ранговой корреляции Спирмена находится по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n} \qquad (10.11)$$

где r_i и s_i - ранги *i*-го объекта по переменным X и Y n – число пар наблюдений.

Если ранги всех объектов равны ($r_i = s_i, \quad i = 1, 2, \dots, n$) то $\rho = 1$, т.е. при полной прямой связи *коэффициент ранговой корреляции Спирмена* равен единице.

При полной обратной связи, когда ранги объектов по двум переменным расположены в обратном порядке, можно показать, что $\sum_{i=1}^{n}(r_i - s_i)^2 = (n^3 - n)/3$ и по формуле (10.11) $\rho = -1$. Во всех остальных случаях $|\rho| < 1$.

При ранжировании данных иногда сталкиваются со случаями, когда невозможно найти существенные различия между объектами по величине проявления рассматриваемого признака. Такие объекты, как говорят, оказываются *связанными*. Связанным объектам приписывают одинаковые средние ранги, такие, чтобы сумма всех рангов осталась такой же, как и при отсутствии связанных рангов. Например, если у четырех объектов рассматриваемые признаки оказались одинаковыми и невозможно определить, какие из четырех рангов (4, 5, 6, 7)приписать этим объектам, то каждому объекту приписывается средний ранг, равный (4+5+6+7)/4=5.5. При наличии *связанных рангов* вычисляется модифицированный коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{\sum_{i=1}^{n}(r_i - s_i)^2}{\frac{1}{6}(n^3 - n) - (T_r + T_s)} \qquad (10.12)$$

Здесь

$$T_r = \frac{1}{12}\sum_{i=1}^{m_r}(t_r^3 - t_r); \qquad T_s = \frac{1}{12}\sum_{i=1}^{m_s}(t_s^3 - t_s); \qquad (10.13)$$

m_r, m_s –число групп неразличимых рангов у переменных X и Y;

$t_r,\ t_s$ – число рангов, входящих в группу неразличимых рангов переменных X и Y.

При проверке значимости p воспользуемся тем, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными при $n > 10$ статистика

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \qquad (10.14)$$

имеет $t-$ распределение Стьюдента с $k = n - 2$ степенями свободы. Поэтому ρ значим на уровне α, если фактически наблюдаемое значение t будет больше критического (по абсолютной величине), т.е. $|t| > t_{1-\alpha;n-2}$, где $t_{1-\alpha;n-2}$ — табличное значение t — критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $k = n - 2$.

ЛЕКЦИЯ №11

ЛИНЕЙНЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ (ПРОДОЛЖЕНИЕ)

Пример 1. *КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА*

По двум дисциплинам A и B тестировались 10 студентов. На основе набранных баллов вычислены соответствующие ранги. Необходимо вычислить ранговый коэффициент Спирмена и проверить его значимость на уровне $\alpha = 0.05$.

Решение. Вычисленные ранги, разности рангов и их квадраты представим в виде таблицы:

Ранги по дисциплинам		Студент, i										Всего
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
A	r_i	2	4	5	1	7.5	7.5	7.5	7.5	3	10	55
B	s_i	2.5	6	4	1	2.5	7	8	9.5	5	9.5	55
$r_i - s_i$		-0.5	-2	1	0	5	0.5	-0.5	-2	-2	0.5	-
$(r_i - s_i)^2$		0.25	4	1	0	25	0.25	0.25	4	4	0.25	39

По формуле (10.11) $\rho = 1 - \frac{6\cdot 39}{10^3-10} = 0.764$. Однако эта формула не учитывает наличие связанных рангов.

По дисциплине A имеем $m_r = 1$ —одну группу неразличимых рангов с $t_r = 4$; по дисциплине B – $m_s = 2$ — две группы неразличимых рангов по $t_s = 2$ ранга. Поэтому по формуле (10.13) $T_r = \frac{1}{12}(4^3 - 4) = 5$, $T_s = \frac{1}{12}[(2^3 - 2) + (2^3 - 2)] = 1$. Теперь вычислим модифицированный коэффициент корреляции Спирмена (10.12):

$$\rho = 1 - \frac{39}{\frac{1}{6}(10^3 - 10) - (5 + 1)} = 0.755$$

Для проверки значимости ρ по формуле (10.14) вычислим $t = 0.755\frac{\sqrt{10-2}}{\sqrt{1-0.755^2}} = 3.26$. По таблицам распределения Стьюдента найдем $t_{0.95;8} = 2.31$, т.к. $t > t_{0.95;8}$, то ранговый коэффициент корреляции значим на 5% -ном уровне. Значит связь между оценками двух дисциплин достаточно тесная.

Коэффициент ранговой корреляции Кендалла вычисляется по формуле:

$$\tau = 1 - \frac{4K}{n(n-1)}, \qquad (11.1)$$

где K – статистика Кендалла.

Для определения K необходимо ранжировать объекты по одной переменной в порядке возрастания рангов $(1, 2, \dots, n)$ и определить соответствующие им ранги $(r_1, r_2, \dots, r_n)$ по другой переменной. Статистика K равна общему числу *инверсий* (нарушений порядка, когда большее число стоит слева от меньшего) в ранговой последовательности (*ранжировке*) $r_1, r_2, \dots, r_n$. При полном совпадении двух ранжировок получим $K = 0$ и $\tau = 1$; при полной противоположности можно показать, что $K = n(n-1)/2$ и $\tau = -1$. Во всех остальных случаях $|\tau| < 1$.

При проверке значимости τ исходят из того, что в случае справедливости нулевой гипотезы об отсутствии корреляционной связи между переменными (при $n > 10$) τ имеет приближенно нормальный закон распределения с математическим ожиданием, равным нулю, и средним квадратическим отклонением $S_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$. Поэтому τ значим на уровне α если значение статистики

$$t = \frac{\tau - 0}{S_\tau} = \tau\sqrt{\frac{9n(n-1)}{2(2n+5)}} \qquad (11.2)$$

больше критического $t_{1-\alpha}$, где $\Phi(t_{1-\alpha}) = 1 - \alpha$.

Пример 2. *КОЭФФИЦИЕНТ РАНГОВОЙ КОРРЕЛЯЦИИ КЕНДАЛЛА*

Проведено исследование 10 важнейших видов оборудования, используемого судоводителями во время вахты. Вычислены ранги по важности оборудования *X* и по частоте его использования *Y*. Полученные результаты приведены в таблице. Необходимо вычислить ранговый коэффициент Кендалла и оценить его значение на уровне $\alpha = 0.05$.

Решение. Дополним таблицу рангов числом инверсий в ранжировках по переменной *Y* для различных рангов по переменной *X*.

Ранг	Тип оборудования										Всего
	А	Б	В	Г	Д	Е	Ж	З	И	К	
X, n	1	2	3	4	5	6	7	8	9	10	-
Y, r_i	1	4	2	6	3	9	10	8	7	5	-
Число инверсий	0	2	0	2	0	3	3	2	1	0	$K = 13$

Покажем, как считается число инверсий рангов r_i при ранге $n = 4$ по переменной X. Тогда соответствующий ранг по переменной Y равен $r_4 = 6$. Последующие ранги переменной Y равны: 3, 9, 10, 8, 7, 5, а из них только два ранга меньше $r_4 = 6$. Таким образом, при ранге $n = 4$, число инверсий рангов r_4 равно двум. Аналогичным образом подсчитаны все инверсии, сумма которых равна 13. Теперь по формуле (11.1) $\tau = 1 - \frac{4 \cdot 13}{10 \cdot 9} = 0.422$. Оценим значимость τ. Вычислим по формуле (11.2) значение статистики $t = 0.422\sqrt{\frac{9 \cdot 10 \cdot (10-1)}{2(2 \cdot 10+5)}} = 1.6985$. Так как $t < t_{0.95} = 1.96$, то ранговый коэффициент корреляции Кендалла не значим на 5%-ном уровне.

Замечания

- *Ранговые коэффициенты корреляции ρ и τ могут быть использованы и для оценки тесноты связи между обычными количественными переменными, измеряемыми в интервальных шкалах.*
- *Преимущества ρ и τ здесь заключается в том, что нахождение этих коэффициентов не требует нормального распределения переменных, линейной связи между ними (хотя и предполагает монотонность функции регрессии, отражающей эту связь).*
- *Однако необходимо учитывать, что при переходе от первоначальных значений переменных к их рангам происходит определенная потеря информации.*

ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Корреляционный анализ позволяет установить степень взаимосвязи двух и более случайных величин. Однако наряду с этим желательно иметь модель этой связи, которая дала бы возможность предсказать значения одной случайной величины по конкретным значениям другой. Например, корреляционный анализ данных (смотри пример в предыдущей лекции) установил значимую линейную связь между ростом и массой студентов.

Логичен следующий шаг: конкретизировать эту связь так, чтобы по данному росту можно было бы предсказать массу студента. Методы решения подобных задач носят название "***регрессионный анализ***".

Рассмотрим простой случай двух случайных величин (коррелированных). В этой задаче случайная величина $X -$ рост студента, а $Y -$ масса студента. Линейная связь между двумя случайными величинами означает, что прогноз значения величины y по данному значению x имеет вид:

$$\tilde{y} = A + Bx. \tag{11.3}$$

Здесь A и B отрезок оси ординат, отсекаемый прямой и её наклон соответственно.

Если исследуемые данные связаны идеальной линейной зависимостью $(r_{xy} = \pm 1)$, то предсказанное значение $\tilde{y}_i$ будет в точности соответствовать наблюденному значению y_i при любом x_i. Однако на практике идеальная линейная зависимость между данными, обычно, отсутствует. Случайные внешние воздействия приводят к разбросу данных, и, кроме того, возможны искажения за счет присутствия нелинейных эффектов.

Т. о. $\tilde{y}_i$ не обязательно совпадает с наблюденным значением y_i, соответствующим данному значению x_i, однако оно будет равно среднему значению всех таких наблюденных значений.

Общепринята такая процедура определения коэффициентов A и B, при которой минимизируется сумма квадратов отклонений наблюденных значений от предсказанных значений $\tilde{y}$. Эта процедура называется ***методом наименьших квадратов*** **(МНК)*.***

Оценим отклонение наблюденных значений от предсказанных:

$$y_i - \tilde{y}_i = y_i - (A + Bx_i) = \varepsilon_i. \tag{11.4}$$

Тогда сумма квадратов отклонений равна:

$$Q = \sum_{i=1}^{N} (y_i - A - Bx_i)^2. \tag{11.5}$$

Следовательно, лучшее согласие в смысле наименьших квадратов обеспечивают следующие условия:

$$\frac{\partial Q}{\partial A} = \frac{\partial Q}{\partial B} = 0. \tag{11.6}$$

На практике мы имеем ограниченную выборку из N пар наблюденных значений x и y. Это значит, что уравнение (11.6) даст лишь оценки A и B, которые обозначим a и b соответственно. Подставляя (11.5) в уравнение (11.6) и решив его относительно оценок величин A и B, получим:

$$\frac{\partial Q}{\partial A} = 2\sum_{i=1}^{N} (y_i - A - Bx_i) \cdot 1 = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} A - B\sum_{i=1}^{N} x_i = 0;$$

$$a = \bar{y} - b\bar{x}. \qquad (11.7)$$

$$\frac{\partial Q}{\partial B} = 2\sum_{i=1}^{n}\left(y_i - A - Bx_i\right)\cdot x_i = \sum_{i=1}^{N} y_i x_i - A\sum_{i=1}^{N} x_i - b\sum_{i=1}^{N} x_i^2 =$$

$$= \sum_{i=1}^{N} y_i x_i - (\bar{y} - b\bar{x})\cdot \bar{x}N - b\sum_{i=1}^{N} x_i^2 = 0;$$

$$b(N\cdot\bar{x}^2 - \sum_{i=1}^{N} x_i^2) = N\cdot\bar{x}\cdot\bar{y} - \sum_{i=1}^{N} y_i x_i;$$

$$b = \frac{\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\cdot\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\cdot\bar{x}^2} = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y}\cdot\frac{S_y}{S_x} = r_{xy}\cdot\frac{S_y}{S_x}. \qquad (11.8)$$

Теперь эти оценки можно использовать для построения модели, позволяющей предсказывать значения y по данному значению x:

$$\hat{y} = a + bx = (\bar{y} - b\bar{x}) + bx = \bar{y} + b(x - \bar{x}). \qquad (11.9)$$

Уравнение (11.9) называется ***прямой линейной регрессии*** y на x. Поменяем местами зависимую и независимую переменные в уравнениях (11.7) и (11.8). Тогда получим прямую регрессии x на y:

$$\hat{x} = \bar{x} + b'(y - \bar{y});$$

$$b' = \frac{\sum_{i=1}^{N} x_i y_i - N\cdot\overline{xy}}{\sum_{i=1}^{N} y_i^2 - N\cdot\bar{y}^2} = r_{xy}\cdot\frac{S_x}{S_y}. \qquad (11.10)$$

Сравнивая (11.7, 11.8) и (11.10) видим, что наклоны прямых регрессии y на x и x на y связаны с выборочным коэффициентом корреляции соотношением:

$$r_{xy} = \left[bb'\right]^{\frac{1}{2}}. \qquad (11.11)$$

Таким образом, по выборке ограниченного объема мы построили ***эмпирическое уравнение регрессии*** (11.9, 11.10) и получили оценки неизвестных параметров. Называемые ***эмпирическими коэффициентами регрессии***. Следовательно, в конкретном случае

$$y_i - \hat{y}_i = e_i, \qquad (11.12)$$

где *отклонение* (*остаток*) e_i– оценка теоретического случайного отклонения ε_i.

Проведенные рассуждения и формулы (11.4 – 11.10) позволяют сделать ряд выводов:

1. Оценки МНК являются функциями от выборки, что позволяет их легко рассчитывать.

2. Оценки МНК являются точечными оценками теоретических коэффициентов регрессии.

3. Согласно (11.9) и (11.10) эмпирическая прямая регрессии обязательно проходит через точку $(\bar{x}, \ \bar{y})$.

4. Эмпирическое уравнение регрессии построено таким образом, что сумма отклонений $\sum e_i$, а также среднее значение отклонения $\bar{e}$ равны нулю.

Действительно, из (11.7)

$$-2\sum(y_i - a - bx_i) = 0 \Rightarrow 2\sum e_i = 0 \Rightarrow \sum e_i = 0 \Rightarrow \frac{1}{n}\sum e_i = 0 \Rightarrow \bar{e} = 0.$$

5. Случайные отклонения e_i не коррелированны с наблюдаемыми значениями y_i зависимой переменной Y.

Показано, что ковариация между Y и e равна нулю:

$$S_{ye} = \frac{1}{n}\sum(y_i - \bar{y})\,(e_i - \bar{e}) = \frac{1}{n}\sum(y_i - \bar{y})e_i = 0.$$

ЛЕКЦИЯ №12

Анализ точности определения оценок коэффициентов регрессии

В силу случайного отбора элементов данных в выборку, случайными являются также оценки a и b коэффициентов A и B теоретического уравнения регрессии. Их математические ожидания при выполнении предпосылок об отклонении ε_i равны соответственно $M[a] = A, \ M[b] = B$. При этом оценки тем надежнее, чем меньше их разброс вокруг A и B, т.е. чем меньше дисперсии $D[a]$ и $D[b]$ оценок. Очевидно, надежность полученных оценок тесно связана с дисперсией случайных отклонений ε_i. Фактически $D[\varepsilon_i]$ является дисперсией $D[Y|X = x_i]$ переменной Y относительно линии регрессии (дисперсией Y, очищенной от влияния X). Полагая все измерения равноточными, считаем, что все эти дисперсии равны между собой $D[\varepsilon_i] = \sigma_\varepsilon^2 = \sigma^2$.

Покажем связь дисперсий коэффициентов $D[a]$ и $D[b]$ с дисперсией случайных отклонений ε_i. С этой целью представим зависимости коэффициентов a и b (формулы (11.7) и (11.8)) в виде линейных функций относительно значений зависимой переменной Y:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\bar{y}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \Rightarrow$$

Так как $\sum(x_i - \bar{x}) = 0$, и введя обозначение $c_i = \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$, окончательно получим:

$$b = \sum c_i y_i. \tag{12.1}$$

Аналогично:

$$a = \bar{y} - b\bar{x} = \frac{\sum y_i}{n} - \sum c_i y_i \bar{x} = \sum \left(\frac{1}{n} - c_i \bar{x}\right) y_i.$$

Обозначим $d_i = \frac{1}{n} - c_i \bar{x}$, тогда окончательно получим:

$$a = \sum d_i y_i \tag{12.2}$$

Полагая, что дисперсия Y постоянная и не зависит от значений X, можно рассматривать c_i и d_i как некоторые постоянные. Следовательно,

$$D[b] = D\left[\sum c_i\, y_i\right] = \sigma^2 \sum c_i^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}. \tag{12.3}$$

$$D[a] = D\left[\sum d_i\, y_i\right] = \sigma^2 \sum d_i^2 = \sigma^2 \sum \left(\frac{1}{n} - c_i \bar{x}\right)^2 =$$

$$= \sigma^2 \sum \left(\frac{1}{n^2} - \frac{2c_i \bar{x}}{n} + {c_i}^2 \bar{x}^2\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right). \tag{12.4}$$

Из (12.3) и (12.4) можно сделать ряд выводов.

- Дисперсии a и b прямо пропорциональны дисперсии случайного отклонения σ^2.
- Чем больше дисперсия независимой (объясняющей) переменной (разброс значений $\sum (x_i - \bar{x})^2$), тем меньше дисперсия оценок коэффициентов.

Ввиду того, что случайные отклонения ε_i по выборке определены быть не могут, при анализе надежности оценок коэффициентов регрессии они заменяются отклонениями $e_i = y_i - a - bx_i$ значений y_i переменной Y от оцененной линии регрессии. Дисперсия случайных отклонений $D[\varepsilon_i] = \sigma^2$ заменяется ее несмещенной оценкой.

$$S^2 = \frac{1}{n-2} \sum (y_i - a - bx_i)^2 = \frac{\sum e_i^2}{n-2}. \tag{12.5}$$

Тогда

$$D[b] \approx S_b^2 = \frac{S^2}{\sum (x_i - \bar{x})^2}, \tag{12.6}$$

$$D[a] \approx S_a^2 = S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right). \tag{12.7}$$

В этих выражениях S^2 – необъясненная дисперсия (мера разброса зависимой переменной относительно линии регрессии). Корень квадратный

из необъясненной дисперсии, т.е. $S = \sqrt{\frac{\sum e_i^2}{n-2}}$, называется *стандартной ошибкой оценки* (*стандартной ошибкой регрессии*). Стандартные отклонения случайных величин a и b $\quad S_a = \sqrt{S_a^2}, \quad S_b = \sqrt{S_b^2}$

называются *стандартными ошибками коэффициентов регрессии*.

Проверка гипотез относительно коэффициентов линейной регрессии

Эмпирическое уравнение регрессии определяется на основе конечного числа статистических данных. Поэтому коэффициенты эмпирического уравнения регрессии являются СВ, изменяющимися от выборки к выборке. При проведении статистического анализа перед исследователем зачастую возникает необходимость сравнения эмпирических коэффициентов регрессии a и b с некоторыми теоретически ожидаемыми значениями A и B этих коэффициентов.

Данный анализ производится в рамках статистической проверки параметрических гипотез.

Показано, что в предположении нормальности распределения y при данном значении x, оценки a и b являются несмещенными оценками A и B соответственно. Их выборочные распределения связаны с $t-$ распределением (Стьюдента), которое имеет $\nu = n - 2$ степени свободы.

На первом этапе анализа наиболее важной является задача установления линейной зависимости между переменными Y и X. С этой целью сформулируем гипотезы:

H_0: $\quad b = 0, -$ линейная зависимость отсутствует, коэффициент угла наклона прямой незначимо отличается от нуля;

H_1: $\quad b \neq 0, -$ линейная зависимость значительная и коэффициент угла наклона не равен нулю.

При проверке гипотезы воспользуемся $t -$ статистикой:

$$t = \frac{b - B}{S_b} = \frac{b}{\sqrt{S_b^2}} \qquad (12.8)$$

Аналогичным образом проверяется гипотеза о статистической значимости нулю коэффициента регрессии a (свободный член линейного уравнения равен нулю):

$$t = \frac{a - A}{S_a} = \frac{a}{\sqrt{S_a^2}} \qquad (12.9)$$

Интервальные оценки коэффициентов линейной регрессии

Как указывалось выше, коэффициенты регрессии a и b являются нормально распределенными СВ, с соответствующими дисперсиями, т.е. $a \sim N(A, D[A])$, $b \sim N(B, D[B])$. Тогда следующие статистики

$$t_a = \frac{a - A}{S_a}, \qquad t_b = \frac{b - B}{S_b} \tag{12.10}$$

имеют распределение Стьюдента с числом степеней свободы $\nu = n - 2$. Тогда, для построения доверительного интервала с заданной доверительной вероятностью β найдем по статистическим таблицам критические значения:

$$P\left(|t| < t_{1-\frac{\alpha}{2};n-2}\right) = \beta = 1 - \alpha \tag{12.11}$$

С учетом (12.10) получим:

$$P\left(t_{\frac{\alpha}{2};n-2} < \frac{a - A}{S_a} < t_{1-\frac{\alpha}{2};n-2}\right) = \beta = 1 - \alpha,$$

$$P\left(t_{\frac{\alpha}{2};n-2} < \frac{b - B}{S_b} < t_{1-\frac{\alpha}{2};n-2}\right) = \beta = 1 - \alpha. \tag{12.12}$$

Если разрешить неравенства в формулах (12.12) относительно неизвестных коэффициентов регрессии A и B, то получим соответствующие доверительные интервалы

$$I_\beta(A) = \left(a - t_{1-\frac{\alpha}{2};n-2} \cdot S_a; a + t_{1-\frac{\alpha}{2};n-2} \cdot S_a\right),$$

$$I_\beta(B) = \left(b - t_{1-\frac{\alpha}{2};n-2} \cdot S_b; b + t_{1-\frac{\alpha}{2};n-2} \cdot S_b)\right), \tag{12.13}$$

Которые с доверительной вероятностью $1 - \alpha$ накрывают определяемые параметры (теоретические коэффициенты регрессии).

Особый интерес представляет выборочное распределение $\hat{y}$ при конкретном значении $x = x_0$. Так как $\hat{y}$ ведет себя как СВ, распределенная по нормальному закону, для нее тоже можно построить доверительный интервал. Соответствующая статистика имеет вид:

$$t_{\tilde{y}} = \frac{\hat{y} - \tilde{y}}{S_{y|x}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_0 - \bar{x})^2}\right)^{\frac{1}{2}}}. \tag{12.14}$$

В выражении (12.14) величина $S_{y|x}$ – это выборочное стандартное отклонение наблюденного значения y_i от предсказанного $\hat{y} = a + bx_i$, равное

$$S_{y|x} = \left[\frac{\sum(y_i - \hat{y}_i)^2}{n-2}\right]^{\frac{1}{2}} = \left[\frac{n-1}{n-2} \cdot S_y^2(1 - r_{xy}^2)\right]^{\frac{1}{2}}. \qquad (12.15)$$

Т.о. формулы (12.13 – 12.15) дают возможность построить доверительные интервалы для неизвестных параметров A, B и $\tilde{y}$, по оценкам a, b и $\hat{y}$.

Пример 1. *ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ*

Имеется, выборка $N = 25$ пар чисел $x -$ рост студента (сантиметры), $y -$ вес (масса) (килограммы).

Задание:

1. Определим прямую регрессию, задающую линейный прогноз средней массы студента по его росту.

2. Найдем также 95% доверительный интервал для средней массы студентов, имеющих рост 178 см.

Решение

По формуле (11.8) вычислим b:

$$b = \frac{344493 - 25 \cdot 179.4 \cdot 76.64}{806105 - 25 \cdot 179.4^2} = 0.51.$$

По формуле (11.7) находим a: $\quad a = 76.64 - 0.51 \cdot 179.4 = -14.854$.

Т.о. прямая регрессии, оценивающая среднюю массу студента по его росту, имеет вид:

$$\hat{y} = -14.854 + 0.51x$$

Отсюда, для роста $x = 178$ получим $\hat{y} = 75.9$. Теперь для построения доверительного интервала для средней массы $\tilde{y}$ по оценке $\hat{y} = 75.9$ вычислим $S_{y|x}$:

$$S_{y|x} = \left[\frac{1}{23}\left(2075.76 - \frac{762.6^2}{1496}\right)\right]^{\frac{1}{2}} = 8.56.$$

Теперь по формуле (12.14) Вычислим 95% доверительный интервал:

$$\hat{y} \pm S_{y|x} \cdot t_{N-2;\alpha/2} \cdot \left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x}_0)^2}\right]^{\frac{1}{2}} =$$

$$75.9 \pm 8{,}56 \cdot t_{23;0{,}025} \cdot \left[\frac{1}{25} + \frac{(178 - 179.4)^2}{1496}\right]^{\frac{1}{2}} = 75.9 \pm 3.6$$

$$I_{0.95}(\hat{y} = 75.9) = [72.3; 79.5]$$

ЛЕКЦИЯ №13

Проверка качества уравнения регрессии

Оценим, насколько хорошо модель линейной регрессии описывает данную систему наблюдений. В качестве этой оценки воспользуемся <u>коэффициентом детерминации</u>.

Составим следующие суммы квадратов отклонений:

$S^2 = \sum(y_i - \bar{y})^2$ —фактических значений от их среднего арифметического;

$\hat{S}^2 = \sum(\hat{y}_i - \bar{y})^2$ —выравненных значений от их среднего арифметического фактических значений;

$S_R^2 = \sum(y_i - \hat{y}_i)^2$ —фактических от выравненных значений.

Можно показать, что справедливо равенство:

$$S^2 = \hat{S}^2 + S_R^2. \qquad (13.1)$$

Действительно,

$$S^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 =$$

$$= \sum_{i=1}^{N}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 +$$

$$+2\sum_{i=1}^{N}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

А последнее слагаемое представим:

$\sum_{i=1}^{N}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{N}(y_i - \hat{y}_i)\, b(x_i - \bar{x}) = b\sum_{i=1}^{N}\big(y_i - \bar{y} - b(x_i - \bar{x})\big)(x_i - \bar{x}) = b\sum_{i=1}^{N}\big(y_i - \bar{y} - b(x_i - \bar{x})\big)\cdot x_i - \bar{x}\, b\sum_{i=1}^{N}(y_i - \bar{y} - b(x_i - \bar{x}))$.

Учитывая (11.8) получим, что первая сумма равна нулю, а вторую сумму представим:

$$\bar{x}b\sum_{i=1}^{N}(y_i - \bar{y} - b(x_i - \bar{x})) = \bar{x}b\sum_{i=1}^{N} e_i \equiv 0.$$

Коэффициент детерминации – это отношение объясненной части вариации ко всей вариации в целом:

$$d = \frac{\hat{S}^2}{S^2} = \frac{S^2 - S_R^2}{S^2} = 1 - \frac{S_R^2}{S^2} \qquad (13.2)$$

Т.о. чем "ближе" этот коэффициент к 1, тем лучше модель описывает эмпирические данные, разумеется, если при этом модель методически правильна.

Проверка значимости (качества) уравнения регрессии производится на основе дисперсионного анализа.

Дисперсионный анализ – самостоятельный инструмент (метод) математической статистики будет подробно рассмотрен в дальнейшем. Пока же кратко рассмотрим схему дисперсионного анализа, представленную в виде таблицы.

Компоненты дисперсии	**Сумма квадратов**	**Число степеней свободы**	**Дисперсия**
Регрессия	$\hat{S}^2 = \sum(\hat{y}_i - \bar{y})^2$	$m-1$	$D[\hat{y}] = \frac{\hat{S}^2}{m-1}$
Остаточная	$S_R^2 = \sum(y_i - \hat{y}_i)^2$	$n-m$	$D[e] = \frac{S_r^2}{n-m}$
Общая	$S^2 = \sum(y_i - \bar{y})^2$	$n-1$	

Уравнение регрессии значимо на уровне α, если фактически наблюдаемое значение статистики удовлетворяет соотношению

$$F = \frac{\hat{S}^2(n-m)}{S_R^2(m-1)} = \frac{D[\hat{y}]}{D[e]} > F_{\alpha;k_1;k_2} \qquad (13.3)$$

Здесь $F_{\alpha;k_1;k_2}$ – критическое значение критерия Фишера – Снедекора при $k_1 = m-1$ и $k_2 = n-m$ степенях свободы; m –число оцениваемых параметров уравнения регрессии; n –число наблюдений.

В случае линейной парной регрессии $m = 2$ и уравнение регрессии значимо на уровне α, если

$$F = \frac{\hat{S}^2(n-2)}{S_R^2} > F_{\alpha;1;n-2} \qquad (13.4)$$

Оценка остатков

Остатками называются разности наблюдаемых величин и подогнанных или прогнозируемых с помощью модели.

При анализе остатков следует учитывать ряд существенных факторов:

- Если модель подобрана правильно, то остатки будут вести себя достаточно хаотично, в известном смысле они будут напоминать белый шум.
- В остатках не будет систематической составляющей, резких выбросов, в чередовании их знаков не будет никаких закономерностей, остатки будут независимы друг от друга.

Согласно общим предположениям регрессионного анализа, остатки должны вести себя как независимые одинаково распределенные случайные величины. Независимость остатков проверяется с помощью критерия Дарбина – Уотсона. Исследование остатков полезно начинать с изучения их графика. Он может показать наличие какой-либо зависимости, не учтенной в модели.

Поведение остатков e_i, $(i = 1, 2, \dots, n)$ должно имитировать поведение ошибок ε_i, $(i = 1, 2, \dots, n)$. Иначе говоря, поскольку предполагается, что ошибки ε_i, $(i = 1, 2, \dots, n)$ — независимые в совокупности случайные величины, имеющие стандартное нормальное распределение $N(0, 1)$, то ожидаем, что поведение последовательности остатков должно имитировать поведение последовательности независимых в совокупности случайных величин с распределением $N(0, 1)$.

Исходя из этих предположений, проанализируем представленный реальный график (рис.1).

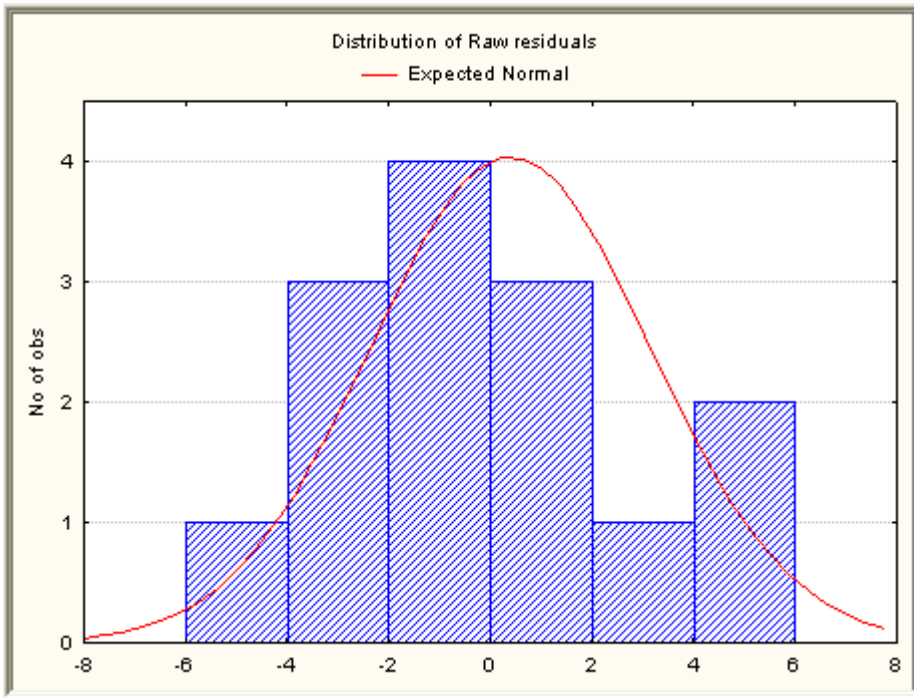


Рис. 1. Сравнение стандартизованных остатков с N (0,1)

Гистограмма остатков "почти" симметрична относительно нуля, т.е. количество отрицательных значений равно количеству положительных. Как и в нормальном распределении, количество малых остатков (разностей между наблюденными результатами и данными модели) велико, а большие остатки малы.

Рассмотрим еще одно графическое представление остатков (см. рис.2). Из графика остатков на нормальной вероятностной бумаге видно, что они достаточно хорошо ложатся на прямую, которая соответствует стандартному нормальному распределению.

Исходя из построенных графических результатов, можно сделать вывод о том, что предположение о нормальности распределения ошибок – достаточно корректно.

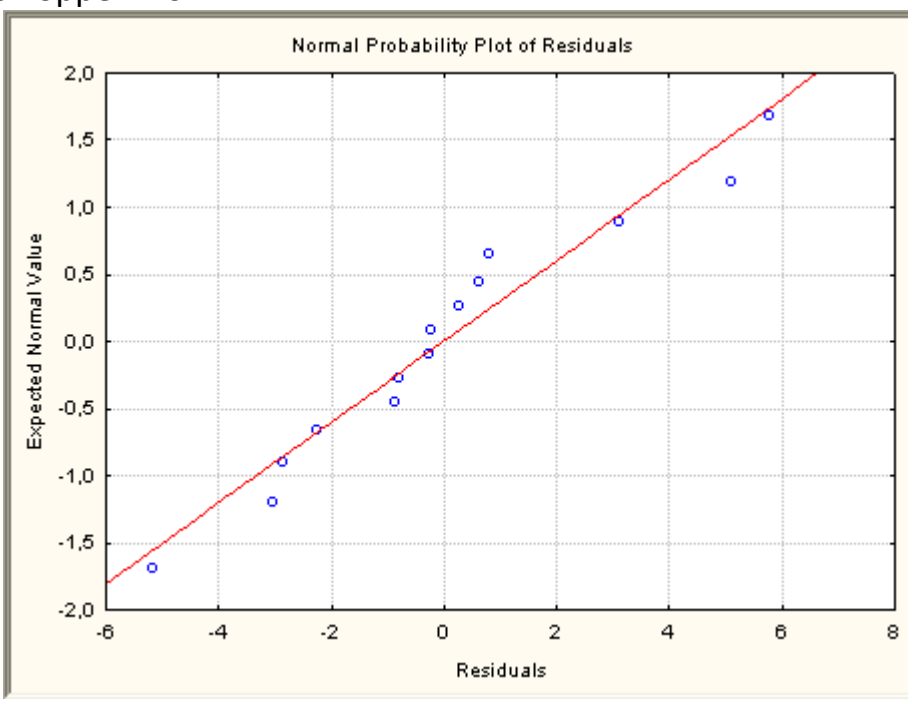


Рис. 2. График остатков на нормальной вероятностной бумаге

Критерий Дарбина – Уотсона (Durbin - Watson)

Оценивая качество уравнения регрессии, мы предполагаем, что реальная взаимосвязь переменных линейна. Отклонения от регрессионной прямой являются случайными, независимыми друг от друга величинами с нулевым математическим ожиданием и постоянной дисперсией. Если эти предположения не выполняются, то оценки коэффициентов регрессии не обладают свойствами *несмещенности*, *эффективности* и *состоятельности*. В этом случае анализ значимости полученных оценок будет неточным.

Статистика **Дарбина—Уотсона** используется для проверки гипотезы о том, что остатки построенной регрессионной модели некоррелированы (корреляции равны нулю), против альтернативы: остатки связаны авторегрессионной зависимостью (первого порядка) вида:

$$\varepsilon_i = \rho\, \varepsilon_{i-1} + \delta_i$$

На практике для анализа коррелированности отклонений вместо коэффициента корреляции используют тесно с ним связанную статистику Дарбина—Уотсона, рассчитываемую по формуле

$$d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}. \qquad (13.5)$$

Действительно,

$$\sum(e_i - e_{i-1})^2 = \sum(e_i^2 - 2e_ie_{i-1} + e_{i-1}^2) =$$
$$= \sum e_i^2 - 2\sum 2e_ie_{i-1} + \sum e_{i-1}^2 \approx 2\sum e_i^2 - 2\sum e_ie_{i-1}.$$

Здесь сделано допущение, что при больших значениях n выполняется соотношение $\sum e_i^2 \approx \sum e_{i-1}^2$.

Тогда

$$d = \frac{2(\sum e_i^2 - \sum e_ie_{i-1})}{\sum e_i^2} = 2(1 - r_{e_ie_{i-1}}). \qquad (13.6)$$

Нетрудно заметить, что если $e_i = e_{i-1}$, то $r_{e_ie_{i-1}} = 1$ и $d = 0$. Если $e_i = -e_{i-1}$ то $r_{e_ie_{i-1}} = -1$ и $d = 4$. Во всех других случаях $0 < d < 4$.

Критические точки статистики Дарбина—Уотсона табулированы для различных α, n, m. При проверке гипотезы об отсутствии автокорреляции остатков используется числовой отрезок, на котором отложены d_l – нижняя граница статистики и d_u –верхняя граница:

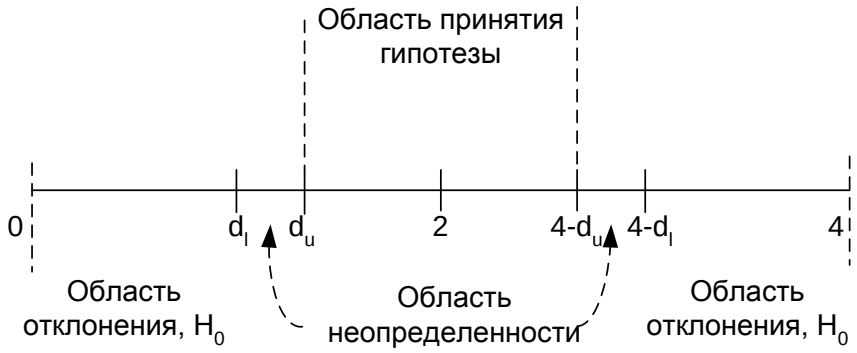


Рис. 3. Статистика Дарбина—Уотсона

Проверка гипотезы проводится по схеме:

- Если $d < d_l$, то гипотеза H_0 отклоняется, принимается H_1 – значительная положительная автокорреляция остатков;
- Если $d > 4 - d_l$, , то гипотеза H_0 отклоняется, принимается H_1 – значительная отрицательная автокорреляция остатков;
- Если $d_u < d < 4 - d_u$, то гипотеза H_0 об отсутствии автокорреляции остатков принимается;
- Если $d_l < d < d_u$, или $4 - d_u < d < 4 - d_l$, то гипотеза об отсутствии автокорреляции не может быть ни принята, ни отклонена.

Не обращаясь к таблице критических точек Дарбина—Уотсона можно воспользоваться "грубым" правилом и считать, что автокорреляция остатков отсутствует, если $1.5 < d < 2.5$. Для более надежных выводов необходимо воспользоваться статистическими таблицами.

ЛЕКЦИЯ №14

Множественная линейная регрессия

Обобщением линейной регрессионной модели с двумя переменными является многомерная регрессионная модель (или модель множественной регрессии). Уравнение множественной регрессии может быть представлено в виде

$$Y = f(\beta, X) + \varepsilon, \qquad (14.1)$$

где $X = (X_1,\ X_2, \dots,\ X_m)$ — вектор *независимых* (*объясняющих*) *переменных*; β — вектор *параметров* (подлежащих определению); ε — *случайная ошибка* (*отклонение*); Y — *зависимая* (*объясняемая*) *переменная.*

Рассмотрим самую употребляемую и наиболее простую модель множественной регрессии – модель множественной линейной регрессии.

Теоретическое линейное уравнение регрессии имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_m X_m + \varepsilon, \qquad (14.2)$$

или для индивидуальных наблюдений $i,\ \ i = 1, 2, \dots, n$,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_m x_{im} + \varepsilon_i. \qquad (14.3)$$

Здесь $\beta = (\beta_0, \beta_1, \dots, \beta_m)$ —вектор размерности $(m+1)$ неизвестных параметров. $\beta_j,\ \ j = 1, 2, \dots m$, называется j-м теоретическим коэффициентом регрессии (частичным коэффициентом регрессии). Он отражает влияние на условное математическое ожидание $M[Y|x_1, x_2, \dots x_m]$ зависимой переменной Y объясняющей переменной X_j при условии, что все другие объясняющие переменные модели остаются постоянными.

β_0 — *свободный член*, определяющий значение Y в случае, когда все объясняющие переменные X_j равны нулю.

Если число наблюдений $n < m + 1$, то существует бесконечно много различных векторов параметров, при которых линейная формула (14.3) связи между X и Y будет выполняться абсолютно точно. Если число наблюдений $n = m + 1$, то вектор $\boldsymbol{в}$ рассчитывается единственным образом. При $n > m + 1$ возникает необходимость оптимизации, т.е. оценивания параметров $\beta_0, \beta_1, \dots, \beta_m$, при которых формула (14.3) дает наилучшее приближение для имеющихся наблюдений.

В данном случае число $\nu = n - m - 1$ называется *числом степеней свободы*.

Наиболее распространенным методом оценки параметров уравнения множественной регрессии является *метод наименьших квадратов* (МНК).

Требования МНК

1. *Математическое ожидание случайного отклонения* ε_i *равно нулю для всех наблюдений*:
$$M[\varepsilon_i] = 0, \quad i = 1, 2, \dots, n.$$
2. *Гомоскедастичность* (*постоянство дисперсии отклонений*):
$$D[\varepsilon_i] = D[\varepsilon_j] = \sigma^2 \text{ для любых наблюдений } i \text{ и } j.$$
3. *Отсутствие автокорреляции.*
Случайные отклонения ε_i и ε_j являются независимыми друг от друга для всех i и j.
$$\sigma_{\varepsilon_i \varepsilon_j} = cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & \text{если} \quad i \neq j, \\ \sigma^2, & \text{если} \quad i = j. \end{cases}$$
4. *Случайное отклонение должно быть независимо от объясняющих переменных.*
$$\sigma_{\varepsilon_i x_i} = 0.$$
5. *Модель является линейной относительно параметров.*

<u>Для случая множественной линейной регрессии существенными являются еще два требования.</u>

6. *Отсутствие мультиколлинеарности*.
Между объясняющими переменными отсутствует строгая (сильная) линейная зависимость.
7. *Ошибки* ε_i, $i = 1, 2, \dots, n$ б *имеют нормальное распределение* $(\varepsilon_i \sim N(0, \sigma)$.
Выполнение данного требования важно для проверки статистических гипотез и построения интервальных оценок.

Представим выражение (14.3) в матричной форме:
$$Y = X\beta + \varepsilon. \tag{14.4}$$
Здесь $Y = (y_1, y_2, \dots, y_n)^T$ – вектор-столбец значений зависимой переменной, T – символ транспонирования, $\beta = (\beta_0, \beta_1, \dots, \beta_m)^{\mathrm{T}}$ – вектор-столбец (размерности m+1) неизвестных коэффициентов регрессии, $\varepsilon = (\varepsilon_1, \ \varepsilon_2, \dots, \ \varepsilon_n,)^T$ –вектор-столбец случайных отклонений, X –матрица размерности $n \times (m + 1)$:
$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}.$$
В этой матрице -я строка $(i = 1, 2, \dots, n)$ представляет наблюдение вектора значений независимых переменных $X_1, \ X_2, \dots, \ X_m$; единица соответствует переменной при свободном члене β_0.

Оценка коэффициентов регрессии

По аналогии с парной регрессией построим оценку $\hat{\beta}$ для вектора β так, чтобы вектор оценок зависимой переменной $\hat{Y} = X\hat{\beta}$ минимально (в смысле квадрата нормы разности) отличался от вектора Y заданных значений:

$$\left\|Y - \hat{Y}\right\|^2 = \left\|Y - X\hat{\beta}\right\|^2 \longrightarrow \min \text{ по } \hat{\beta} \qquad (14.5)$$

Решением условия (14.5), если ранг матрицы X равен $m + 1$, является оценка

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (14.6)$$

Нетрудно проверить, что эта оценка несмещенная. Ковариационная (дисперсионная) матрица оценки равна

$$D[\hat{\beta}] = \left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^T \qquad (14.7)$$

Доказана справедливость **теоремы Гаусса - Маркова**.

В условиях справедливости требований МНК (*п.*3) *оценка* (14.6) *является наилучшей* (*в смысле минимума дисперсии*) *оценкой в классе линейных несмещенных оценок.*

Оценка дисперсии σ^2 ошибок

Обозначим

$$e = Y - \hat{Y} = Y - X\hat{\beta} = [I - X(X^T X)^{-1} X^T]Y = BY. \qquad (14.8)$$

e —вектор остатков (или невязок); $B = I - X(X^T X)^{-1} X^T$ — матрица. Можно проверить, что $B^2 = B$. Для остаточной суммы квадратов $\|e\|^2$ справедливо соотношение

$$M\|e\|^2 = M[\sum e_i^2] = (n - m - 1)\sigma^2,$$

откуда следует, что несмещенной оценкой для σ^2 является

$$S^2 = \frac{\sum e_i^2}{n - m - 1} = \frac{Y^T BY}{n - m - 1}. \qquad (14.9)$$

Если справедливо требование МНК (п.7), т.е. $\varepsilon_i \sim N(0,\ \sigma)$, то справедливы следующие свойства оценок:

- $(n - m - 1)\frac{S^2}{\sigma^2}$ —имеет распределение хи-квадрат χ^2_{n-m-1} с $n - m - 1$ степенями свободы;
- оценки $\hat{\beta}$ и S^2 независимы.

Как и в случае парной регрессии, справедливо соотношение:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2.$$

По аналогии с (13.1) запишем

$$S^2 = \hat{S}^2 + S_R^2. \qquad (14.10)$$

В векторном виде:

$$\|Y - \bar{Y}\|^2 = \left\|Y - \hat{Y}\right\|^2 + \left\|\hat{Y} - \bar{Y}\right\|^2. \qquad (14.11)$$

Для проверки качества уравнения множественной регрессии, как и в случае парной регрессии, воспользуемся *коэффициентом детерминации*:

$$R^2 = 1 - \frac{\left\|Y - \hat{Y}\right\|^2}{\|Y - \bar{Y}\|^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}. \qquad (14.12)$$

Коэффициент R^2 показывает качество подгонки регрессионной модели к наблюдённым значениям y_i. Если $R^2 = 0$, то регрессия Y на $x_1, x_2, \dots x_m$ не улучшает качество предсказания y_i по сравнению с тривиальным предсказанием $\hat{y}_i = \bar{y}$. Другой крайний случай $R^2 = 1$ означает точную подгонку: все $e_i = 0$, т.е. все точки наблюдений лежат на регрессионной плоскости.

Однако значение R^2 возрастает с ростом числа переменных (регрессоров) в многомерной регрессии, что не означает улучшения качества предсказания, и поэтому вводится *скорректированный* (*adjusted*) *коэффициент детерминации*:

$$R_{adj}^2 = 1 - \frac{\left\|Y - \hat{Y}\right\|^2 / (n - m - 1)}{\|Y - \bar{Y}\|^2 / (n - 1)} =$$

$$= 1 - (1 - R^2)\frac{(n - 1)}{(n - m - 1)}. \qquad (14.13)$$

Его использование более корректно для сравнения регрессий при изменении числа переменных (регрессоров).

Доверительные интервалы для коэффициентов регрессии

Стандартной ошибкой оценки $\hat{\beta}_j$ является величина $\sigma\sqrt{z_{jj}}$, оценка для которой

$$S_j = S\sqrt{z_{jj}}, \qquad j = 0, 1, \dots, m \qquad (14.14)$$

Здесь z_{jj} —диагональный элемент матрицы $Z = (X^T X)^{-1}$. Если ошибки распределены нормально $\varepsilon_i \sim N(0,\ \sigma)$, то, статистика

$$t = \frac{\left(\hat{\beta}_j - \beta_j\right)/\sigma\sqrt{z_{jj}}}{S/\sigma} = \frac{\hat{\beta}_j - \beta_j}{S_j}, \qquad (14.15)$$

распределена по закону Стьюдента с $k = (n - m - 1)$ степенями свободы. Тогда при доверительной вероятности $P = 1 - \alpha$ соответствующий

доверительный интервал вычисляется по формуле:

$$I_{\alpha,\beta_j} = \hat{\beta}_j \pm t_{1-\frac{\alpha}{2};k} \cdot S_j \qquad (14.16)$$

Проверка гипотезы о нулевых значениях коэффициентов регрессии

Для проверки гипотезы H_0 об отсутствии какой бы то ни было линейной связи между y и совокупностью факторов, H_0: $\beta_1 = \beta_2 = \cdots = \beta_m = 0$, т.е. об одновременном равенстве нулю всех коэффициентов при независимых переменных, кроме коэффициента β_0, используется статистика

$$F = \frac{R^2/m}{(1-R^2)/(n-m-1)} = \frac{\sum(y_i - \bar{y})^2}{\sum e_i^2} \cdot \frac{(n-m-1)}{m}. \qquad (14.17)$$

Статистика F —статистика Фишера – Снедекора при $k_1 = m$ и $k_2 = n - m - 1$ степенях свободы; m —число оцениваемых параметров уравнения регрессии; n —число наблюдений. Если $F < F_{\alpha;k_{1;k_2}}$, то верна гипотеза H_0 – линейная связь между зависимой и независимыми переменными отсутствует. Если $F > F_{\alpha;k_{1;k_2}}$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1 —линейная связь значима на уровне α. Здесь $F_{\alpha;k_1;k_2}$ – критическое значение критерия Фишера – Снедекора.

Замечание

Для выбора наиболее существенных объясняющих переменных можно предложить следующий практический подход. Строятся различные модели многомерной линейной регрессии (*с различным набором переменных*). *Затем можно сравнить скорректированные коэффициенты детерминации* (14.13) *и принять тот вариант регрессии, для которого* R^2_{adj} *максимален.*

ЛЕКЦИЯ №15

Выполнение многомерного регрессионного анализа в пакете *STATISTICA*

Рассмотрим пример построения регрессионной модели в пакете Statistica 6.0. Для этих целей обычно используется модуль **Multiple Regressions** (Множественная регрессия), который позволяет предсказать зависимую переменную по нескольким независимым переменным.

В стартовом диалоговом окне этого модуля (рис.1) при помощи кнопки **Variables** указываются зависимая (dependent) и независимые(ая) (independent) переменные. В поле **Input file** указывается тип файла с данными:

- **Raw Data** - данные в виде строчной таблицы;
- **Correlation Matrix** - данные в виде корреляционной матрицы.

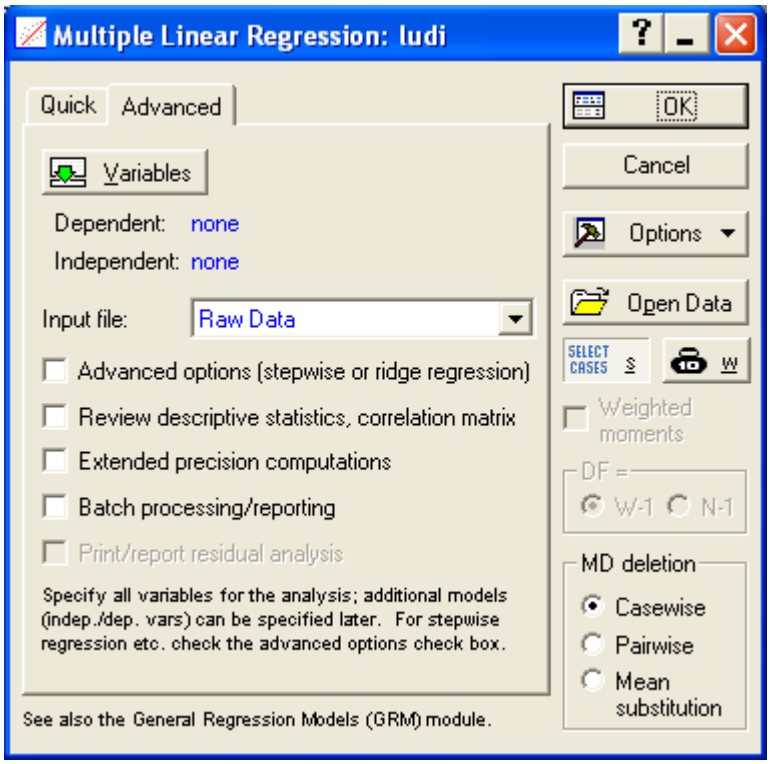


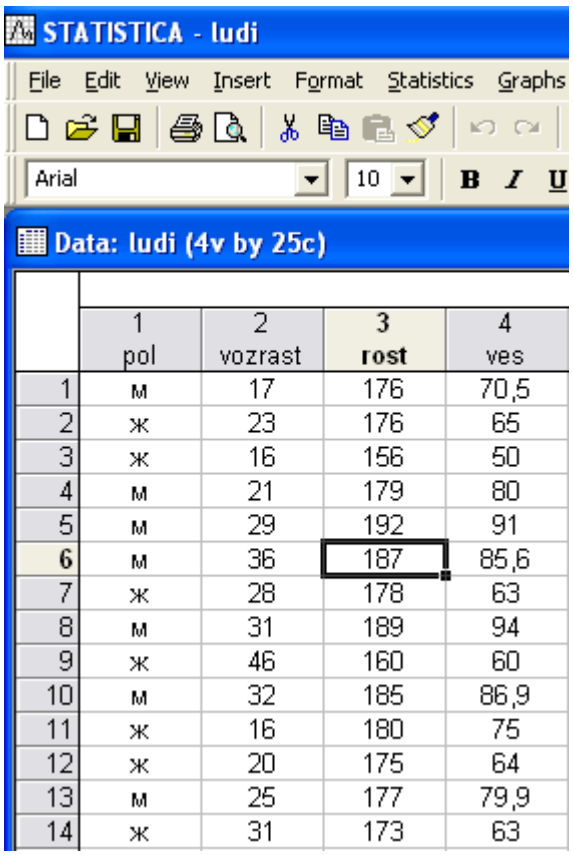
Рис.1. Модуль Multiple Regression

В поле **MD deletion** указывается способ исключения из обработки недостающих данных:

- **Casewise - игнорируется вся строка, в которой есть хотя бы одно пропущенное значение;**
- **Mean Substitution - взамен пропущенных данных подставляются средние значения переменной;**
- **Pairwise - попарное исключение данных с пропусками из тех переменных, корреляция которых вычисляется.**

Рассмотрим проведение регрессионного анализа на конкретном примере. Имеются результаты измерения физических данных 25 людей (мужчин и женщин). В файле данных (рис.2) 4 переменные:

1	pol	Пол обследуемого(ж – женщина; м – мужчина)
2	vozrast	Возраст обследуемого, лет
	rost	Рост обследуемого, см
	ves	Вес обследуемого, кг



	1 pol	2 vozrast	3 rost	4 ves
1	м	17	176	70,5
2	ж	23	176	65
3	ж	16	156	50
4	м	21	179	80
5	м	29	192	91
6	м	36	187	85,6
7	ж	28	178	63
8	м	31	189	94
9	ж	46	160	60
10	м	32	185	86,9
11	ж	16	180	75
12	ж	20	175	64
13	м	25	177	79,9
14	ж	31	173	63

Рис. 2. Окно файла данных

Так как в файле данных содержится информация о мужчинах и женщинах, а мы хотим провести исследования только для мужчин, то воспользовавшись кнопкой Select cases (рис. 1) можно в анализ включить только те случаи, для которых первая переменная (pol) равна "м".

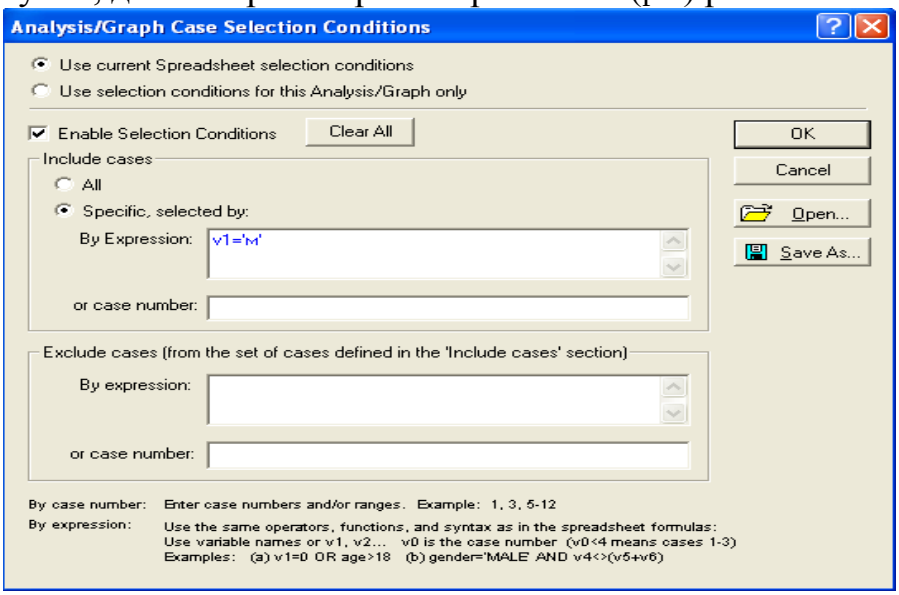


Рис. 3. Окно включения (исключения) данных в анализ

На первом этапе исследований учтем, что при наличии одной зависимой переменной (rost) и двух независимых переменных (vozrast и rost) можно предложить различные модели линейной регрессии:

№ Модели	Вид зависимости	Комментарии
1	**rost=$a_0 + a_1 \cdot$ vozrast**	одномерная

2	**rost=**$\boldsymbol{c_0 + c_1 \cdot}$ **ves**	одномерная
3	**rost=**$\boldsymbol{b_0 + b_1 \cdot}$ **vozrast** $\boldsymbol{+\cdot b_2 \cdot}$ **ves**	многомерная

О качестве предложенной модели регрессии будем судить по величине *коэффициента детерминации*.

Модель №1 описывает 69% данных, модель №2 только 41% данных, а третья модель 73% данных.

Если в качестве критерия оптимизации выбрать простоту модели (одномерная) – выберем модель №1 или №2, но если добавить ещё один критерий – максимальный % описания данных, то из этих двух моделей выбираем модель №1. Теперь в качестве главного критерия оптимизации выбираем максимальный процент описания данных и сравниваем модели №1 и №3. Нужно сказать, что модель №3 – многомерная, а модель №1 – одномерная. Таким образом, на первом этапе можно сказать, что многомерная модель №3 более адекватна и лучше описывает исходные данные. Естественно предположить, что и предсказания по модели №3 будут более надежными (точными).

Теперь более подробно рассмотрим последовательность действий создания модели и анализ полученных результатов.

После выбора всех опций стартового диалогового окна регрессионного анализа и нажатия кнопки **ОК** появляется окно результатов регрессионного анализа Multiple Regressions Results (см. рис. 4). Детально проанализируем полученные результаты регрессионной модели.

В верхней части окна приведены наиболее важные параметры полученной регрессионной модели:

- **Multiple R** - коэффициент множественной корреляции, который характеризует тесноту линейной связи между зависимой и всеми независимыми переменными. Может принимать значения от 0 до 1.
- $\mathbf{R^2}$ **-** коэффициент детерминации. Численно выражает долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения. Чем больше $\mathbf{R^2}$, тем большую долю вариации объясняют переменные, включенные в модель.
- **adjusted R** - скорректированный коэффициент множественной корреляции. Включение новой переменной в регрессионное уравнение увеличивает $\mathbf{R^2}$ не всегда, а только в том случае, когда частный **F**-критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение $\mathbf{R^2}$ и **adjusted R**.

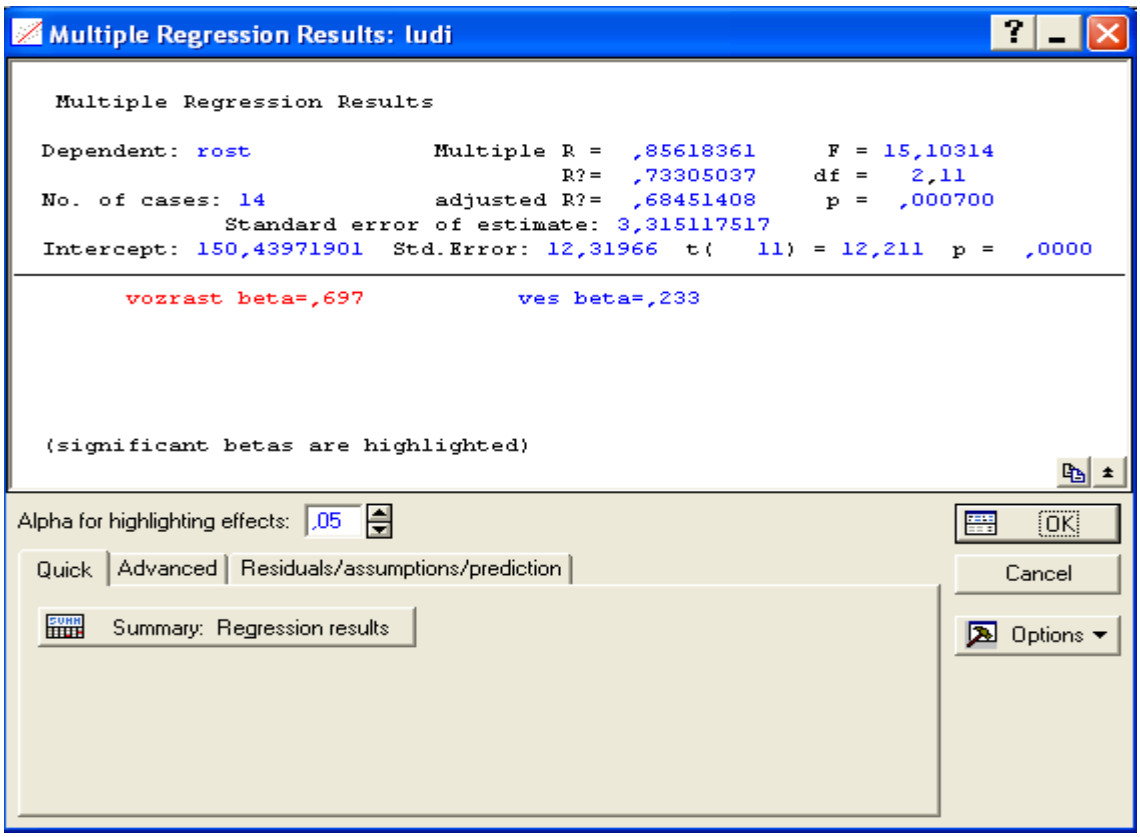


Рис. 4. Результаты регрессионного анализа

- **F** - F-критерий используется для проверки значимости регрессии. В данном случае в качестве нулевой гипотезы проверяется гипотеза: между зависимой и независимыми переменными нет линейной зависимости;
- **df** - числа степеней свободы для F-критерия;
- **p** - вероятность нулевой гипотезы для F-критерия;
- **Standard error of estimate** - стандартная ошибка оценки (уравнения); Эта оценка является мерой рассеяния наблюденных значений относительно регрессионной прямой;
- **Intercept** – оценка свободного члена уравнения;
- **Std.Error** - стандартная ошибка оценки свободного члена уравнения;
- **t** - t-критерий для оценки свободного члена уравнения;
- **p** - вероятность нулевой гипотезы для свободного члена уравнения.
- **Beta** - β-коэффициенты уравнения. Это стандартизированные регрессионные коэффициенты, рассчитанные по стандартизированным значениям переменных. По их величине можно оценить значимость зависимых переменных. Коэффициент показывает, на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной, при условии постоянства остальных независимых переменных. Свободный член в таком уравнении равен 0.

Нажатие кнопки

- в окне результатов (см рис. 4) позволяет получить основные результаты регрессионной модели (рис. 5), часть из которых уже была описана: **B** - коэффициенты уравнения регрессии; **St. Err. of B** - стандартные ошибки коэффициентов уравнения регрессии; **t** (11) - t-критерий для коэффициентов уравнения регрессии; **p-level** - вероятность нулевой гипотезы для коэффициентов уравнения регрессии.

N=14	Regression Summary for Dependent Variable: rost (ludi) R= ,85618361 R?= ,73305037 Adjusted R?= ,68451408 F(2,11)=15,103 p<,00070 Std.Error of estimate: 3,3151					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(11)	p-level
Intercept			150,4397	12,31966	12,21135	0,000000
vozrast	0,696927	0,194002	0,6050	0,16842	3,59237	0,004226
ves	0,232612	0,194002	0,2081	0,17360	1,19902	0,255713

Рис. 5. Параметры уравнения регрессии

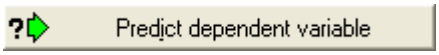
В результате проведенного анализа было получено следующее уравнение:

rost = 150,4397 + 0,605*vozrast + 0,2081*ves.

Это уравнение объясняет 73,3% ($R^2 = 0.733$) вариации зависимой переменной. Полученные результаты свидетельствуют о том что коэффициент $\boldsymbol{b_3}$ при переменной **ves** незначимо отличается от нуля, однако включение этой переменной в регрессионную модель увеличивает на 4 % процент исходных данных, корректно описанных регрессионным уравнением.

Проверка качества уравнения регрессии осуществлялась с помощью статистики $F = 15.10314$. По статистическим таблицам Фишера – Снедекора с данными степенями свободы $(\mathrm{df} = 2{,}11)$ гипотезу H_0 (линейная зависимость отсутствует) можно принять с вероятностью $(p = 0.000700)$; при уровне значимости $\alpha = 0.05$ принимаем альтернативную гипотезу – линейная зависимость значима.

Одновременно проверялась статистическая значимость коэффициентов множественной регрессии (критерий Стьюдента). Видно (см. рис. 5), что коэффициенты $\boldsymbol{b_0}$ и $\boldsymbol{b_1}$ значимо отличаются от нуля, коэффициент $\boldsymbol{b_2}$ незначимо отличается от нуля.

Для расчета по полученному регрессионному уравнению значений зависимой переменной по значениям независимых переменных воспользуемся кнопкой 

(раздел Residuals/assumptions/prediction) (рис.6).

Зададим значения возраста (vozrast = 23) и веса (ves = 65). Учтем, что в пакете Statistica приводится как точечная, так и интервальная оценка (рис. 7).

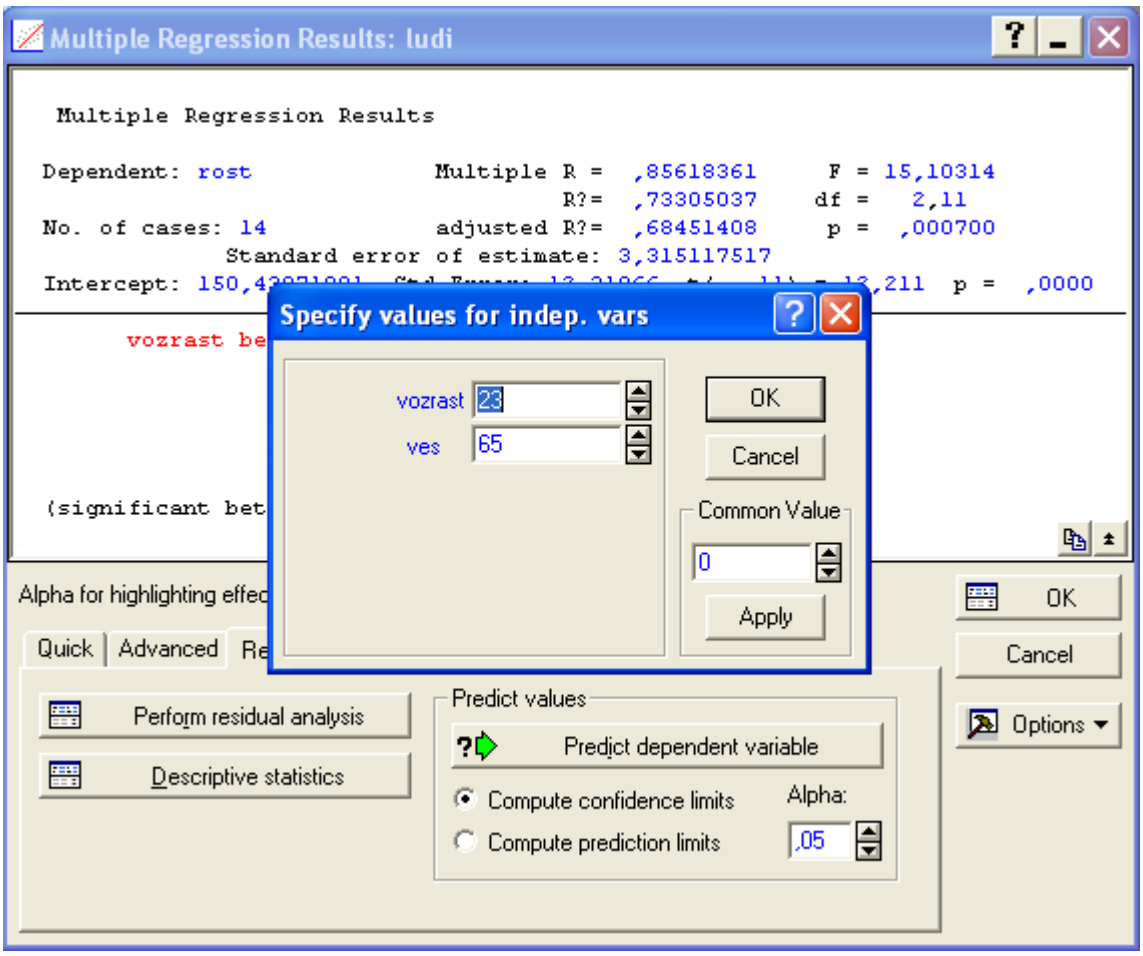


Рис. 6. Окно задание значений независимых переменных

Variable	Predicting Values for (ludi) variable: rost		
	B-Weight	Value	B-Weight * Value
vozrast	0,605038	23,00000	13,9159
ves	0,208146	65,00000	13,5295
Intercept			150,4397
Predicted			177,8851
-95,0%CL			171,4087
+95,0%CL			184,3615

Рис. 7. Предсказанные точечные и интервальные значения

О полученных результатах можно сказать следующее: rost = 177,8851 – это точечная оценка. 95% доверительный интервал равен (171.4; 184,4).

При нажатии на кнопку Summary: Residuals & predicted можно оценить величины остатков и специальных критериев (см. рис. 8).

В таблицу включены все случаи (м), приведены исходные данные (Observed), данные модели (Predicted) и остатки (Residual). Остатки – это разность исходных и предсказанных данных.

STATISTICA - Workbook_posl* - [Predicted & Residual Values (ludi)] - [Workbook_posl* - Predicted & Residu

File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help

Add to Workbook Add to Report

Arial 10

Predicted & Residual Values (ludi)
Dependent variable: rost

Case No.	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	176,0000	175,3997	0,60031	-1,75846	0,18108	2,032792	3,959432	0,96204	0,010555
4	179,0000	179,7972	-0,79723	-0,88823	-0,24048	1,229301	0,858995	-0,92433	0,003563
5	192,0000	186,9271	5,07286	0,52271	1,53022	1,482563	1,671421	6,34107	0,243913
6	Predicted & Residual Values (ludi)		03841	1,13840	-0,91653	1,548817	1,908993	-3,88679	0,100015
8	189,0000	188,7617	0,23834	0,88574	0,07190	1,801455	2,910202	0,33821	0,001024
10	185,0000	187,8889	-2,88886	0,71302	-0,87142	1,102226	0,508527	-3,24790	0,035369
13	177,0000	182,1966	-5,19656	-0,41342	-1,56753	1,004908	0,265962	-5,72238	0,091262
15	181,0000	181,2348	-0,23485	-0,60374	-0,07084	1,312575	1,109378	-0,27851	0,000369
16	193,0000	189,8941	3,10588	1,10985	0,93688	1,987440	3,743760	4,84847	0,256260
18	189,0000	183,2387	5,76129	-0,20719	1,73788	0,907928	0,046526	6,22847	0,088257
19	177,0000	176,2129	0,78714	-1,59754	0,23744	1,898845	3,336482	1,17148	0,013656
20	180,0000	182,2756	-2,27557	-0,39779	-0,68642	1,900110	3,342165	-3,38888	0,114434
22	190,0000	190,8724	-0,87241	1,30344	-0,26316	1,630399	2,215796	-1,15074	0,009715
24	185,0000	185,2620	-0,26196	0,19319	-0,07902	0,942573	0,122362	-0,28500	0,000199
Minimum	176,0000	175,3997	-5,19656	-1,75846	-1,56753	0,907928	0,046526	-5,72238	0,000199
Maximum	193,0000	190,8724	5,76129	1,30344	1,73788	2,032792	3,959432	6,34107	0,256260
Mean	184,2857	184,2857	-0,00000	-0,00000	-0,00000	1,484424	1,857143	0,07180	0,069185
Median	185,0000	184,2503	-0,24841	-0,00700	-0,07493	1,515690	1,790207	-0,28176	0,024513

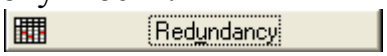
Рис. 8. Таблица остатков

Для выделения имеющихся в регрессионных остатках выбросов предложен ряд дополнительных показателей:

- Расстояние Кука (Cook's Distance) - принимает только положительное значение и показывает расстояние между коэффициентами уравнения регрессии после исключения из обработки i-ой точки данных. Большое значение показателя Кука указывает на сильно влияющий случай (выброс).

В нашем случае Case № 5, 16 и 20 смещают оценки коэффициентов регрессии.

- Расстояние Махаланобиса (Mahalns. Distance) - показывает насколько каждый случай или точка в p-мерном пространстве независимых переменных отклоняется от центра статистической совокупности.

Кнопка 

(раздел Advanced) предназначена для поиска выбросов. Выбросы – это остатки, которые значительно превосходят по абсолютной величине остальные. Выбросы показывают опытные данные, которые являются не типичными по отношению к остальным данным, и требует выяснения причин их возникновения. Выбросы должны исключаться из обработки, если они вызваны ошибками регистрации, измерения и т.п.

ЛЕКЦИЯ №16

Нелинейная регрессия

Связь между признаком x и y может быть нелинейной, например, в виде полинома:

$$y = P_k(x) + \varepsilon. \qquad (16.1)$$

Здесь $P_k = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$, $k -$ степень полинома, $\varepsilon -$ случайная составляющая, $M[\varepsilon] = 0$, $D[\varepsilon] = \sigma^2$.

Для имеющихся данных $(x_i,\ y_i)$ $i = 1, 2, \dots, n$ можно записать

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k, \quad (i = \overline{1, n}). \qquad (16.2)$$

По аналогии с (14.4) в матричной форме получим:

$$Y = X\beta + \varepsilon, \qquad (16.3)$$

где $X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ \dots & & \dots & & \dots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}$.

Таким образом, получили формулы аналогичные многомерной регрессии. Слово “линейный” в названии “линейный регрессионный анализ” означает линейность относительно параметров β_j, но не относительно факторов x_j. Кроме полиномиальной формы широко используются, например, следующие модели:

- *логарифмическая*; если зависимость $y = a_0 x^{a_1}$, то после логарифмирования получим:
$$\ln y = \ln a_0 + a_1 \ln x = \beta_0 + \beta_1 \cdot \ln x;$$
- *гиперболическая;* (при обратной зависимости, т.е. при увеличении x признак y уменьшается):
$$y = \beta_0 + \frac{\beta_1}{x}.$$
- *тригонометрическая:*
$$y = \beta_0 + \beta_1 \cdot \sin\omega x + \beta_2 \cdot \cos\omega x, \text{ и многие другие.}$$

Правильный выбор вида модели является отправной точкой для качественного ее анализа. На практике неизвестно, какая модель является верной. Поэтому зачастую подбирают такую модель, которая наиболее точно соответствует реальным данным. Учитывая, что идеальной модели не существует, попытаемся сформулировать критерии, позволяющие остановить свой выбор на качественной модели

Обычно для построения "хорошей" работоспособной модели и сравнения ее с другими возможными моделями необходимо учитывать следующие свойства (критерии).

Скупость (*простота*). Модель должна быть максимально простой. Это свойство обусловлено тем фактом, что модель не отражает действительность

идеально, а является ее упрощением. Поэтому из двух моделей, примерно одинаково отражающих реальность, предпочтительнее та, которая содержит меньшее число объясняющих переменных.

Единственность. Для любого набора статистических данных определяемые параметры (коэффициенты) должны вычисляться однозначно.

Максимальное соответствие. Уравнение (модель) тем лучше, чем большую часть разброса зависимой переменной оно может объяснить. Поэтому стремятся построить модель с максимально возможным скорректированным коэффициентом детерминации R^2_{adj}.

Согласованность с теорией. Никакое уравнение не может быть признано качественным, если оно не соответствует известным теоретическим предпосылкам.

Прогнозные качества. Модель может быть признана качественной, если полученные на ее основе прогнозы подтверждаются реальностью.

Таким критерием прогнозных качеств оцененной модели регрессии может служить отношение:

$$V = \frac{S}{\bar{y}}, \qquad (16.4)$$

где $S = \sqrt{\frac{\sum e_i^2}{n-m-1}}$ – стандартная ошибка регрессии, $\bar{y}$ – среднее значение зависимой переменной уравнения регрессии. Если величина мала (а она определяет относительную ошибку прогноза в процентах) и отсутствует автокорреляция (критерий Дарбина – Уотсона), то прогнозные качества модели высоки.

Если уравнение регрессии используется для прогнозирования, то величина обычно рассчитывается не для того периода (данных), на котором оценивались параметры уравнения, а для некоторого следующего за ним временного интервала, для которого известны значения зависимой объясняющей переменных.

Поскольку не существует какого либо единого правила построения регрессионных моделей, анализ перечисленных свойств и практический опыт, опирающийся на глубокие знания теории и статистического анализа, позволяют строить более качественные модели.

Рассмотрим практический пример: Как и в предыдущих задачах исследования будем проводить в пакете Statistica.

Все исходные данные сведены в таблицу.

	Y	X
1	8,475	0,1
2	9,21	0,2
3	10,181	0,3
4	11,386	0,4
5	12,851	0,5
6	14,877	0,6
7	16,834	0,7
8	19,634	0,8
9	22,973	0,9
10	27,139	1
11	31,773	1,1
12	37,678	1,2
13	44,99	1,3
14	54,074	1,4
15	65,106	1,5

Вначале построим линейную зависимость (модель)

$$Y = B_0 + B_1 \cdot X.$$

Заполним все опции стартового диалогового окна Multiple Regressions и выполним анализ. Проанализируем параметры модели и ее качество (рис.1).

	Regression Summary for Dependent Variable: Y (regressanal.s R= ,93729158 R?= ,87851551 Adjusted R?= ,86917055 F(1,13)=94,010 p<,00000 Std.Error of estimate: 6,3623					
N=15	Beta	Std.Err. of Beta	B	Std.Err. of B	t(13)	p-level
Intercept			-3,6805	3,45703	-1,0646	0,30640
X	0,93729	0,09666	36,8658	3,80222	9,6958	0,00000

Рис.1. Основные результаты регрессионного анализа

Построенная модель имеет вид: $Y = -3.681 + 36.866 \cdot X$. Это уравнение объясняет 87.9% ($R^2 = 0.879$) вариации зависимой переменной Y, скорректированный коэффициент детерминации равен $R^2_{adj} = 0.869$. Что касается качества модели, то линейная зависимость существенна ($F = 94.01$). Относительно коэффициентов уравнения – выводы неоднозначны: свободный член уравнения B_0 незначимо отличается от нуля, а коэффициент B_1 значимо отличен от нуля. При подробном анализе модели необходимо тщательно проанализировать остатки (графически и аналитически). Мы же ограничимся только графиком, на котором представлены предсказанные данные (сплошная линия) и исходные данные (рис.2).

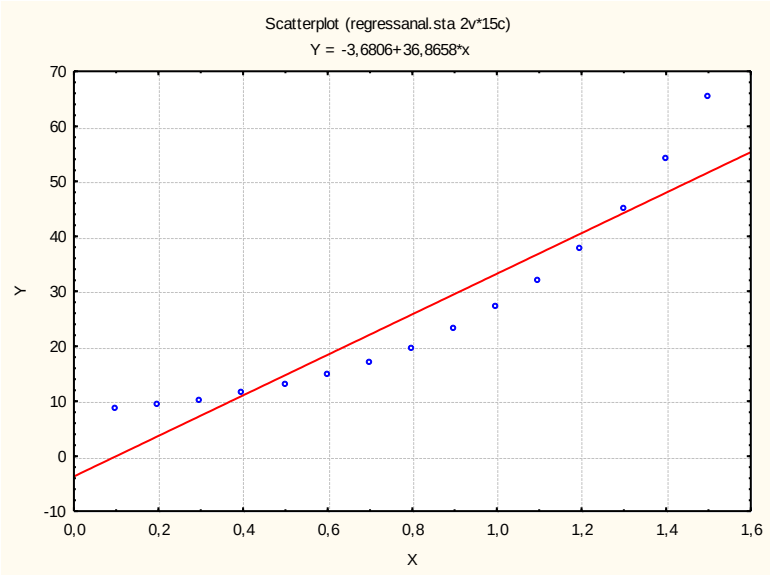


Рис. 2. Сравнение исходных данных с результатами линейной модели

Анализ графика исходных данных свидетельствует об их нелинейной зависимости. Выскажем предположение, что искомая функциональная зависимость имеет экспоненциальный вид:

$$Y = c + \exp(b_0 + b_1 \cdot x + \cdots).$$

Для получения уравнения экспоненциальной регрессии воспользуемся пунктом **Exponential growth regression** в меню Statistics → Advanced Linear/Nonlinear Models (см. рис.3.)

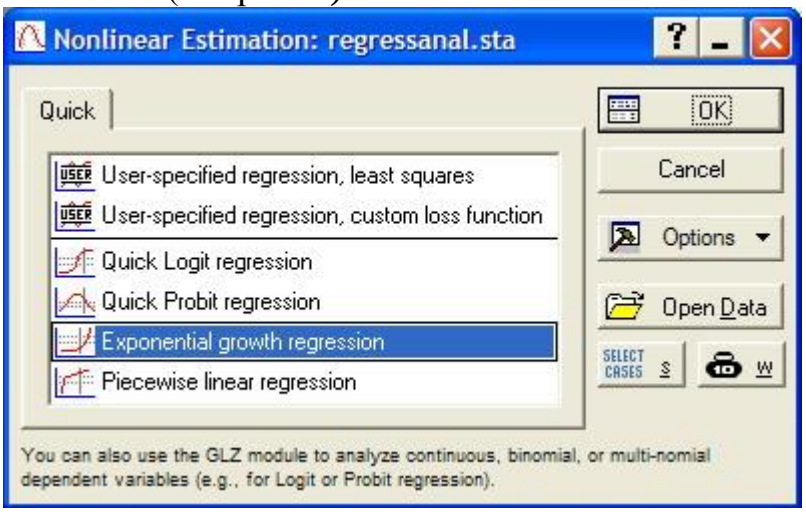


Рис. 3. Модуль экспоненциальной зависимости

Выбрав зависимую и независимую переменные, перейдем к настройкам проведения регрессионного анализа (рис.4). Выберем квази-ньютоновский метод для вычисления параметров модели. Квази-ньютоновский метод вычисляет значения функции в различных точках для оценивания первой (тангенс угла наклона графика функции в конкретной точке) и второй (скорость изменения угла наклона) производной, используя эти данные для определения направления изменения параметров и минимизации функции потерь. Кроме метода вычислений укажем максимальное число итераций и критерий сходимости.

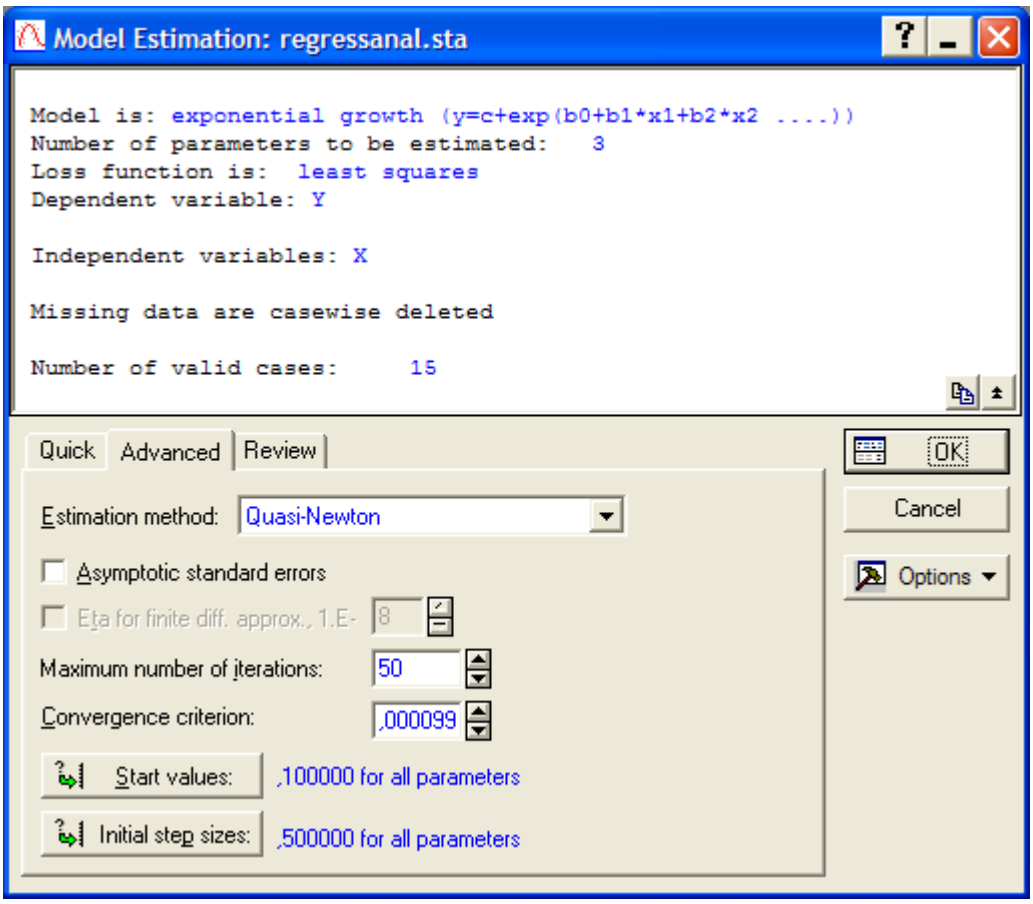


Рис.4. Выбор метода оценки коэффициентов экспоненциальной регрессии

Проведем регрессионный анализ, для чего нажимаем «ОК».

Для получения полной информации о коэффициентах регрессионного уравнения выбираем пункт Summary: Parameter estimates и получим результат (рис. 5).

	Model is: exponential growth (y=c+exp(b0+b1*x1+b2*x2 ....)) (re Dependent variable: Y Loss: Least squares Final loss: ,152868627 R=,99998 Variance explained: 99,996%		
N=15	Const.C	Const.B0	X
Estimate	4,794573	1,092713	2,003539

Рис.5. Значения коэффициентов экспоненциальной модели.

Построенная модель имеет вид:

$$y = 4.7946 + \exp(1.0927 + 2.0035 \cdot x).$$

О качестве модели можно судить хотя бы по тому, что она описывает 99.996 % зависимой переменной. При проведении исследований часто полезным бывает использование диаграммы рассеяния наблюдаемых и предсказанных значений - Fitted 2D function & observed values (рис.6). Вспомним, что для линейной модели аналогичные результаты приведены на рис. 2.

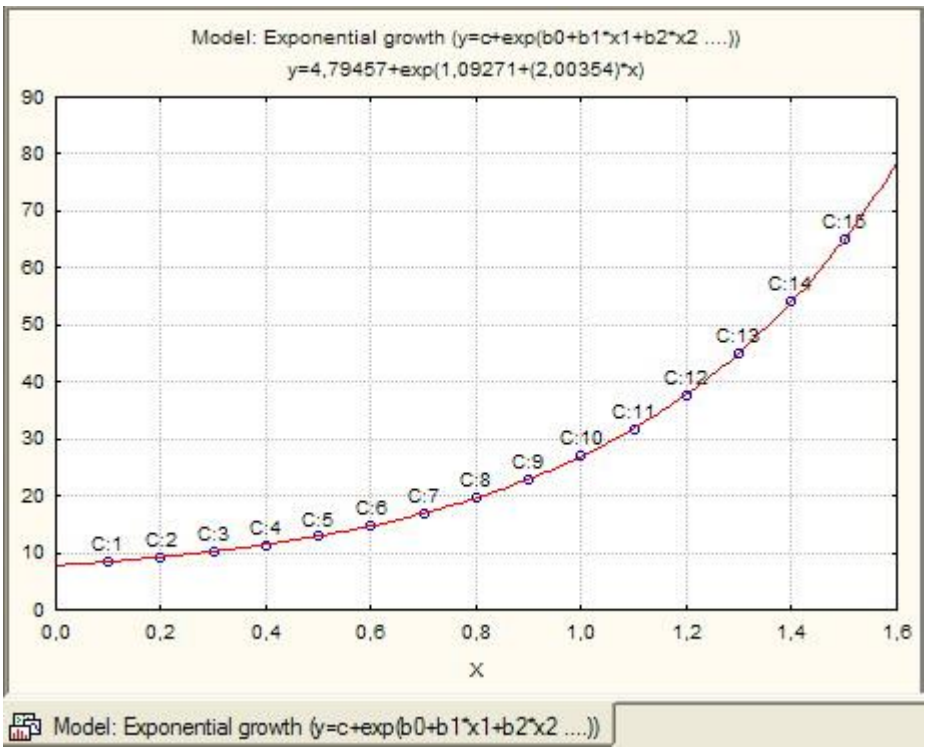


Рис. 6. Сравнение исходных данных с результатами экспоненциальной модели

Как свидетельствует построенный график, предложенная экспоненциальная модель идеально описывает исходные данные (все исходные точки легли на теоретическую кривую).

ЛЕКЦИЯ №17

ФАКТОРНЫЙ АНАЛИЗ

Ранее была рассмотрена проверка значимости различия выборочных средних двух совокупностей. На практике часто возникает необходимость обобщения задачи, т.е. проверки существенности различия выборочных средних m совокупностей $(m > 2)$. Например, требуется оценить влияние различных режимов плавок на механические свойства металла, свойств сырья на показатели качества продукции, количества вносимых удобрений на урожайность и т.п.

Для эффективного решения таких задач нужен новый подход, который реализуется в рамках *теории ранговых критериев* и *в дисперсионном анализе*.

Ранговые критерии (*непараметрические критерии*) – *предназначены, в первую очередь, для задач проверки согласия и сравнения двух и более выборок*.

Ранговые критерии применяются не только при отсутствии информации о виде распределения, но и тогда, когда наблюдения могут быть только упорядочены, как это часто бывает в социологии, медицине и т.п.

Дисперсионный анализ *– это статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для последующего планирования аналогичных экспериментов.*

Первоначально (1918 г.) дисперсионный анализ был разработан английским математиком – статистиком Р.А. Фишером. По числу факторов, влияние которых исследуется, различают однофакторный и многофакторный анализ.

Задача однофакторного анализа

При исследовании зависимостей одной из наиболее простых является ситуация, когда можно указать только один фактор, влияющий на конечный результат, и этот фактор может принимать лишь конечное число значений (уровней). Такие задачи, называемые ***задачами однофакторного анализа***, весьма часто встречаются на практике. Типичный пример задач однофакторного анализа – сравнение по достигаемым результатам нескольких различных способов действия, направленных на достижение одной цели, скажем, нескольких школьных учебников или нескольких лекарств.

Терминология

Фактором (***факторами***) называют то, что, как мы считаем, должно оказывать влияние на конечный результат. В приведенных выше примерах факторами являются понятия "школьный учебник" или "лекарство".

Уровень фактора или ***способ обработки*** **–** это конкретная реализация фактора (например, определённый школьный учебник, или выбранное лекарство).

Откликом часто называют значения измеряемого признака (т.е. величину результата).

Для сравнения влияния фактора (факторов) на результат необходим определённый статистический материал. Обычно его получают следующим образом: каждый из K способов обработки применяют несколько раз (не обязательно одно и то же число раз) к исследуемому объекту и регистрируют отклик (результат). Итогом подобных испытаний являются K выборок, вообще говоря, разных объемов. В зависимости от количества влияющих факторов (в нашем случае один фактор), говорят, что данные сведены в таблицу, с одним, двумя и т. д. входами.

Таблица 1

	Уровни фактора				
	1	**2**	**3**	...	k

Отклики (результаты измерений)	x_{11} x_{21} $\dots$ $x_{n_1 1}$	x_{12} x_{22} $\dots$ $x_{n_2 2}$	...	...	x_{1k} x_{2k} $\dots$ $x_{n_k k}$

Здесь $n_1, \dots, n_k$ – объёмы выборок при соответствующем уровне фактора, $N = n_1 + n_2 + \dots + n_k$ – общее число наблюдений.

Прежде, чем судить о количественном влиянии фактора на измеряемый признак, полезно вначале решить, а есть ли такое влияние вообще. Нельзя ли объяснить расхождение наблюденных значений в опыте для различных уровней факторов действием чистой случайности? Ведь присущая исследуемому явлению внутренняя изменчивость уже привела к тому, что результаты оказываются различными даже при неизменном значении фактора (т.е. в табл. 1 значения в одном столбце различны). Может той же причиной (изменчивость, случайность) можно объяснить и различие между столбцами?

Выдвинем статистическое предположение – все данные принадлежат одному и тому же распределению. Это предположение называется нулевой гипотезой H_0. Если она считается справедливой, то анализ закончен. В противном случае возникает задача оценки величины эффектов обработки и выяснения качества полученных оценок.

РАНГОВЫЙ ОДНОФАКТОРНЫЙ АНАЛИЗ

Если ничего не известно о распределении неизвестных, то использовать для проверки гипотезы H_0 количественные наблюдения x_{ij} становится затруднительно. В этих случаях проще основывать свои выводы на отношениях "больше – меньше" между наблюдениями, т.к. они не зависят от вида распределения. Теперь вся информация, включенная в табл. 1, содержится в тех рангах, что, получают числа x_{ij} при упорядочении всей совокупности (направление упорядочивания **min→max**, или **max→min** – не существенно).

Тогда таблица 1 преобразуется в таблицу 2:

Таблица 2

	Обработки (Уровни фактора)				
	1	**2**	**3**	...	k

Ранги результатов измерений	r_{11} r_{21} $\dots$ $r_{n_1 1}$	r_{12} r_{22} $\dots$ $r_{n_2 2}$	...	...	r_{1k} r_{2k} $\dots$ $r_{n_k k}$

Критерий Краскела – Уоллиса

Заменим все наблюдения x_{ij} их рангами r_{ij}, упорядочивая всю совокупность $\|x_{ij}\|$ в порядке возрастания. Затем для каждой обработки (фактора) j вычисляем:

$$R_j = \sum_{i=1}^{n_j} r_{ij}, \text{ и } \overline{R}_j = \frac{1}{n_j} \cdot \sum_{i=1}^{n_j} r_{ij}, \tag{17.1}$$

$\overline{R}_j$ – это средний ранг по столбцу. Если по столбцам нет систематических различий, средние ранги $\overline{R}_j$ $(j = 1, \dots, k)$ не должны значительно отличатся от среднего ранга, рассчитанного по всей совокупности $\|r_{ij}\|$. Ясно, что последний равен $(N+1)/2$. Поэтому величины $\left(\overline{R}_1 - \frac{N+1}{2}\right)^2, \dots, \left(\overline{R}_k - \frac{N+1}{2}\right)^2$ при H_0 в совокупности должны быть небольшими. В качестве оценки меры отступления от чистой случайности и для учета различия в числе наблюдений для разных обработок предложена ***статистика Краскела – Уоллеса***:

$$H = \frac{12}{N(N+1)} \cdot \sum_{j=1}^{k} \left(\overline{R}_j - \frac{N+1}{2}\right)^2 = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_i^2}{n_j} - 3(N+1). \tag{17.2}$$

Статистика H имеет распределение χ^2 с $(k-1)$ степенями свободы.

Замечание:*Если в измерениях много совпадающих значений, то используется модифицированная статистика* H':

$$H' = \frac{H}{1 - \frac{\sum_{j=1}^{g} T_j}{N^3 - N}}, \tag{17.3}$$

Здесь g –число групп, совпадающих наблюдений, $T_j = (t_j^3 - t_j)$, t_j – число совпадающих наблюдений в группе с номером j.

Критерий Кронкхиера
(альтернативы с упорядочиванием)

Нередко известно, что имеющиеся группы результатов упорядочены по возрастанию (убыванию) влияния фактора. Пусть для определённости, первый столбец соответствует **min** фактору (см. табл. 2), а последний – **max**. В таких случаях можно использовать ***статистику Джонкхиера***, более чувствительную (более мощную) против альтернатив об упорядоченном влиянии фактора. Разумеется, против других альтернатив свойства этого критерия могут оказаться хуже свойств ***критерия Краскела – Уоллиса***.

Статистика Джонкхиера

Для простоты сначала рассмотрим статистику этого критерия в случае, когда сравниваются только 2 способа обработки (два уровня фактора). Тогда в нашей таблице (см. выше) будет всего два столбца. Фактически здесь идет речь о проверке однородности двух выборок. Вспомним, что для решения этой задачи используется ***статистика Манна – Уитни***. А именно, пусть имеются две выборки: $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$.

Положим:

$$\varphi(x_i, y_j) = \begin{cases} 1, & x_i < y_j; \\ 1/2, & x_i = y_j; \\ 0, & x_i > y_j. \end{cases} \tag{17.4}$$

Статистика Манна – Уитни это:

$$U = \sum_{\substack{i=1,\ldots,m \\ j=1,\ldots,n}} \varphi(x_i, y_j). \tag{17.5}$$

Теперь, в случае, когда сравниваются k способов обработки (уровней), поступим следующим образом. Для каждой пары чисел u и v, где $1 \le u \le v \le k$, составляем по выборкам с номерами u и v статистику Манна – Уитни:

$$U_{u,v} = \sum_{\substack{i=1,\ldots,m_u \\ j=1,\ldots,n_v}} \varphi(x_{i_u}, y_{j_v}). \tag{17.6}$$

Определим статистику Джонкхиера I как

$$I = \sum_{1 \le u \le v \le k} U_{u,v}. \tag{17.7}$$

Свидетельством в пользу альтернативы упорядоченности эффектов (против гипотезы однородности) служат большие значения статистики I, полученные в эксперименте.

Статистика I – табулирована для малых выборок с небольшим значением k. Для больших выборок в отношении I действует нормальная аппроксимация:

$I \sim N(m_I, D_I)$, где m_I и D_I соответственно равны:

$$m_I = \frac{1}{4}\left(N^2 - \sum_{j=1}^{k} n_j^2\right); \tag{17.8}$$

$$D_I = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right]. \tag{17.9}$$

Свидетельством против гипотезы однородности служат большие (сравнительно с процентными точками нормального распределения) значения статистики $(I - m_I)/\sqrt{D_I}$.

Замечание *В результате проверки рангового критерия (критериев) можно однозначно (статистически) решить, если верна гипотеза H_0, то влияние фактора не значимо и на этом однофакторный анализ закончен. Если же верна альтернативная гипотеза H_1 –влияние фактора значительное, то хотелось бы продолжить однофакторный анализ и перейти от качественной оценки к количественной.*

ЛЕКЦИЯ №18

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Для описания данных в большинстве случаев оказывается приемлема ***аддитивная модель***. Она предполагает, что значение отклика x_{ij} можно представить в виде суммы вклада (воздействия) фактора и независимой от вкладов факторов случайной величины. Иначе говоря, каждое наблюдение x_{ij} является суммой вида:

$$x_{ij} = a_j + \varepsilon_{ij}; \quad j = 1, \ldots, k; \quad i = 1, \ldots, n. \tag{18.1}$$

Здесь $a_1, a_2, \ldots, a_k$ – неизвестные неслучайные величины, являющиеся результатом действия соответствующих обработок; ε_{ij} – независимые одинаково распределённые случайные величины, отражающие внутреннюю, присущую наблюдениям, изменчивость.

Если в рассматриваемой модели известно, что величины $\varepsilon_{ij} \sim N(0, \sigma^2)$, то это позволяет использовать в модели однофакторного анализа более сильные методы, как для проверки гипотез, так и для оценки параметров. Совокупность этих методов носит название ***однофакторного дисперсионного анализа***.

Это название связано с тем, что анализ модели (18.1) основан на сопоставлении двух оценок дисперсий σ^2. Одна из них действует вне зависимости от того, верна или нет гипотезы $H_0: a_1 = a_2 = \ldots = a_k$. Другая

оценка существенно использует это предположение, она дает близкий к σ^2 результат, только в том случае, если гипотеза верна. Сопоставляя эти две оценки, мы можем заключить, что H_0 следует отвергнуть, если дисперсии оказываются заметно (значимо) различны.

Т.к. каждая однородная группа (столбец) дает оценку σ^2, то для каждого столбца найдем выборочную сумму квадратов отклонений от выборочного среднего $\bar{x}_j$ (по фактору). Тогда получим:

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}, \quad j = 1,\ldots,k; \tag{18.2}$$

и далее вычисляем $\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2$. Показано, что такую сумму квадратов можно представить в виде произведения $\sigma^2\chi^2$, где случайная величина χ^2 имеет распределение χ^2 с $n_j - 1$ степенями свободы. В связи с тем, что данные в разных столбцах получены независимо, то объединенная сумма квадратов $\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2$ имеет распределение $\sigma^2\chi^2$ с $N-k$ степенями свободы. Отсюда получаем основную оценку:

$$S^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2. \tag{18.3}$$

Следует обратить внимание, что при выводе оценки (18.3) мы не упоминали о гипотезе H_0. Следовательно, $S^2 \approx \sigma^2$ независимо от того, верна гипотеза H_0 или нет.

Теперь получим другую оценку σ^2, для этого опять обратимся к столбцам (факторам). Полагаем, что

$$\bar{x}_j \sim N\left(a_j, \sigma^2/n_j\right). \tag{18.4}$$

Отметим, что $\bar{x}_j$ и $\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2$ статистически независимы. Найдем центр совокупности (18.4) с учетом весов средних значений n_j, т.е. найдем, при каких значениях z достигается минимум выражения:

$$\sum_{j=1}^{k}\left(\bar{x}_j - z\right)^2 \cdot n_j \to \min_z . \tag{18.5}$$

Минимум (18.5) достигается при $z = \bar{x}$:

$$\bar{x} = z = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\,. \qquad (18.6)$$

Если верна гипотеза H_0, то значение выражения (18.5) при $z = \bar{x}$ имеет распределение $\sigma^2\chi^2_{k-1}$, где $\chi^2_{k-1}-$ распределение χ^2 с $k-1$ степенями свободы. Отсюда следует вторая оценка для дисперсии:

$$S_1^2 = \frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2. \qquad (18.7)$$

Учитывая, что $\bar{x}_j$ независима от S^2, то это справедливо и для их комбинаций. Поэтому оценки (18.3) и (18.7) являются независимыми.

Если гипотеза H_0- не верна (нарушена), то оценка (18.7) имеет тенденцию к возрастанию, тем большему, чем больше отклонение от H_0.

Поскольку для оценки σ^2 мы получили две независимые оценки S_1^2 и S^2, имеющие при гипотезе H_0 распределение хи-квадрат, их частное должно иметь $F-$ распределение Фишера – Снедекора с $(k-1, N-k)$ степенями свободы:

$$F_{\text{набл.}} = \frac{S_1^2}{S^2} = \frac{\dfrac{1}{k-1}\sum_{j=1}^{k} n_j \cdot \left(\bar{x}_j - \bar{x}\right)^2}{\dfrac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2}. \qquad (18.8)$$

Замечания:

- При большом (неправдоподобно большом) значении $F_{\text{набл.}}$ гипотеза H_0 отвергается и принимается гипотеза H_1. Аналогично, если вероятность того, что $P\left(F \geq F_{\text{набл.}}\right)-$ мала, то гипотезу H_0 следует отвергнуть.
- $F-$ распределение (распределение Фишера–Снедекора) обозначается обычно F_{k_1,k_2}, где k_1 и k_2 числа степеней свободы. При $n \to \infty$ $F-$ распределение приближается к нормальному закону.

Оценка эффектов обработки в нормальной модели (Доверительные интервалы)

Если гипотеза H_0 оказалась несовместимой с наблюдениями, то есть основание для обсуждения параметров $a_1, a_2, \ldots, a_k$. Ранее было показано, что их оценками могут служить внутригрупповые средние $\bar{x}_j$, которые

имеют распределение $N\left(a_j,\ \sigma^2 / n_j\right)$ и статистически независимы от оценки дисперсии (18.3)$\Rightarrow S^2$. Поэтому отношение:

$$t = \frac{\bar{x}_j - a_j}{S} \cdot \sqrt{n_j} \tag{18.9}$$

подчиняется распределению Стьюдента с $N-k$ степенями свободы. Теперь с помощью выражения (18.9) можно вычислить доверительный интервал для фактора a_j с произвольной доверительной вероятностью $1-\alpha$:

$$P\left\{\left|\sqrt{n_j} \cdot \frac{\bar{x}_j - a_j}{S}\right| < t_{1-\alpha/2;N-k}\right\} = 1-\alpha. \tag{18.10}$$

Отсюда

$$\left|\bar{x}_j - a_j\right| < \frac{S}{\sqrt{n_j}} \cdot t_{1-\alpha/2;N-k}. \tag{18.11}$$

Таким образом, доверительный интервал для оценки a_j будет равен:

$$a_j = \bar{x}_j \pm \frac{S}{\sqrt{n_j}} \cdot t_{N-k,1-\alpha/2}. \tag{18.12}$$

Пример *ОДНОФАКТОРНЫЙ АНАЛИЗ*

Проверим гипотезу об отсутствии влияния денежного вознаграждения на число решенных сотрудниками фирмы задач.

Таблица 1

Уровни фактора					
1	**2**	**3**	**4**	**5**	**6**
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Решение

1. **Ранговые критерии**

- *Критерий Краскела – Уоллеса*

  В связи с наличием в таблице совпадений, применим средние ранги. После пересчета таблица примет вид:

Таблица 2

Уровни фактора						
	1	**2**	**3**	**4**	**5**	**6**
	5.5	2	9	9	27.5	23.5
	7	5.5	20	14	17	21.5
	3.5	17	13	17	26	30

	11.5	11.5	3.5	17	21.5	29
	1	9	17	23.5	25	27.5
$R_j = \sum r_{ij}$	**28.5**	**45**	**62.5**	**80.5**	**117**	**131.5**
$\overline{R}_j$	**5.7**	**9**	**12.5**	**16.1**	**23.4**	**26.3**

По формуле (17.2) вычислим статистику H, учитывая, что $N = 30, n_j = 5\,(j = 1, \ldots 6)$. Поставив все значения из табл. 4, окончательно получим $H = \dfrac{17682}{155} - 93 = 21.077$.

Величина H имеет асимптотическое распределение χ^2 с $5 \to (K-1)$ степенями свободы. По таблице найдем уровень значимости, соответствующий вычисленному значению, ≈ 0,001. С учетом повторов, можно пересчитать H'. Тогда, ввиду малости вероятности значения H', гипотезу H_0 можно отвергнуть.

- *Критерий Джонкхиера*

Предполагая монотонную зависимость количества решенных задач от материального стимула, считаем применение критерия оправданным. Для этого найдем статистику Манна – Уитни для всех пар: таких что $u < v, (u, v = \overline{1,6})$;

$U_{1,2} = 17;$ $U_{2,3} = 17;$ $U_{3,4} = 16.5;$ $U_{4,5} = 22;$ $U_{5,6} = 18;$

$U_{1,3} = 18.5;$ $U_{2,4} = 20.5;$ $U_{3,5} = 23.5;$ $U_{4,6} = 23.5;$

$U_{1,4} = 24;$ $U_{2,5} = 24.5;$ $U_{3,6} = 25;$

$U_{1,5} = 25;$ $U_{2,6} = 25;$

$U_{1,6} = 25;$

$$I = \sum_{\substack{u=\overline{1,6} \\ v=\overline{1,6} \\ u<v}} U_{u,v} = 325.$$

По формулам (17.8 и 17.9) вычислим:

$$m_I = 187{,}5;\ D_I = 782{,}083 \Rightarrow I^* = \frac{(I - m_I)}{\sqrt{D_I}}.$$

Следовательно, $I^* \approx \dfrac{325 - 187{,}5}{27{,}97} \approx 4.916.$ По таблице нормального распределения получим $\alpha \approx 1 \cdot 10^{-6}$, т.е. гипотезу H_0 следует отвергнуть.

2. Оценка дисперсионного анализа

Используя формулы (18.3, 18.7 и 18.8), вычислим:

$$S_1^2 \approx 118{,}16; \quad S^2 \approx 9{,}35; \quad \Rightarrow \quad F_{набл} \approx 12{,}637; \qquad p \approx 4 \cdot 10^{-6}.$$

Т.о., значение F очень велико и могло бы быть "не значимо" с вероятностью ~ 0, значит гипотезу H_0 отвергаем.

Оценим теперь параметры модели (18.1) по формуле (18.2) и заодно вычислим 95% доверительный интервал для каждого параметра. Все данные сведем в таблицу:

Таблица 3

Уровни фактора	$a_j = \bar{x}_j$	S_j	$x_{\min,0.95}$	$x_{\max,0.95}$
1	10.0	1.00	7.177	12.823
2	11.8	1.3565	8.977	14.623
3	13.6	1.4353	10.777	16.423
4	15.6	1.1225	12.777	18.423
5	20.0	1.4142	17.177	22.823
6	22.6	1.7493	19.777	25.423

$$S_j = \sqrt{\frac{1}{n_j - 1} \cdot \sum_{j=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2}.$$

ЛЕКЦИЯ №19

ОДНОФАКТОРНЫЙ АНАЛИЗ В СИСТЕМЕ Statistica 6.0

Рассмотрим типичную задачу ***однофакторного анализа***, реально возникшую на производстве.

Пример. На заводе разработаны две новые технологии T_1 и T_2. Чтобы оценить, как изменится дневная производительность при переходе на новые технологии, завод в течение 10 дней работал по каждой, включая существующую T_0. Дневная производительность в условных единицах (***отклики***) и способ обработки (***уровни фактора***, соответствующие технологиям T_0, T_1, и T_2) приведены в табл. 1.

Таблица 1.

Number	Technology	Power	Number	Technology	Power	Number	Technology	Power
1	T0	46	11	T1	74	21	T2	52
2	T0	48	12	T1	82	22	T2	63
3	T0	73	13	T1	64	23	T2	72
4	T0	52	14	T1	72	24	T2	64
5	T0	72	15	T1	84	25	T2	48
6	T0	44	16	T1	68	26	T2	70
7	T0	66	17	T1	76	27	T2	78
8	T0	46	18	T1	88	28	T2	68

9	T0	60	19	T1	70	29	T2	70
10	T0	48	20	T1	60	30	T2	54

Выдвигаем нулевую гипотезу H_0 – отклики (производительность) принадлежат одному и тому же распределению. То есть, влияние фактора (технологии) не существенно.

Сначала воспользуемся более мощными, свободными от распределений ***ранговыми критериями***. И только в том случае, если при проверке ранговых критериев нулевая гипотеза будет отвергнута в пользу альтернативной гипотезы H_1 – влияние фактора существенное, воспользуемся методами ***дисперсионного анализа*** для количественной оценки влияния фактора.

Критерий Краскела – Уоллиса

Для проверки данного критерия последовательно выбираем программу (модуль):

***Statistics→Nonparametrics→Comparing Multiple Indep. Samples* (*Groups*)**

рис.1.

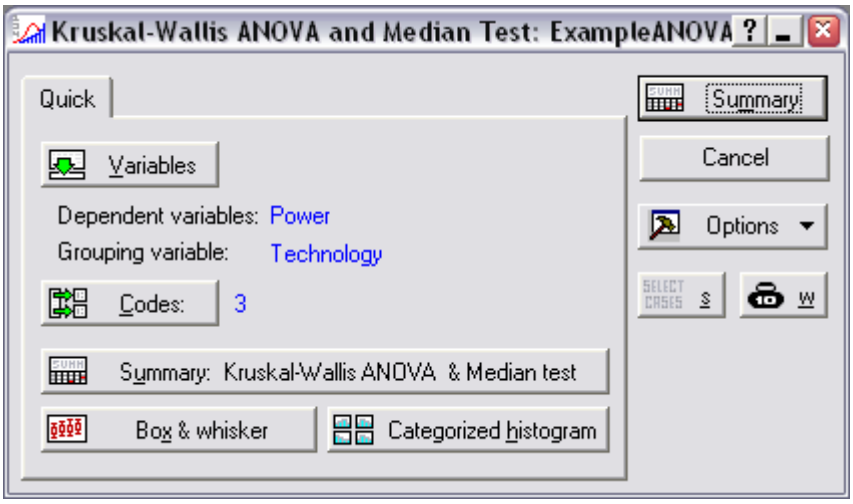


Рис.1. Тест Краскела – Уоллиса

Для выбора переменных нажимаем клавишу **Variables** и выбираем *Dependent variable* – **отклики** (*Power*) и Indep.(grouping) variable – **уровни фактора** (*Technology*) рис.2.

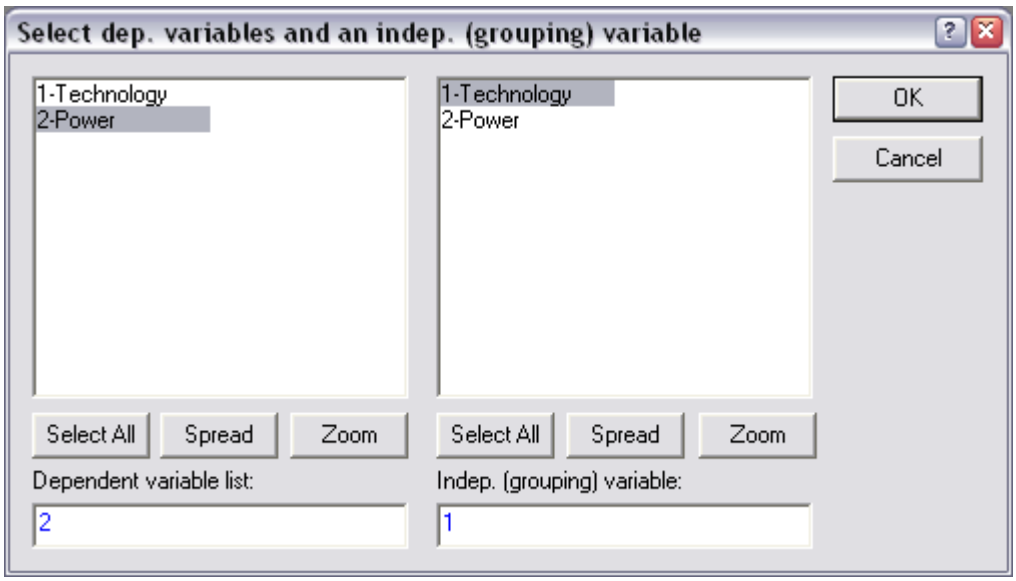


Рис. 2. Окно выбора переменных

В закладке Для выбора группы факторов, в нашем примере переменная *Technology* разбита на группы T0, T1 и T2, выбираем пункт меню **Code** (см.рис.1). Для выбора всех групп нажимаем кнопку **All.**

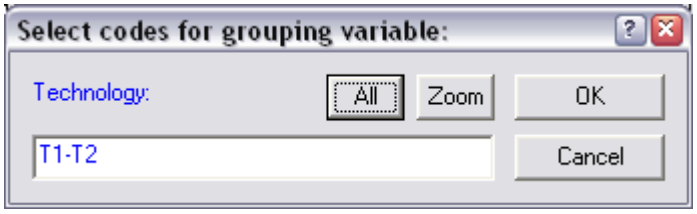


Рис. 3. Выбор групп факторов для анализа

Этот параметр необходим для корректной работы пакета Statistica, т.к. исходные данные в файле могут идти в случайном порядке, программа объединит их в соответствующие группы рис. 3.

По умолчанию при запуске теста факторы объединяются на максимально возможное количество групп. Выполнив все необходимые действия, нажимаем кнопку **Summary**, после чего получим результат тестирования. Полный результат теста Краскела - Уоллиса представлен на рис.4.

Kruskal-Wallis ANOVA by Ranks; Power (ExampleANOVA)
Independent (grouping) variable: Technology
Kruskal-Wallis test: H (2, N= 30) =10,07747 p =,0065

Depend.: Power	Code	Valid N	Sum of Ranks
T1	101	10	219,5000
T0	102	10	95,0000
T2	103	10	150,5000

Рис. 4. Результаты теста Краскела - Уоллиса

<u>В приведенных результатах приняты следующие обозначения:</u>

Codes – уникальный код группы (число);
Valid N – число значений в группе;
Sum of Ranks – сумма рангов;
H – статистика Краскела - Уоллиса;
p – вероятность принятия гипотезы H_0.

Анализируя суммы рангов, представленные в результирующем отчете можно говорить о влиянии уровня фактора на производительность. Из результатов видно что лучшая производительность обеспечивается технологией T1, а худшая – существующей технологией T0.
Вспомним, что в статистике Краскела – Уоллиса вычисляется сумма квадратов разностей средних рангов в группе и среднего ранга по всей выборке. Тогда, если верна гипотеза H_0 и влияние фактора незначимо, то значение статистики мало. В нашем примере $H = 10.077$ и нулевую гипотезу можно принять с вероятностью $p = 0.0065$. Поскольку заданный нами уровень значимости много больше $\alpha = 0.05$, то нулевую гипотезу следует отвергнуть в пользу альтернативной гипотезы H_1 – влияние фактора существенное.

***Медианный тест* (*критерий*)**

Как и в критерии Краскела – Уоллиса заменим все наблюдения x_{ij} их рангами r_{ij}, упорядочивая всю совокупность $\|x_{ij}\|$ в порядке возрастания. Учтем, что медиана объединенной выборки равна $\frac{1}{2}(N+1)$, здесь N –объем объединенной выборки. Введем обозначение:

$$A_j = \sum_{i \in s_j} \frac{1}{2}[\mathrm{sign}(R_i - \frac{1}{2}(N+1)) + 1], \;\; j = 1, \dots, k. \qquad (19.1)$$

То есть A_j – это число наблюдений из j-й выборки (уровня фактора), больших, чем медиана объединенной выборки. Мы прибавляем ½ к этому числу в том и только том случае, если N нечетно и эта медиана принадлежит j-й выборке. Тогда можно ввести статистику медианного критерия

$$Q = 4 \cdot \sum_{j=1}^{k} \frac{1}{n_j}\left[A_j - \frac{1}{2}n_j\right]^2 = 4 \cdot \sum_{j=1}^{k} \frac{1}{n_j} {A_j}^2 - N. \qquad (19.2)$$

Показано, что статистика Q при гипотезе H_0 асимптотически подчиняется χ^2 распределению с $k-1$ степенями свободы.
Результат медианного теста в системе **Statistica** представлен на рис. 5.

Dependent: Power	Median Test, Overall Median = 67,0000; Power (ExampleANOVA) Independent (grouping) variable: Technology Chi-Square = 7,200000, df = 2, p = ,0273			
	T1	T0	T2	Total
<= Median: observed	2,00000	8,00000	5,00000	15,00000
expected	5,00000	5,00000	5,00000	
obs.-exp.	-3,00000	3,00000	0,00000	
> Median: observed	8,00000	2,00000	5,00000	15,00000
expected	5,00000	5,00000	5,00000	
obs.-exp.	3,00000	-3,00000	0,00000	
Total: observed	10,00000	10,00000	10,00000	30,00000

Рис. 5. Результаты медианного теста

В верхней части таблицы приведены количества рангов в группах, которые были меньше или равны медиане. В нижней части таблицы – аналогичные значения, превышающие значение медианы.

Проанализируем полученные результаты на качественном уровне. По значению разности предсказанных и полученных значений можно сделать следующие выводы:

- верхняя половина таблицы – максимальное значение указывает на худшую технологию;
- нижняя половина таблицы – максимальное значение указывает на лучшую технологию.

Количественная оценка статистики $Q = 7.2$ свидетельствует о том, что нулевую гипотезу можно принять с вероятностью $p = 0.0273$, что меньше уровня значимости, следовательно, принимается гипотеза H_1.

В системе **Statistica** при проведении рангового однофакторного анализа предлагаются дополнительные графические возможности. В исходном окне модуля (см. рис.1) выбор пункта **Categorized Histogram** позволяет посмотреть и оценить виды распределения выборок (групп). В данном случае нас интересует распределение исходных данных о производительности по группам (технологиям).

Выбираем переменную для гистограмм рис. 6, построенные гистограммы приведены на рис. 7.

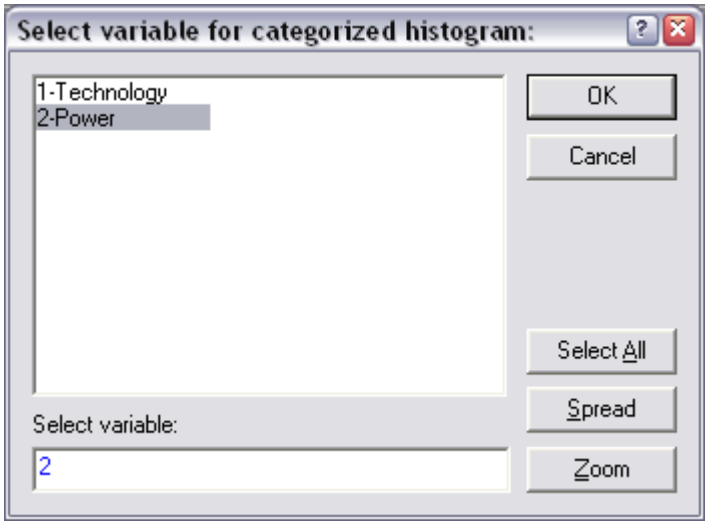


Рис. 6. Выбор переменной для гистограммы

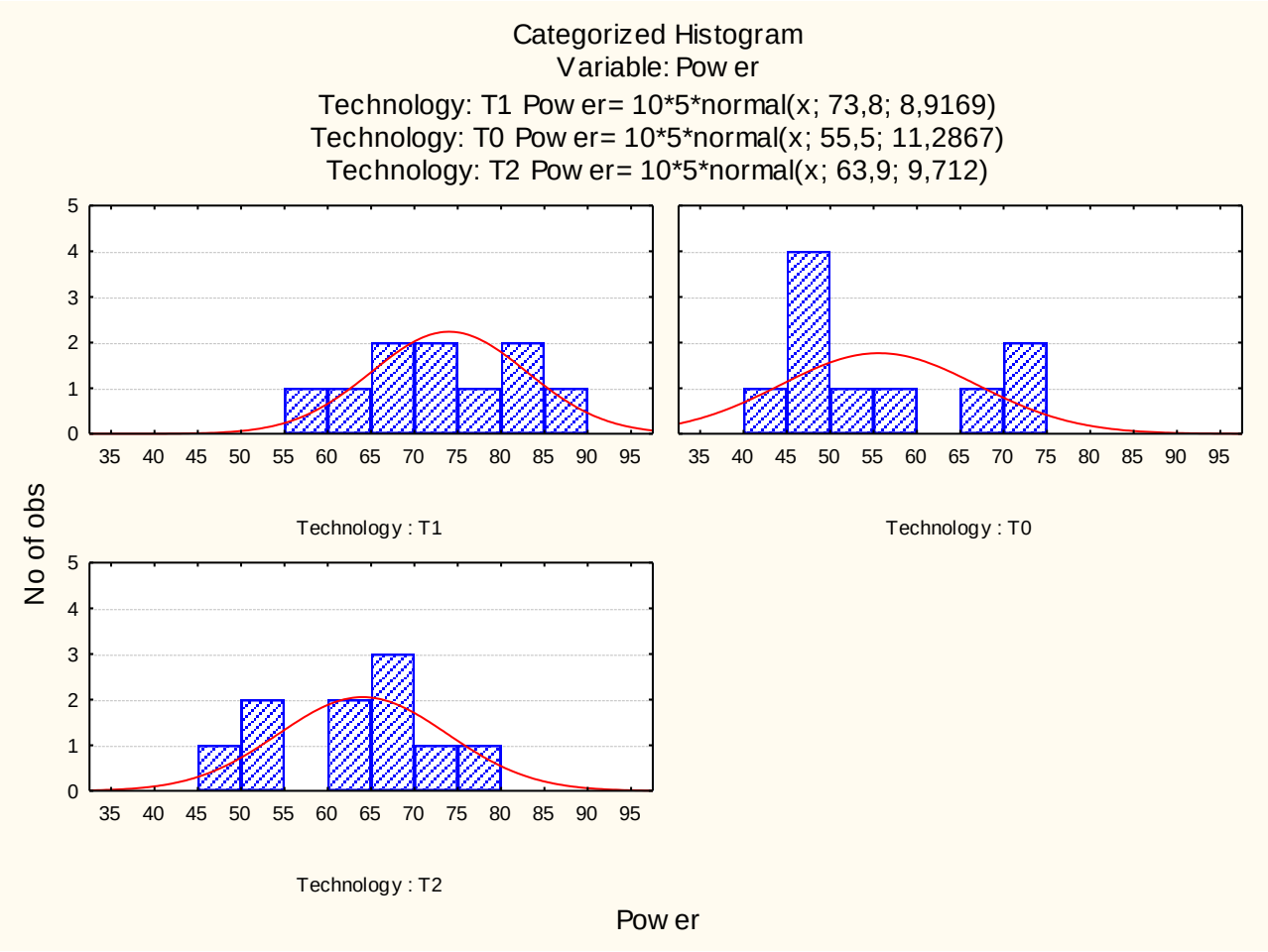


Рис. 7. Гистограммы распределения производительности

На построенных гистограммах сплошной линией проведены гауссовы распределения с соответствующими параметрами. Визуальный анализ подтверждает, что лучшая технология T1, т.к. при этой технологии минимальная и максимальная производительности больше, чем при технологиях T0 и T2, Эта же технология обеспечивает 70 % значений производительности в интервале
[65, 85], что значительно лучше, чем в других группах.

ЛЕКЦИЯ №20

ОДНОФАКТОРНЫЙ АНАЛИЗ В СИСТЕМЕ Statistica 6.0 (продолжение)

Критерий Кронкхиера

Если известно, что имеющиеся группы результатов упорядочены по возрастанию (убыванию) влияния фактора, то в таких случаях можно использовать ***статистику Джонкхиера***, более чувствительную (более мощную) против альтернатив об упорядоченном влиянии фактора.
Так как в системе **Statistica** ***статистика Джонкхиера*** не реализована, воспользуемся критерием проверки двух выборок, в нашей задаче (с тремя

уровнями фактора) можно выделить три пары выборок. Для сравнения двух способов обработки (два уровня фактора) воспользуемся ***статистикой Манна – Уитни***, реализованной в данной системе.

В критерии Манна – Уитни сформулируем нулевую гипотезу H_0 – исходные две выборки – однородны, соответственно гипотеза H_1 утверждает, что выборки не однородны, т. е. влияние фактора значимо.

Данный критерий проверяется в модуле:

***Statistics→Nonparametrics→Comparing Two Independent Samples* (*Groups*)** (рис. 1).

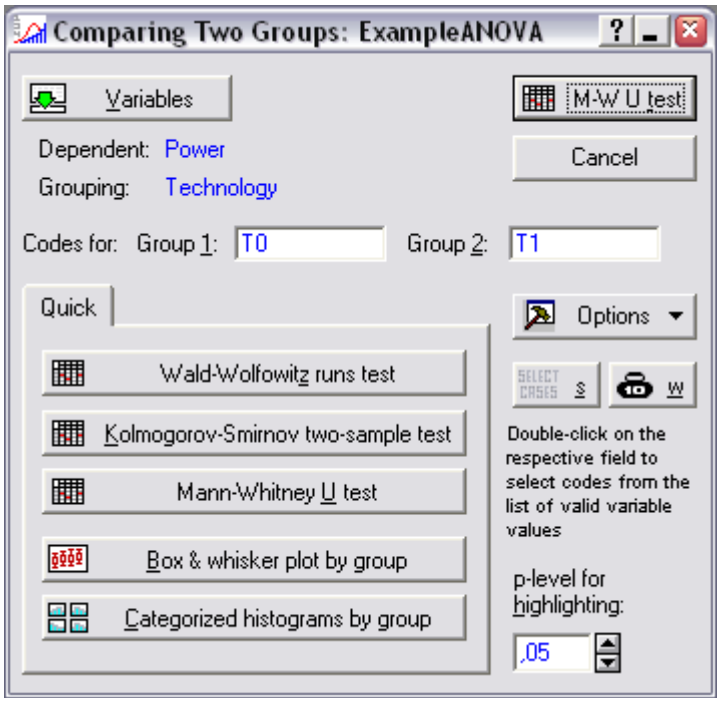


Рис. 1. Тест Манна-Уитни

С помощью клавиши **Variables** выбираем *Dependent variable* – **отклики** (*Power*) и Indep.(grouping) variable – **уровни фактор**. Далее выбираем группы факторов (технологии), которые будут анализироваться (*Group* 1 и *Group* 2). После данных действий, нажав на кнопку **Mann-Whitney U test** либо **M-W U Test**, можно получить результат тестирования. Результаты теста Манна - Уитни для всех возможных пар выборок представлены на рис. 2.

Mann-Whitney U Test (ExampleANOVA)
By variable Technology
Marked tests are significant at p <,05000

variable	Rank Sum T0	Rank Sum T1	U	Z	p-level	Z adjusted	p-level	Valid N T0	Valid N T1	2*1sided exact p
Power	67,00000	143,0000	12,00000	-2,87253	0,004072	-2,87686	0,004017	10	10	0,002879

Mann-Whitney U Test (ExampleMann)
By variable Technology
Marked tests are significant at p <,05000

variable	Rank Sum T0	Rank Sum T2	U	Z	p-level	Z adjusted	p-level	Valid N T0	Valid N T2	2*1sided exact p
Power	83,00000	127,0000	28,00000	-1,66304	0,096305	-1,66807	0,095303	10	10	0,105122

variable	Mann-Whitney U Test (ExampleMann) By variable Technology Marked tests are significant at p <,05000									
	Rank Sum T1	Rank Sum T2	U	Z	p-level	Z adjusted	p-level	Valid N T1	Valid N T2	2*1sided exact p
Power	131,5000	78,50000	23,50000	2,003212	0,045155	2,008504	0,044591	10	10	0,043257

Рис. 2. Результаты теста Манна - Уитни

<u>В приведенных таблицах приняты следующие обозначения:</u>

Rank Sum T_i – сумма рангов выборки T_i;
U –статистика Манна - Уитни для малых выборок;

p - level – вероятность принятия гипотезы H_0;

p - level – скорректированная вероятность принятия гипотезы H_0;

Rank Sum T_j – сумма рангов выборки T_j;
Z – нормальная аппроксимация статистики Манна - Уитни для больших выборок;
Z adjusted – скорректированная нормальная аппроксимация статистики Манна - Уитни;

Valid N – объем выборки;

2*1 sided exact p – здесь вероятность **p** равна 1 минус кумулятивная односторонняя вероятность соответствующей статистики Манна – Уитни.

<u>Анализ результатов:</u>

1. Для двух технологий (уровней факторов) T0-T1 статистика U достаточно велика и нулевую гипотезу можно принять с вероятностью
   p = 0,004072. При 5% уровне значимости гипотезу H_0 следует признать ложной и принять гипотезу H_1 – влияния фактора значительное.
2. Сравнивая технологии T0 и T2, мы убедились, что изменение откликов незначимо и две выборки можно признать однородными.
3. Что касается двух новых технологий T1 и T2, то нулевую гипотезу можно принять с вероятностью p = 0,045155, что меньше уровня значимости. На основании этого нулевую гипотезу отвергаем в пользу альтернативной – влияния фактора значительное.

Как и в случае теста Краскела - Уоллиса в закладке **Categorized Histogram By Group** можно посмотреть и оценить виды распределения отдельных выборок.

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Так как предварительный ранговый однофакторный анализ подтвердил гипотезу о значимом влиянии фактора, попробуем оценить это влияние количественно в рамках дисперсионного анализа.

Проверяем нулевую гипотезу H_0 – влияние фактора на распределение данных не существенно. Дисперсионный анализ проведем в модуле: ***Statistics→Basic Statistics→Breakdown&One-way ANOVA*** (рис. 3).

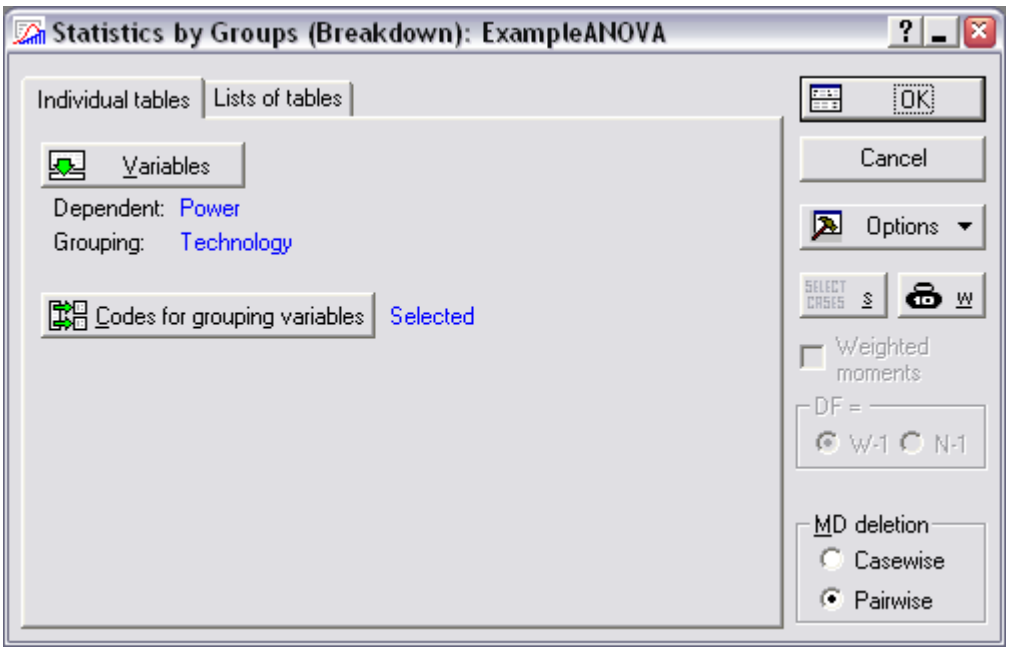


Рис. 3. Модуль дисперсионного анализа данных

В закладке **Variables** выбираются *Dependent variable* и *Grouping variables*. В закладке **Codes for grouping variables** выбираются группы факторов, в нашем примере переменная *Technology* разбивается на группы T0, T1 и T2.

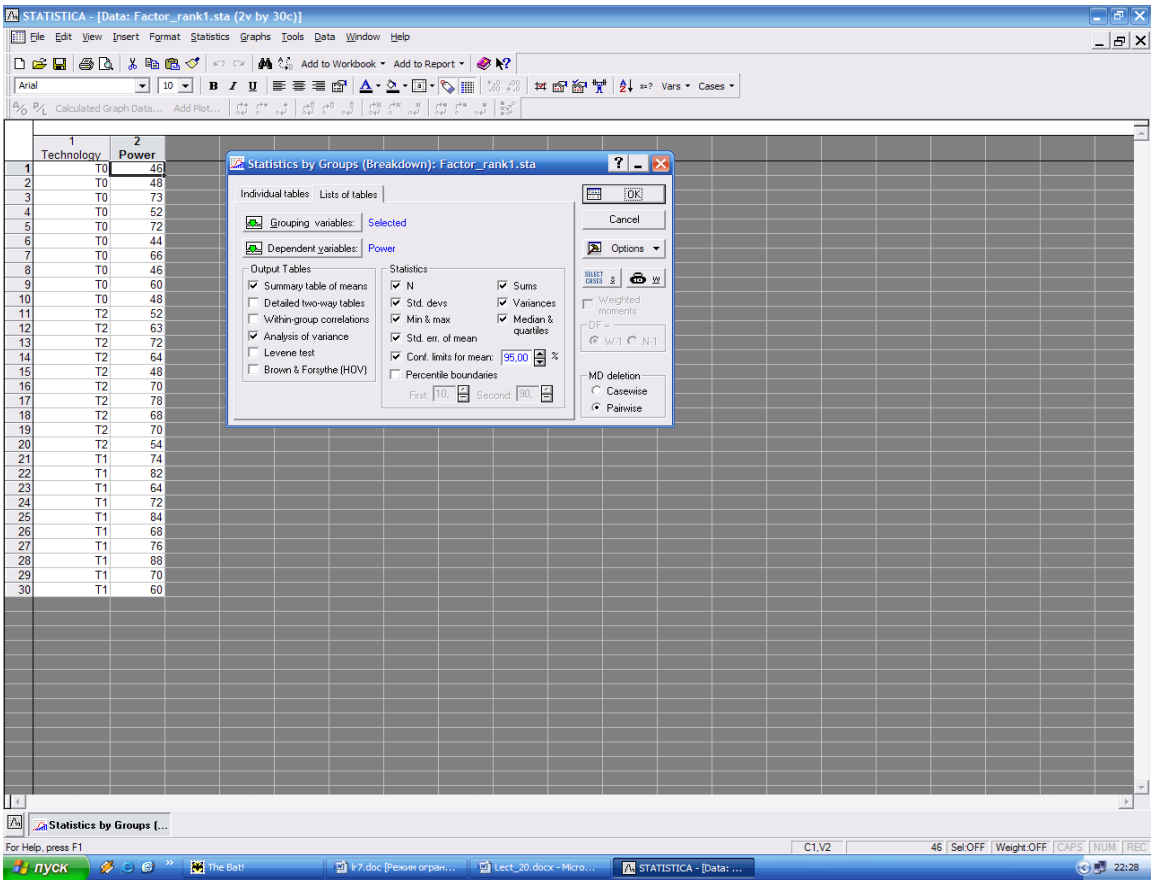


Рис. 4. Настройка таблицы результатов дисперсионного анализа

Выбор всех групп осуществляем нажатием на кнопку **All**. В окне настройки результирующих таблиц выбираем группирующие переменные *Grouping Variables*, ставим галочки в позиции **Summary table of means** и в позиции **Analysis of variance**, аналогично галочкой отмечаем те статистические

данные, которые хотим более подробно проанализировать (см. рис. 4). После нажатия клавиши *OK* получим таблицу (рис. 5).

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Analysis of Variance (Factor_rank1.sta) Marked effects are significant at p < ,05000								
Power	1678,20	2	839,100	2711,00	27	100,407	8,35695	0,00149

Рис.5. Результаты дисперсионного анализа

Поясним принятые в таблице обозначения:

SS (Sum of Squares) Effect – сумма квадратов факторов (вторая оценка дисперсии (18.7)) умноженная на k - 1;

MS (Mean Square) Effect – средний квадрат фактора;

df Error – число степеней свободы равная $N - k$;

F – значение статистики Фишера;

df Effect – число степеней свободы фактора;

SS Error – сумма квадратов (оценка дисперсии (18.3)) умноженная на $N - k$;

MS Error – оценка дисперсии (18.3);

p –вероятность принятия гипотезы H_0.

Статистика Фишера $F = 8.356953$ незначимо отличается от единицы с вероятностью $p = 0.001496$, что значительно меньше уровня значимости. Следовательно, нулевую гипотезу следует отвергнуть в пользу альтернативной гипотезы H_1 – влияние фактора существенно.

Так как наряду с дисперсионным анализом в системе **Statistica** мы позаботились о получении оценок эффектов обработки **Summary table of means** рис. 6, проанализируем полученные результаты.

Technology	Power Means	Confidence -95,000%	Confidence +95,000%	Power N	Power Sum	Power Std.Dev.
Breakdown Table of Descriptive Statistics (Factor_rank1.sta) N=30 (No missing data in dep. var. list)						
T0	55,5000	47,4260	63,5740	10	555,00	11,2866
T2	63,9000	56,9524	70,8475	10	639,00	9,7119
T1	73,8000	67,4212	80,1787	10	738,00	8,9169
All Grps	64,4000	59,8061	68,9938	30	1932,00	12,3025

Technology	Power Variance	Power Minimum	Power Maximum	Power Q25	Power Median	Power Q75
Breakdown Table of Descriptive Statistics (Factor_rank1.sta) N=30 (No missing data in dep. var. list)						
T0	127,388	44,0000	73,0000	46,0000	50,0000	66,0000
T2	94,322	48,0000	78,0000	54,0000	66,0000	70,0000
T1	79,511	60,0000	88,0000	68,0000	73,0000	82,0000
All Grps	151,351	44,0000	88,0000	52,0000	67,0000	72,0000

Рис. 6. Влияние технологии (уровня фактора) на производительность (отклики)

Полученные результаты (средние) свидетельствуют о существенном различии точечных характеристик для различных групп. Отметим, что наряду со средними значениями мы можем проанализировать такие групповые параметры как дисперсия, медиана, нижний и верхний квартили, минимальное и максимальное значения.

Наряду с точечными оценками в результирующей таблице приведены 95 % доверительные интервалы для каждого параметра (группового среднего). Фактически исследования в рамках однофакторного анализа можно считать законченными, но анализируя последние данные (см. рис. 6) возникает вопрос. Какие технологии можно считать значимо различными? Для ответа на этот вопрос вернемся в окно *Statistics by Groups* (*Breakdown*)→*OK*→*Sheffe test*.

Результат сравнения средних по методу Шеффе для различных пар уровней приведен на рис. 7.

		Scheffe Test; Variable Power(Fa Marked differences are significa		
Technology		{1} M=55,500	{2} M=63,900	{3} M=73,800
T0	{1}		0,191749	0,001514
T2	{2}	0,191749		0,106123
T1	{3}	0,001514	0,106123	

Рис. 7. Тест Шеффа

В результата проверки гипотезы о незначимом различии средних, только для пары T1 – T0 вероятность нулевой гипотезы равная 0.001514 много меньше уровня значимости. Поэтому нулевая гипотеза отклоняется, влияние фактора значительное.

ЛЕКЦИЯ №21

ДВУХФАКТОРНЫЙ АНАЛИЗ

Бывает, что в рамках однофакторной модели влияние интересующего нас фактора не проявляется, хотя логические соображения указывают, что такое влияние должно быть. Иногда это влияние проявляется, но точность выводов о количественной оценке этого влияния недостаточна. Причиной этого может быть большой внутригрупповой разброс, на фоне которого действие фактора остается незаметным или почти незаметным. Очень часто такой разброс вызывается не только случайными причинами, но также действием еще одного фактора. Если мы в состоянии указать такой фактор, можно попытаться включить его в модель, чтобы уменьшить статистическую неоднородность наблюдений и благодаря этому выявить действие на отклик закономерных причин. Конечно, не всегда удается поправить дело введением одного "мешающего" фактора и переходом к двухфакторным схемам. Иногда приходится рассматривать и трех–, и многофакторные модели.

К задачам двухфакторного или многофакторного анализа часто приводят исследования по оптимизации технологических процессов. Иногда факторы разделяют на важные и мешающие, но это совершенно не обязательно. В ряде задач факторы для экспериментатора содержательно равноправны.

Рассмотрим ***таблицу двухфакторного анализа*** (см. табл. 1). Назовем главный фактор фактором A, а мешающий фактор – фактором B. Пусть фактор A принимает k, а фактор B – n – различных значений. Фактор B разбивает все объекты наблюдения на n блоков. Каждый блок образует наблюдения, проведенные при одном уровне фактора B. В блоке отклики могут значимо различаться только за счет применения к ним различных *обработок* (уровнях фактора A). Уровни фактора A отображаются в таблице по столбцам, а уровни фактора B – по строкам.

Отличие таблицы двухфакторного анализа от аналогичной таблицы однофакторного анализа заключается в том, что наблюдения в одном столбце не являются однородными, т.е. могут не образовывать выборки (если влияние мешающего фактора значимо).

Таблица 1

	Уровни фактора A				
Блоки	**1**	**2**	**3**	...	k
1	x_{11}	x_{12}			x_{1k}
2	x_{21}	x_{22}	...	...	x_{2k}
...	...	...			...
n	$x_{n_1 1}$	$x_{n_2 2}$			$x_{n_k k}$

Замечание: *На практике часто рассматриваются таблицы с повторными измерениями, т.е. в каждой клетке табл.* 1 *могут содержаться несколько наблюдений, что не сможет существенно усложнить анализ.*

Аддитивная модель данных двухфакторного эксперимента при независимом действии факторов

В аддитивной двухфакторной модели каждое наблюдение x_{ij} представляется в виде:

$$x_{ij} = b_i + t_j + \varepsilon_{ij}, \quad (i = 1, \ldots, n;\ j = 1, \ldots, k). \qquad (21.1)$$

В этой модели (21.1) числа $b_1, b_2, \ldots, b_n$ – являются результатом воздействия на отклик мешающего фактора B, действие которого разбивает все данные на блоки. Поэтому величины $b_1, b_2, \ldots, b_n$ называют *эффектами блоков*. Числа $t_1, t_2, \ldots, t_k$ отражают действие на отклик интересующего нас фактора A и именуются *эффектами обработки*.

Относительно ε_{ij} – предполагается, что они одинаково распределены и независимы в совокупности. Существуют различные методы двухфакторного анализа, которые требуют от их распределения (ε_{ij}) либо только непрерывности, либо принадлежности к нормальному

распределению $\sim N(0,\sigma^2)$.

Очевидно, что даже при справедливости выражения (21.1), величины вкладов факторов b_i и t_j не могут быть восстановлены однозначно. Действительно, увеличение всех b_i на одну и ту же const и одновременное уменьшение всех t_j на эту же const, оставляет выражение (21.1) неизменным. Для однозначной определенности вкладов факторов удобно перейти к представлению наблюдений в ином виде:

$$x_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, \quad (i = 1,\ldots,n;\ j = 1,\ldots,k). \tag{21.2}$$

Здесь полагаем, что $\sum_{i=1}^{n}\beta_i = 0;\ \sum_{j=1}^{k}\tau_j = 0$. При этом параметр μ интерпретируется как среднее значение $(m = \bar{x})$, присущее всем величинам x_{ij}, а β_i и τ_j – как отклонения от μ в результате действия факторов B и A.

Гипотеза Как и в однофакторном анализе, вначале целесообразно проверить гипотезу о значимости эффектов обработки (уровней факторов A). Сформулируем нулевую гипотезу $H_0:\ \tau_1 = \tau_2 = \ldots = \tau_k = 0$. Т.е. предполагаем, что влияние фактора A отсутствует.

На первом этапе рассмотрим ранговые критерии, которые не требуют знания вида распределения случайной величины.

РАНГОВЫЕ КРИТЕРИИ

Критерий Фридмана

В этом критерии полагаем, что значения ε_{ij} – непрерывны, а сами величины ε_{ij} – независимы в совокупности. В двухфакторном анализе ранжирование проводится не по всей совокупности x_{ij}, а поблочно, т.е. (см. табл. 1) рассматривается каждая отдельная строка. При этом устраняется влияние "мешающего" фактора B, значение которого для каждой строки постоянно.

Итак, подсчитываем ранги: $x_{ij} \Rightarrow r_{ij}$. При этом известно, что r_{ij} изменяется от 1 до k, а каждая строка представляется соответствующей перестановкой чисел от 1 до k. Для простоты, считаем, что одинаковые значения отсутствуют, в противном случае берем средние ранги. При истинности нулевой гипотезы $H_0:\ \tau_1 = \tau_2 = \ldots = \tau_k = 0$; каждая строка рангов – случайная перестановка k чисел. Причем все $k!$ перестановок – равновероятны. Введем средние ранги (по столбцу):

$$\bar{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}\ . \tag{21.3}$$

При гипотезе H_0 все перестановки рангов равновозможны, поэтому значения $\bar{r}_j$ не должны сильно отличаться от *общего среднего ранга таблицы*:

$$\bar{r} = \frac{k+1}{2}. \qquad (21.4)$$

Статистика Фридмана имеет вид:

$$S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} (\bar{r}_j - \bar{r})^2 = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} (\sum_{i=1}^{n} r_{ij})^2 \right] - 3n(k+1). \qquad (21.5)$$

Если гипотеза $H_0 -$ верна, то величины $\left(\bar{r}_j - \bar{r}\right)^2$ сравнительно малы для всех j, тогда и $S -$ не велико. Если значение $S -$ велико, то гипотеза H_0 отвергается.

Замечание *При больших n для выбора критических значений пользуются аппроксимацией, основанной на том, что при справедливости гипотезы H_0 и $n \to \infty$ статистика Фридмана S асимптотически распределена как χ^2 с $(k-1)$ степенями свободы. Т.о. гипотеза H_0 принимается, если $S < \chi^2_{(k-1),(1-\alpha)}$, в противном случае принимается альтернативная гипотеза H_1.*

Критерий Пейджа
(альтернативы с упорядочением)

Данный критерий применяется в тех случаях, когда хотим определить преимущества различной обработки. Здесь, как и в критерии Фридмана, проверяем нулевую гипотезу $H_0: \ \tau_1 = \tau_2 = \ldots = \tau_k = 0$, против альтернативы $H_1: \ \tau_1 \le \tau_2 \le \ldots \le \tau_k$. Причем в альтернативной гипотезе H_1 хотя бы одно из неравенств – строгое. Статистика Пейджа имеет вид:

$$R_j = \sum_{i=1}^{n} r_{ij} \quad \Rightarrow \quad L = \sum_{j=1}^{k} j \cdot R_j = R_1 + 2R_2 + \ldots + k \cdot R_k. \qquad (21.6)$$

Нулевая гипотеза отклоняется с достоверностью α, если $L > L_\alpha(k,n)$, где $L_\alpha(k,n) -$ критические значения статистики Пейджа. При $n > 10$ справедлива аппроксимация

$$L^* = \frac{L - M[L]}{\sqrt{D[L]}}. \qquad (21.7)$$

Здесь

$$M[L] = \frac{nk(k+1)^2}{4}; \qquad D[L] = \frac{n(k^3-k)^2}{144(k-1)}. \qquad (21.8)$$

При $L^* > u_{1-\alpha}$ нулевая гипотеза отклоняется ($u_{1-\alpha}$ – квантиль стандартного нормального распределения).

ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Если предположить, что ε_{ij} имеет распределение $N(0,\sigma^2)$, то можно предложить более мощный критерий для гипотезы $H_0: \quad \tau_1 = \ldots = \tau_k = 0$ и построить более эффективные оценки параметров μ, τ_j и β_i. Отметим, что при этом используются методы аналогичные тем, что использовались при однофакторном дисперсионном анализе.

Как и в однофакторном анализе, получим две различные оценки дисперсии: оценка S^2 – не зависит от того, верна ли гипотеза H_0 или нет; оценка S_1^2 – справедлива лишь при истинности гипотезы H_0.

Оптимальная в классе несмещенных оценок оценка S^2 получается методом наименьших квадратов. Для этого сначала оценим неизвестные параметры μ, β_i и τ_j в модели (21.2). Найдем значения μ, β_i и τ_j такие, при которых достигается min функции:

$$\sum_{i,j}\left(x_{ij} - \mu - \beta_i - \tau_j\right)^2, \tag{21.9}$$

при условии, что $\sum_{i=1}^{n}\beta_i = \sum_{j=1}^{k}\tau_j = 0$. Решая задачу (21.9) получим:

$$\begin{aligned} \mu &= \bar{x} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}; \\ \beta_i &= \bar{x}_{i.} - \bar{x} = \frac{1}{k}\sum_{j=1}^{k} x_{ij} - \bar{x}; \\ \tau_j &= \bar{x}_{.j} - \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} - \bar{x}. \end{aligned} \tag{21.10}$$

Полученные оценки параметров имеют следующие распределения:

$$\mu \sim N(\mu, \frac{\sigma^2}{nk}); \qquad \beta_i \sim N(\beta_i, \frac{\sigma^2(n-1)}{nk}); \qquad \tau_j \sim N(\tau_j, \frac{\sigma^2(k-1)}{nk}).$$

Оценка S^2 имеет распределение $\sigma^2\chi^2$ с числом степеней свободы равным $nk-(k-1)-(n-1)-1=(n-1)(k-1)$. Сама оценка S^2 равна:

$$S^2 = \frac{1}{(n-1)(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x} \right)^2 . \qquad (21.11)$$

Вторую оценку S_1^2 получим, воспользовавшись тем, что СВ $\bar{x}_{.1}, \ldots, \bar{x}_{.k}$ являются средними значениями по соответствующим столбцам таблицы двухфакторного анализа. Эти СВ при нулевой гипотезе независимы и одинаково распределены по нормальному закону $N(\mu, \sigma^2 / n)$. На основе этого можно считать, что оценка S_1^2 имеет распределение $\sigma^2 \chi^2$ с $(k-1)$ числом степеней свободы и равна:

$$S_1^2 = \frac{n}{k-1} \sum_{j=1}^{k} \left(\bar{x}_{.j} - \bar{x} \right)^2 . \qquad (21.12)$$

Как и в однофакторном анализе, составляем $F-$ отношение двух оценок дисперсий:

$$F = \frac{\frac{n}{k-1} \sum_{j=1}^{k} \left(\bar{x}_{.j} - \bar{x} \right)^2}{\frac{1}{(n-1)(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x} \right)^2} . \qquad (21.13)$$

$F-$ распределение (распределение Фишера–Снедекора) с числом степеней свободы $(k-1)$ и $(n-1)(k-1)$. Критерий проверки гипотезы H_0 имеет вид:

- гипотеза H_0 отвергается на уровне значимости α, если $F \geq F_{1-\alpha; f_1; f_2}$;
- гипотеза H_0 принимается на уровне значимости α, если $F < F_{1-\alpha; f_1; f_2}$.

Здесь $f_1 = (k-1)$ и $f_2 = (n-1)(k-1)$ – число степеней свободы распределения Фишера–Снедекора.

ЛЕКЦИЯ №22

ДВУХФАКТОРНЫЙ АНАЛИЗ (ПРОДОЛЖЕНИЕ)

Пример ДВУХФАКТОРНЫЙ АНАЛИЗ

Исследуем зависимость частоты самопроизвольного дрожания мышц рук (тремора) от тяжести специального браслета, одеваемого на запястье. Полученные результаты приведены в табл.1, причем каждое значение – среднее из 5 экспериментальных измерений.

Таблица 1

Обработка	**1**	**2**	**3**	**4**	**5**
	Вес браслета (фунты)				
Испытуемый (блоки)	**0**	**1.25**	**2.5**	**5**	**7.5**
1	3.01	2.85	2.62	2.63	2.58

2	3.47	3.43	3.15	2.83	2.70
3	3.35	3.14	3.02	2.71	2.78
4	3.10	2.86	2.58	2.49	2.36
5	3.41	3.32	3.08	2.96	2.67
6	3.07	3.06	2.85	2.50	2.43

Заменив в табл. 1 полученные результаты соответствующими рангами, представим их в табл. 2. В двух последних строках таблицы приведены суммы рангов по каждому столбцу и средние суммы рангов по столбцам.

Таблица 2

Обработка	**1**	**2**	**3**	**4**	**5**
Испытуемый (блоки)					
1	5	4	2	3	1
2	5	4	3	2	1
3	5	4	3	1	2
4	5	4	3	2	1
5	5	4	3	2	1
6	5	4	3	2	1
r_j	30	24	17	12	7
$\overline{r_j}$	5	4	2.8333	2	1.1667

Подставив данные из табл. 2 в формулу (21.5), вычислим статистику Фридмана S (здесь $n = 6, k = 5$):

$$S = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} r_j^2 \right] - 3n(k+1) = 22.5333.$$

При уровне значимости $\alpha = 0.05$ определим, что $\chi^2_{4,0.95} = 9{,}487729$. Значит, гипотезу H_0 следует отвергнуть (ее можно было бы принять только на уровне значимости $\alpha \cong 0{,}000157$).

Поскольку имеются основания считать, что частота тремора уменьшается с увеличением веса браслета, обосновано применение критерия Пейджа. Чтобы непосредственно применить формулу (21.6), необходимо перенумеровать столбцы в табл.3 в обратном порядке. Статистика Пейджа соответственно равна:

$$L = \sum_{j=1}^{k} j \cdot r_j = 7 + 2 \cdot 12 + 3 \cdot 17 + 4 \cdot 24 + 5 \cdot 30 = 328.$$

Из таблицы критических значений статистики Пейджа для уровня значимости $\alpha = 0.01$ $L(0.01, 5, 6) = 299$.

Оценим приближенно значение минимального уровня значимости критерия Пейджа. Для этого воспользуемся нормальной аппроксимацией статистики $L^* \sim N(0,1)$. Здесь

$$L^* = \frac{L - nk(k+1)^2/4}{\left[n(k^3-k)^2/(144(k-1))\right]^{1/2}} = \frac{328 - 6 \cdot 5 \cdot (5+1)^2/4}{\left[6(125-5)^2/(144 \cdot 4)\right]^{1/2}} \approx 4.75.$$

Отсюда максимальный уровень значимости, при котором гипотеза H_0 могла бы быть принята, равен $\alpha \approx 0.000001$. Это свидетельствует о том, что для упорядоченных альтернатив критерий Пейджа более мощный, чем критерий Фридмана.

ДВУХФАКТОРНЫЙ АНАЛИЗ В СИСТЕМЕ Statistica 6.0

Решим поставленную задачу (см. таб. 1) в статистическом пакете. Следует отметить, что формат таблицы в точности соответствует форме ввода данных в систему Statistica.

Двухфакторный ранговый анализ

Критерий Фридмана

Данный критерий проверяется в модуле (рис. 1):
Statistics→Nonparametrics→Comparing Multiple dep. Samples (variables)

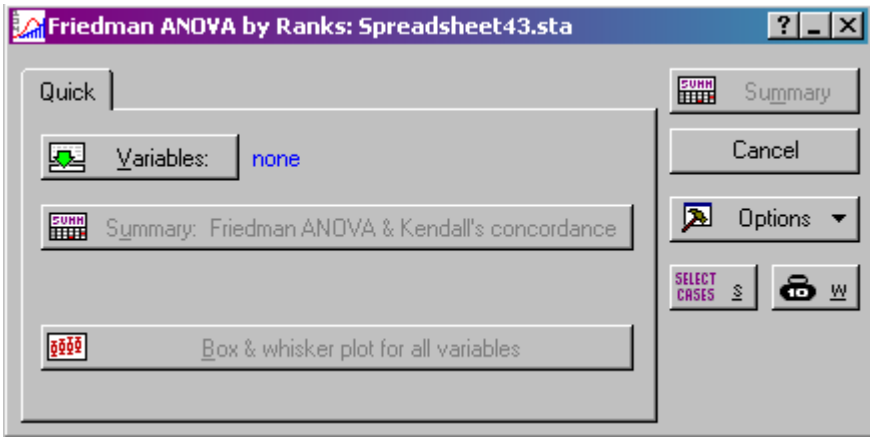


Рис. 1. Тест Фридмана

В закладке **Variables** выбираем все переменные, нажимая на кнопку **Select All**, т.к. проводим анализ для всех групп (столбцов) значений влияющего фактора рис.2.

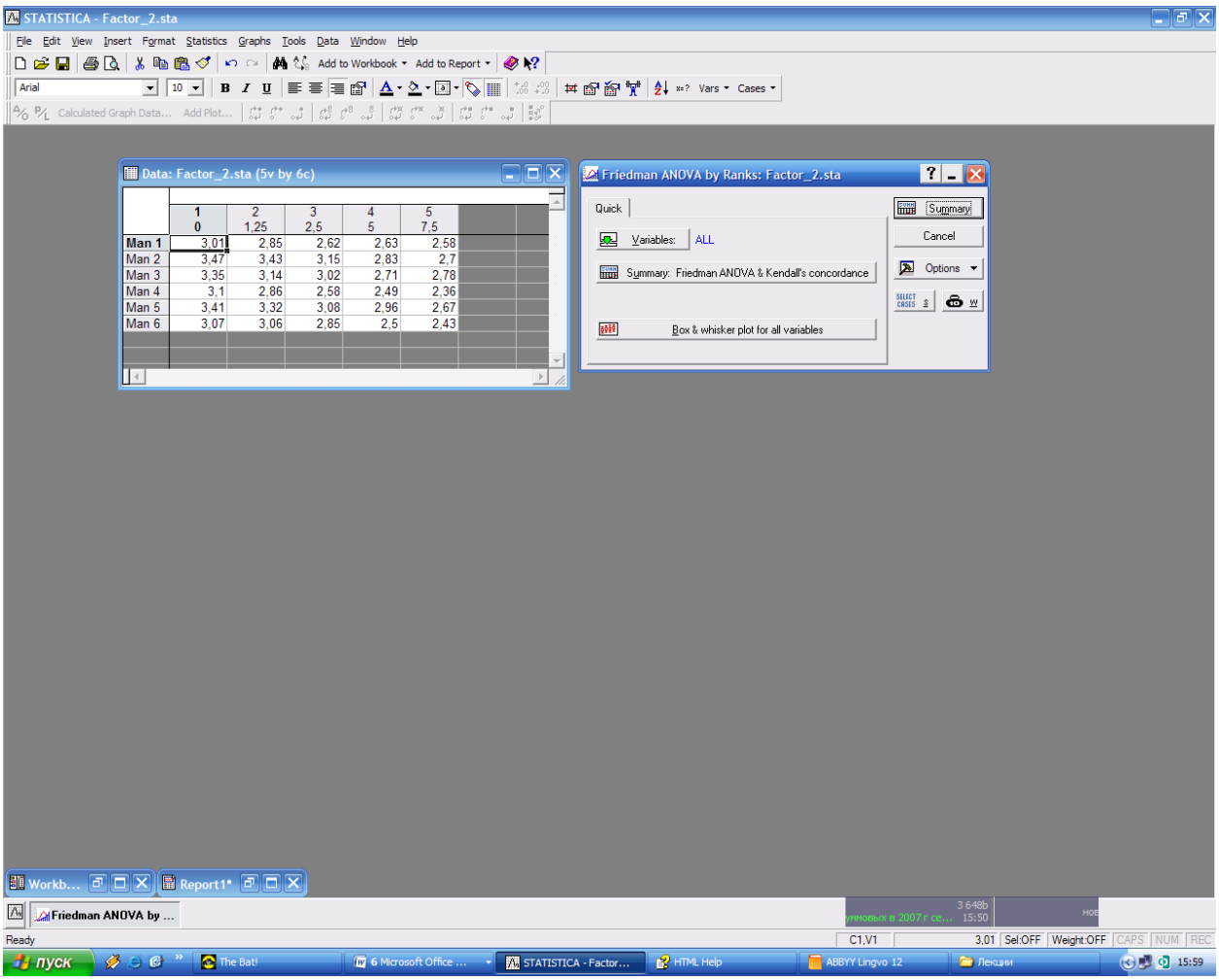


Рис. 2. Таблица данных и модуль теста Фридмана с выбранными переменными

Нажав кнопку *Summary*, получим результат теста Фридмана, представленный на рис. 3.

Friedman ANOVA and Kendall Coeff. of Concordance (Spreadsheet43.sta)
ANOVA Chi Sqr. (N = 6, df = 4) = 22,53333 p < ,00016
Coeff. of Concordance = ,93889 Aver. rank r = ,92667

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
0	5,000000	30,00000	3,235000	0,197560
1,25	4,000000	24,00000	3,110000	0,236643
2,5	2,833333	17,00000	2,883333	0,241219
5	2,000000	12,00000	2,686667	0,185760
7,5	1,166667	7,00000	2,586667	0,163177

Рис. 3 Результаты теста Фридмана

Поясним принятые в таблице обозначения:

Average Rank – средний ранг;
Coeff. of Concordance – коэффициент ранговой корреляции Кендалла;
df – число степеней свободы равное k - 1;
Sum of Ranks – сумма рангов по столбцу;
Std. Dev – среднеквадратическое отклонение;
Mean – среднее значение по столбцу.

Анализируя средние ранги (суммы рангов), можно говорить о качестве влияния фактора. Очевидно, что величина тремора обратно пропорциональна весу браслета.

В модуле теста Фридмана (рис. 1) нажмем закладку Box & Whisker plot, получим график интересующих нас характеристик по группам (рис.5)

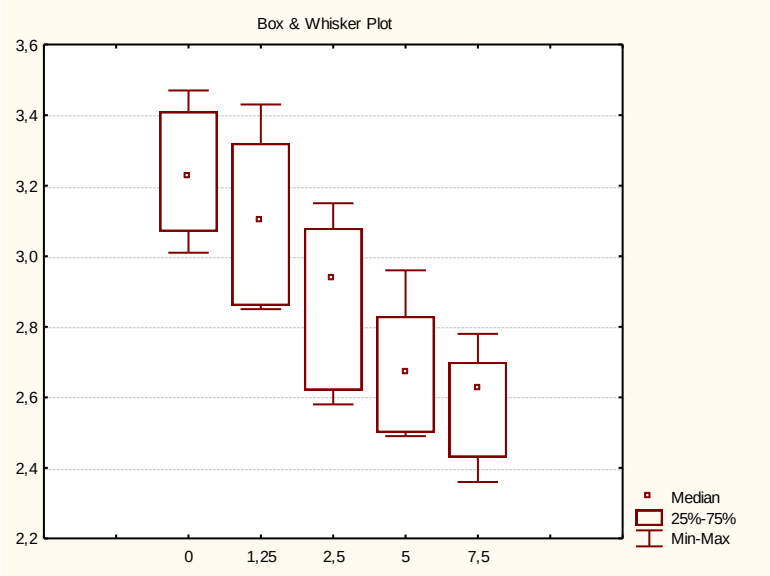


Рис. 5. Характеристики по столбцам при одном уровне главного фактора

Количественно статистика Фридмана показала, что нулевая гипотеза может быть принята с вероятностью $p < 0.00016$. Так как критерий значимости, по умолчанию, был принят равным 0.05, то нулевую гипотезу отвергнем и примем альтернативную – влияние фактора значимо.

Двухфакторный дисперсионный анализ

При проведении двухфакторного дисперсионного анализа выдвигается гипотеза H_0, говорящая о том, что влияние фактора A при наличии мешающего фактора B не существенно. Для проведения анализа исходные данные (табл. №1) были перегруппированы (рис.6):

	1 Vibration	2 Weight	3 Man
1	3,01	0	1
2	3,47	0	2
3	3,35	0	3
4	3,1	0	4
5	3,41	0	5
6	3,07	0	6
7	2,85	1,25	1
8	3,43	1,25	2
9	3,14	1,25	3
10	2,86	1,25	4
11	3,32	1,25	5
12	3,06	1,25	6
13	2,62	2,5	1
14	3,15	2,5	2
15	3,02	2,5	3
16	2,58	2,5	4
17	3,08	2,5	5
18	2,85	2,5	6
19	2,63	5	1
20	2,83	5	2
21	2,71	5	3
22	2,49	5	4
23	2,96	5	5
24	2,5	5	6
25	2,58	7,5	1
26	2,7	7,5	2
27	2,78	7,5	3
28	2,36	7,5	4
29	2,67	7,5	5
30	2,43	7,5	6

Рис. 6. Перегруппированные данные из таб. 1

Дисперсионный анализ проведем в модуле:

Statistics→Anova→Main effects ANOVA + Quick specs dialog (рис. 7).

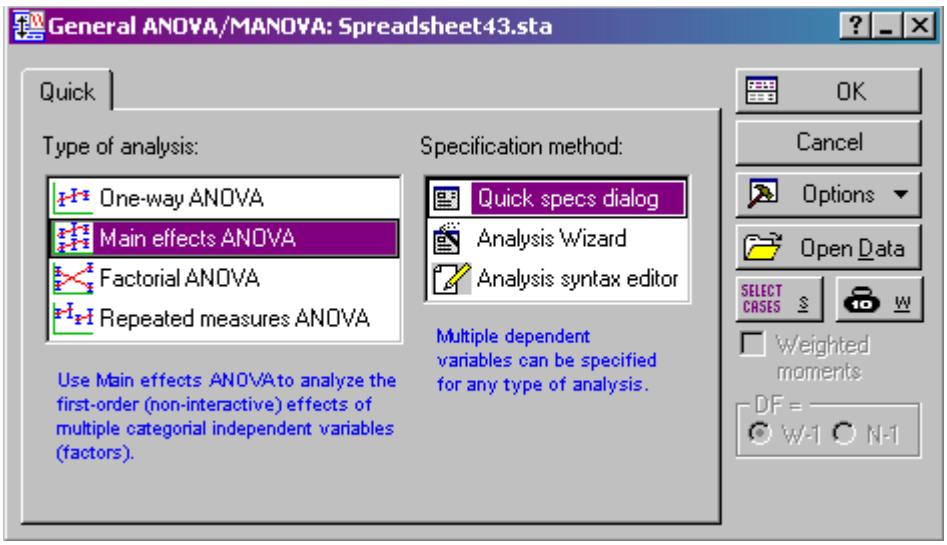


Рис. 7. Дисперсионный анализ данных

Нажав кнопку *ОК,* получим следующее окно рис. 8:

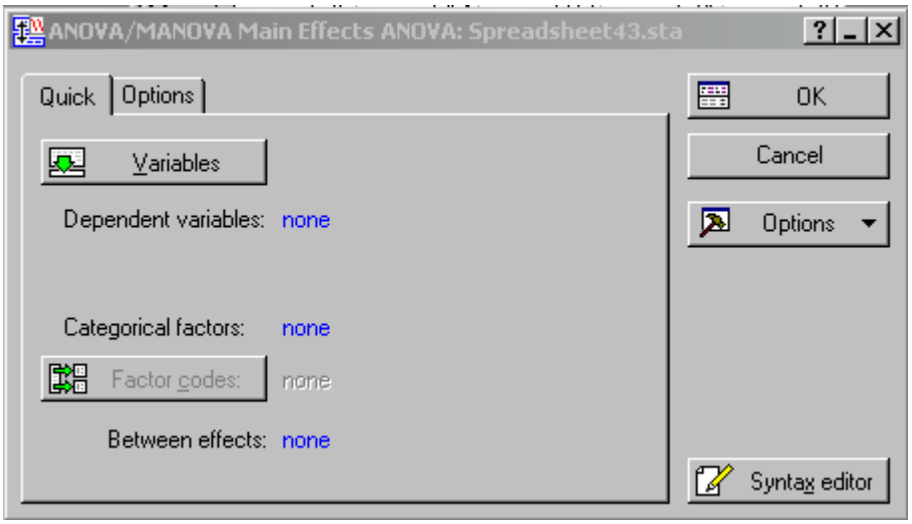


Рис. 8 Окно многофакторного анализа.

В закладке **Variables** выбираем *Dependent variable* и *Categorical predictors* (*factors*) рис. 9:

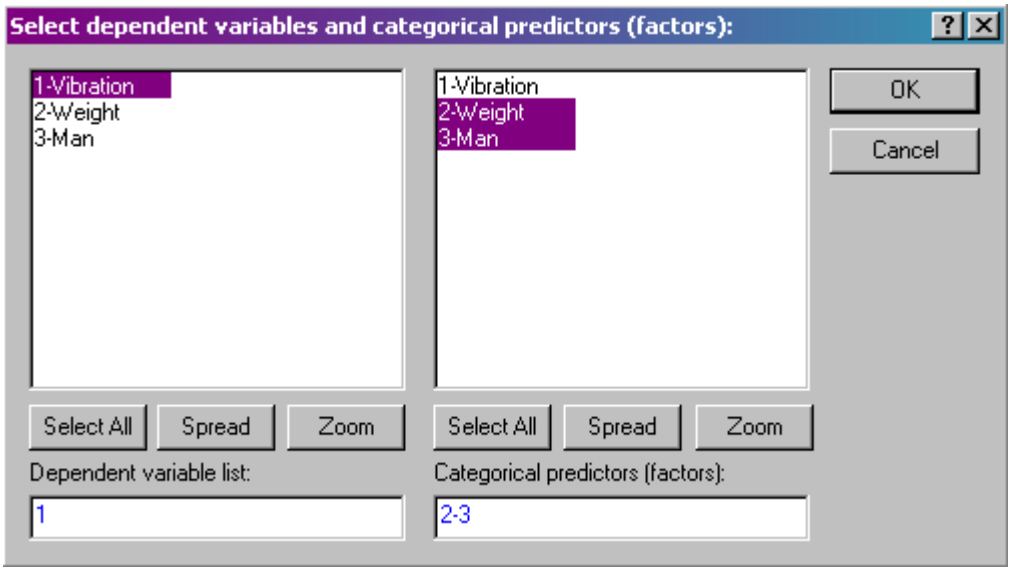


Рис. 9. Окно выбора переменных

После нажатия кнопки *OK*, в окне результатов анализа (рис. 10) выберем пункт ***All effects***:

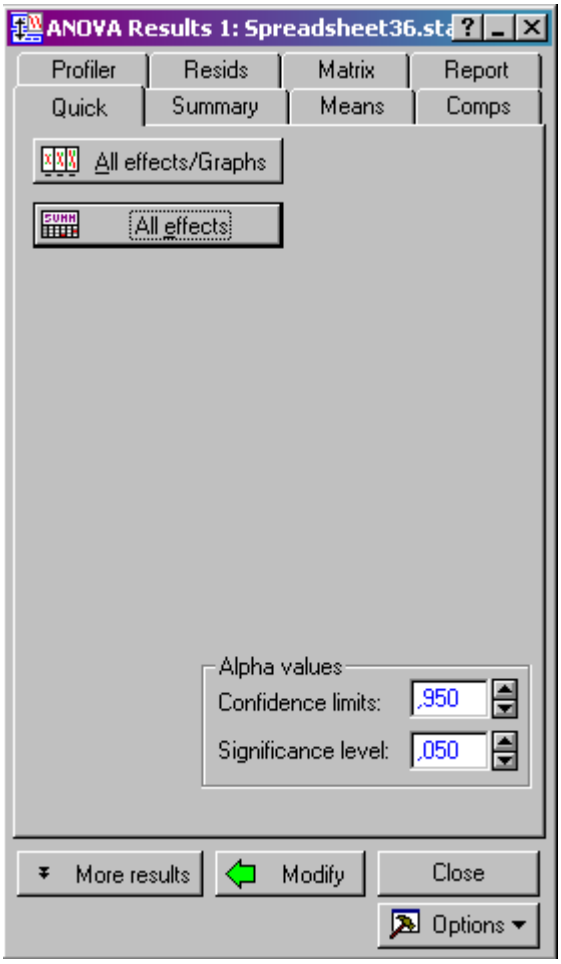


Рис.10. Окно анализа

Результаты анализа представлены на рис. 11.

Effect	Univariate Tests of Significance for Vibration (Spreadsheet36.sta) Sigma-restricted parameterization Effective hypothesis decomposition				
	SS	Degr. of Freedom	MS	F	p
Intercept	252,3580	1	252,3580	30683,07	0,000000
Weight	1,8017	4	0,4504	54,77	0,000000
Man	0,9073	5	0,1815	22,06	0,000000
Error	0,1645	20	0,0082		

Рис. 11. Результаты дисперсионного анализа

<u>В таблице приняты следующие обозначения:</u>

SS (Sum of Squares) Effect – сумма квадратов факторов;
Degr. of Freedom – число степеней свободы фактора;
MS (Mean Square) Effect – средний квадрат фактора;
F – значение критерия Фишера; **p** – вероятность нулевой гипотезы.
Intersept – учитывает в исходной модели (21.2) влияние двух факторов;
Error – остатки, обусловленные разностью исходных данных и модели (21.2).

В полученных результатах минимальное значение статистики Фишера равно 22.06, вероятность нулевой гипотезы при этом равно нулю. Следовательно верна гипотеза H_1 – вес браслета влияет на тремор.

ЛЕКЦИЯ №23

КЛАСТЕРНЫЙ АНАЛИЗ

Если процедура факторного анализа сжимает в малое число количественных переменных данные, описанные количественными переменными, то кластерный анализ сжимает данные в классификацию объектов. Синонимами термина "*кластерный анализ*" являются "*автоматическая классификация объектов без учителя"* и "*таксономия*".

Кластер - это множество объектов, близких между собой по некоторой мере сходства.

Задача классификации – отнесение объекта к определенной группе.

Если данные понимать как точки в признаковом (многомерном) пространстве, то задача кластерного анализа формулируется как выделение "сгущений точек", разбиение совокупности на однородные подмножества объектов. В пространстве переменных кластеры представляют собой скопления точек (объектов) различной формы, рис.1.

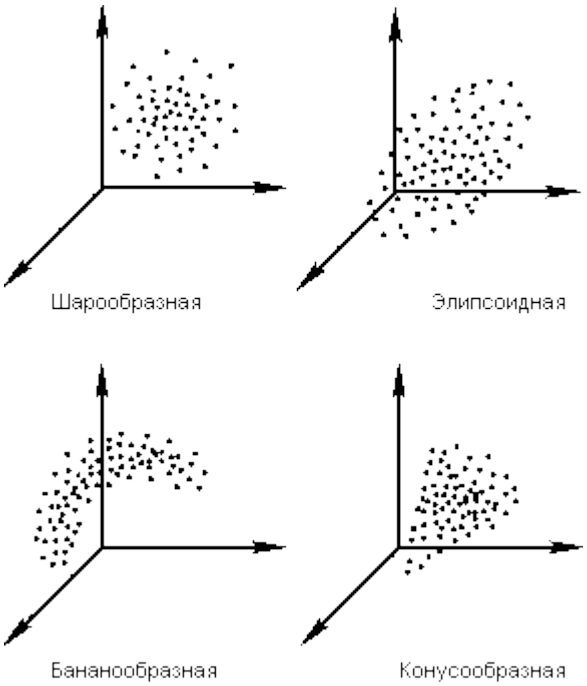


Рис. 1. Наиболее широко распространенные формы скоплений

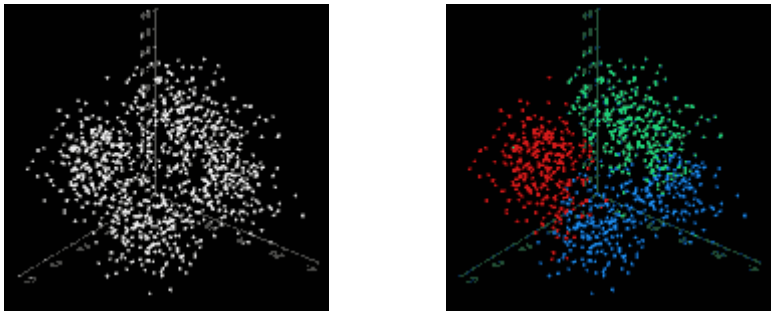
При проведении кластерного анализа обычно определяют расстояние на множестве объектов; алгоритмы кластерного анализа формулируют в терминах этих расстояний.

Кластеризация - это процесс разбиения множества объектов на кластеры (группы объектов, близких по мере сходства). Методы кластеризации делятся на две группы: *классификация с обучением* и *классификация без обучения*.

Классификация с обучением означает, что категории установлены до отнесения объектов к классам.

В *классификации без обучения*: классификационная схема имеет целью определение естественных популяций на основе параметрических или

непараметрических критериев. Примерами классификации без обучения являются: иерархическая классификация и метод ISODATA рис.2.



а) Объекты до кластеризации

б) Объекты после кластеризации

Рис. 2. Кластеризация объектов

На рис.2а изображено скопление "белых", неклассифицированных точек. Кластеризованные данные (см. рис.2б) окрашены в разные цвета, причем каждый кластер образуют облако точек одного цвета.

Для изучения полученного разбиения объектов на однородные группы применяют математические характеристики кластеров рис. 3.

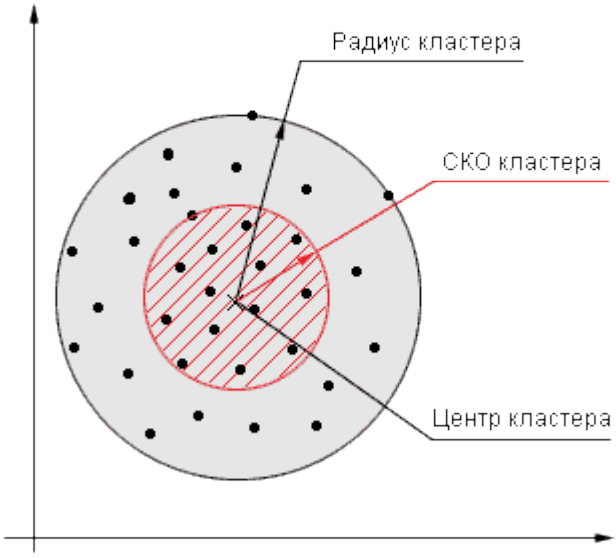


Рис. 3. Графическая иллюстрация основных характеристик кластера

Центр кластера – это среднее геометрическое место точек в пространстве переменных.

Дисперсия кластера – это мера рассеяния точек в пространстве относительно центра кластера.

Радиус кластера – максимальное расстояние точек от центра кластера.

Методы иерархической классификации

Численная классификация или ***численная таксономия*** не занимается распределением объектов по известным классам, а устанавливает классификацию либо не существующую ранее, либо если это желательно, игнорирующую предшествующие работы и пересматривающую данные заново.

Ее цель – почти всегда состоит в упрощении матрицы данных, слишком обширной для непосредственного анализа человеком. Не существует, однако, единственно "правильной" классификации, какого либо набора данных. Различные численные стратегии, как правило, приводят к совершенно разным результатам. Следовательно, необходима помощь специалиста – выбрать тип стратегии.

Итак, исходная информация может быть представлена в форме матрицы X – "объект - свойство":

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(p)} & \cdots & \cdots & x_n^{(p)} \end{bmatrix}. \qquad (23.1)$$

Здесь $x_i^{(j)}$ – значение j -го признака на i-м статистически обследованном объекте. Таким образом, i -й столбец этой матрицы $X_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(p)}\right)'$ характеризует объект O_i, т.е. представляет результат его статистического обследования по всем p анализируемым параметрам (переменным).

Исходная информация, также, может быть задана в форме матрицы ρ попарно взаимных расстояний (близостей) объектов:

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \cdots & \cdots & \rho_{nn} \end{bmatrix}. \qquad (23.2)$$

Здесь ρ_{ij} – характеризует взаимную отдаленность или близость объектов O_i и O_j. В общем случае понятие однородности объектов правилом вычисления ρ_{ij}, характеризующей либо расстояние $d(O_i, O_j)$; $(i, j = 1, 2, \ldots, n)$, либо степень близости (сходства $r(O_i, O_j)$) тех же объектов.

Следует помнить: $r(O_i, O_j) = r(O_j, O_i)$. Требование максимального сходства объекта с самим собой $r(O_i, O_i) = \max r(O_i, O_j)$, и монотонное требование: из $d(O_k, O_l) \geq d(O_i, O_j) \Longrightarrow r(O_k, O_l) \leq r(O_i, O_j)$.

Расстояние между классами и мера близости классов

При кластеризации целесообразно ввести понятие расстояния между целыми группами объектов, так же, как и меру близости двух групп объектов. Введем обозначения:

S_i – i-й кластер.

n_j –число объектов образующих S_i –кластер.

$\bar{X}(i)$ –среднее арифметическое векторных наблюдений, т.е. $\bar{X}(i)$ – центр тяжести i -го кластера.

$\rho(S_i, S_m)$ – расстояние между кластерами S_i и S_m.

Рассмотрим наиболее употребительные и наиболее общие расстояния и меры близости между классами объектов.

Расстояние, измеряемое по принципу "**ближнего соседа**" (Nearest neighbor):

$$\rho_{min}(S_l,\ S_m) = \min_{X_i \in S_l;\ X_j \in S_m} d(X_i, X_j) \qquad (23.3)$$

Расстояние, измеряемое по принципу "**дальнего соседа**" (Furthest neighbor):

$$\rho_{max}(S_l,\ S_m) = \max_{X_i \in S_l;\ X_j \in S_m} d(X_i, X_j) \qquad (23.4)$$

Расстояние, измеряемое по принципу "дальнего соседа" (Furthest neighbor):

Расстояние, измеряемое по "**центрам тяжести групп**" (Centroid clustering):

$$\rho(S_l,\ S_m) = d\big(\bar{X}(l), \bar{X}(m)\big) \qquad (23.5)$$

Примеры расстояний

<u>Обычное евклидово расстояние:</u>

$$d_E(X_i, X_j) = \sqrt{\left(X_i^{(1)} - X_j^{(1)}\right)^2 + \left(X_i^{(2)} - X_j^{(2)}\right)^2 + \cdots + \left(X_i^{(p)} - X_j^{(p)}\right)^2} \qquad (23.6)$$

<u>"Взвешенное" евклидово расстояние:</u>

$$d_E(X_i, X_j) = \sqrt{\omega_1 \left(X_i^{(1)} - X_j^{(1)}\right)^2 + \omega_2 \left(X_i^{(2)} - X_j^{(2)}\right)^2 + \cdots} =$$

$$= \sqrt{\cdots + \omega_\rho \left(X_i^{(p)} - X_j^{(p)}\right)^2}\ . \qquad (23.7)$$

Определение весов ω_i, как правило, связано с дополнительными исследованиями.

<u>Стандартизация.</u>

Непосредственное использование переменных в анализе может привести к тому, что классификацию будут определять переменные, имеющие наибольший разброс значений. Поэтому применяются различные виды стандартизации, одним из которых являются **Z-шкалы** (Z-Scores). Из значений переменных вычитается их среднее значение, и эти значения делятся на стандартное отклонение. Данная стандартизация приводит все переменные к единому диапазону значений от -3 до +3.

<u>Общий вид метрики махаланобисского типа.</u>

В общем случае зависимых компонент $x^{(1)}, x^{(2)}, \cdots, x^{(p)}$ вектора наблюдений X и их различной значимости в решении задачи классификации пользуются обобщенным ("взвешенным") расстоянием махаланобисского типа:

$$d_0(X_i, X_j) = \sqrt{(X_i - X_j)' \Lambda' \Sigma^{-1} \Lambda (X_i - X_j)} \qquad (23.8)$$

Здесь $\sum$ –ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения X_i;

Λ –некоторая симметричная неотрицательно определенная матрица "весовых" коэффициентов λ_{mq}, которая чаще всего выбирается диагональной.

<u>Хеммингово расстояние.</u>

Это расстояние используется как мера различия объектов, задаваемых дихотомическими признаками:

$$d_H(X_i, X_j) = \sum_{S=1}^{p} \left| X_i^{(S)} - X_j^{(S)} \right|. \qquad (23.9)$$

Следовательно, это расстояние равно числу ν_{ij} несовпадений значений соответствующих признаков в рассматриваемых –м и j–м объектах.

Стратегия объединения (*агломеративные системы*)

- Для всех систем вычисляются все $n(n-1)/2$ мер различия и пара индивидов с наименьшей мерой объединяется в одну группу.
- Далее необходимо определить подходящую меру различия между этой группой и остальными $n-2$ индивидами.

Стратегия объединения определяется именно мерой различия между группами.

Рассмотрим комбинаторные решения

Пусть первоначально задана матрица различий (расстояний). Имеются две группы i и j с n_i и n_j элементами соответственно. Мера различия между этими группами обозначается d_{ij} и пусть это минимальная мера из всех оставшихся. Обозначим новую группу через $k \Rightarrow n_k = n_i + n_j$ элементов. Рассмотрим теперь некоторую h группу из $n-1$ оставшихся. В группе $h \Rightarrow n_h$ элементов. Перед объединением известны следующие значения: $d_{hi}, d_{hj}, d_{ij}, n_h, n_j, n_i$. Положим:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma \left| d_{hi} - d_{hj} \right| \qquad (23.10)$$

Параметры $\alpha_i, \alpha_j, \beta$ и $\gamma -$ определяют сущность стратегии.

ЛЕКЦИЯ №24

КЛАСТЕРНЫЙ АНАЛИЗ (ПРОДОЛЖЕНИЕ)

Монотонность

Для графического представления процесса объединения все индивиды (группы) размещаются в соответствующем порядке на оси абсцисс. Последовательность объединений (иерархия или дендрограмма) требует, чтобы каждое объединение было связано с некоторым значением ординаты (обычно для этого используют меру различия).

∃ множество стратегий объединения:

- стратегия "**ближнего соседа**" $\alpha_i = \alpha_j = 1/2$, $\beta = 0, \gamma = -1/2$.
  Это монотонная стратегия, сильно сжимающая пространство.
- стратегия " **дальнего соседа**" (монотонная сильно растягивающая стратегия) $\alpha_i = \alpha_j = 1/2$, $\beta = 0, \gamma = 1/2$.
- гибкая стратегия (применима для любой меры различия и определяется четырьмя ограничениями):

$$\alpha_i + \alpha_j + \beta = 1, \quad \alpha_i = \alpha_j, \qquad \beta < 1, \qquad \gamma = 0.$$

Стратегия монотонная, если $\beta = 0$, то стратегия сохраняет метрику. Если $\beta > 0$, то стратегия сжимает пространство, а если $\beta < 0$, то растягивает. На практике обычно используют $\beta = 1/4$.

Замечание. *Разделяющие* (*дивизионные*) *стратегии здесь не рассматриваются.*

Пример 1: Имеются 5 объектов, для которых заданы меры различия d_{ij}, образующие матрицу D:

Таблица 1.

	1	**2**	**3**	**4**	**5**
1	-	0.227	0.250	0.422	0.897
2	0.227	-	0.492	0.387	0.917
3	0.250	0.492	-	0.356	1.000
4	0.422	0.387	0.356	-	0.773
5	0.897	0.917	1.000	0.773	-

Шаг 1. Т.к. $\min D = d_{12} = 0.227$, то объекты 1 и 2 объединяются в группу 6. Затем вычислим d_{36}, d_{46}, d_{56}. Для вычисления воспользуемся гибкой стратегией:

$$\beta = -0.25;\ \alpha_i = \alpha_j = 0.625;\ \gamma = 0.$$

Согласно (23.10) запишем: $d_{nk} = \alpha_i d_{ni} + \alpha_j d_{nj} + \beta d_{ij};\ \ i\ j\ \Rightarrow k.$

В результате вычислений получим:

$d_{36} = 0.407, d_{46} = 0.449, d_{56} = 1.077.$

Новая матрица будет иметь вид:

Таблица 2.

	6	**3**	**4**	**5**
6	-	0.407	0.449	1.077
3	0.407	-	0.356	1.000
4	0.449	0.356	-	0.773
5	1.077	1.000	0.773	-

Шаг 2. $\mathbf{min\,D} = \mathbf{d}_{34} = 0.356$**,** т.е. на втором шаге объединим группы 3 и 4, новую группу обозначим номером 7.

Шаг 3. Т.к. $\mathbf{min\,D} = \mathbf{d}_{67} = 0.446$**,** то на третьем шаге объединяем группы 6 и 7, новую группу обозначим номером 8.

Шаг 4. На последнем шаге объединяем оставшиеся две группы на уровне $\mathbf{d}_{58} = 1.198$**.** Новую группу обозначим номером 9

Результаты иерархической классификации наблюдений представлены на рис.1.

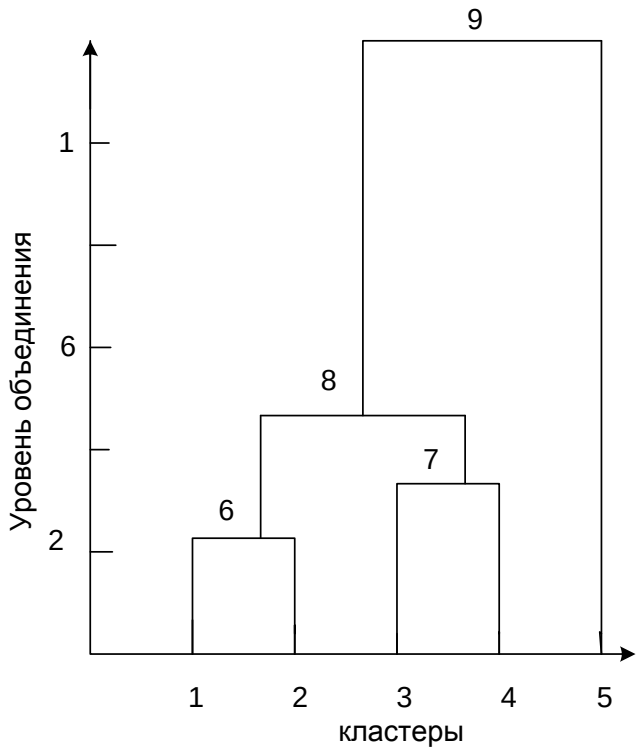


Рис.1. Дендрограмма наблюдений

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

Анализ временных рядов представляет собой самостоятельную, весьма обширную и одну из наиболее интенсивно развивающихся областей математической статистики.

***Временным рядом* (*динамическим рядом*)** в технике и экономике называется последовательность наблюдений некоторого признака (случайной

величины) X в последовательные равноотстоящие моменты времени. Отдельные наблюдения называются *уровнями ряда*, которые будем обозначать
x_t $(t = 1, 2, \dots, n)$, где n – число уровней.

Анализ временных рядов используется, в частности, для решения следующих задач:

- для построения математической модели процесса, представленного временным рядом;
- для исследования структуры временного ряда, например для выявления изменения среднего уровня значений (тренда) и обнаружения периодических колебаний;
- для прогнозирования будущего развития процесса, представленного временным рядом.

Для решения этих и других задач анализа временных рядов исследователями предложено большое количество различных методов:

- методы *корреляционного анализа*, позволяют выбрать наиболее существенные периодические зависимости и их *лаги* (задержки) в одном процессе (автокорреляция) или между несколькими процессами (кросскорреляция);
- методы *спектрального анализа* позволяют находить периодические и квазипериодические зависимости в данных;
- методы *сглаживания и фильтрации* предназначены для преобразования временных рядов с целью удаления из них высокочастотных или сезонных колебаний;
- методы *авторегрессии и скользящего среднего* оказываются особенно полезными для описания и прогнозирования процессов, проявляющих однородные колебания вокруг среднего значения.

Таким образом, важнейшей классической задачей при исследовании временных рядов является выявление и статистическая оценка основной тенденции развития изучаемого процесса и отклонений от нее.

Пример 2. Рассмотрим простейший пример временного ряда. В табл. 3 приведены данные, отражающие цену и спрос (усл. ед.) на некоторый товар за восьмилетний период, т.е. два временных ряда – цена товара x_t и спроса y_t на него. Отметим, что при анализе временных рядов, на первом этапе исследования изучается графическое представление и описание поведения временного ряда (см. рис. 2).

Таблица 3.

Год, t	1	2	3	4	5	6	7	8
Цена, x_t	492	462	350	317	340	351	368	381
Спрос, y_t	213	171	291	309	317	362	351	361

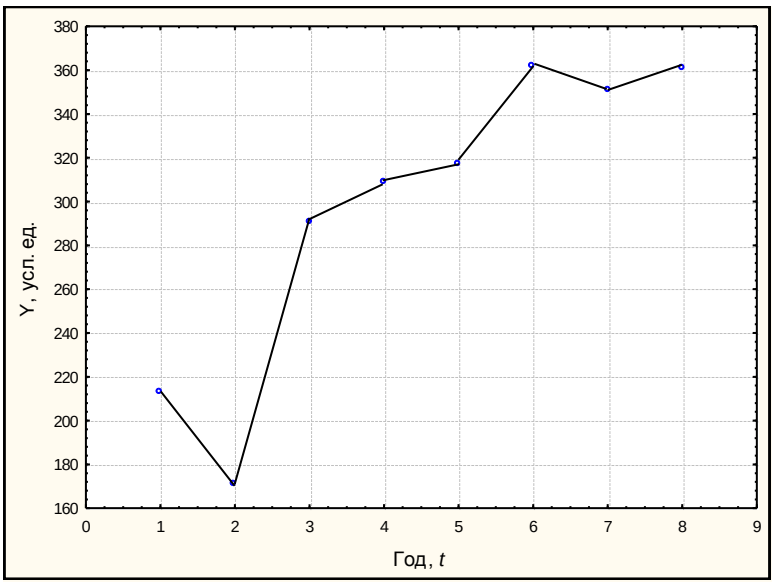


Рис. 2. Динамика спроса за 8 лет

В общем виде при исследовании экономического временного ряда x_t выделяются несколько составляющих:

$$x_t = u_t + v_t + c_t + \varepsilon_t\,; \qquad (t = 1, 2, \dots, n), \qquad (24.1)$$

где u_t – *тренд*, плавно меняющаяся компонента, описывающая чистое влияние долговременных факторов, т.е. длительную ("вековую") тенденцию изменения признака, (например, рост населения, изменение структуры потребления и т.п.);

v_t – сезонная компонента, отражающая повторяемость экономических процессов в течение не очень длительного периода (года, иногда месяца, недели и т.п. например, объем продаж товаров или перевозок пассажиров в различное время года);

c_t – *циклическая компонента*, отражающая повторяемость экономических процессов в течение длительных периодов (например, влияние демографических "ям", циклов солнечной активности и т.п.);

ε_t – *случайная компонента*, отражающая влияние не поддающихся учету и регистрации случайных факторов.

Следует обратить внимание на то, что в отличие от ε_t первые три составляющие (компоненты) u_t, v_t, c_t являются *закономерными, неслучайными*.

Несмотря на кажущуюся схожесть последовательности наблюдений (вариационного ряда) $x_1, x_2, \dots, x_i, \dots x_n$ и временного ряда x_t $(t = 1, 2, \dots, n)$, они имеют принципиальные отличия:

- во-первых, в отличие от элементов выборки члены временного ряда, как правило, не являются статистически независимыми;
- во-вторых, члены временного ряда не являются одинаково распределенными.

Стационарные временные ряды и автокорреляционная функция

Большое значение в анализе временных рядов имеют с*тационарные* временные ряды, вероятностные свойства которых не изменяются во времени.

Временной ряд x_t $(t = 1, 2, \dots, n)$ называется *строго стационарным* (или *стационарным в узком смысле*), если совместное распределение вероятностей n наблюдений $x_1, x_2, \ \dots, x_i, \dots x_n$ такое же, как и n наблюдений $x_{1+\tau}, x_{2+\tau}, \ \dots, x_{i+\tau}, \dots x_{n+\tau}$ при любых n, t и τ. Другими словами, свойства строго стационарных рядов x_t не зависит от момента t, т.е. закон распределения и его числовые характеристики не зависят от t. Следовательно, математическое ожидание $a_x(t) = a$, среднее квадратическое отклонение $\sigma_x(t) = \sigma$ могут быть оценены по наблюдениям x_t $(t = 1, 2, \dots, n)$:

$$\bar{x}_t = \frac{1}{n}\sum_{t=1}^{n} x_t, \tag{24.2}$$

$$S_t^2 = \frac{1}{n}\sum_{t=1}^{n} (x_t - \bar{x}_t)^2. \tag{24.3}$$

Степень тесноты связи между последовательностями наблюдений временного ряда $x_1, x_2, \dots x_n$ и $x_{1+\tau}, x_{2+\tau}, \ \dots, x_{n+\tau}$ (сдвинутых относительно друг друга на τ единиц, или, как говорят, с *лагом* τ) может быть определена с помощью коэффициента корреляции

$$\rho(\tau) = \frac{M[(x_t - M[x_t])(x_{t+\tau} - M[x_{t+\tau}])]}{\sigma_x(t)\sigma_x(t+\tau)} = \\ = \frac{M[(x_t - a)(x_{t+\tau} - a)]}{\sigma^2}. \tag{24.4}$$

Так как коэффициент $\rho(\tau)$ измеряет корреляцию между членами одного и того же ряда, его называют *коэффициентом автокорреляции*, а зависимость $\rho(\tau)$ – *автокорреляционной функцией*. Для стационарного временного ряда автокорреляционная функция $\rho(\tau)$ зависит только от лага τ, причем $\rho(-\tau) = \rho(\tau)$, т.е. при изучении автокорреляционной функции можно ограничиться рассмотрением только положительных значений τ.

Статистической оценкой $\rho(\tau)$ является выборочный коэффициент автокорреляции r_τ, определяемый по формуле коэффициента корреляции (10.7), в которой $x_i = x_t, \ y_i = x_{t+\tau}$, а n заменяется на $n - \tau$:

$$r_\tau = \frac{(n-\tau)\sum_{t=1}^{n-\tau} x_t x_{t+\tau} - \sum_{t=1}^{n-\tau} x_t \sum_{t=1+\tau}^{n} x_{t+\tau}}{\sqrt{(n-\tau)\sum_{t=1}^{n-\tau} x_t^2 - \left(\sum_{t=1}^{n-\tau} x_t\right)^2} \cdot \sqrt{(n-\tau)\sum_{t=1+\tau}^{n} x_t^2 - \left(\sum_{t=1+\tau}^{n} x_{t+\tau}\right)^2}}. \quad (24.5)$$

Функцию r_τ называют *выборочной автокорреляционной функцией*, а ее график – *коррелограммой*.

Для стационарного временного ряда с увеличением лага τ взаимосвязь членов временного ряда x_t и $x_{t+\tau}$ ослабевает и автокорреляционная функция $\rho(\tau)$ должна убывать (по абсолютной величине).

Пример 3. По данным примера 2 для временного ряда y_i вычислим среднее значение, среднее квадратическое отклонение и коэффициент автокорреляции (для лага $\tau = 1, 2, 3$).

РЕШЕНИЕ. По формуле (24.2) вычислим:

$$\bar{y}_t = \frac{213+171+\cdots+361}{8} = 296.88 \text{ (усл. ед.)}.$$

Для вычисления дисперсии (среднего квадратического отклонения) вспомним свойство дисперсии: $S_t^2 = \overline{y_t^2} - \bar{y}_t^2 = 4343.61$, $S_t = 65.31$ (усл. ед.).

Коэффициент автокорреляции для $\tau = 1$, равен коэффициенту корреляции между последовательностями семи пар наблюдений y_t и y_{t+1} $(t = 1, 2, \dots, 7)$:

y_t	213	171	291	309	317	362	351
y_{t+1}	171	291	309	317	362	351	361

Теперь по формуле (24.5) получим: $r_1 = 0.725$. Аналогично вычислим $r_2 = 0.84$, $r_3 = 0.91$.

Знание автокорреляционной функции r_τ может оказать существенную помощь при подборе модели анализируемого временного ряда и статистической оценке ее параметров.

Замечания. *При расчете r_τ следует помнить, что с увеличением τ число $n - \tau$ пар наблюдений x_t, $x_{t+\tau}$* уменьшается, поэтому лаг *τ должен быть таким, чтобы число $n - \tau$ было достаточным для определения r_τ.*

Для выборочного коэффициента автокорреляции r_τ, особенно при небольшом числе пар наблюдений $n - \tau$, свойство монотонного убывания (*по абсолютной величине*) *при возрастании τ может нарушаться* (*см. пример* 2).

ЛЕКЦИЯ №25

Сглаживание временного ряда (выделение неслучайной компоненты)

Одной из важнейших задач исследования временного ряда является выявление *основной тенденции* изучаемого процесса, выраженной неслучайной составляющей $f(t)$ (тренда, либо тренда с циклической или/и сезонной компонентой).

Для решения этой задачи необходимо выбрать модель (*вид функции*) $f(t)$. Наиболее часто используются следующие функции:

- линейная $- \ f(t) = b_0 + b_1 t;$
- полиномиальная $- \ f(t) = b_0 + b_1 t + b_2 t^2 + \cdots + b_n t^n;$
- экспоненциальная $- \ f(t) = e^{b_0 + b_1 t};$
- логистическая $- \ f(t) = \dfrac{a}{1 + be^{-ct}};$
- Гомперца $- \log_c f(t) = a - br^t$, где $0 < r < 1$.

Из двух функций предпочтение, обычно, отдается той, для которой сумма квадратов отклонений фактических данных от расчетных – минимальна. Однако этот принцип нельзя доводить до абсурда, т.к. для любого ряда из n точек можно подобрать полином $(n-1)$-й степени, проходящий через все точки. В этом случае сумма квадратов отклонений будет равна нулю, однако при этом бессмысленно говорить о выделении основной тенденции, учитывая случайный характер эмпирических данных (точек). На практике, при прочих равных условиях предпочтительно использовать более простые функции.

Значения временного ряда x_t или y_t рассматриваются как зависимая переменная, а время t как объясняющая:

$$y_t = f(t) + \varepsilon_t, \tag{25.1}$$

где ε_t — возмущения, представляющие независимые и одинаково распределенные случайные величины, распределение которых предполагается нормальным. Как и в регрессионном анализе, для определения коэффициентов (параметров) модели воспользуемся **методом наименьших квадратов**. Представим функцию $f(t)$ в виде линейной комбинации базисных функций, являющихся линейно независимыми степенными функциями. Учтем, что во временных рядах в качестве аргумента x_i берем t, а $n_i = 1$. Тогда для линейной модели матрица Грама примет вид:

$$\begin{bmatrix} n & \sum_{t=1}^{n} t & \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} t & \sum_{t=1}^{n} t^2 & \sum_{t=1}^{n} y_t\, t \end{bmatrix}. \qquad (25.2)$$

Для квадратичной (параболической) функции $\tilde{y}_t = b_0 + b_1 t + b_2 t^2$ получим матрицу Грама:

$$\begin{bmatrix} n & \sum_{t=1}^{n} t & \sum_{t=1}^{n} t^2 & \sum_{t=1}^{n} y_t \\ \sum_{t=1}^{n} t & \sum_{t=1}^{n} t^2 & \sum_{t=1}^{n} t^3 & \sum_{t=1}^{n} y_t\, t \\ \sum_{t=1}^{n} t^2 & \sum_{t=1}^{n} t^3 & \sum_{t=1}^{n} t^4 & \sum_{t=1}^{n} y_t\, t^2 \end{bmatrix}. \qquad (25.3)$$

Учитывая, что значения переменной $t = 1, 2, \dots, n$ образуют натуральный ряд чисел от 1 до n суммы из (25.2) и (25.3) можно выразить через число членов ряда:

$$\sum_{t=1}^{n} t = \frac{n(n+1)}{2}; \qquad \sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}; \qquad (25.4)$$

$$\sum_{t=1}^{n} t^3 = \frac{n^2(n+1)^2}{4}; \ \sum_{t=1}^{n} t^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}; \quad (25.5)$$

Пример 1. По данным (Пример 2, Лекция 24) найдем уравнение неслучайной составляющей (тренда) для временного ряда y_t, полагая тренд линейным.

РЕШЕНИЕ. По формуле (25.4) вычислим: $\sum_{t=1}^{8} t = \frac{8\cdot 9}{2} = 36$; $\sum_{t=1}^{8} t^2 = \frac{8\cdot 9\cdot 17}{6} = 204$; $\sum_{t=1}^{8} y_t = 213 + 171 + \dots + 361 = 2375$; $\sum_{t=1}^{8} y_t\, t = 213\cdot 1 + 171\cdot 2 + \dots + 361\cdot 8 = 11766$.

Система нормальных уравнений имеет вид:

$$\begin{cases} 8b_0 + 36b_1 = 2375, \\ 36b_0 + 204b_1 = 11766. \end{cases}$$

Отсюда $b_0 = 181.32$; $b_1 = 25.679$ и уравнение тренда: $\tilde{y}_t = 181.32 +$

$25.679t$, т.е. спрос ежегодно увеличивается, в среднем, на 25.7 усл.ед. Исходный ряд и модель линейного тренда приведена на рис.1.

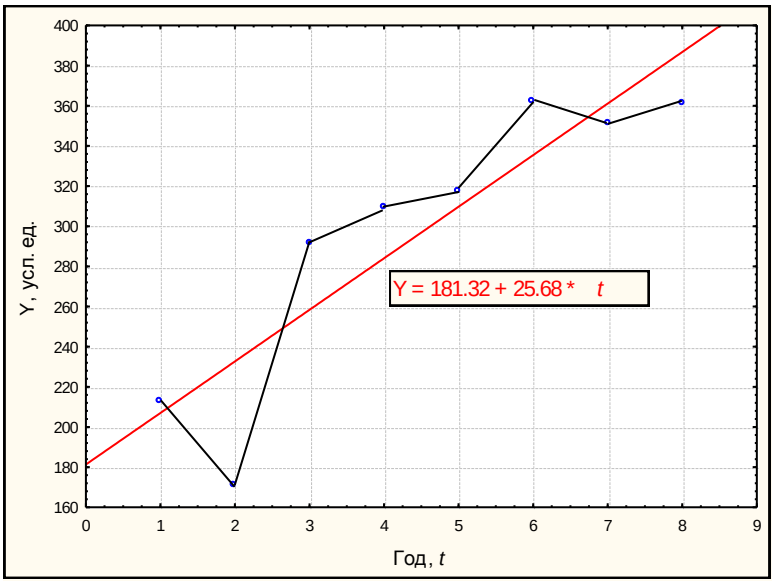


Рис. 1. Линейное сглаживание временного ряда

Проверим качество линейной модели тренда по F – критерию на 5%-ном уровне значимости. Для этого вычислим суммы квадратов:

- обусловленную регрессией – $\hat{S}^2 = \sum(\hat{y}_i - \bar{y})^2 = 27695.3$;
- общую – $S^2 = \sum(y_i - \bar{y})^2 = 34748.9$;
- остаточную – $S_R^2 = S^2 - \hat{S}^2 = \sum(y_i - \hat{y}_i)^2 = 7053.6$.

По формуле (13.3) значение статистики

$$F = \frac{\hat{S}^2(n-2)}{S_R^2} = \frac{27695.3 \cdot 6}{7053.6} = 23.56 > F_{0.05;1;6} = 5.99.$$

Следовательно, принимаем альтернативную гипотезу H_1 – уравнение тренда *значимо*.

Модель скользящего среднего

Другим методом выравнивания (сглаживания) временного ряда, т.е. выделения неслучайной составляющей, является *метод скользящего среднего*. Метод основан на переходе от исходных эмпирических данных к средним значениям на интервале времени, длина которого определена заранее. При этом сам выбранный интервал времени "скользит" вдоль ряда. Получаемый таким образом ряд скользящих средних ведет себя более гладко, чем исходный ряд, из-за усреднения отклонений ряда. Действительно, если индивидуальный разброс значений члена временного ряда x_t около своего среднего (сглаженного) значения a характеризуется дисперсией σ^2, то разброс средних из m членов временного ряда $(x_1 + x_2 + \cdots x_m)/m$ около того же значения a будет характеризоваться существенно меньшей величиной дисперсии, равной σ^2/m. Для усреднения могут быть

использованы средняя арифметическая (простая и с некоторым весом), медиана и др.

Пример 2. По данным (Пример 2, Лекция 24) найдем уравнение неслучайной составляющей (тренда) для временного ряда y_t методом скользящего среднего. В качестве усреднения используем простую среднюю арифметическую оценку с интервалом сглаживания $m = 3$ года.

РЕШЕНИЕ. Скользящие средние находим по формуле:

$$\hat{y} = \frac{1}{m} \sum_{i=t-p}^{t+p} y_i.$$

Здесь $m = (2p + 1) -$ нечетное число; при $m = 3$ $p = 1$. В результате получим сглаженный ряд:

t	1	2	3	4	5	6	7	8
$\hat{y}$	-	225.0	241.0	305.7	329.3	336.3	358.0	-

Полученный результат представлен на рис. 2.

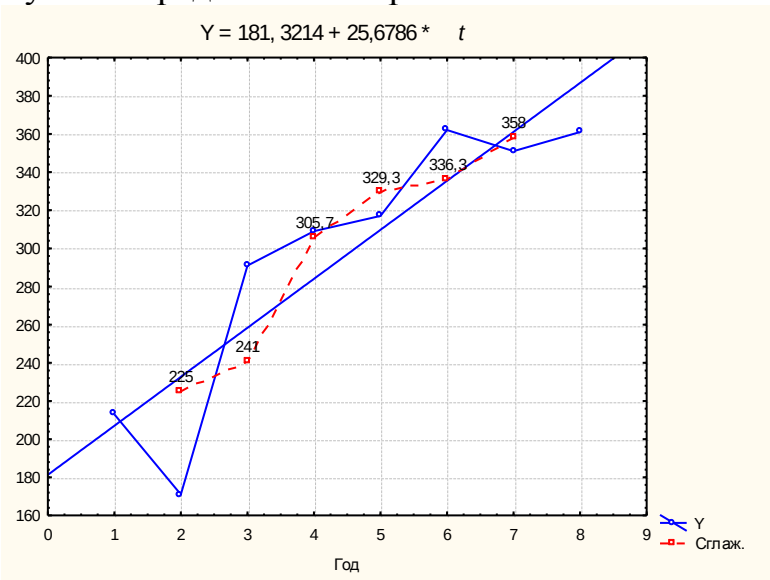


Рис. 2. Сглаживание временного ряда (сплошная ломанная – исходный ряд; прямая – линейная модель; пунктирная ломанная – скользящее среднее m=3)

Прогнозирование временных рядов

Одна из важнейших задач анализа временного (*динамического*) ряда состоит в ***прогнозировании*** на его основе развития изучаемого процесса. При этом исходят из того, что тенденция развития, установленная в прошлом, может быть распространена (*экстраполирована*) на будущий период.

Задача ставится так: имеется временной (динамический) ряд y_t $(t = 1, 2, \dots, n)$, необходимо дать прогноз уровня этого ряда на момент времени $n + \tau$.

Если рассматривать временной ряд как регрессионную модель изучаемого признака по переменной "время", то к нему могут быть применены рассмотренные выше методы анализа. Следует, однако, вспомнить, что одна из основных предпосылок регрессионного анализа

состоит в том, что возмущения ε_t $(t = 1, 2, \dots, n)$ представляют собой независимые случайные величины с математическим ожиданием (средним значением) равным нулю. А при работе с временными рядами такое допущение оказывается в большинстве случаев неверным.

Действительно, если вид функции тренда выбран неудачно, то вряд ли можно говорить о том, что отклонения от нее (возмущения ε_t), являются независимыми. В этом случае наблюдается заметная концентрация положительных и отрицательных возмущений, и можно предполагать их взаимосвязь. Если последовательные значения ε_t коррелируют между собой, то говорят об *автокорреляции возмущений* (*остатков*, *ошибок*).

Метод наименьших квадратов, вообще говоря, и в случае автокорреляции возмущений дает несмещенные и состоятельные оценки параметров, однако их интервальные оценки могут содержать грубые ошибки. В случае выявления автокорреляции возмущений целесообразно вновь вернуться к проблеме спецификации уравнений регрессии (выбора функции тренда), пересмотреть набор включенных в него переменных и т.п.

Наиболее простым и достаточно надежным критерием определения автокорреляции возмущений является *критерий Дарбина – Уотсона*. С помощью этого критерия проверяется гипотеза об отсутствии автокорреляции между соседними остаточными членами ряда e_t и e_{t-1} (для лага $\tau = 1$), где $e_t -$ выборочная оценка ε_t. Статистика критерия достаточно подробно рассмотрена в лекции № 13, формулы (13.5 – 13.6).

Пример 3. По данным (Пример 2, Лекция 24) выявим на уровне значимости $\alpha = 0.05$ наличие автокорреляции возмущений для временного ряда y_t. В случае отсутствия значимой автокорреляции возмущений методами регрессионного анализа найдем точечную и интервальную оценки уровней ряда с надежностью (доверительной вероятностью) $\beta = 0.95$. Прогноз среднего и индивидуального значений спроса на некоторый товар необходимо дать на момент $t = 9$ (девятый год).

РЕШЕНИЕ. В примере 1 было получено уравнение тренда: $\tilde{y}_t = 181.32 + 25.679t$. Все расчеты, необходимые для вычисления $d -$ статистики (Дарбина – Уотсона) сведены в таблицу:

t	y_t	$\tilde{y}_t$	$e_t = y_t - \tilde{y}_t$	e_{t-1}	$e_t e_{t-1}$	e_t^2
1	213	207.0	6.0	–	–	–
2	171	232.7	-61.7	6.0	-370.2	3806.9
3	291	258.4	32.6	-61.7	-2011.4	1062.8
4	309	284.0	25.0	32.6	815.0	625.0
5	317	309.7	7.3	25.0	182.5	53.3
5	362	335.4	26.6	7.3	194.2	707.6
7	351	361.1	-10.1	26.6	-268.7	102.0
8	361	386.8	-25.8	-10.1	260.6	665.6
$\sum_{t=1}^{8}$	–	–	–	–	–1198.0	7023.2

Теперь по формуле (13.6) статистика $d \approx 2\left(1+\frac{1198.0}{7023.2}\right) = 2.34$. Так как в статистических таблицах минимальное значение $n = 15$, у нашего же ряда $n = 8$, поэтому воспользуемся "грубым" правилом: считаем, что автокорреляция остатков отсутствует, если $1.5 < d < 2.5$. Таким образом, для рассматриваемого временного ряда спроса на уровне значимости $\alpha = 0.05$ гипотеза об отсутствии автокорреляции возмущений принимается.

Вычислим условное математическое ожидание, оценкой для которого является групповая средняя:

$M_{t=9}[Y] = \bar{y}(9) \cong \tilde{y}_{t=9} = 181.32 + 25.679 \cdot 9 = 412.4$ (усл. ед.).

Теперь вычислим оценку S^2 дисперсии σ^2: $S^2 = \frac{\sum_{t=1}^{8} e_t^2}{n-2} = \frac{7059.2}{8-2} = 1176.5$. Вычислим оценку дисперсии групповой средней, учитывая (12.14):

$$S^2_{\bar{y}_{t=9}} = S^2\left(\frac{1}{n} + \frac{(t_0-\bar{t})^2}{\sum(t_i-\bar{t})^2}\right) = 1176.5\left(\frac{1}{8} + \frac{(9-4.5)^2}{42}\right) = 714.3;$$

$S_{\bar{y}_{t=9}} = 26.73$; По таблицам получим $t_{0.95;6} = 2{,}447$. Теперь по формуле (12.14) получим интервальную оценку среднего спроса:

$$412.4 - 2.447 \cdot 26.73 \le \bar{y}(9) \le 412.4 + 2.447 \cdot 26.73;$$

Для нахождения интервальной оценки прогноза индивидуального значения $y^*(9)$ вычислим дисперсию оценки:

$S^2_{y_{t_0=9}} = S^2\left(1+\frac{1}{n}+\frac{(t_0-\bar{t})^2}{\sum(t_i-\bar{t})^2}\right) = 1176.5\left(1+\frac{1}{8}+\frac{(9-4.5)^2}{42}\right) = 1890.8$; $S_{y_{t_0=9}} = 43.48$.

Теперь получим интервальную оценку индивидуального значения:

$$412.4 - 2.447 \cdot 43.48 \le y^*(9) \le 412.4 + 2.447 \cdot 43.48;$$

Итак, с надежностью 0.95 *среднее значение спроса на товар на девятый год будет находиться в пределах от* 347 *до* 477.8 (*усл. ед.*) *а его индивидуальное значение* – от 306 *до* 518.8 (*усл. ед.*).

Авторегрессионная модель

Для конкретного временного ряда далеко не всегда удается подобрать адекватную модель вида (25.1), для которой ряд возмущений ε_t будет удовлетворять основным предпосылкам регрессионного анализа, в частности, не будет автокоррелирован. В этом случае целесообразно попытаться использовать другие модели, например, авторегрессионные, учитывающие влияние предыдущих уровней фактора.

Авторегрессионная модель p*-го порядка* имеет вид:

$$x_t = b_0 + b_1 x_{t-1} + b_2 x_{t-2+\cdots+} b_p x_{t-p} + \varepsilon_t \quad (t = 1, 2, \ldots n), \qquad (25.6)$$

где $b_0, b_1, \dots, b_p$ – некоторые константы. Модель описывает изучаемый процесс в момент t в зависимости от его значений в предыдущие моменты $t-1, t-2, \dots, t-p$.

Если исследуемый процесс x_t в момент t определяется лишь его значениями в предшествующий период $t-1$, то рассматривают *авторегрессионную модель* 1-*го порядка* (*марковский случайный процесс*):

$$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t \quad (t = 1, 2, \dots n), \qquad (25.7)$$

Пример 4. В таблице, приведенной ниже, представлены данные, отражающие динамику курса акций некоторой компании ($).

t	1	2	3	4	5	6	7	8	9	10	11
y_t	971	1166	1044	907	957	727	752	1019	972	815	823
t	12	13	14	15	17	17	18	19	20	21	22
y_t	1112	1386	1428	1364	1241	1145	1351	1325	1226	1189	1213

Необходимо построить авторегрессионную модель 1-го порядка, дать точечный и интервальный прогнозы среднего и индивидуального значения курса акций в момент t = 23.

РЕШЕНИЕ. Попытка построить линейную и полиномиальную модель вида (25.1) оказалась бесполезной. Построим авторегрессионную модель тренда вида (25.7). Неизвестные коэффициента получим, решив матрицу Грама (25.2):

$$\tilde{y}_t = 284.0 + 0.7503 y_{t-1}.$$

Найденное уравнение регрессии значимо на 5%-ном уровне (критерий Фишера – Снедекора), так как полученное значение статистики $F = 24.32 > F_{0.05;1;20} = 4.35$.

Проверка критерия Дарбина – Уотсона свидетельствует о незначимой автокорреляции возмущений $e_t = y_t - \tilde{y}_t$.

Вычислим точечный прогноз $\tilde{y}_{t=23} = 284.0 + 0.7503 \cdot 1213 = 1194.1$. Интервальные оценки для среднего и индивидуального значения на уровне значимости 0.05 вычислим по формулам (см. пример 3). Получим окончательные оценки: $1046.6 \le \bar{y}_{t=23} \le 1341.6$; $879.1 \le y^*_{t_0=23} \le 1509.1$.

Итак, с доверительной вероятностью 0.95 *среднее значение курса акций данной компании на момент* t = 23 *будет заключен в пределах от* 1046.6 *до* 1341.6 ($), *а его индивидуальное значение – от* 879.1 *до* 1509.1 ($).