МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Томский государственный университет систем управления и радиоэлектроники»
Кафедра автоматизированных систем управления

ПРИКЛАДНАЯ МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Практические работы

2019

ПРИКЛАДНАЯ МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Практические работы

Составитель А.А. Мицель
Томск: Томский государственный университет систем управления и радиоэлектроники. – 2019. – 81 с.

В пособии приводится описание практических занятий по дисциплине «Прикладная математическая статистика» по следующим темам: генерация случайных чисел с заданным законом распределения, методы получения непрерывных случайных величин на основе равномерного и нормального датчиков, критерии проверки гипотезы о законе распределения выборочных данных, дисперсионный и корреляционный анализ зависимостей, регрессионный анализ.
Пособие подготовлено для магистрантов направления 09.04.01. – «информатика и вычислительная техника» и для магистрантов направления 01.04.02 – «прикладная математика и информатика». Представляет интерес для инженеров, аспирантов, преподавателей, ученых, занимающихся вопросами обработки данных.

ОГЛАВЛЕНИЕ

Тема 1. Генерация случайных чисел с заданным законом распределения

Содержание занятия:

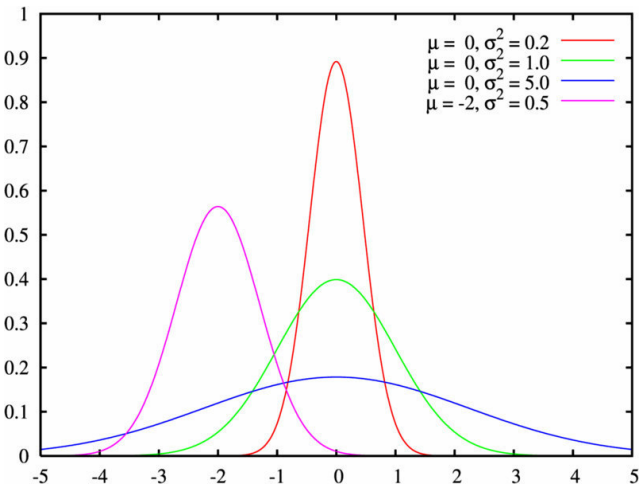
1. Законы распределения случайных величин: нормальное, равномерное, χ^2-распределение (распределение Пирсона), распределение Стьюдента, экспоненциальное, биномиальное, Пуассона.
2. Построение выборок с заданным законом распределения
3. С помощью пакета Mathcad получение выборок случайных чисел с заданным законом распределения, вычисление выборочных моментов.
4. Вычисление вероятностей попадания случайных величин с заданными законами распределения в данный интервал $P(a \le X \le b) = \alpha$, с использованием пакета Mathcad

1.1. Некоторые непрерывные законы распределения случайных величин

1.1.1 Нормальное распределение

Определение. Случайная величина ξ имеет нормальное распределение $N(\mu,\sigma^2)$, где $\sigma > 0$, $\mu \in R$, если ее плотность распределения имеет вид:

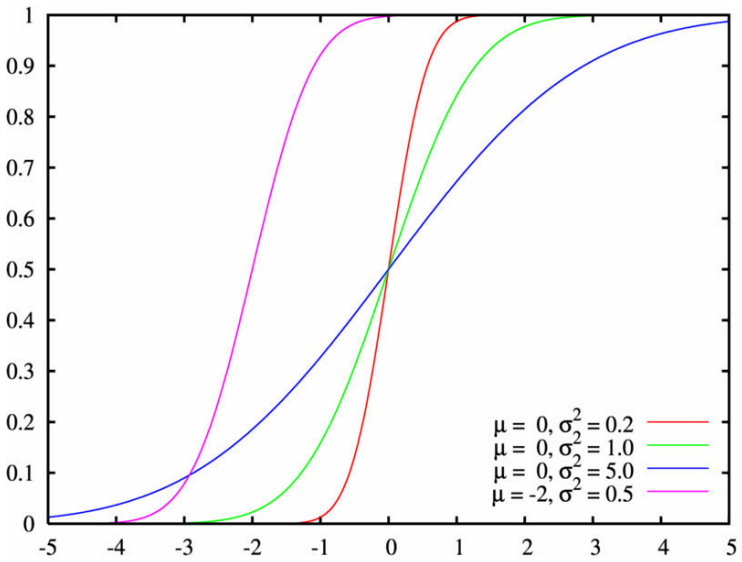
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in (-\infty,\infty).$$



Числовые характеристики

- математическое ожидание $M(\xi)=\mu$;
- дисперсия $D(\xi)=\sigma^2$;
- коэффициент асимметрии $A=\dfrac{M(\xi-\mu)^3}{\sigma^3}=0$;
- коэффициент эксцесса $E=\dfrac{M(\xi-\mu)^4}{\sigma^4}-3=0$;
- медиана $med=F^{-1}(0,5)=\mu$;
- мода $\mathrm{mod}=\arg(\max f(x))=\mu$.

Функция распределения $F(x)=\int\limits_{-\infty}^{x}\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$.



Применение.

Простейшие методы моделирования основываются на центральной предельной теореме. Именно, если сложить много независимых одинаково распределённых величин с конечной дисперсией, то сумма будет распределена *примерно* нормально. Например, если сложить 12 независимых базовых случайных величин, получится грубое приближение стандартного нормального распределения. Тем не менее, с увеличением слагаемых распределение суммы стремится к нормальному.

Пример 1. При сетевом планировании выполнения комплекса работ в условиях, когда продолжительности работ являются случайными величинами, то продолжительность пути при достаточно большом количестве работ может рассматриваться как нормальная случайная величина, среднее значение которой равно сумме средних значений отдельных работ, а дисперсия равна сумме дисперсий отдельных работ.

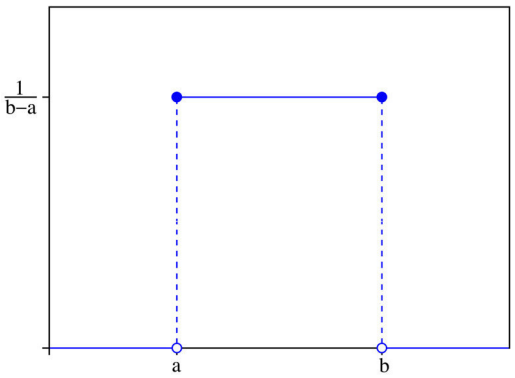
Пример 2. Нормальное распределение и связанное с ним логнормальное очень популярны в классической теории финансовых рынков. Они применяются для моделирования доходностей рисковых активов, таких как, напр., акции или валюты, распределения цены опционных контрактов и т.д. Считается, что на

длинных временных интервалах – квартальных, годовых и т.д. «свечках» распределение реальных логдоходностей довольно неплохо приближается нормальным распределением.

1.1.2 Равномерное распределение

Определение. Говорят, что случайная величина ξ имеет непрерывное равномерное распределение $U(a,b)$ на отрезке $[a,b]$, где $a,b \in R$, если её плотность $f(x)$ имеет вид:

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

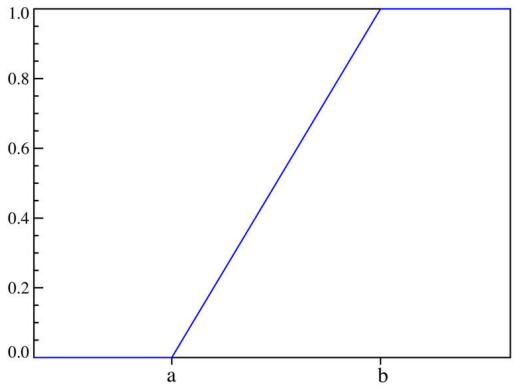


Пишут: $\xi \sim U(a,b)$, или $\xi \sim U_{a,b}$.

Числовые характеристики

- математическое ожидание $M(\xi) = \dfrac{a+b}{2}$;
- дисперсия гамма-распределения $D(\xi) = \dfrac{(b-a)^2}{12}$;
- коэффициент асимметрии $A = 0$;
- коэффициент эксцесса $E = -6/5$;
- медиана $med = F^{-1}(0,5) = \dfrac{a+b}{2}$;
- мода $\mathrm{mod} = x \in [a,b]$ – любое число из отрезка [a,b].

Функция распределения $F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\le x\le b\\ 1, & x>b\end{cases}$.



Применение

Равномерное распределение можно использовать при расчетах по сетевым графикам работ.

С помощью стандартного равномерного генератора можно построить генератор выборки любого непрерывного распределения с помощью метода обратной функции.

1.1.3 Распределение "хи-квадрат"

Определение. Пусть $\xi_1,\xi_2,...,\xi_k$ - независимые случайные величины распределенные по стандартному нормальному закону $N(0,1)$. Тогда случайная величина

$$x=\sum_{j=1}^{k}\xi_j^2$$

имеет распределение "хи-квадрат" с k степенями свободы и обозначают χ_k^2. (Саму случайную величину также часто обозначают χ_k^2). Распределение хи-квадрат является частным случаем гамма-распределения.

Плотность распределения χ_k^2:

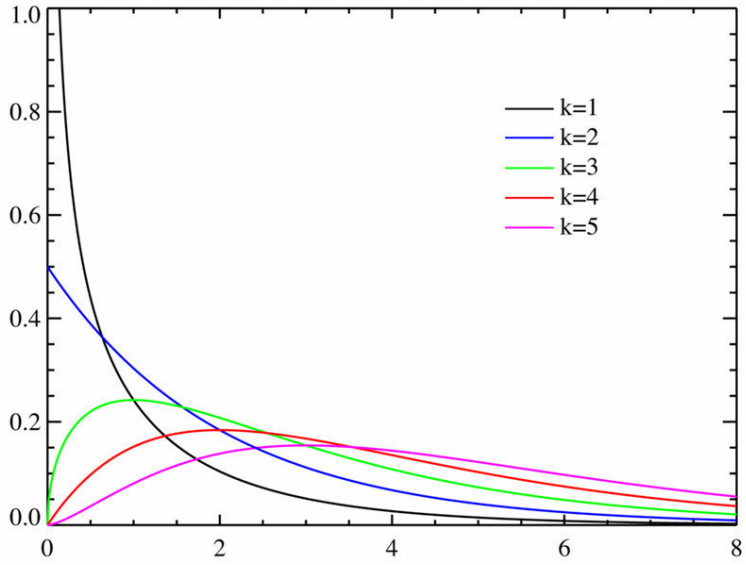
$$f(x)=\frac{1}{2^{k/2}\Gamma(k/2)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}},\quad x>0,$$

а основные числовые характеристики равны:

- математическое ожидание $M(\chi_k^2)=k$;
- дисперсия $D(\chi_k^2)=2k$;
- коэффициент асимметрии $A=\dfrac{M(\xi-k)^3}{(2k)^{3/2}}=\sqrt{8/k}$;
- коэффициент эксцесса $E=12/k$;

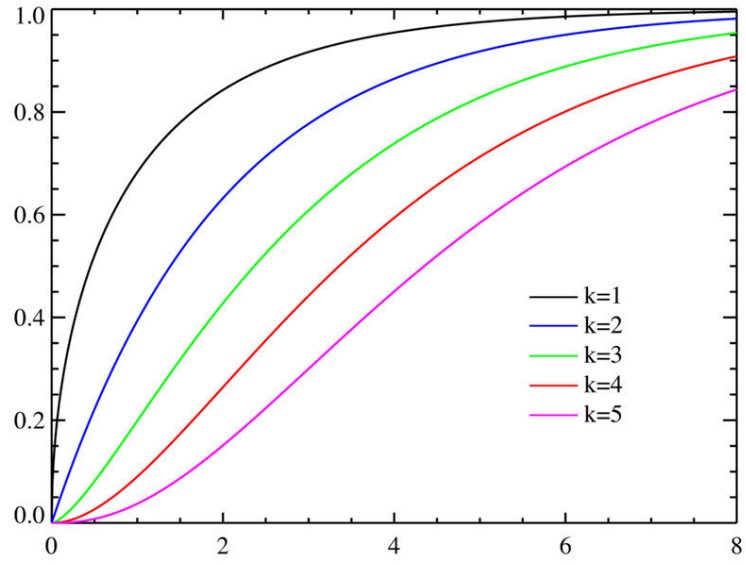
- медиана $med = F^{-1}(0,5) = k - 2/3$;
- мода $\mathrm{mod} = (k-2), \quad k \geq 2$.

График плотности вероятностей для различных степеней свободы k приведен на рисунке



Если случайные величины ξ и η независимы и $\xi \in \chi^2_k$, $\eta \in \chi^2_m$, то, очевидно, их сумма $\xi + \eta \in \chi^2_{k+m}$.

Функция распределения $F(x) = \int_0^x \frac{t^{\frac{1}{2}-1} e^{-t/2}}{2^{k/2} \cdot \Gamma(1/2)} dt$



Применение

Критерий χ² (Хи-квадрат) применяется для проверки гипотезы о законе распределения. Во многих практических задачах точный закон распределения неизвестен, то есть является гипотезой, которая требует статистической проверки.

Обозначим через X исследуемую случайную величину. Пусть требуется проверить гипотезу H_0 о том, что эта случайная величина подчиняется закону распределения $F(x)$. Для проверки гипотезы произведём выборку, состоящую из n независимых наблюдений над случайной величиной X. По выборке можно построить эмпирическое распределение $F^*(x)$ исследуемой случайной величины. Сравнение эмпирического распределения $F^*(x)$ и теоретического распределения $F(x)$, соответствующего гипотезе H_0, производится с помощью критерия согласия χ^2.

1.1.4. Распределение Стьюдента

Определение. Пусть ξ и $\xi_1, \xi_2, ..., \xi_n$ - независимые случайные величины распределенная по закону $N_{0,1}$. Тогда распределение величины

$$t_n = \frac{\xi}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}\xi_j^2}}$$

называется распределением Стьюдента с n степенями свободы и обозначают T_n.

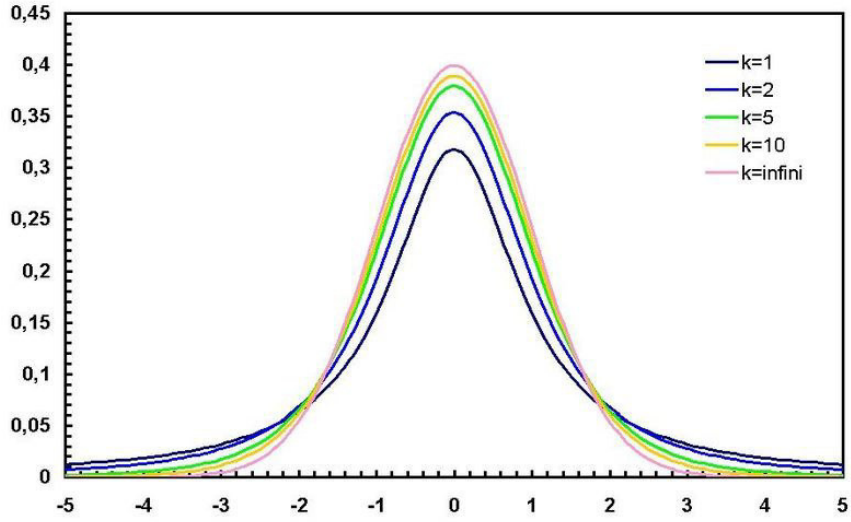
Плотность распределения Стьюдента:

$$f(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \; x \in R$$

Числовые характеристики

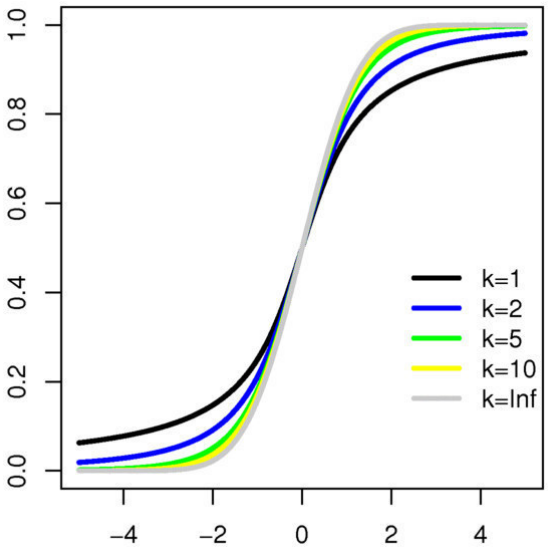
- $M(t_n) = 0, \; n > 1;$
- $D(t_n) = \frac{n}{n-2}, \; n > 2;$
- коэффициент асимметрии $A = 0, \; n > 3;$
- коэффициент эксцесса $E = \frac{6}{n-4}, \; n > 4.$
- медиана $\text{med} = 0;$
- мода $\text{mod} = 0.$

Распределение Стьюдента симметрично относительно $M(t_n) = 0$.



Так как при $n \to \infty$, согласно закону больших чисел, $\frac{\chi_n^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^2 \xrightarrow{p} M\left(\xi^2\right) = M\left(\chi_1^2\right) = 1$, то при $n \to \infty$ $t_n \Rightarrow N_{0,1}$.

Функция распределения $F(x) = \int_0^x \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt$



Применение распределения Стьюдента

Распределение Стьюдента используется в статистике для точечного оценивания, построения доверительных интервалов и тестирования гипотез, касающихся неизвестного среднего статистической выборки из нормального распределения. В частности, пусть $X_1, \ldots, X_n$ независимые случайные величины, такие что

$X_i \sim \mathrm{N}(\mu, \sigma^2),\ i = 1, \ldots, n$. Обозначим $\bar{X}$ выборочное среднее этой выборки, а S^2 её выборочную дисперсию. Тогда

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \mathrm{t}(n-1).$$

1.1.5 Показательное распределение

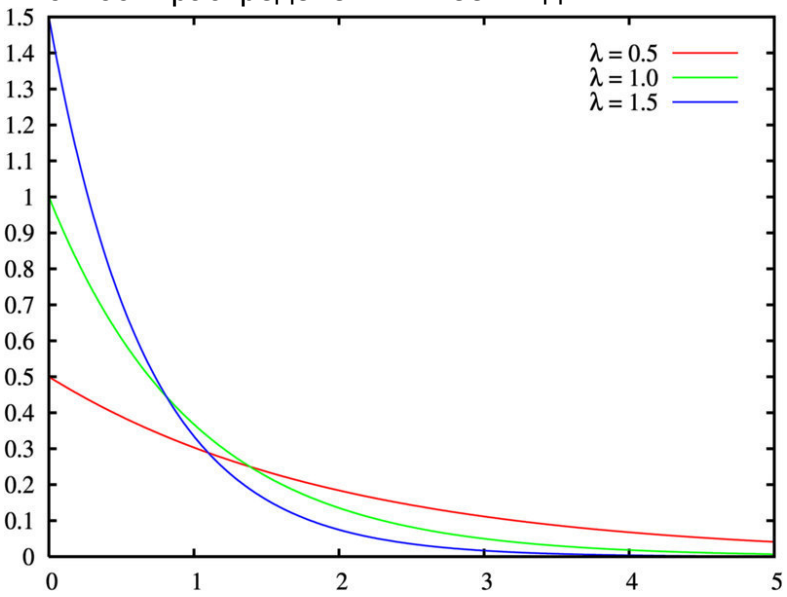
Определение. Случайная величина ξ имеет показательное распределение $\Pi(\lambda)$ с параметром $\lambda > 0$ если ее плотность распределения имеет вид:

$$f(x) = \lambda e^{-\lambda x},\ x \in [0,\infty).$$

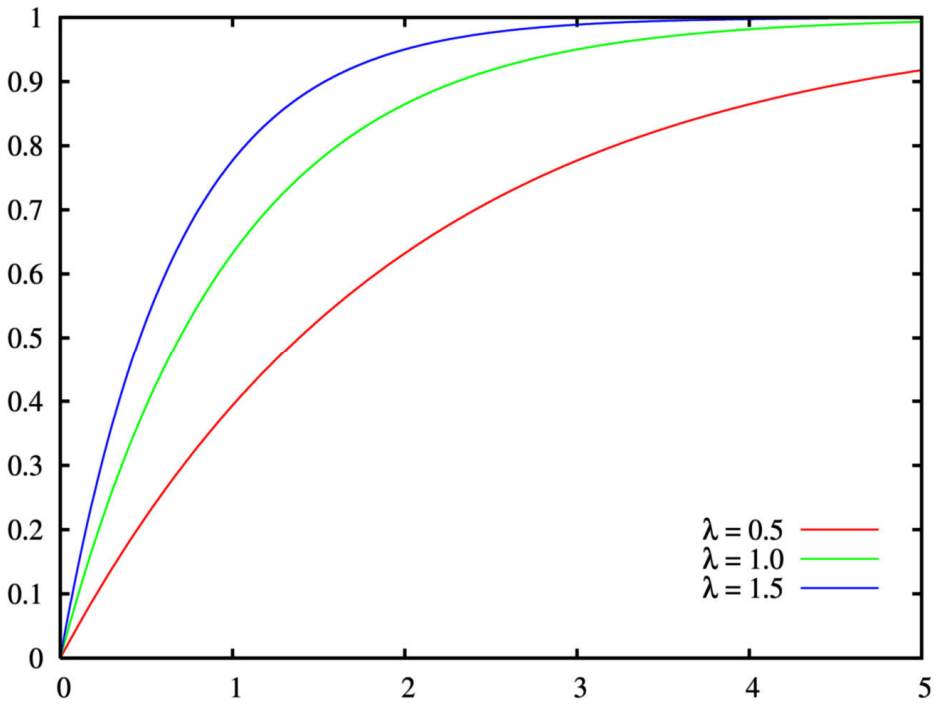
Числовые характеристики

- $M(\xi) = 1/\lambda$;
- $D(\xi) = 1/\lambda^2$;
- коэффициент асимметрии $A = \dfrac{M(\xi - 1/\lambda)^3}{(1/\lambda)^3} = 2$;
- коэффициент эксцесса $E = \dfrac{M(\xi - 1/\lambda)^4}{(1/\lambda)^4} - 3 = 6$;
- медиана $\mathrm{med} = \ln 2/\lambda$;
- мода $\mathrm{mod} = 0$.

Плотность распределения имеет вид



Функция распределения $F(x) = 1 - e^{-\lambda x}$ показана на следующем рисунке



Применение.

Экспоненциальный закон распределения применяют для моделирования следующих явлений в системах массового обслуживания:

- времени поступления заказа на предприятие;
- посещения покупателями магазина супер-маркета;
- времени телефонных разговоров;
- срока службы деталей и узлов в компьютере.

1.2 Дискретные распределения

1.2.1 Биномиальное распределение

Биномиальное распределение — распределение количества «успехов» в последовательности из n независимых случайных экспериментов, таких что вероятность «успеха» в каждом из них постоянна и равна p.

Определение. Пусть $X_1, X_2, ..., X_n$ — конечная последовательность независимых случайных величин с распределением Бернулли, то есть

$$X_i = \begin{cases} 1 & p \\ 0 & q = 1 - p \end{cases}, \quad i = 1, 2, ..., n$$

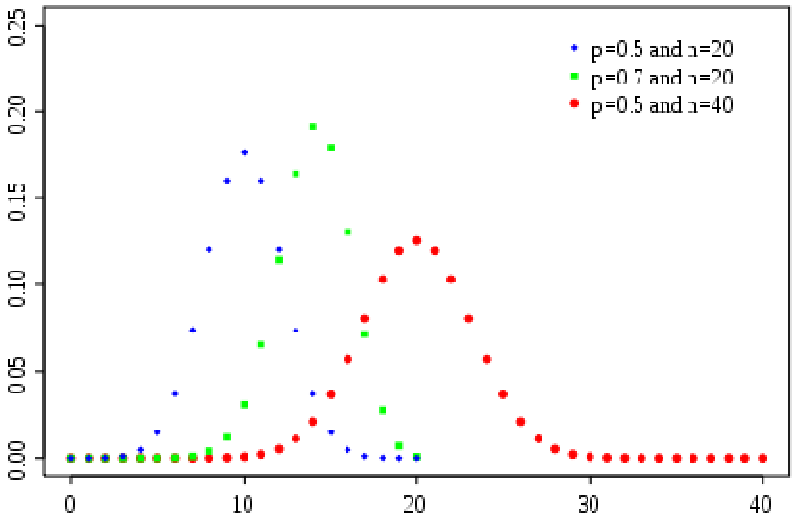
Построим случайную величину Y:

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

Тогда Y – число единиц (успехов) в последовательности $X_1, X_2, ..., X_n$, имеет биномиальное распределение с n степенями свободы и вероятностью «успеха» p. Пишем: $Y \sim B(n, p)$. Распределения вероятности задаётся формулой:

$$p(k) \equiv P(Y = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k}, \ k = 0,1,...,n$$

где $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ — биномиальный коэффициент.



Функция распределения

Функция распределения биномиального распределения может быть записана в виде суммы:

$$F(y) \equiv P(Y \le y) = \sum_{k=0}^{[y]} \binom{n}{k} \cdot p^k \cdot q^{n-k}, \ y \in R,$$

где $[y]$ обозначает наибольшее целое, не превосходящее число y.

Моменты

Производящая функция моментов биномиального распределения имеет вид:
$M_Y(t) = E(e^{tY}) = (pe^t + q)^n$, откуда
$E[Y] = np,$
$E[Y^2] = np(q + np)$,
а дисперсия случайной величины $D[Y] = npq$.

Коэффициент асимметрии $A = \frac{q - p}{\sqrt{npq}}$,

коэффициент эксцесса $E = \frac{1 - 6pq}{npq}$,

медиана $med = \begin{cases} [np]-1 \\ [np] \\ [np]+1 \end{cases}$,

мода $\mathrm{mod} = [(n+1)p]$.

Справка. $E[Y] = \frac{d}{dt} M_Y(t)\Big|_{t=0}$, $E[Y^2] = \frac{d^2}{dt^2} M_Y(t)\Big|_{t=0}$

Свойства биномиального распределения

Пусть $Y_1 \sim B(n, p)$ и $Y_2 \sim B(n, 1-p)$. Тогда $p_{Y_1}(k) = p_{Y_2}(n-k)$.

Пусть $Y_1 \sim B(n_1, p)$ и $Y_2 \sim B(n_2, p)$. Тогда $Y_1 + Y_2 \sim B(n_1 + n_2, p)$.

Связь с другими распределениями

Если $n = 1$, то получаем распределение Бернулли

Если n большое, то в силу центральной предельной теоремы $B(n, p) \sim N(np, npq)$, где $N(np, npq)$ – нормальное распределение с математическим ожиданием np и дисперсией npq.

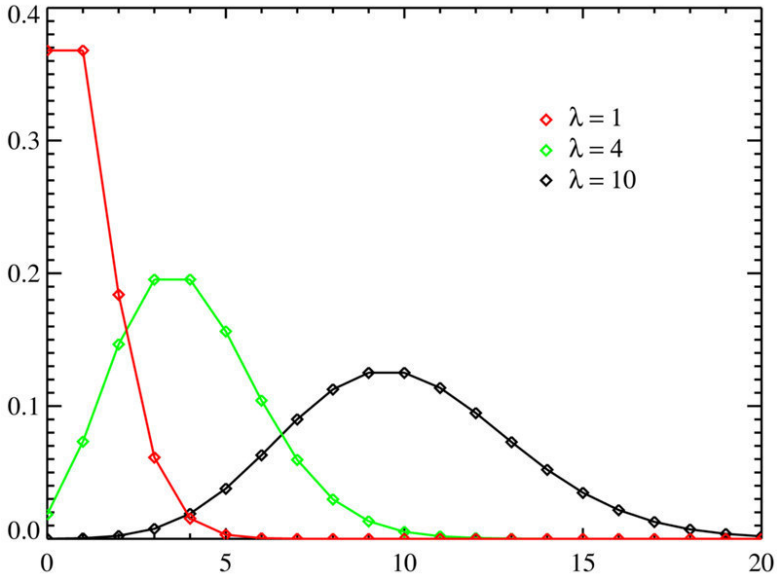
Если n большое, а λ – фиксированное число, то $B(n, \lambda/n) \simeq P(\lambda)$, где $P(\lambda)$ – распределение Пуассона с параметром λ.

1.2.2 Распределение Пуассона

Распределение Пуассона моделирует случайную величину, представляющую собой число событий, произошедших за фиксированное время, при условии, что данные события происходят с некоторой фиксированной средней интенсивностью и независимо друг от друга.

Определение. Говорят, что случайная дискретная величина Y имеет распределение Пуассона $P(\lambda)$ с параметром λ, если функция вероятности задается следующей формулой

$$p(k) = P(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda},\ \lambda > 0;\ k = 0, 1, 2, \ldots.$$

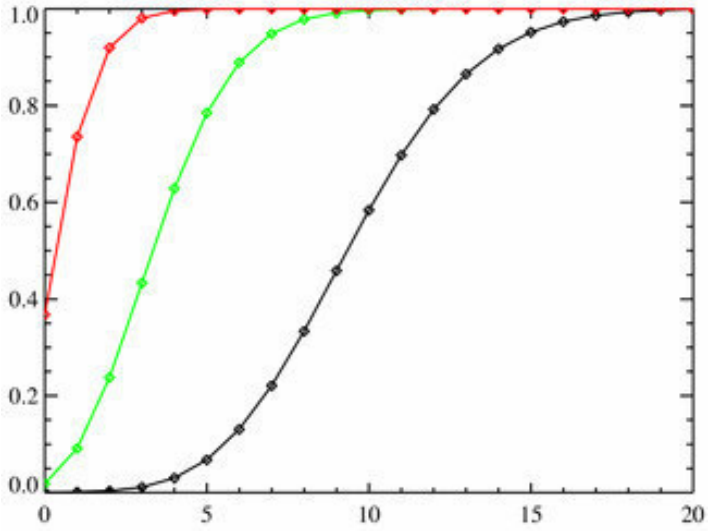


Функция распределения

Функция распределения биномиального распределения может быть записана в виде суммы:

$$F(y) \equiv P(Y \le y) = \frac{\Gamma(k+1,\lambda)}{k!},$$

где $[y]$ обозначает наибольшее целое, не превосходящее число y.



Моменты

Производящая функция моментов биномиального распределения имеет вид:

$$M_Y(t) = E(e^{tY}) = e^{\lambda(e^t-1)},$$

откуда

$$E[Y] = \lambda,$$

$$E[Y^2] = \lambda^2 + \lambda,$$

а дисперсия случайной величины $D[Y]=\lambda$. Коэффициент асимметрии $A=\lambda^{-1/2}$, коэффициент эксцесса $E=\lambda^{-1}$, медиана $med=$.

Применение

Изначально распределение Пуассона было предложено для моделирования потока входящих телефонных звонков на коммутатор. Примеры других ситуаций, которые можно смоделировать, применив это распределение: поломки оборудования, длительность исполнения ремонтных работ стабильно работающим сотрудником, ошибка печати и др.

1.3 Практическое задание

Построение выборок с заданным законом распределения

1) С помощью пакета Mathcad или Excel получить выборку из нормального распределения с параметрами $\mu=1$ и $\sigma=1$ объемом 100 значений. Построить график и рассчитать выборочные моменты распределения.
2) С помощью пакета Mathcad или Excel получить выборку из биномиального распределения с параметрами $n=20$ и $p=0,2$ объемом 100 значений. Построить график и рассчитать выборочные моменты распределения.
3) Вычисление вероятностей попадания нормальных случайных величин в данный интервал $[a,b]$ $P(a\le X\le b)=\alpha$, с использованием пакета Mathcad или Excel.
4) Вычисление вероятностей попадания биномиальной случайной величины в данный интервал $[a,b]$ $P(a\le X\le b)=\alpha$, с использованием пакета Mathcad или Excel.

Тема 2. Методы получения непрерывных случайных чисел на основе равномерного и нормального датчиков

Содержание занятия:

1. Метод обратной функции.
2. Метод композиции случайных величин
3. С помощью пакета Mathcad получение выборок из показательного распределения, распределения χ^2, распределения Стьюдента.

2.1. Метод обратной функции

Для генерации случайных чисел распределенных по законам, которые отсутствуют в генераторе случайных чисел пакета EXCEL, можно использовать следующие приемы.

1. Воспользуемся следующей теоремой из курса “Теория вероятностей”:

Если случайная величина ξ имеет непрерывную и строго монотонную функцию распределения $F_\xi(x)$, а величина $z \in U_{0,1}$, то $\xi = F_\xi^{-1}(z)$, где $F_\xi^{-1}(x)$ - функция обратная к $F_\xi(x)$.

Итак, если $y = F(x)$ - функция распределения непрерывной случайной величины, для которой можно найти обратную $x = F^{-1}(y)$, то генерируя случайную величину z, равномерно распределенную на $[0,1]$, получим величину $\xi = F^{-1}(z)$ с требуемой функцией распределения $F(x)$. Данный прием можно использовать для генерации случайных величин, распределенных по показательному закону, по закону Коши и др.

Пример 1. Распределение Коши имеет плотность $f(x) = \frac{1}{\pi \cdot \theta} \cdot \frac{1}{1 + ((x-a)/\theta)^2}$, $x \in R$.

Функция распределения Коши

$$F(x) = \int_0^x \frac{1}{\pi \cdot \theta} \cdot \frac{1}{1 + ((x-a)/\theta)^2} dx = \frac{1}{\pi} arctg\left(\frac{x-a}{\theta}\right) + \frac{1}{2}.$$

Обратная функция $F^{-1}(x) = a + \theta \cdot tg\left(\pi \cdot (x - 1/2)\right)$. Генерируя случайную величину z, равномерно распределенную на $(0,1)$, получим случайную величину $x = a + \theta \cdot tg\left(\pi \cdot (z - 1/2)\right)$ с плотностью распределения

$$f(x) = \frac{1}{\pi \cdot \theta} \cdot \frac{1}{1 + ((x-a)/\theta)^2}.$$

Пример 2. Задан показательный закон с плотностью $f(x) = \lambda e^{-\lambda x}$. Функция распределения $F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{x} \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$. Обратная функция $F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x)$. Генерируем равномерную случайную величину $z \in U_{0,1}$ и вычисляем искомый результат $x = -\frac{1}{\lambda}\ln(1-z)$.

Пример 3. Задан закон распределения Лапласа с плотностью $f(x) = \frac{1}{\sigma\sqrt{2}} e^{-\frac{|x-a|\sqrt{2}}{\sigma}}$, $x \in R$.

Функция распределения Лапласа имеет вид

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2}} e^{-\frac{|x-a|\sqrt{2}}{\sigma}} dx = \begin{cases} \frac{1}{2} e^{\frac{\sqrt{2}}{\sigma}(x-a)}, & x \le a, \\ 1 - \frac{1}{2} e^{-\frac{\sqrt{2}}{\sigma}(x-a)}, & x > a \end{cases}.$$

Моделирование случайной величины, распределенной по закону Лапласа проводим по формулам

$$x = \begin{cases} a + \frac{\sigma}{\sqrt{2}}\ln(2z), & \text{если } z \le 0{,}5 \\ a - \frac{\sigma}{\sqrt{2}}\ln(2(1-z)), & \text{если } z > 0{,}5 \end{cases},$$

где z – равномерная случайная величина из интервала (0,1).

Пример 4. Задан закон распределения Релея с плотностью $f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$, $x > 0$.

Функция распределения Рэлея $F(x) = \int_{0}^{x} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx = 1 - e^{-\frac{x^2}{2\sigma^2}}$. Моделирование случайной величины, распределенной по закону Релея проводим по формуле

$x = \sqrt{-2\sigma^2 \ln(1-z)}$.

Пример 5. Задан закон распределение с плотностью $f(x) = \frac{2a}{(1+ax)^3}$, $x \in [0, \infty)$.

Функция распределения $F(x) = \int_0^x \frac{2a}{(1+ax)^3} dx = 1 - \frac{1}{(1+ax)^2}$. Случайную величину моделируем по формуле $x = \frac{1}{a}\left(\frac{1}{\sqrt{1-z}} - 1\right)$.

2.2. Метод композиции случайных величин

Если случайная величина есть композиция других случайных величин, то генерируем эти величины и строим из них искомую величину. Данный прием можно использовать, например, для получения случайных величин распределенных по законам χ^2, Стьюдента. Так, случайной величиной, имеющей распределение "хи-квадрат" с k степенями свободы называют величину равную сумме квадратов k независимых стандартных нормальных случайных величин, т.е. $\chi^2 = \sum_{n=1}^{k} \xi_n^2$, $\xi \in N_{0,1}$. Случайной величиной t, имеющей распределение Стьюдента с k степенями свободы называют величину равную $t_k = \frac{\xi}{\sqrt{\chi^2/k}}$, где ξ - случайная величина распределенная по закону $N_{0,1}$, а χ^2 - независимая от нее случайная величина распределенная по закону хи-квадрат с k степенями свободы.

2.3 Практическое задание

Получение выборок

1) С помощью пакета Mathcad получить выборку из показательного распределения с параметром $\lambda = 1$, построить график и рассчитать выборочные моменты.
2) С помощью пакета Mathcad получить выборку из распределения χ^2 с $k = 5$ степенями свободы, построить график и рассчитать выборочные моменты
3) С помощью пакета Mathcad получить выборку из распределения Стьютента с $k = 5$ степенями свободы, построить график и рассчитать выборочные моменты

Тема 3. Критерии проверки гипотезы о законе распределения выборочных данных

Цель занятия:

Оценка закона распределения генеральной совокупности на основе выборочных данных

Содержание занятия:

1) Критерии, основанные на сравнении теоретической плотности распределения и эмпирической гистограммой
2) Критерии, основанные на сравнении теоретической и эмпирической функций распределения вероятностей
3) Критерии нормальности распределения
4) Критерий проверки экспоненциальности распределения

3.1. Критерии, основанные на сравнении теоретической плотности распределения и эмпирической гистограммой

Критерий χ^2 (Пирсона) для простой гипотезы

Пусть $\{X_1, X_2, \ldots, X_n\}$ выборка из генеральной совокупности F. Проверяется гипотеза $H_0 : F = F_1$ против альтернативы $H_1 : F \neq F_1$.

Представим выборку в виде сгруппированного ряда, разбив предполагаемую область значений случайной величины на m интервалов. Пусть n_i - число элементов выборки попавших в i-ый интервал, а p_i - теоретическая вероятность попадания в этот интервал при условии истинности H_0. Составим статистику $\rho(\vec{X}) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$, которая характеризует сумму квадратов отклонения наблюдаемых значений n_i от ожидаемых np_i по всем интервалам группирования.

Теорема Пирсона. Если H_0 верна, то при фиксированном m и $n \to \infty$

$$\rho(\vec{X}) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \Rightarrow \chi^2_{m-1}. \qquad (1)$$

Таким образом, $\rho(\vec{X})$ можно использовать в качестве статистики критерия согласия для проверки гипотезы о виде закона распределения, который будет иметь вид:

$$F(\vec{X}) = \begin{cases} H_0, & \rho(\vec{X}) < \tau_{1-\alpha} \\ H_1, & \rho(\vec{X}) \geq \tau_{1-\alpha} \end{cases}, \qquad (2)$$

где $\tau_{1-\alpha}$ - квантиль распределения χ^2_{m-1}.

Данный критерий называется критерием χ^2 или критерием согласия Пирсона.

Замечание. Критерий не состоятелен для альтернатив, для которых $\tilde{p}_i = p_i$ для всех $i \in \{1,2,\ldots,m\}$. Поэтому, следует стремиться к как можно большему числу интервалов группирования. Однако, с другой стороны, сходимость к χ^2 величины $\frac{(n_i - np_i)^2}{np_i}$ обеспечивается ЦПТ, то есть ожидаемое значение np_i для каждой ячейки не должно быть слишком мало. Поэтому обычно число интервалов выбирают таким образом, чтобы $np_i \geq 5$.

Критерий χ^2 (Пирсона) для сложной гипотезы

Пусть $\{X_1, X_2, \ldots, X_n\}$ выборка из генеральной совокупности F. Проверяется сложная гипотеза $H_0 : F = F_\theta$, где θ - неизвестный параметр распределения F (или вектор параметров), против альтернативы $H_1 : F \neq F_\theta$.

Пусть выборка по прежнему представлена в виде группированного ряда и n_i - число элементов выборки попавших в i-ый интервал, $i \in \{1,2,\ldots,m\}$. Статистику (1) мы не можем в этом случае использовать для построения критерия Пирсона, так как не можем вычислить теоретические значения вероятностей p_i, которые зависят от неизвестного параметра θ. Пусть θ^* - оценка параметра θ, а $p_i^*(\theta^*)$ - соответствующие ей оценки вероятностей p_i. Составим статистику

$$\rho(\vec{X}) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*}.$$

Теорема Пирсона. Если H_0 верна, и l - число компонент вектора θ (число неизвестных параметров распределения), то при фиксированном m и $n \to \infty$

$$\rho(\vec{X}) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*} \Rightarrow \chi^2_{m-l-1}. \qquad (3)$$

Таким образом, критерий Пирсона для параметрической гипотезы будет иметь вид:

$$\delta(\vec{X}) = \begin{cases} H_0, & \rho(\vec{X}) < \tau_{1-\alpha} \\ H_1, & \rho(\vec{X}) \geq \tau_{1-\alpha} \end{cases}, \qquad \rho(\vec{X}) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*}, \qquad (4)$$

где $\tau_{1-\alpha}$ - квантиль распределения χ^2_{m-l-1}.

Замечание. Вообще говоря, оценки, используемые для построения статистики критерия хи-квадрат, должны быть определены из условия минимума статистики $\rho(\vec{X})$. Поэтому желательно уточнить оценки, найденные другим способом (методом максимального правдоподобия или методом моментов) путем минимизации $\rho(\vec{X})$.

Пример 1. Имеем ряд выборочных значений случайной величины ($n = 100$):

43	76	84	91	95	101	105	114	122	129
54	77	84	91	96	101	106	114	122	132
56	77	85	91	96	101	107	115	122	134
57	78	85	91	96	103	107	116	123	136
61	78	86	92	97	103	107	116	124	136
64	79	87	92	97	104	108	116	124	138
67	79	87	93	98	104	111	117	125	143
73	82	87	93	98	104	112	118	125	113
74	82	88	93	*99*	104	113	118	125	145
76	83	89	95	101	105	114	119	126	150

Необходимо проверить критерием χ^2 гипотезу о том, что распределение случайной величины не противоречит нормальному закону с параметрами $\mu = 101$ и $\sigma = 16$ на уровне значимости $\alpha = 0{,}1$.

Решение. Сначала примем решение, на какое количество классов следует разбить гистограмму эмпирического распределения.

$m = 4 \cdot [0{,}75 \cdot (n-1)^2]^{1/5} = 4 \cdot [0{,}75 \cdot 99^2]^{1/5} = 24$; $m = 1 + 3{,}32 \cdot \lg(100) = 8$.

Учитывая, что первая рекомендация эффективна при $n \geq 200$, и исходя из ограничения $m \leq \frac{n}{5} = 20$, примем $m = 8$.

Продемонстрируем теперь технику вычисления теоретических вероятностей p_i. Пусть a_i и a_{i+1} – границы i-го класса разбиения. Тогда теоретическая

вероятность попадания случайной величины в этот интервал рана

$$p_i = F\left(\frac{a_{i+1}-\mu}{\sigma}\right) - F\left(\frac{a_i-\mu}{\sigma}\right).$$

Например, для интервала $[90;100]$ имеем

$$p_i = F\left(\frac{90-101}{16}\right) - F\left(\frac{90-101}{16}\right) = F\left(\frac{-1}{16}\right) - F\left(\frac{-11}{16}\right) =$$

$$= 1 - F\left(\frac{1}{16}\right) - 1 + F\left(\frac{11}{16}\right) = F\left(\frac{11}{16}\right) - F\left(\frac{1}{16}\right) = 0,229.$$

Выберем границы классов и условия равномерного разбиения диапазона изменения случайной величины на 8 классов, с условием попадания в крайние классы не менее 5 наблюдений. Результаты сведем в таблицу:

i	a_i	n_i	$F(a_{i+1})$	$F(a_i)$	p_i	np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
1	< 70	7	0,0263	0,0000	0,0263	2,6300	19,0969	7,2610
2	70 - 80	10	0,0945	0,0263	0,0682	6,8200	10,1124	1,4830
3	80 - 90	13	0,2458	0,0945	0,1513	15,1300	4,5369	0,2998
4	90 -100	18	0,4751	0,2458	0,2293	22,9300	24,3049	1,0600
5	100 -110	17	0,7131	0,0475	0,2380	2,3,8000	46,2400	1,9428
6	110 -120	14	0,8827	0,7131	0,1696	16,9600	8,7616	0,5166
7	120 -130	12	0,9650	0,8827	0,0824	8,2400	14,1317	1,7157
8	> 130	9	1,0000	0,9650	0,0350	3,5000	30,2500	8,6428
		100			1,0	10,197	13,651	$\chi^2 = 22,92$

Итак, значение статистики критерия $\chi^2 = 22,92$. Теперь необходимо найти критическое значений статистики, равное $\chi^2_{1-\alpha}(v = m-1)$. По таблице процентных точек χ^2 - распределения находим . Можно использовать аппроксимацию

$\chi^2_{1-\alpha}(v) = v \cdot \left(1 - \frac{2}{9v} + u_{1-\alpha}\sqrt{\frac{2}{9v}}\right)^3$, $u_{1-\alpha}$ – квантиль нормального стандартного распределения. В нашем случае $u_{0,9} = 1,28$, в результате получим

$$\chi^2_{0,9}(7) = 7 \cdot \left(1 - \frac{2}{9 \cdot 7} + 1,28\sqrt{\frac{2}{9 \cdot 7}}\right)^3 = 11.98 \approx 12.$$

Так как $\chi^2 = 22,92 > 12$, нулевая гипотеза отклоняется.►

3.2 Критерии, основанные на сравнении теоретической и эмпирической функций распределения вероятностей

Критерий Колмогорова-Смирнова

Пусть $F_n(x)$ – эмпирическая функция распределения случайной величины x, представленной выборкой $x_1, x_2, ..., x_n$:

$$F_n(x) = \begin{cases} 0, & x < x_1; \\ \dfrac{i}{n}, & x_i \le x \le x_{i+1},\ 1 \le i \le n-1; \\ 1, & x \ge x_n. \end{cases}$$

Для проверки нулевой гипотезы $H_0 : F_n(x) = F(x)$, где $F(x)$ – полностью определенная (с точностью до параметров) теоретическая функция распределения, рассматривается расстояние между эмпирической и теоретической функциями распределения

$$D_n = \sup_{|x|<\infty} \left| F_n(x) - F(x) \right|;\ D_n^+ = \sup_{|x|<\infty} \left(F_n(x) - F(x) \right);\ D_n^- = -\inf_{|x|<\infty} \left(F_n(x) - F(x) \right).$$

Здесь $\sup$, $\inf$ – точные верхняя и нижняя границы соответствующих разностей.

Для практического применения используются формулы

$$D_n^+ = \max_{1 \le i \le n} \left(\frac{i}{n} - F(x_i) \right);\ D_n^- = \max_{1 \le i \le n} \left(F(x_i) - \frac{i-1}{n} \right);\ D_n = \max\left(D_n^+, D_n^- \right).$$

Критические значения разностей рассчитываются по приближенным формулам

$$D_n(\alpha) = \left\{ \frac{1}{2n} \ln \frac{2}{1-\alpha} \right\}^{1/2}.$$

Если $D_n > D_n(\alpha)$, то гипотеза согласия H_0 отклоняется на уровне значимости α.

При $n \ge 20$ полезна аппроксимация

$$\chi^2 = \frac{1}{9n} \left(6n D_n^{+(-)} + 1 \right)^2,$$

Распределение которой описывается распределением χ^2 с $\nu = 2$ степенями свободы.

При $n \geq 10$ необходимо использовать более точное приближение

$$D_n^{+(-)}(\alpha) = \left\{ \frac{1}{2n}\left(y - \frac{2y^2 - 4y - 1}{18n} \right) \right\}^{1/2} - \frac{1}{6n} \approx \left(\frac{y}{2n} \right)^{1/2} - \frac{1}{6n},$$

где $y = -\ln \alpha$ для $D_n^{+(-)}(\alpha)$ и $y = -\ln(\alpha/2)$ для D_n, при $0{,}01 \leq \alpha \leq 0{,}2$ и $0{,}005 \leq \alpha$.

Стефенс предложил следующие преобразования статистик $D_n^{+(-)}$, D_n

$\widetilde{D}_n = D_n\left(\sqrt{n} + 0{,}275 - \frac{0{,}04}{\sqrt{n}} \right)$ — для нижней процентной точки;

$\widetilde{D}_n = D_n\left(\sqrt{n} + 0{,}12 + \frac{0{,}11}{\sqrt{n}} \right)$ — для верхней процентной точки;

$\widetilde{D}_n^{+(-)} = D_n\left(\sqrt{n} + 0{,}12 + \frac{0{,}11}{\sqrt{n}} \right)$.

Критические значения статистик Стефенса приведены в табл. 1.

Таблица 1. Процентные точки статистик $\widetilde{D}_n$ и $\widetilde{D}_n^{+(-)}$

α	00,150	0,100	0,050	0,025	0,010
$\widetilde{D}_n$	0,973	1,073	1,224	1,358	1,518
$\widetilde{D}_n^{+(-)}$	1,138	1,224	1,358	1,480	1,628

Критерий Колмогорова-Смирнова применяется при $n \geq 50$.

Пример 2. Проверить на уровне значимости $\alpha = 0{,}1$ нормальность распределения выборки x_i: 4, 7, 8, 9, 12, 19, 21, 25, 30 при условии, что $F(x) = N(10;5)$ (т. е. гипотетическим распределением является нормальное распределение с параметрами $\mu = 10$ и $\sigma = 5$).

Задача является демонстрационной — на практике критерий Колмогорова-Смирнова применяется при $n \geq 50$. Для вычисления значений функции нормального распределения $F(x)$ можно использовать либо таблицы, либо аппроксимации. Результаты расчетов сведем в таблицу:

i	x_i	z_i	$F(z_i)$	i/n	$(i-1)/n$	$i/n - F(z_i)$	$F(z_i) - (i-1)/n$
1	4	-1,20	0,1151	0,10	0,00	-0,0151	0,11510
2	7	-0,60	0,2743	0,12	0,10	-0,0743	0,17430
3	8	-0,40	0,3446	0,30	0,20	-0,0446	0,14460
4	9	-0,20	0,4207	0,40	0,30	-0,0207	0,12070
5	12	0,40	0,6554	0,50	0,40	-0,1554	0,25540
6	18	1,00	0,9452	0,00	0,50	-0,3452	0,44520
7	19	1,80	0,9641	0,70	0,60	-0,2641	0,36410
8	21	2,20	0,9866	0,80	0,70	-0,1866	0,18660
9	25	3,00	0,9986	0,90	0,80	-0,0986	0,19860
10	30	4,00	0,9996	1,00	0,90	0,00005	0,09996

Напомним $F(-z_i) = 1 - F(z_i),\ z_i = \dfrac{x_i - \mu}{\sigma}$.

Из таблицы следует, что

$$D_{10}^{+} = \max_{1 \le i \le n}\left(\frac{i}{n} - F(x_i)\right) = 0,00005;\ D_{10}^{-} = \max_{1 \le i \le n}\left(F(x_i) - \frac{i-1}{n}\right) = 0,4452;$$

$$D_{10} = \max\left(D_{10}^{+}, D_{10}^{-}\right) = 0,4452.$$

Критическое значение равно $D_{10}(0,1) = \left\{\dfrac{1}{2 \cdot 10}\ln\dfrac{2}{0,9}\right\}^{1/2} = 0,1998$.

Так как $D_{10} = 0,4452 > D_{10}(\alpha) = 0,1998$, гипотеза нормальности отклоняется на уровне значимости = 0,1.

Более точное приближение вычисляется по формуле

$$\chi^2 = \frac{1}{9 \cdot 10}\left(6 \cdot 10 \cdot 0,4452 + 1\right)^2 = 8,5328.$$

Критическое значение $\chi^2(\alpha)$ при $v = 2$ степенях свободы равно 4,605.

Так как $\chi^2 = 8,53 > \chi^2 = 4,605$, гипотеза H_0 отклоняется.

Рассмотрим более точную аппроксимацию

$y = -\ln(0,1) = 2,302$;

$$D_n^-(\alpha) = \left\{ \frac{1}{20}\left(2,302 - \frac{2 \cdot 2,302^2 - 4 \cdot 2,302 - 1}{18 \cdot 10} \right) \right\}^{1/2} - \frac{1}{6 \cdot 10} = 0,3224 .$$

Так как $D_n^- = 0,4452 > D_n(\alpha) = 0,3224$, гипотеза H_0 отклоняется.

Далее, находим статистику $\widetilde{D}_n^- = 0,4452\left(\sqrt{10} + 0,12 + \frac{0,11}{\sqrt{10}} \right) = 1,445$. Ее критическое значение равно 1,224 (см. табл. 1 при $\alpha = 0,1$). Поскольку, $\widetilde{D}_n^- = 1,445 > \widetilde{D}_n^-(\alpha) = 1,224$, гипотеза H_0 отклоняется.►

Критерий Крамера-фон Мизеса

Статистика критерия имеет вид

$$w^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left\{ F(x_i) - \frac{2i-1}{2n} \right\}^2 ,$$

где $F(x)$ – теоретическая функция распределения.

Необходимо помнить, что теоретическая функция распределения должна быть известна с точностью до параметров. Распространенная ошибка — использование в качестве $F(x)$ функции распределения с параметрами, оцениваемыми по выборке приводит к уменьшению величины критического значения статистики. т.е. к увеличению количества ошибок второго рода.

При объеме выборки $n > 40$ можно использовать приведенные в табл. 2 квантили распределения w^2, которые следуют из его предельного распределения (α – уровень значимости, принятый для проверки H_0).

Таблица 2. Квантили распределения w^2

α	0,900	0,950	0,990	0,995	0,999
$w^2(\alpha)$	0,3473	0,4614	0,7435	0,8694	1,1679

При $n < 40$ можно использовать аппроксимацию

$$(w^2)' = \left(w^2 - \frac{0,4}{n} + \frac{0,6}{n^2} \right) \cdot \left(1 + + \frac{1}{n} \right).$$

Пример 3. В условиях примера 1 проверить нулевую гипотезу нормальности распределения случайных величин критерием w^2.

Решение. Вычисления сводим в таблицу

i	x_i	z_i	$F(z_i)$	$(2i-1)/n$	$F(z_i)-(2i-1)/n$	$\{F(z_i)-(2i-1)/n\}^2$
1	4	-1,20	0,1151	0,1	0,0151	$2{,}28\cdot10^{-4}$
2	7	-0,60	0,2743	0,3	-0,0257	$6{,}60\cdot10^{-4}$
3	8	-0,40	0,3446	0,5	-0,1554	0,02415
4	9	-0,20	0,4207	0,7	-0,2793	0,0780
5	12	0,40	0,6554	0,9	- 0,2446	0,0598
6	18	1,00	0,9452	1,1	-0,1548	0,0240
7	19	1,80	0,9641	1,3	-0,3359	0,1183
8	21	2,20	0,9866	1,5	-0,5134	0,2636
9	25	3,00	0,9986	1,7	-0,7014	0,4919
10	30	4,00	0,9996	1,9	-0,9000	0,8100
						1,8706

Имеем $w^2=\dfrac{1}{12\cdot10}+1{,}8706=1{,}8789$. При $\alpha=0{,}9$ критическое значение равно $w^2(0{,}9)=0{,}3473$. Так как $w^2=1{,}8789>w^2(0{,}9)=0{,}3473$, гипотеза нормальности отклоняется.

Вычислим более точный критерий

$$(w^2)'=\left(1{,}8789-\frac{0{,}4}{10}+\frac{0{,}6}{100}\right)\cdot\left(1++\frac{1}{10}\right)=2{,}029.$$

Видим, что результат тот же – H_0 отклоняется. ►

3.3 Критерии нормальности распределения

Модифицированный критерий χ^2

Пусть дана выборка $x_1,x_2,...,x_n$ данных из распределения $F(x)$. После оценки параметров $\bar{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i,\ s=\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ распределения совокупность выборочных данных разбивается на m равновероятных интервалов ($p_i=\dfrac{1}{m}=const$) и статистика критерия подсчитывается по формуле

$$\chi^2=\frac{m}{n}\sum_{i=1}^{m}n_i^2-n,$$

где n – объем выборки; n_i – количество членов выборки, попавшие в i-й интервал.

Границы интервалов определяются как

$$a_i = \bar{x} + c_i s \ \ (i = 0,1,...,m).$$

Значения коэффициентов c_i приведены в табл. 3. Следует отметить, $c_0 = -\infty$ и $c_m = \infty$. Так как c_i симметричны относительно нуля, то недостающие значения c_i можно найти из соотношений

$c_{\frac{1}{2}(m-1)+i} = -c_{\frac{1}{2}(m-1)-i}$, $(i = 1,...,\frac{m-1}{2})$ – для нечетных m;

$c_{\frac{1}{2}m+i} = -c_{\frac{1}{2}m-i}$, $(i = 1,...,\frac{m-2}{2})$ – для четных m.

Таблица 3. Значения коэффициентов c_i модифицированного χ^2-критерия нормальности для $m \in [3;15]$

m	c_1	c_2	c_3	c_4	c_5	c_6	c_8
3	-0,4307						
4	-0,6745	0					
5	-0,8416	-0,2533					
6	-0,9074	-0,4307	0				
7	-1,0676	-0,5659	-0,1800				
8	-1,1503	-0,6745	-0,3180	0			
9	-1, 2206	-0,7647	-0,4307	-0,1397			
10	-1,2816	-0,8416	-0,5244	-0,2533	0		
11	-1,3352	-0,9085	-0,6040	-0,3488	0,1142		
12	-1,3830	-0,9674	-0,6745	-0,4307	-0,2194	0	
13	-1,4201	-1,0201	-0,7303	-0,5024	-0,2934	-0,0966	
14	-1,4652	-1,0676	-0,7916	-0,5660	-0,3661	-0,1800	0
15	-1,5011	-1,1108	-0,8416	-0,6229	-0,4307	-0,2533	-0,0837

Если $\chi^2 > d_m(\alpha)$, где $d_m(\alpha)$ – критическое значение статистики критерия на уровне значимости α, то гипотеза нормальности отклоняется. Критические значения $d_m(\alpha)$ приведены в табл. 4

Таблица 4. Критические значения $d_m(\alpha)$ модифицированного χ^2 - критерия нормальности

m	α	m	α

	0,10	0,05	0,01		0,10	0,05	0,01
3	2,371	3,248	5,418	10	12,384	14,438	18,852
4	3,928	5,107	7,917	11	13,694	15,843	20,431
5	5,442	6,844	10,075	12	14,988	17,226	21,977
6	6,905	8,479	12,021	13	16,267	19,589	23,495
7	8,322	10,038	13,837	14	17,535	19,937	24,990
8	9,703	11,543	15,567	15	18,792	21,270	26,464
9	11,055	13,007	17,234				

Пример 4. Для данных примера 1 проверить модифицированным критерием χ^2 на уровне значимости $\alpha = 0,1$ гипотезу нормальности распределния при оценке ее параметров по негруппированным данным.

Решение. Имеем $\bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i = 100,77;\ s = \sqrt{\frac{1}{100}\sum_{i=1}^{100}\left(x_i - \bar{x}\right)^2} = 21,583.$

Из табл. 3 находим коэффициенты разбиения (принимаем $m = 10$)

$c_1 = -1,2816;\ c_4 = -0,2533;\ c_7 = 0,5244;$

$c_2 = -0,8416;\ c_5 = 0;\qquad c_8 = 0,8416;$

$c_3 - 0,5244;\quad c_6 = 0,2533;\ \ c_9 = 1,2816$

Результаты расчетов сведем в таблицу:

i	Границы интервалов	n_i	n_i^2
1	$-\infty \div$ 72,79	7	49
2	72,79 $\div$ 82,40	12	144
3	82,40 $\div$ 89,32	11	121
4	89,32 $\div$ 95,24	11	121
5	95,24 $\div$ 100,77	8	64
6	100,77 $\div$ 106,30	13	169
7	106,30 $\div$ 112,22	6	36
8	112,22 $\div$ 119,14	12	144
9	119,14 $\div$ 128,40	10	100
10	128,40 $\div \infty$	10	100
	Σ	100	1048

Статистика критерия равна $\chi^2 = \frac{m}{n}\sum_{i=1}^{m} n_i^2 - n = \frac{10}{100}\cdot 1048 - 100 = 4,8.$

Из табл. 4 находим критическое значение статистики для $m = 10$ и $\alpha = 0,1$: $d_{10}(0,1) = 12,38$. Так как $\chi^2 = 4,8 < 12,38$, гипотеза нормальности исходного распределения вероятностей не отклоняется.►

Критерий типа Колмогорова – Смирнова

Рассмотрим применение критерия Колмогорова-Смирнова (см. раздел 3.2.2) для проверки нормальности распределения в ситуации, когда оба его параметра оцениваются по выборке. Алгоритм проверки нулевой гипотезы H_0 для этого случая сохраняется. При этом используется модифицированная статистика

$$D_n^{и} = D_n\left(\sqrt{n} - 0{,}01 + \frac{0{,}85}{\sqrt{n}}\right).$$

Критические значения $D_n^{и}(\alpha)$ (α – уровень значимости) приведены в табл. 5

Таблица 5. Критические значения статистики Колмогорова – Смирнова, модифицированной для проверки нормальности распределения

α	0,15	0,10	0,05	0,03	0,01
$D_n^{и}(\alpha)$	0,775	0,819	0,895	0,955	1,035

Применим критерий согласия w^2 (см. раздел 3.2.2) для задачи проверки гипотезы нормальности распределения вероятностей случайных величин. Алгоритм вычисления статистики критерия в этом случае не меняется — меняются только критические значения статистики проверки гипотезы. Для различных ситуаций. когда параметры гипотетического распределения оцениваются непосредственно по самой выборке, критические значения статистики w^2 приведены в табл. 6.

Таблица 6, Критические значения статистики w^2 *для* проверки нормальности распределения ($1-\alpha$ – уровень значимости)

Исходные условия	α				
	0,90	0,95	0,99	0,995	0,999
Параметры (μ и σ) известны заранее	0,3473	0,4614	0,7435	0,8694	1,1679
Параметр σ известен, а параметр μ оценивается по выборке $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	0,1344	0,1653	0,2380	0,2698	0,3443
Параметр μ известен, а параметр σ оценивается по выборке $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$	0,2370	0,4418	0,7245	0,8506	1,1490

Параметры (μ и σ) оцениваются по выборке	0,1035	0,1260	0,1788	0,2018	0,2559

Пример 5. Для данных примера 2 проверит гипотезу нормальности распределения случайных величин критерием типа Колмогорова-Смирнова с оценкой параметров распределения по выборке.

Решение. Находим $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 15,3;\ s = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^2} = 8,149$.

Критерий Колмогорова-Смирнова

Имеем $z_i = \frac{x_i - 15,3}{8,149}$. Результаты расчетов сведем в таблицу

i	x_i	z_i	$F(z_i)$	i/n	$(i-1)/n$	$i/n - F(z_i)$	$F(z_i) - (i-1)/n$
1	4	-1,386	0,0823	0,10	0,00	0,0179	0,0823
2	7	-1,018	0,1535	0,12	0,10	0,0465	0,0535
3	8	-0,896	0,1841	0,30	0,20	0,1159	-0,0159
4	9	-0,773	0,2207	0,40	0,30	0,1793	-0,0793
5	12	-0,405	0,3446	0,50	0,40	0,1554	-0,0554
6	18	0,331	0,6293	0,00	0,50	0,0293	-0,1293
7	19	0,454	0,6753	0,70	0,60	0,0247	-0,1293
8	21	0,699	0,7580	0,80	0,70	0,0420	- 0,0420
9	25	1,190	0,8830	0,90	0,80	0,0170	-0,0170
10	30	1,804	0,9640	1,00	0,90	0,0360	-0,0360

Из таблицы следует, что $D_{10}^{+} = \max\left(i/n - F(z_i)\right) = 17,93$;

$D_{10}^{-} = \max\left(F(z_i) - (i-1)/n\right) = 0,1293$;

$D_{10} = \max\left(D_{10}^{+}, D_{10}^{-}\right) = 0,1793$.

Далее $D_n^{н} = 1,1793\left(\sqrt{10} - 0,01 + \frac{0,85}{\sqrt{10}}\right) = 0,613$. Из табл. 5 имеем $D_n^{н}(0,1) = 0,819$. Поскольку $D_n^{н} = 0,613 < D_n^{н}(0,1) = 0,819$, то гипотеза о нормальности распределения принимается.

Критерий w^2

Результаты расчетов представлены в таблице

i	x_i	z_i	$F(z_i)$	$(2i-1)/n$	$F(z_i)-(2i-1)/n$	$\{F(z_i)-(2i-1)/n\}^2$
1	4	-1,386	0,0823	0,1	- 0,0177	$3{,}13\cdot10^{-4}$
2	7	-1,018	0,1535	0,3	-0,1465	0,0214
3	8	-0,896	0,1841	0,5	- 0,3159	0,0998
4	9	-0,773	0,2207	0,7	- 0,4793	0,2297
5	12	-0,405	0,3446	0,9	-0,5554	0,3085
6	18	0,331	0,6293	1,1	- 0,4707	0,2215
7	19	0,454	0,6753	1,3	-0,6247	0,3902
8	21	0,699	0,7580	1,5	-0,7420	0,5506
9	25	1,190	0,8830	1,7	- 0,8170	0,6675
10	30	1,804	0,9640	1,9	-0,9360	0,8761
						Σ = 3,3656

Находим $w^2 = \frac{1}{12\cdot 10} + 3{,}3656 = 3{,}374$. Из табл. 6 имеем $w^2(1-\alpha) = w^2(0{,}1) = 0{,}1035$. Видим, что $w^2 = 3{,}374 > w^2(0{,}9) = 0{,}1035$, поэтому нулевая гипотеза нормальности распределения отклоняется.►

3.4 Критерий проверки экспоненциальности распределения

Критерии типа Колмогорова –Смирнова

Предположим, имеет место гипотетически закон распределения вероятностей $F(x) = 1 - \exp\left(-\frac{(x-\mu)}{v}\right)$, где μ и v – неизвестные параметры, оценки которых по выборке могут быть найдены по формулам (отметим, что выборка упорядочена, т.е. $x_1 \le x_2 \le .... \le x_n$)

$$\hat{v} = \frac{n(\bar{x} - x_1)}{n-1};\ \hat{\mu} = x_1 - \frac{\hat{v}}{n}.$$

Обозначим $z_i = \frac{x_i - \hat{\mu}}{\hat{v}}$ и перейдем к нормированному экспоненциальному распределению $F(z_i) = 1 - \exp(-z_i)$, для которого можно применить следующие критерии согласия

- Критерий Колмогорова-Смирнова

$$D_n^+ = \max\left[\frac{1}{n} - F(z_i)\right];\ D_n^- = \max\left[F(z_i)_i - \frac{i-1}{n}\right];\ D_n \max\left(D_n^+, D_n^-\right);$$

- Критерий Смирнова-Крамера-фон Мизеса $w^2 = \sum_{i=1}^{n}\left(F(z_i) - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}$.

Критерий Фишера

Критерий Фишера имеет вид $F = \frac{\sum_{i=1}^{n} x_i}{(n-1)x_1}$.

Эта статистика имеет F-распределение с $v_1 = 2n-2$ и $v_2 = 2$ степенями свободы. Если $\frac{\sum_{i=1}^{n} x_i}{(n-1)x_1} > F_\alpha(2n-2, 2)$, то нулевая гипотеза отклоняется. Здесь $F_\alpha(v_1, v_2)$ – α - критическое значение F-статистики с v_1 и v_2 степенями свободы.

Для случая проверки экспоненциальности распределения с неизвестными параметрами критические значения для различных уровней значимости приведены в табл. 7. Критические значения F-статистики при уровне значимости $\alpha = 0,05$ приведены в таблице (см. табл. 7 статистические таблицы).

Таблица 7. Критические значения статистик критериев согласия типа Колмогорова-Смирнова для проверки экспоненциальности распределения с неизвестными параметрами

n	Уровень значимости α (верхние процентные точки)					
	0,25	0,15	0,10	0,05	0,025	0,01
	Статистика $\sqrt{n}D_n$					
5	0,683	0,749	0,793	0,865	0,921	0,992
10	0,753	0,833	0,889	0,977	1,048	1,119
15	0,771	0,865	0,912	1,002	1,079	1,163
20	0,786	0,872	0,927	1,021	1,099	1,198
25	0,792	0,878	0,936	1,033	1,115	1,215
50	0,813	0,879	0,960	1,061	1,149	1,257
100	0,824	0,911	0,972	1,072	1,171	1,278
∞	0,840	0,927	0,995	1,094	1,184	1,298
	Статистика w^2					

5	0,083	0,102	0,117	0,141	0,166	0,197
10	0,097	0,122	0,142	0,176	0,211	0,259
15	0,103	0,130	0,151	0,188	0,229	0,281
20	0,106	0,133	0,157	0,195	0,237	0,293
25	0,107	0,135	0,160	0,199	0,247	0,301
50	0,111	0,141	0,166	0,209	0,256	0,319
100	0,113	0,144	0,170	0,215	0,263	0,328
∞	0,116	0,148	0,175	0,222	0,271	0,338

Пример 6.Имеется ряд наблюдений $x_i: 1, 2, 4, 5, 9, 11, 18, 21, 29, 35$.
Проверить гипотезу экспоненциальности распределения вероятностей случайных величин на уровне значимости $\alpha = 0,05$ критериями типа Колмогорова-Смирнова.

Решение. Имеем $\bar{x} = 13,5;\ \hat{v} = \dfrac{10 \cdot (13,5-1)}{9} = 13,889;\ \hat{\mu} = 1 - \dfrac{13,889}{10} = -0,3889$.

Критерий Колмогорова-Смирнова

Вычисления сведем в таблицу

i	$\frac{i}{n}$	z_i	$F(z_i)$	$\frac{i}{n} - F(z_i)$	$\frac{i-1}{n}$	$F(z_i) - \frac{i-1}{n}$
1	0,1	0,100	0,0952	0,0048	0,0	0,0952
2	0,2	0,172	0,1580	0,0420	0,1	0,0580
3	0,3	0,136	0,2709	0,0291	0,2	0,0709
4	0,4	0,388	0,3216	0,0784	0,3	0,0216
5	0,5	0,676	0,4913	0,0087	0,4	0,0913
6	0,6	0,820	0,5596	0,0404	0,5	0,0956
7	0,7	1,324	0,7339	-0,0339	0,6	0,1339
8	0,8	1,540	0,7856	0,0144	0,7	0,0856
9	0,9	2,116	0,8795	0,0205	0,8	0,0795
10	1,0	2,548	0,9218	0,0782	0,9	0,0218

Из таблицы видим, что

$D_n^+ = \max\left[\dfrac{1}{n} - F(z_i)\right] = 0,0784;\ D_n^- = \max\left[F(z_i)_i - \dfrac{i-1}{n}\right] = 0,1339;$

$D_n \max\left(D_n^+, D_n^-\right) = 0,1339$. Значение статистики критерия равно

$\sqrt{n}D_n = \sqrt{10} \cdot 0,1339 = 0,423$. Из табл. 7 для $\alpha = 0,05$ и $n = 10$ находим критическое значение статистики $0,977$. Так как $\sqrt{n}D_n = 0,423 < 0,977$, гипотеза экспоненциальности не отклоняется.

Критерий Смирнова-Крамера-фон Мизеса

Вычисляем статистику критерия

$$w^2 = \left(0{,}0952 - \frac{1}{20}\right)^2 + \left(0{,}158 - \frac{3}{20}\right)^2 + \left(0{,}2709 - \frac{5}{20}\right)^2 + \ldots$$

$$+ \left(0{,}9218 - \frac{19}{20}\right)^2 + \frac{1}{120} = 0{,}0234$$

Из табл. 7 имеем критическое значение статистики $w^2(\alpha) = 0{,}176$ (для $\alpha = 0{,}05$ и $n = 10$).

Поскольку $w^2 = 0{,}0234 < w^2(\alpha) = 0{,}176$, гипотеза экспоненциальности принимается.

Критерий Фишера

Имеем $n = 10,\ x_1 = 1,\ \sum_{i=1}^{10} x_i = 135$. $F = \frac{135}{9} = 15$. Из таблицы распределения Фишера для $F_{0{,}05}(18,2) = 19{,}4$ при уровне значимости $\alpha = 0{,}05$. Так как $F = 15 < F_{0{,}05}(18,2) = 19{,}4$, гипотеза экспоненциальности принимается.►

3.5 Критерии согласия для равномерного распределения

Критерии типа Колмогорова-Смирнова

Приведем модифицированные формы критериев Колмогорова-Смирнова для задачи проверки распределения порядковой статистики $U_1 < U_2 < \ldots < U_n$.

$$D^+ = \max_i \left(U_i - \frac{i}{n+1}\right);\ D^- = \max_i \left(\frac{i}{n+1} - U_i\right);\ D = \max\left(D^+, D^-\right).$$

Распределения указанных статистик быстро сходятся к предельному, если использовать их модификации:

$$\widetilde{D}^+ = \left(D^+ + \frac{0{,}4}{n}\right) \cdot \left(\sqrt{n} + 0{,}2 + \frac{0{,}68}{\sqrt{n}}\right);\ \widetilde{D}^- = \left(D^- + \frac{0{,}4}{n}\right) \cdot \left(\sqrt{n} + 0{,}2 + \frac{0{,}68}{\sqrt{n}}\right);$$

$$\widetilde{D} = \left(D + \frac{0{,}4}{n}\right) \cdot \left(\sqrt{n} + 0{,}2 + \frac{0{,}68}{\sqrt{n}}\right).$$

Критические значения для модифицированных статистик приведены в табл. 8.

Таблица 8. Критические значения $\widetilde{D}^+$, $\widetilde{D}^-$, $\widetilde{D}$ критериев равномерности

n	Уровень значимости α				
	0,01	0,025	0,05	0,1	0,15
$\widetilde{D}^+$	1,518	1,358	1,224	1,073	0,973
$\widetilde{D}^-$	1,518	1,358	1,224	1,073	0,973
$\widetilde{D}$	1,628	1,480	1,358	1,224	1,138

Пример 7. Имеется ряд наблюдений над случайной величиной:

$U_i: 0,047;\ 0,05;\ 0,15;\ 0,18;\ 0,28;\ 0,48;\ 0,52;\ 0,61;\ 0,72;\ 0,91.$

Проверить гипотезу равномерности распределения на уровне значимости $\alpha = 0,05$.

Решение. Результаты вычислений представлены в таблице:

i	U_i	$\frac{i}{n+1}$	$U_i - \frac{i}{n+1}$	$\frac{i}{n+1} - U_i$
1	0,047	0,1111	-0,0641	0,0641
2	0,050	0,1818	-0,0404	0,1318
3	0,150	0,2727	-0,1227	0,1227
4	0,180	0,3636	-0,1836	0,1836
5	0,290	0,4545	0,1645	0,1645
6	0,480	0,5454	-0,0654	0,0654
7	0,520	0,6363	-0,1163	0,1163
8	0,610	0,7272	-0,1172	0,1172
9	0,720	0,8181	-0,0981	0,0981
10	0,910	0,9090	0,0010	-0,0010

Вычисляем

$D^+ = 0,001;\ D^- = 0,1836;\ D = 0,1836.$

$$\widetilde{D}^+ = \left(0,001 + \frac{0,4}{10}\right) \cdot \left(\sqrt{10} + 0,2 + \frac{0,68}{\sqrt{10}}\right) = 0,1466;$$

$$\widetilde{D}^- = \left(0,1836 + \frac{0,4}{10}\right) \cdot \left(\sqrt{10} + 0,2 + \frac{0,68}{\sqrt{10}}\right) = 0,799;$$

$$\widetilde{D} = \left(0,1836 + \frac{0,4}{10}\right) \cdot \left(\sqrt{10} + 0,2 + \frac{0,68}{\sqrt{10}}\right) = 0,799.$$

Из табл. 8 находим для $\alpha = 0,05$ $\widetilde{D}(0,05) = 1,358$.

Так как $\widetilde{D} = 0,799 < \widetilde{D}(0,05) = 1,358$, то гипотеза равномерности распределения принимается.►

3.6 Практическое задание

Проверка гипотез

1) С помощью пакета Mathcad получить выборку из показательного распределения с параметром $\lambda = 1$.
2) Используя критерий Пирсона проверить гипотезу о принадлежности выборки показательному распределению. Использовать простую и сложную гипотезу.
3) С помощью пакета Mathcad получить выборку из нормального распределения с параметрами $\mu = 1,\ \sigma = 1$.

4) Используя критерий Пирсона проверить гипотезу о принадлежности выборки нормальному распределению. Использовать простую и сложную гипотезу.

Тема 4. Дисперсионный анализ данных

Цель занятия:

Практическое проведение дисперсионного анализа выборочных данных

Содержание занятия:

1) Однофакторный параметрический анализ
2) Однофакторный непараметрический анализ

сравнении теоретической и эмпирической функций распределения вероятностей

3) Двухфакторный параметрический дисперсионный анализ
4) Двухфакторный непараметрический дисперсионный анализ

4.1. Однофакторный параметрический анализ

Рассмотрим влияние фактора *A* на исследуемый процесс X , принимающего k различных значений — уровней фактора. На каждом i-м уровне производится n_i наблюдений, результаты которых занесены в таблицу 4.1.

Результат каждого наблюдения может быть представлен в виде модели:

$$x_{ji} = \mu + \alpha_i + e_{ji},\ i = 1,...,n, \qquad (1)$$

где μ – суммарный эффект во всех опытах; α_i – эффект фактора A на i-м уровне; e_{ji} – ошибка определения x_{ji} на i-м уровне.

Таблица 4.1. Форма представления экспериментальных данных однофакторной модели

Номер Наблюдения	Уровни фактора A			
	A_1	A_2	A_i	A_k
1	x_{11}	x_{12}	x_{1i}	x_{1k}
2	x_{21}	x_{22}	x_{2i}	x_{2k}
….	….	….	….	….
j	x_{j1}	x_{j2}	x_{ji}	x_{jk}
….	….	…..	…..	…..
n	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_i i}$	$x_{n_k k}$

Средние значения по уровням	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_i$	$\bar{x}_k$

Предположим, что наблюдения на i-м фиксированном уровне фактора нормально распределены относительно среднего значения $(\mu + \alpha_i)$ с общей дисперсией σ^2.

Общее число опытов

$$N = \sum_{i=1}^{k} n_i . \qquad (2)$$

Следует установить, оказывает ли влияние фактор A на исследуемый процесс X. Сформулируем гипотезу H_0 о том, что расхождение наблюдений в сериях опытов для различных уровней факторов можно объяснить только случайными причинами. На статистическом языке это предположение означает, что все данные таблицы x_{ij} принадлежат одному и тому же распределению.

Осуществим проверку нулевой гипотезы равенства средних значений на различных уровнях фактора A:

$$H_0 : m_1 = m_2 = ... = m_k = m .$$

Наиболее часто расчет проводится при равном числе опытов на каждом уровне A, т.е. $n_1 = n_2 = ... = n_k = n$. При этом общее число наблюдений $N = k \times n$.

Среднее значение результатов наблюдений на i-м уровне:

$$\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ji}}{n} = \frac{A_i}{n}. \qquad (3)$$

Общее среднее значение для всей выборки из N наблюдений:

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji} . \qquad (4)$$

Выборочная дисперсия на каждом уровне:

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_{ji}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_{ji} \right)^2 \right]. \qquad (5)$$

Общая выборочная дисперсия:

$$s_0^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n} (x_{ji} - \bar{x})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}^2 - \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji} \right)^2 \right]. \qquad (6)$$

Если между дисперсиями s_i^2 нет значимых различий (однородность дисперсий s_i^2 определяется по критерию Кохрена), то для оценки дисперсии, характеризующей фактор случайности, можно использовать выборочную дисперсию

$$s_{сл}^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{k(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ji}\right)^2\right] \quad (7)$$

с числом степеней свободы $v = k(n-1) = N - k$.

Введем теперь оценку дисперсии s_A^2, характеризующей изменение средних $\bar{x}_i$, связанное с влиянием фактора A:

$$s_A^2 = \frac{n}{k-1}\sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 . \quad (8)$$

с числом степеней свободы $v = k - 1$.

Если дисперсия s_A^2 значимо отличается от $s_{сл}^2$ то нулевая гипотеза $H_0 : m_1 = m_2 = ... = m_k = m$ отвергается и влияние фактора A считается существенным. Так как альтернативой к $H_0 : s_{сл}^2$ является неравенство $H_1 : \sigma_A^2 > s_{сл}^2$, для проверки гипотезы применяется односторонний критерий Фишера: влияние A считается значимым, если

$$F = \frac{s_A^2}{s_{сл}^2} > F_q(v_1, v_2),\ v_1 = k-1;\ v_2 = N - k = k(n-1) . \quad (9)$$

Если отношение $\frac{s_A^2}{s_{сл}^2} \le F_q(v_1, v_2)$, то влияние фактора A следует считать незначимым. При этом общая дисперсия s_0^2 связана только с фактором случайности.

При значимости влияния фактора A, т.е. при значимости различия между $m_1 = m_2 = ... = m_k = m$, можно выяснить, какие именно средние m_i различны. Для этого используют критерии Стьюдента или ранговый критерий Дункана.

При интерпретации результатов дисперсионного анализа со случайными уровнями фактора обычно интересуются не проверкой гипотез относительно средних, а оценкой компонент дисперсии.

Пример 1.

Имеем наблюдения — оценки успеваемости студентов за выполнение лабораторных работ, которые проставляются с точностью до одного знака после запятой. Число лабораторных работ равно 5, число студентов — 8 (подгруппа). Требуется установить, влияет ли номер лабораторной работы на оценки студентов, т.е. одинаковы ли по сложности лабораторные работы.

Решение. Предположим, что плотность распределения оценок соответствует нормальному закону распределения. Исходные данные и расчеты приведены ниже (табл. 4.2).

Таблица 4.2

Исходные данные к примеру 1

Наблюдения (студенты)	Уровни фактора *A* (лабораторные работы)				
	1-я	2-я	3-я	4-я	5-я
1	4,0	3,5	4,3	4,0	5,0
2	4,5	4,6	5,0	4,7	4,0
3	3,0	3,5	4,0	3,6	3,0
4	4,3	4,0	4,4	4,5	4,0
5	5,0	4,5	4,0	4,5	5,0
6	3,5	3,3	3,0	4,0	3,5
7	2,0	3,0	2,5	3,5	3,0
8	4,7	4,0	4,0	4,5	4,2
$\bar{x}_i$	3,88	3,80	3,90	4,16	3,96
s_i^2	0,99	0,33	0,63	0,21	0.61

Выдвинем нулевую гипотезу о том, что фактор *A* (номер лабораторной работы) не влияет на оценки, т.е. $H_0 : m_1 = m_2 = ... = m_k = m$. Сделаем необходимые вычисления:

- среднее значение всей выборки определим по формуле (4):

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{n}\bar{x}_i = 3,94;$$

- дисперсию, характеризующую фактор случайности, по формуле (7):

$$s_{сл}^2 - \frac{1}{k}\sum_{i=1}^{n} s_i^2 = 0,55;$$

- дисперсию фактора *A* — по формуле (11):

$$s_A^2 = \frac{n}{n-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 = 0,15;$$

Так как расчетное значение критерия

$$F = \frac{s_A^2}{s_{сл}^2} = 0,27 < F_{0,05}(4,35) = 2,65,$$

то гипотеза H_0 не отвергается, следовательно, номер лабораторной работы не влияет на оценки.►

4.2 Однофакторный непараметрический анализ

Анализ на основе критерия Краскела-Уоллеса (произвольные альтернативы)

Этот метод используется, когда невозможно сказать что-либо определенное об альтернативах H_0, так как он свободен от распределения. Заменим наблюдения x_{ji} их рангами r_{ji}, упорядочивая всю совокупность $\{x_{ji}\}$ в порядке возрастания. Затем для каждой обработки i (уровня фактора, столбца таблицы) надо вычислить суммарный и средний ранги:

$$R_i = \sum_{j=1}^{n_i} r_{ji} \text{ и } \overline{R}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ji} . \quad (13)$$

Если между столбцами нет систематических различий, то средние ранги $\overline{R}_i$, не должны значительно отличаться от среднего, рассчитанного по всей совокупности $\{r_{ji}\}$. Значение последнего $\hat{R} = \frac{N+1}{2}$. Здесь N — общее число наблюдений.

$$N = \sum_{i=1}^{k} n_i .$$

Вычислим величины дисперсий $(\overline{R}_i - \hat{R})^2$ для каждого уровня фактора $\left(\overline{R}_i - \frac{M+1}{2}\right),...,\left(\overline{R}_k - \frac{M+1}{2}\right)$.

Эти значения при H_0 в совокупности должны быть небольшими. Составляя общую характеристику, разумно учесть различия в числе наблюдений для разных обработок (уровней факторов) и взять в качестве меры отступления от чистой случайности величину

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{n} n_i \left(\overline{R}_i - \frac{N+1}{2}\right)^2 . \quad (14)$$

Эта величина называется статистикой *Краскела-Уоллеса*

Множитель $\frac{12}{N(N+1)}$ присутствует в качестве нормировочного для обеспечения сходимости распределения статистики H и χ^2 с числом степеней свободы $v = k-1$. Гипотеза H_0 отвергается при уровне значимости q; если $H_{набл} > \chi_q^2(v)$, то фактор считается значимым.

Если среди x_{ji} есть совпадающие значения, то при ранжировании и переходе к r_{ji} надо использовать средние ранги (например, если 2 значения (5 и 5) занимают ранги 11, 12, то средний ранг (11,5) надо присвоить им обоим). Если совпадений много, рекомендуется использовать модифицированную форму статистики H':

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{m} T_j / (N^3 - N) \right)}, \qquad (15)$$

где m — число групп совпадающих наблюдений; $T_j = t_j^3 - t_j$ (t_j — число совпадающих наблюдений в группе j).

♦ Пример 2.

Для выяснения влияния денежного стимулирования на производительность труда шести однородным группам из 5 человек были предложены задания одинаковой трудности. Задания предлагались каждому испытуемому независимо от остальных. Группы отличались величиной денежного вознаграждения за решаемую задачу. Данные (число решаемых задач) приведены в таблице 3.

Таблица 4.3

Исходные данные к примеру 2 (первая цифра – это наблюдение x_{ji}, вторая цифра (маленькая)– это порядковый номер).

Наблю	Уровни					
	Гр. 1	Гр. 2	Гр. 3	Гр. 4	Гр. 5	Гр. 6
1	10-5	8-2	12-9	12-10	24-27	19-24
2	11-7	10-6	17-20	15-14	16-19	18-22
3	9-3	16-15	14-13	16-17	22-26	27-30
4	7-1	13-12	9-4	16-18	18-21	25-29
5	13-11	12-8	16-16	19-23	20-25	24-28

Решение. Проверим гипотезу H_0 об отсутствии эффектов обработки (отсутствии влияния денежного вознаграждения). Поскольку закон распределения x_{ji} неизвестен, воспользуемся ранговыми критериями.

В связи с наличием совпадений необходимо воспользоваться средними рангами. Так, $x_{ji} = 10$ встречается дважды и при упорядочении x_{ji} занимает 5-е и 6-е места. Поэтому средний ранг $x_{ji} = 10$ равен 5,5. В результате ранжирования получаем таблицу (табл. 4.4).

Таблица 4.4

Ранжированные данные

Наблюдения	Уровни					
	Гр. 1	Гр. 2	Гр.3	Гр. 4	Гр. 5	Гр. 6
1	5,5	2	9	9	27,5	23,5
2	7	5,5	20	14	17	21,5
3	3,5	17	13	17	26	30

4	1	11,5	3,5	17	21,5	29
5	11,5	9	17	23,5	25	27,5
R_i	28,5	45	62,5	80,5	117	131,5
$\overline{R}_i$	5,7	9	12,5	16,1	23,4	26,3

В двух нижних строках приведены суммы рангов R_i и средние ранги $\overline{R}_i$ по столбцам. Вычислим статистику Краскела - Уоллеса при общем числе наблюдений $N = 30$, числе опытов при каждом значении фактора $n_j = 5,\ j = 1,2,...,6$. Подставляя эти значения, получим

$$H = \frac{12}{30 \cdot (30+1)} \sum_{i=1}^{n} 5 \cdot \left(\overline{R}_i - \frac{31}{2} \right)^2 = 21,077$$

Величина H имеет распределение χ^2. По таблицам распределения χ^2 для степеней свободы $v = k - 1$ находим, что минимальный уровень значимости q чуть больше 0,001, что слишком мало, чтобы принять гипотезу H_0.

Для учета влияния совпадений в $\{x_{ji}\}$ можно воспользоваться статистикой H'.

В нашем случае 8 групп совпадающих наблюдений: 9,9; 10,10; 12,12,12; 13,13; 16,16,16,16,16; 18,18; 19,19; 24,24.

$T_1 = 2^3 - 2 = 6;\ T_2 = 2^3 - 2 = 6;\ T_3 = 3^3 - 3 = 24;$
$T_4 = 6;\ T_5 = 5^3 - 5 = 120;\ T_6 = 6;\ T_7 = 6;\ T_8 = 6.$

$$H' = \frac{21,077}{1 - \left(\sum_{j=1}^{m} T_j / (30^3 - 30) \right)} = 21,219.$$

Так как скорректированное значение H' мало отличается от H, мы можем отвергнуть гипотезу H_0 при минимальном уровне значимости $q = 0,001$.►

Анализ на основе критерия Джонкхиера (альтернативы с упорядочением)

Нередко исследователю заранее известно, что имеющиеся группы результатов упорядочены по возрастанию влияния фактора. Пусть первый столбец таблицы $\{x_{ji}\}$ соответствует наименьшему уровню, а последний — наибольшему. В таких случаях критерий Джонкхиера более чувствителен (более мощный) в сравнении с упорядоченным влиянием фактора.

Рассмотрим сначала случай, когда сравниваются только 2 способа обработки (2 уровня фактора). Фактически речь идет тогда об однородности двух выборок. Для проверки этой гипотезы рассмотрим статистику Манна-Уитни.

Пусть имеем 2 выборки: $x_1, x_2, ..., x_m$ и $y_1, y_2, ..., y_n$. Положим

$$\varphi(x_i, y_j) = \begin{cases} 0, & если\, x_i > y_j; \\ \frac{1}{2}, & если\, x_i = y_j; \\ 1, & если\, x_i < y_j. \end{cases} \quad (16)$$

Статистика Манна-Уитни:

$$U = \sum_{i=1}^{m} \sum_{i=1}^{n} \varphi(x_i, y_j). \quad (17)$$

Обратившись теперь к общему случаю, когда сравниваются k способов обработки (k уровней), поступим следующим способом. Для каждой пары уровней u и v, где $1 \le u < v \le k$, составим по выборкам с номерами u и v статистики Манна-Уитни:

$$U(u,v) = \sum_{i=1}^{m} \sum_{i=1}^{n} \varphi(x_i, y_j). \quad (18)$$

Получим $U(1,2), U(1,3),...,U(1,k), U(2,3),...,U(2,k),...,U(k-1,k)$.

Определим статистику Джонкхиера I как $I = \sum U(u,v)$ для $1 \le u < v \le k$. Свидетельством против H_0 (в пользу альтернативы) служат большие значения статистики I, полученные в эксперименте. Для больших объемов выборок в отношении статистики I действует нормальное распределение $I \sim N(MI, DI)$, с математическим ожиданием и дисперсией

$$MI = \frac{1}{4}\left(N^2 - \sum_{i=1}^{k} n_j \right),$$

$$DI = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3) \right], \quad (19)$$

где n_j – количество наблюдений в каждом уровне;

N — общий объем наблюдений;

Свидетельством против H_0 (в пользу альтернативы) служат большие значения статистики $I^* = \frac{I - MI}{\sqrt{DI}}$, полученные в эксперименте, в сравнении с P-процентными точками нормального распределения $\Phi(I^*) = P$ (табличные значения нормированной функции Лапласа). Тогда $q - 1 - P$ — уровень значимости, уровень принятия гипотезы H_0.

Пример 3

Для условия, изложенного в примере 2, решим задачу, используя критерий Джонкхиера.

Решение. Заметим, что в данном примере предлагается монотонное изменение стимулирования для оценки влияния на производительность. Поэтому оправдано применение критерия Джонкхиера. Выберем в качестве альтернативы к H_0

утверждение, что чем выше уровень стимулирования, тем выше производительность. Для вычисления статистики I найдем значения статистик Манна-Уитни [для всех комбинаций u и v, где $1 \le u < v \le k$.

Результаты расчета:

$$U(1,2) = \varphi(x_4, y_1) + \big(\varphi(x_1, y_2) + \varphi(x_3, y_2) + \varphi(x_4, y_2)\big) +$$
$$+ \big(\varphi(x_1, y_3) + \varphi(x_2, y_3) + \varphi(x_3, y_3) + \varphi(x_4, y_3) + \varphi(x_5, y_3)\big) +$$
$$+ \big(\varphi(x_1, y_4) + \varphi(x_2, y_4) + \varphi(x_3, y_4) + \varphi(x_4, y_4) + \varphi(x_5, y_4)\big) +$$
$$+ \big(\varphi(x_1, y_5) + \varphi(x_2, y_5) + \varphi(x_3, y_5) + \varphi(x_4, y_5)\big) = 17.$$

$$U(1,3) = \big(\varphi(x_1, y_1) + \varphi(x_2, y_1) + \varphi(x_3, y_1) + \varphi(x_4, y_1)\big) +$$
$$+ \big(\varphi(x_1, y_2) + \varphi(x_2, y_2) + \varphi(x_3, y_2) + \varphi(x_4, y_2) + \varphi(x_5, y_2)\big) +$$
$$+ \big(\varphi(x_1, y_3) + \varphi(x_2, y_3) + \varphi(x_3, y_3) + \varphi(x_4, y_3) + \varphi(x_5, y_3)\big) +$$
$$+ \big(\varphi(x_3, y_4) + \varphi(x_4, y_4)\big) +$$
$$+ \big(\varphi(x_1, y_5) + \varphi(x_2, y_5) + \varphi(x_3, y_5) + \varphi(x_4, y_5) + \varphi(x_5, y_5)\big) = 20{,}5.$$

$$U(1,4) = \big(\varphi(x_1, y_1) + \varphi(x_2, y_1) + \varphi(x_3, y_1) + \varphi(x_4, y_1)\big) +$$
$$+ \big(\varphi(x_1, y_2) + \varphi(x_2, y_2) + \varphi(x_3, y_2) + \varphi(x_4, y_2) + \varphi(x_5, y_2)\big) +$$
$$+ \big(\varphi(x_1, y_3) + \varphi(x_2, y_3) + \varphi(x_3, y_3) + \varphi(x_4, y_3) + \varphi(x_5, y_3)\big) +$$
$$+ \big(\varphi(x_1, y_4) + \varphi(x_2, y_4) + \varphi(x_3, y_4) + \varphi(x_4, y_4) + \varphi(x_5, y_4)\big) +$$
$$+ \big(\varphi(x_1, y_5) + \varphi(x_2, y_5) + \varphi(x_3, y_5) + \varphi(x_4, y_5) + \varphi(x_5, y_5)\big) = 24.$$

$$U(1,5) = \big(\varphi(x_1, y_1) + \varphi(x_2, y_1) + \varphi(x_3, y_1) + \varphi(x_4, y_1) + \varphi(x_5, y_1)\big) +$$
$$+ \big(\varphi(x_1, y_2) + \varphi(x_2, y_2) + \varphi(x_3, y_2) + \varphi(x_4, y_2) + \varphi(x_5, y_2)\big) +$$
$$+ \big(\varphi(x_1, y_3) + \varphi(x_2, y_3) + \varphi(x_3, y_3) + \varphi(x_4, y_3) + \varphi(x_5, y_3)\big) +$$
$$+ \big(\varphi(x_1, y_4) + \varphi(x_2, y_4) + \varphi(x_3, y_4) + \varphi(x_4, y_4) + \varphi(x_5, y_4)\big) +$$
$$+ \big(\varphi(x_1, y_5) + \varphi(x_2, y_5) + \varphi(x_3, y_5) + \varphi(x_4, y_5) + \varphi(x_5, y_5)\big) = 25.$$

$$U(1,6) = \big(\varphi(x_1, y_1) + \varphi(x_2, y_1) + \varphi(x_3, y_1) + \varphi(x_4, y_1) + \varphi(x_5, y_1)\big) +$$
$$+ \big(\varphi(x_1, y_2) + \varphi(x_2, y_2) + \varphi(x_3, y_2) + \varphi(x_4, y_2) + \varphi(x_5, y_2)\big) +$$
$$+ \big(\varphi(x_1, y_3) + \varphi(x_2, y_3) + \varphi(x_3, y_3) + \varphi(x_4, y_3) + \varphi(x_5, y_3)\big) +$$
$$+ \big(\varphi(x_1, y_4) + \varphi(x_2, y_4) + \varphi(x_3, y_4) + \varphi(x_4, y_4) + \varphi(x_5, y_4)\big) +$$
$$+ \big(\varphi(x_1, y_5) + \varphi(x_2, y_5) + \varphi(x_3, y_5) + \varphi(x_4, y_5) + \varphi(x_5, y_5)\big) = 25.$$

$U(2,3) = 17;$ $U(3,4) = 16{,}5;$ $U(4,5) = 22;$ $U(5,6) = 18;$

$U(2,4) = 20{,}5;$ $U(3,5) = 23{,}5;$ $U(4,6) = 23{,}5;$

$U(2,5) = 24{,}5;$ $U(3,6) = 25;$

$U(2,6) = 25$

;

Отсюда $I = \sum U(u,v) = 327$.

Для нахождения минимального уровня значимости воспользуемся нормальной величиной I^*:

$$MI = \frac{1}{4}\left(30^2 - \sum_{i=1}^{k} 5 \right) = 217{,}5.$$

$$DI = \frac{1}{72}\left[30^2 (2 \cdot 30 + 3) - \sum_{j=1}^{k} 5^2 \cdot (2 \cdot 5 + 3) \right] = 760{,}4.$$

$$I^* = \frac{327 - 217.5}{\sqrt{760{,}4}} = 4{,}3$$

$$\Phi(4{,}3) = 0{,}99998;\ \mathrm{q} = 1\text{-}0{,}99998 = 2 \cdot 10^{-5}$$

Заметим, что мы получили более значительный результат ($\mathrm{q} = 2 \cdot 10^{-5}$) по сравнению с критерием Краскела-У ол-леса ($q = 1 \cdot 10^{-3}$), так как минимальный уровень значимости понизился на 2 порядка.►

4.3. Двухфакторный анализ

Иногда в однофакторной модели влияние интересующего нас фактора не проявляется, хотя такое влияние должно быть. Причиной этого может быть большой внутригрупповой разброс, на фоне которого действие фактора остаётся незаметным. Очень часто этот разброс вызван не только случайными причинами, но и действиями еще одного фактора. Если мы в состоянии указать такой фактор, то можно попытаться включить его в модель, чтобы уменьшить статистическую неоднородность наблюдений. Конечно, не всегда удается поправить дело введением *мешающего* фактора и переходом к двухфакторной схеме. Иногда приходится рассматривать трехфакторные и более сложные модели. Замысел во всех этих случаях остается прежним.

Назовём фактор A (рис. 1) главным, B — мешающим. Пусть фактор A принимает k значений, а мешающий — n значений. Фактор B разбивает все группы наблюдений (столбцы таблицы $\{x_{ji}\}$) на блоки.

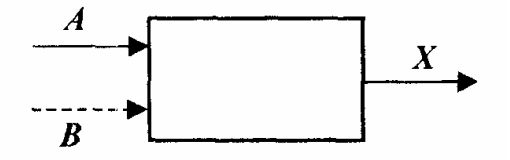


Рис. 6.1. Двухфакторная модель

Каждый блок соответствует определенному уровню фактора B. В частном случае таблица содержит $n \times k$ наблюдений (по одному в клетке). Отличие этой таблицы от однофакторной в том, что наблюдения в любом столбце не являются однородными, если влияние мешающего фактора значимо (табл. 6.3).

Таблица 4.3
Форма представления экспериментальных данных двухфакторной модели

Блоки фактора B	Уровни основного фактора A			
	A_1	A_2	A_i	A_k
B_1	x_{11}	x_{12}	x_{1i}	x_{1k}
B_2	x_{21}	x_{22}	x_{2i}	x_{2k}
….	….	….	….	….
B_j	x_{j1}	x_{j2}	x_{ji}	x_{jk}
….	….	…..	…..	…..
B_n	x_{n1}	x_{n2}	x_{ni}	x_{nk}

Для описания двухфакторного эксперимента обычно применяют аддитивную модель. Она предполагает, что значения отклика x_{ji} являются суммой вкладов соответствующих уровней факторов A и B и независимых случайных факторов: $x_{ji} = a_i + b_j + e_{ji}$. В этой модели величины вкладов A и B не могут быть восстановлены однозначно. Действительно, при одновременном увеличении всех a_i на одну и ту же константу и при уменьшении всех b_j на ту же константу значения x_{ji} не изменяются.

Для однозначности вкладов удобно перейти к представлению в виде:

$$x_{ji} = \eta + \alpha_i + \beta_j, \ при \ \sum_i \alpha_i = 0, \ \sum_j \beta_j = 0.$$

Параметр η интерпретируется как среднее всех x_{ji} (т.е. $\bar{x}$), а α_i и β_j — отклонения от η в результате действия факторов A и B.

4.3.1. Двухфакторный параметрический дисперсионный анализ

Если есть основания предполагать, что случайные величины e_{ji} имеют нормальное распределение с нулевым средним и одинаковой при всех i и j дисперсией σ^2, можно использовать метод, аналогичный однофакторному дисперсионному анализу. Предположим, что влияние фактора A отсутствует. Сформулируем гипотезу $H_0: \alpha_1 = \alpha_2 = ... = \alpha_k$. Проверим гипотезу H_0, также основываясь на сравнении двух независимых оценок: $\sigma_0^2 = \alpha_A^2 + \sigma_{сл}^2$.

Для решения задачи находятся средние дисперсии:

$$\bar{x} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}; \ \bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji}; \ \bar{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{ji};$$

$$s_{сл}^2 = \frac{1}{(n-1)(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ji} - \bar{x}_i - \bar{x}_j + \bar{x} \right)^2 ; \qquad (20)$$

$$s_A^2 = \frac{n}{(k-1)} \sum_{i=1}^{k} \left(\bar{x}_i - \bar{x} \right)^2$$

Если выполняется неравенство

$\left[F_{набл} = \frac{s_A^2}{s_{сл}^2} \right] > F_q \left(v_1 = k-1,\, v_2 = (n-1)(k-1) \right)$, то H_0 отвергается и влияние фактора A считается существенным.

Пример 4

Имеем наблюдения — среднюю по группам успеваемость студентов за выполнение лабораторных работ. Число лабораторных работ равно 5, число групп — 3 (см. табл.4.5).

Таблица 4.5

Исходные данные к примеру 4

Уровни фактора B (группы)	Уровни фактора A (лабораторные работы)					
	1-й	2-й	3-й	4-й	5-й	$\bar{x}_j$
1	4,0	3,9	4,3	4,0	4,4	4,12
2	4,2	4,1	4,0	4,3	4,2	4,16
3	3,3	3,5	3,8	3,6	3,4	3,52
$\bar{x}_i$	3,83	3,83	4,03	3,97	4,00	

Требуется установить, влияет ли номер лабораторной работы (фактор *A)* на оценки студентов, т.е. одинаковы ли по сложности лабораторные.

Решение. Так как наблюдений немного, то предположим, что плотность распределения оценок соответствует нормальному закону. Исходные данные и расчеты приведены ниже (табл. 4.5).

Выдвинем нулевую гипотезу о том, что факторы A не влияют на оценки, т.е. $H_0 : m_1 = m_2 = \ldots = m_k = m$. Сделаем необходимые вычисления:

$$\bar{x} = \frac{1}{3 \cdot 5} \sum_{i=1}^{5} \sum_{j=1}^{3} x_{ji} = 3{,}93 ;$$

$$s_{сл}^2 = \frac{1}{2 \cdot 4} \sum_{i=1}^{5} \sum_{j=1}^{3} \left(x_{ji} - \bar{x}_i - \bar{x}_j + \bar{x} \right)^2 = 0{,}035 ;\ s_A^2 = \frac{3}{4} \sum_{i=1}^{k} \left(\bar{x}_i - 3{,}93 \right)^2 = 0{,}912 ;$$

$$F_A = \frac{0{,}027}{0{,}035} = 0{,}758 < F_{0{,}05}(4,8) = 3{,}84 .$$

Следовательно, H_0 для фактора не отвергается. Таким образом, влияние фактора A (номера лабораторной работы) считается незначимым.►

4.3.2. Двухфакторный непараметрический анализ

Двухфакторный непараметрический анализ используется при проверке гипотезы H_0, если о распределении случайной величины e_{ji} известно только то, что она непрерывна и независима. Рассмотрим решение задачи с использованием критерия Фридмана, который не предъявляет требований к упорядочению уровней факторов.

Двухфакторный непараметрический анализ по критерию Фридмана (произвольные альтернативы)

Критерий основан на идее перехода от значений x_{ji} к их рангам r_{ji}. В отличие от однофакторного анализа ранжирование происходит не по всей таблице $\{x_{ji}\}$, а по блокам, т.е. рассматривается каждая отдельная строка таблицы. При фиксированном j осуществляется ранжирование величин x_{ji} при $i = 1,2,...,k$. Тем самым устраняется влияние мешающего фактора B, значение которого для каждой строки постоянно. Обозначим полученные ранги величин x_{ji} через r_{ji}. Ясно, что r_{ji}. изменяются от 1 до k, а каждая строка представляет перестановку чисел $1,2,...,k$ (при совпадении x_{ji} надо использовать средние ранги). При гипотезе $H_0: \alpha_1 = \alpha_2 = ... = \alpha_k$ все $k!$ перестановок равновероятны. Введем величину $\bar{r}_i = \frac{1}{n}\sum_{j=1}^{n} r_{ji}$ — среднее значение ранга по столбцу. При H_0 значение для каждого столбца не должно сильно отличаться от $\bar{r} = \frac{k+1}{2}$ — среднего ранга всех элементов таблицы.

Статистика *Фридмана* имеет следующий вид

$$S = \frac{12n}{k(k+1)}\sum_{i=1}^{k}\left(\bar{r}_i - \bar{r}\right)^2 . \qquad (21)$$

Гипотеза H_0 отвергается в пользу альтернативы о наличии эффектов в обработке, если $S \geq \chi_q^2(v = k-1)$. Для небольших значений n, k величина критерия Фридмана $S(q,n,k)$ может быть найдена по специальным статистическим таблицам.

♦ Пример 5

Исследованы зависимости дрожания мышц рук от тяжести поднятого груза (тремор). Каждое табличное значение — среднее пяти экспериментальных измерений частоты тремора у испытуемого. Каждая обработка (уровень фактора A) соответствует весу груза. Исходные данные приведены в таблице 4.6.

Таблица 4.6

Исходные данные к примеру 5

Испытуемый	Уровни фактора A (вес груза в кг)

	1-й (0)	2-й (0,5)	3-й (1,0)	4-й (2)	5-й
1	3,01	2,85	2,62	2,63	2,58
2	3,47	3,43	3,15	2,83	2,7
3	3,35	3,14	3,02	2,71	2,78
4	3,1	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	2,5	2,43

Решение. Заменим числовые значения рангами (табл. 4.7).

Таблица 4.7
Ранжированные данные

Испытуемый	Уровни фактора *A* (вес груза в кг)				
	1-й (0)	2-й (0,5)	3-й (1,0)	4-й (2)	5-й (3,0)
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_i	30	24	17	12	7
$\overline{R}_i$	5	40	2,833	2	1,1667

Статистика Фридмана, вычисленная по формуле (21), равна

$$S = \frac{12 \cdot 6}{5 \cdot 6} \sum_{i=1}^{5} \left(\overline{R}_i - \overline{R} \right)^2 = 22,533 .$$

Критическое значение $\chi^2_{q=0,05}(v=4) = 9,488$, следовательно, $S > \chi^2_{0,05}(4)$ и H_0 отвергается. Согласно таблицам χ^2 минимальный уровень значимости, при котором гипотеза H_0 может быть принята, $q = 0,0001$.

Двухфакторный непараметрический анализ по критерию Пейджа (альтернативы с упорядочением)

Если уровни факторов в таблице $\{x_{ji}\}$ упорядочены, то для проверки гипотезы $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k$ против альтернативы $H_1 : \alpha_1 < \alpha_2 < ... < \alpha_k$ используется статистика *Пейджа.*

Введем величину $R_i = \sum_{j=1}^{n} r_{ji}$.

Статистика

$$L=\sum_{i=1}^{k} i\cdot R_i = 1\cdot R_1 + 2\cdot R_2 + 3\cdot R_3 + ... + k\cdot R_k . \qquad (22)$$

H_0 отклоняется в пользу H_1, если $L_{набл} > L_q(k,n)$, где значение $L_{набл} \geq L_\alpha(k,n)$ найдено по специальным статистическим таблицам.

Критические значения $L_\alpha(k,n)$ приведены в табл. 9 (см. статистические таблицы). Для $n>10$ справедлива аппроксимация статистики Пеиджа

$$L^* = \frac{L-M(L)}{\sqrt{D(L)}}, \text{ где } M(L)=\frac{nk(k+1)^2}{4},\ D(L)=\frac{n(k^3-k)^2}{144(k-1)}.$$

При $L^* > u_\alpha$ нулевая гипотеза отклоняется (u_α – квантиль стандартного нормального распределения.

♦ Пример 6

Для условия, изложенного в примере 5, решим задачу, используя критерий Пейджа.

Решение. Статистику Пейджа вычислим по формуле (22) с учетом упорядочения:

$$L = 7 + 2\cdot 12 + 3\cdot 17 + 4\cdot 24 + 5\cdot 30 = 328 .$$

Критическое значение $L_{0,01}(5,6)=299$. Так как $L=328>L_{0,01}(5,6)=299$, то H_0 отвергается. Минимальный уровень для принятия H_0 равен $q=0,000001$, что на два порядка меньше, чем по критерию Фридмана.

Используем теперь аппроксимацию

$$M(L)=\frac{6\cdot 5\cdot (5+1)^2}{4}=45;\ D(L)=\frac{6\cdot (5^3-5)^2}{144\cdot (5-1)}=150;\ L^*=\frac{328-45}{\sqrt{150}}=23,1.$$

Так как $u_{0,99}=2,326$ и $L^*=23,1>u_{0,99}=2,326$, то и в этом случае гипотеза H_0 отклоняется.►

Тема 5. Корреляционный анализ

Цель занятия:

Практическое проведение корреляционного анализа выборочных данных

Содержание занятия:

1) Вычисление параметрических коэффициентов корреляции

2) Вычисление непараметрических коэффициентов корреляции

5.1. Вычисление параметрических коэффициентов корреляции

Выборочный коэффициент корреляции между двумя случайными величинами x и y был впервые введен Пирсоном, поэтому его часто *называют коэффициентом корреляции Пирсона.* В теории разработаны и на практике применяются различные модификации формулы расчета данного коэффициента. Приведем некоторые из них:

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

Используя преобразование, получают

$$r = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{n\sigma_x\sigma_y} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}. \quad (1)$$

или через дисперсии

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{xy}^2}{2\sigma_x^2\sigma_y^2}.$$

При большом объеме выборки выборочный коэффициент корреляции будет приближаться к корреляционному моменту генеральной совокупности ρ, который определяется как

$$\rho = \frac{M\left[(x-m_x)(y-m_y)\right]}{\sigma_x\sigma_y}. \quad (2)$$

Если $r=0$, то величины x и y независимы, а при $r=1$ зависимость между x и y является функциональной. Коэффициент корреляции тесно связан с коэффициентом регрессии:

$$r = b\frac{\sigma_x}{\sigma_y}, \quad (3)$$

где b — коэффициент регрессии в уравнении вида $y_i = a + bx_i$;

σ_x — среднеквадратичное отклонение x;

σ_y — среднеквадратичное отклонение y. Значимость коэффициента корреляции проверяется на основе критерия Стьюдента. При проверке этой гипотезы вычисляется t-статистика:

$$t_{рас} = \sqrt{\frac{r^2(n-2)}{1-r^2}} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}\,. \tag{4}$$

Расчетное значение сравнивается с табличным значением $t_q(v = n-2)$. Если расчетное значение больше табличного, это свидетельствует о значимости коэффициента корреляции, а, следовательно, и о статистической существенности зависимости между x и y.

При большом числе наблюдений $(n > 100)$ используется следующая формула t-статистики:

$$t_{рас} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n}\,. \tag{5}$$

Множественный коэффициент корреляции рассчитывается при наличии линейной связи между результирующим признаком Y и несколькими факторными признаками, а также между парой факторных признаков. Множественный коэффициент корреляции вычисляется по формуле

$$R_{y/x_1,x_2} = \sqrt{\frac{\delta^2}{\sigma^2}} = \sqrt{1-\frac{\sigma^2_{ост}}{\sigma^2}}\,, \tag{6}$$

где δ^2 — дисперсия теоретических значений признака Y, рассчитанная по уравнению множественной регрессии;

σ^2 — общая дисперсия признака Y;

$\sigma^2_{ост}$ — остаточная дисперсия.

В случае оценки связи между результирующим признаком Y и двумя факторными признаками x_1 и x_2 множественный коэффициент корреляции можно определить по формуле:

$$R_{y/x_1,x_2} = \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1-r^2_{x_1x_2}}}\,, \tag{7}$$

где r_{ab} — парные коэффициенты корреляции между признаками.

В общем случае коэффициент множественной корреляции между результирующим признаком Y и m факторными признаками $x_1, x_1, ..., x_m$ определяется по формуле

$$R_{y/x_1,x_2,...,x_m} = \sqrt{1-\frac{|\rho|}{|\rho_1|}}\,, \tag{8}$$

где $|\rho|$ – определитель матрицы парной корреляции

$$\rho = \begin{pmatrix} 1 & \rho_{yx_1} & \rho_{yx_2} & \dots & \rho_{yx_m} \\ \rho_{x_1y} & 1 & \rho_{x_1x_2} & \dots & \rho_{x_1x_m} \\ \rho_{x_2y} & \rho_{x_2x_1} & 1 & \dots & \rho_{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{x_my} & \rho_{x_mx_2} & \rho_{x_mx_3} & \dots & 1 \end{pmatrix}; \quad (9)$$

$|\rho_1|$ – алгебраическое дополнение элемента ρ_{11}.

Множественный коэффициент корреляции изменяется в пределах от 0 до 1. Приближение R к единице свидетельствует о сильной зависимости между признаками. При небольшом числе наблюдений величина множественного коэффициента корреляции, как правило, завышается. В этом случае множественный коэффициент корреляции корректируется:

$$\widetilde{R}_{y/x_1,x_2,\dots,x_m} = \sqrt{1-(1-R^2)\frac{n-1}{n-k-1}}, \quad (8)$$

где $\widetilde{R}$ — скорректированное значение;
n — число наблюдений;
k — число факторных признаков.

Корректировка R не производится при условии, если $\frac{n-k}{k} \geq 20$.

Проверка значимости множественного коэффициента корреляции осуществляется по критерию Фишера:

$$F_{pac} = \frac{\frac{1}{2}R^2_{y/x_1,x_2}}{\frac{1}{3}\left(1-R^2_{y/x_1,x_2}\right)}. \quad (9)$$

Множественный коэффициент корреляции считается значимым, если $F_{pac} > F_q(v_1 = 2,\ v_2 = n-3)$.

На основе приведенных выше формул (1)-(9) произведем вычисление коэффициентов корреляции и проверим их значимость.

♦ Пример 1

Расчет коэффициента корреляции Пирсона.

На основе выборочных данных о деловой активности однотипных коммерческих структур необходимо оценить тесноту связи между прибылью (млн. руб.) y и затратами на 1руб. производства продукции x. Исходные данные и результаты расчета приведены в таблице 1.

Таблица 1

Исходные данные к примеру 1

Наблюдения	y	x	xy	y^2	x^2
1	221	96	21216	48841	9216
2	1070	77	82390	1144900	5926
3	1001	77	77077	1002000	5929
4	606	89	53934	367236	7921
5	779	82	63878	606841	6724
6	789	81	63909	622520	6561
Сумма	4466	502	362404	3792338	42280
Средняя	744,33	83,67	60400,67	632056,33	7046,67

Решение. Вычислим дисперсии

$$\sigma_x^2 = (\overline{x^2}) - (\overline{x})^2 = 7046,67 - 83,67^2 = 46;$$

$$\sigma_y^2 = (\overline{y^2}) - (\overline{y})^2 = 632056 - 744,33^2 = 78029.$$

Рассчитаем коэффициент корреляции:

$$r = \frac{xy - \overline{x}\,\overline{y}}{\sigma_x \sigma_y} = \frac{604400 - 744,33}{\sqrt{46 \cdot 78029}} = -0,99;$$

$$t_{рас} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2} = \frac{0,99}{\sqrt{1-(-0,99)^2}}\sqrt{6-2} = 14,04.$$

Критическое значение $t_{q=0,05}(v = 6-2 = 4) = 2,776$. Так как $t_{рас} = 14,04 > t_{кр} = 2,776$, то коэффициент корреляции считается значимым.►

Пример 2

Расчет множественного коэффициента корреляции.

По выборочным данным о деловой активности коммерческих структур необходимо оценить тесноту связи между прибылью (млн. руб.) y, затратами на 1 руб. производства продукции x_1 и стоимостью основных фондов (млн. руб.) x_2. Исходные данные приведены в таблице 2.

Таблица 2

Исходные данные к примеру 2

Наблюдения	Прибыль y	Затраты на 1 руб. продукции x_1	Стоимость основ. фондов x_2	x_1^2	x_1x_2	yx_1	x_2^2	yx_2	y^2
1	1070	77	5,9	5929	454,3	82390	34,81	6313,0	$1,1449 \cdot 10^6$
2	1001	77	5,9	5929	454,3	*11011*	34,81	5905,9	$1,0020 \cdot 10^6$
3	789	81	4,9	6561	396,9	63909	24,01	3866,1	$6,2252 \cdot 10^5$
4	779	82	4,3	6724	352,6	63878	18,49	3349,7	$6,0684 \cdot 10^5$
5	606	89	3,9	7921	347,1	53934	15,21	2363,4	$3,6723 \cdot 10^5$

6	221	96	4,3	9216	412,8	21216	18,49	950,3	$4{,}8841\cdot 10^4$
Сумма	4466	502	29,2	42280	2418,0	362404	145,82	22748,4	$3.7923\cdot 10^6$
Средняя	744,33	83,67	4,867	7046,67	403	60400,7	24,303	3791,4	632056

Решение. Рассчитаем коэффициенты корреляции:

$$r_{y,x_1}=\frac{\overline{x_1y}-\overline{x_1}\,\overline{y}}{\sqrt{\overline{x_1^2}-\overline{x_1}^2}\cdot\sqrt{\overline{y^2}-\overline{y}^2}}=$$

$$=\frac{60400-83{,}67\cdot 744{,}33}{\sqrt{7046{,}6-(83{,}67)^2}\cdot\sqrt{632056-(744{,}33)^2}}=-0{,}992;$$

$$r_{y,x_2}=\frac{\overline{x_2y}-\overline{x_2}\,\overline{y}}{\sqrt{\overline{x_2^2}-\overline{x_2}^2}\cdot\sqrt{\overline{y^2}-\overline{y}^2}}=$$

$$=\frac{3791{,}4-4{,}867\cdot 744{,}33}{\sqrt{24{,}303-(4{,}867)^2}\cdot\sqrt{632056-(744{,}33)^2}}=0{,}770;$$

$$r_{x_1,x_2}=\frac{\overline{x_1x_2}-\overline{x_1}\,\overline{x_2}}{\sqrt{\overline{x_1^2}-\overline{x_1}^2}\cdot\sqrt{\overline{x_2^2}-\overline{x_2}^2}}=$$

$$=\frac{403-83{,}67\cdot 4{,}867}{\sqrt{7046{,}6-(83{,}67)^2}\cdot\sqrt{24{,}303-(4{,}867)^2}}=-0{,}794$$

Матрица линейных коэффициентов корреляции имеет вид

$$\begin{pmatrix} 1 & -0{,}992 & 0{,}770 \\ & 1 & -0{,}794 \\ & & 1 \end{pmatrix}.$$

Множественный коэффициент корреляции:

$$R_{y/x_1,x_2}=\sqrt{\frac{(-0{,}992)^2+(0{,}770)^2-2\cdot(-0{,}992)\cdot(0{,}770)\cdot(-0{,}794)}{1-(-0{,}794)^2}}=0{,}992\,.$$

Проверим значимость множественного коэффициента корреляции:

$$F_{pac}=\frac{\frac{1}{2}(0{,}992)^2}{\frac{1}{6-3}\left(1-(0{,}992)^2\right)}=92{,}63>F_{q=0{,}05}(2,3)=9{,}55\,,$$

следовательно, коэффициент корреляции значим.►

5.2 Вычисление непараметрических коэффициентов корреляции

Если признаки подчиняются различным (отличным от нормального) законам распределения, то можно рассчитать непараметрические коэффициенты корреляции. Эти коэффициенты могут быть использованы для определения тесноты связи как между количественными, так и между качественными признаками при условии, если их значения упорядочить или проранжировать по степени убывания или возрастания признака.

Среди непараметрических методов оценки тесноты связи наибольшее распространение получили следующие:

1. *Коэффициент ранговой корреляции Спирмана:*

$$\rho_{x/y} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad (10)$$

где d^2 – квадрат разности рангов;
n — число наблюдений (число пар рангов).

Коэффициент Спирмана принимает значения от -1 до 1. Значимость коэффициента проверяется на основе t-критерия Стьюдента. Расчетное значение t-критерия определяется по формуле:

$$t_{pac} = \rho_{x/y}\sqrt{\frac{n-2}{1-\rho_{x/y}^2}}.$$

Значение коэффициента корреляции считается статистически существенным, если $t_{pac} > t_{\alpha}(v = n-2)$.

2. *Коэффициент ранговой корреляции Кендалла*

$$\tau = \frac{2(P-Q)}{n(n-1)} = \frac{2S}{n(n-1)}, \qquad (11)$$

где n — число наблюдений;
S — разность между числом последовательностей и числом инверсий по второму признаку. Расчет коэффициента Кендалла производится в такой последовательности:

- значения X ранжируются в порядке возрастания или убывания;
- значения Y располагаются в порядке, соответствующем значениям X;
- для каждого ранга Y определяется число следующих за ним значений рангов, превышающих его по величине. Суммируя эти числа, определяем величину P (число последовательностей) — меру соответствия последовательностей рангов X и Y;
- для каждого ранга Y определяется число следующих за ним рангов, меньших его величины. Суммируя величины, получаем величину Q (число инверсий);
- определяется разность по всем членам ряда $S = P - Q$ и вычисляется τ.

Связь между признаками можно признать статистически значимой, если значение коэффициента корреляции $|\tau| > \tau_\alpha = u_\alpha \sqrt{\frac{2(2n+5)}{9n(n-1)}}$ (на практике часто полагают связь между признаками значимой, если $\tau > 0{,}5$).

3. *Коэффициент конкордации.* Для определения тесноты связи между несколькими ранжированными признаками применяют *множественный коэффициент ранговой корреляции — коэффициент конкордации,* который вычисляется по формуле:

$$W = \frac{12S_W}{m^2(n^3 - n)}, \tag{12}$$

где m —- количество факторов;

n — число наблюдений;

S_W — отклонение суммы квадратов рангов от среднего квадрата суммы рангов.

$$S_W = \left(\widetilde{R^2} - \frac{\widehat{R}^2}{n} \right), \tag{13}$$

где $\widetilde{R} = \sum_{j=1}^{10} R_j$, $\widetilde{R^2} = \sum_{j=1}^{10} R_j^2$.

Критическое значение коэффициента конкордации равно $W_\alpha = \frac{1}{m(n-1)} \chi_\alpha^2 (n-1)$. Если $W > W_\alpha$, то с вероятностью α корреляция между изучаемыми признаками признается значимой.

Если среди последовательностей рангов есть совпадения, то коэффициент конкордации следует вычислять по формуле

$$W = \frac{12S_W}{m^2(n^2 - 1) - m\sum_{j=1}^{m} T_j}, \tag{14}$$

где $T_j = t_j^3 - t_j$, t_j – количество совпавших рангов в j-й последовательности.

Совпавшим рангам присваиваются средние ранги.

На основе приведенных выше формул (10)—(12) произведем вычисление непараметрических коэффициентов корреляции и проверим их значимости. Для непараметрических методов необходимы дополнительные преобразования, которые покажем на примерах.

♦ Пример 3

Поданным группы предприятий, выставивших акции на чековые аукционы (табл. 3), определить с помощью коэффициентов Спирмана и Кендалла, существует ли связь между величиной уставного капитала и количеством выставленных акций.

Таблица 3

Исходные данные к примеру 3

Наблюдения	Уставной капитал, млн руб., x	Ранг x R_x	Число выставленных акций y	Ранг y R_y	Сравнение рангов $d_i = R_x - R_y$	d_i^2
1	2954	9	856	7	2	4
2	1605	1	930	9	-8	64
3	4102	1.0	1563	10	0	0
4	2350	6	682	5	1	1
5	2625	7	616	3	4	16
6	1795	4	495	2	2	4
7	2813	8	815	6	2	4
8	1751	3	858	8	-5	25
9	1700	2	467	1	1	1
10	2264	5	661	4	1	1

Решение. Проранжируем выборки x, y *и* вычислим квадрат разности рангов (см. табл. 3).

Вычислим коэффициент корреляции Спирмана (5.32):

$$\rho_{x/y} = 1 - \frac{6 \cdot 120}{10 \cdot (100 - 1)} = 0,3 .$$

Следовательно, связь слабая.

Для вычисления коэффициента корреляции Кендалла проведем дополнительные упорядочения в данных, расположив x в порядке возрастания, а y – в порядке, соответствующем x (табл. 4).

Таблица 4

Ранжированные данные

Наблюдения	Уставной капитал, млн руб., x	Число выставленных акций y	Ранжирование			
			x	R_x	y	R_y
1	2954	856	1605	1	930	9
2	1605	930	1700	2	467	1
3	4102	1563	1751	3	858	8
4	2350	682	1795	4	495	2
5	2625	616	2264	5	661	4
6	1795	495	2350	6	682	5
7	2813	815	2625	7	616	3
8	1751	858	2813	8	815	6
9	1700	467	2954	9	856	7
10	2264	661	4102	10	1563	10

Для определения величины P для каждого ранга Y определяется число следующих за ним значений рангов, превышающих его по величине. Суммируя эти числа, определяем число последовательностей P.

P = 1 + 8 + 1+6 + 4 + 3 + 3+2+1= 29.

Затем для каждого ранга Y определяется число следующих за ним рангов, меньших его величины. Суммируя эти величины, получаем величину Q (число инверсий);

Q = 8 + 0 + 6 + 0 +1 + 1 + 0 + 0 + 0 = 16.

Коэффициент корреляции Кендалла (11):

$$\tau = \frac{2 \cdot (29 - 16)}{10 \cdot (10 - 1)} = 0,29 .$$

Это свидетельствует также о практическом отсутствии связи между рассматриваемыми признаками (критическое значение коэффициента корреляции равно $\tau_{0,05} = u_{0,05}\sqrt{\frac{2 \cdot (2 \cdot 10 + 5)}{9 \cdot 10 \cdot (10 - 1)}} = 1,645 \cdot 0,248 = 0,408$).►

♦ Пример 5

Определить с помощью множественного коэффициента ранговой корреляции тесноту связи между величиной уставного капитала, количеством выставленных акций и числом работников, занятых на предприятии (табл. 5).

Таблица 5

Исходные данные к примеру 5

Наблюдения	Уставной капитал, млн руб., x	Число выставленных акций y	Число занятых на предприятии z	Ранг R_x	Ранг R_y	Ранг R_z	Сумма рангов	Квадрат суммы
1	2954	856	119	9	7	1	17	289
2	1605	930	125	1	9	2	12	144
3	4102	1563	132	10	10	3	23	529
4	2350	682	141	6	5	4	15	225
5	2625	6-16	150	7	3	5	15	225
6	1795	495	165	4	2	6	12	144
7	2813	815	178	8	6	7	21	441
8	1751	858	181	3	8	8	19	361
9	1700	467	201	2	1	9	12	144
10	2264	661	204	5	4	10	19	361
Сумм							165	2863

a								

Решение. Проранжируем значения всех переменных и вычислим сумму рангов (см. табл.6). Вычислим коэффициент ранговой корреляции:

$$S_W = 2863 - \frac{165^2}{10} = 140,5$$

$$W = \frac{12 \cdot 140,5}{9 \cdot (1000 - 10)} = 0,19 .$$

Полученное значение свидетельствует о слабой связи между рассматриваемыми признаками (критическое значение

$$W_{0,05} = \frac{1}{3 \cdot (10-1)} \chi^2_{0,05}(9) = \frac{16,9}{27} = 0,626).\blacktriangleright$$

Тема 6. Линейная регрессия

Цель работы:

Оценка уравнения линейной регрессии на основе выборочных данных

Содержание занятия:

1) Построение модели парной регрессии

2) Оценка погрешности регрессии

3) Пример построения уравнения регрессии

6.1. Построение модели регрессии

Рассмотрим линейную по коэффициентам модель регрессии:

$$y = f\left(x;\beta\right)+\varepsilon=\beta_0+\beta_1 f_1(x)+\beta_2 f_2(x)+\ldots+\beta_k f_k(x)+\varepsilon, \quad (1)$$

где ε - случайная величина с математическим ожиданием равным нулю и дисперсией σ^2; $f_j(x),\ j=1,...,k$ – некоторые заданные функции.

Полагая, $x_j = f_j(x), \quad j=\overline{1,k}$ перейдем к модели множественной линейной регрессии:

$$y = f\left(x;\beta\right)+\varepsilon=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k+\varepsilon. \quad (2)$$

Пусть для оценки неизвестных параметров $\beta_j,\ j=\overline{0,k}$ уравнения регрессии (2) взята выборка объемом n из значений величин $(Y, X_1, X_2, \ldots, X_k)$. Тогда

$$Y = X\beta+\varepsilon,$$

где $Y=\left(\ y_1, y_2 ..., y_n\right)^T$ - вектор значений переменной y;

$\beta=\left(\beta_0,\beta_1,\ldots,\beta_k\right)^T$ - вектор параметров модели;

$\varepsilon=\left(\varepsilon_1,\ldots,\varepsilon_n\right)^T$ – вектор ошибок, где $\varepsilon_i \in \mathrm{N}\left(0,\sigma^2\right)$ и независимы;

X - матрица исходных данных переменных X_j размерами $n\times(k+1)$. Первый столбец матрицы X содержит единицы (значения фиктивной переменной x_0), остальные столбцы значения переменных $x_1, x_2, ..., x_k$:

$$X=\begin{pmatrix} 1 & x_1^1 & \cdots & x_k^1 \\ 1 & x_1^2 & \cdots & x_k^2 \\ & & \cdots & \\ 1 & x_1^n & \cdots & x_k^n \end{pmatrix}.$$

Для нахождения оценки β^* вектора параметров $\beta=\left(\beta_0,\beta_1,\ldots,\beta_k\right)^T$ используем метод наименьших квадратов, согласно которому в качестве оценок

$\beta_0^*, \beta_1^*, \ldots, \beta_k^*$ берутся такие, которые минимизируют сумму квадратов Q отклонений значений y_i от $f(\vec{x}_i)$:

$$Q = \sum_{i=1}^{n} \left(y_i - f(\vec{x}_i) \right)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = \left(Y - X\beta \right)^T \left(Y - X\beta \right). \quad (3)$$

Оценка β^* метода наименьших квадратов имеет вид:

$$\beta^* = \left(X^T X \right)^{-1} X^T Y. \quad (4)$$

6.2. Оценка адекватности регрессии

Количественной мерой адекватности является отношение дисперсии S^2, определяемой рассеянием значений y_i вокруг линии регрессии, к дисперсии S_y^2 естественного рассеяния значений y_i вокруг своего среднего $\overline{y}$. Другими словами это можно сформулировать так: ошибки, обусловленные заменой истинной зависимости на выборочную регрессию, находятся на уровне естественного разброса, наблюдаемых случайных величин.

Оценим сначала величину дисперсии модели S^2:

$$S^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} e^2 = \frac{1}{n-k-1} e^T e,$$

где $\hat{y}_i = \beta_0^* + \beta_1^* x_i + \ldots \beta_k^* x_k$.

Оценка дисперсии разброса случайных чисел y_i вокруг своего среднего значения равна $S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2$.

Если $\frac{S^2}{S_y^2} > F_\alpha(v_1 = n-k-1, v_2 = n-1)$, где $F_\alpha(v_1, v_2)$ – квантиль распределения Фишера с v_1 и v_2 степенями свободы, то ошибка в определении регрессии с доверительной вероятностью α признается статистически значимой, а модель неадекватной. Если $\frac{S^2}{S_y^2} < F_\alpha(v_1, v_2)$, то модель можно признать адекватной.

Примечание. *В пакете Excel используется статистика* $F = \dfrac{\frac{1}{k} \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$. *Если* $F > F_\alpha(v_1 = n-k-1, v_2 = n-1)$, *то модель признается адекватной (т.е. отбрасывается гипотеза о том, что коэффициенты модели* $\beta_0, \beta_1, \beta_2$ *равны нулю) (см. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Том 1. – М.: Финансы и статистика, 1986. – 366с. [3], на стр.53).*

Качество модели также можно оценить с использованием оценки коэффициента детерминации: $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$. Чем ближе значения R^2 к 1, тем большую долю дисперсии величины Y объясняет модель регрессии.

6.2.1 Анализ регрессионных остатков

Определенную информацию об адекватности уравнения регрессии дает исследование остатков вида $e_i = \hat{y}_i - y_i$. Если выборочная регрессия $\hat{y}$ удовлетворительно описывает истинную зависимость между y и x, остатки e_i должны быть независимыми нормально распределенными случайными величинами с нулевым средним и в значениях ε_i должен отсутствовать тренд. Нормальность распределения остатков e_i может быть установлена одним из критериев согласия (см. раздел 3.3).

Гипотезу о равенстве $M(e) = 0$ можно проверить любым параметрическим или непараметрическим критерием сравнения среднего с заданным значением (в нашем случае с нулем, см. раздел 4.4).

Независимость в последовательности значений e_i $(i = 1,...,n)$ может быть проверена с помощью сериального коэффициента корреляции Дарбина-Ватсона. Статистика сериального коэффициента корреляции имеет вид

$$D = \frac{\sum_{i=2}^{n}\left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2}.$$

Если $D > D_1(\alpha)$ или $D > 4 - D_1(\alpha)$, то с достоверностью α принимается гипотеза о наличии соответственно отрицательной или положительной корреляции остатков.

Если $D_2(\alpha) > D > D_1(\alpha)$ или $4 - D_1(\alpha) > D > 4 - D_2(\alpha)$, то критерий не позволяет принять решение по гипотезе о наличии или отсутствии корреляции остатков. Если $D_2(\alpha) < D < 4 - D_2(\alpha)$, то гипотеза корреляции остатков отклоняется. Критические значения $D_1(\alpha)$ и $D_2(\alpha)$ для различных α и числа k коэффициентов в регрессии приведены в табл. 11 (см. статистические таблицы).

6.2.2 Доверительный интервал для уравнения регрессии

Доверительный интервал для условного среднего $\hat{y} = M(Y \mid X = x)$ в многомерной точке $X_0 = (1, x_1^0, \ldots, x_k^0)^T$ определяется по формуле:

$\left[\left(X_0^T\beta^*\right) \pm t_{1-\alpha/2} S\sqrt{\left(X_0^T\left[(X^TX)^{-1}\right]X_0\right)}\right]$, где t_α – квантиль распределение Стьюдента с $n-k-1$ степенью свободы. Соответственно доверительный интервал для значений y в точке $X_0 = (1, x_1^0, \ldots, x_k^0)^T$ будет иметь вид:

$\left[\left(X_0^T\beta^*\right) \pm t_{1-\alpha/2} S\left(1+\sqrt{X_0^T\left[(X^TX)^{-1}\right]X_0}\right)\right]$, так как погрешность модели $y = f(x) + \varepsilon$ будет определяться двумя источниками: погрешностью $(\Delta f)^2 = S^2\left(X_0^T\left[(X^TX)^{-1}\right]X_0\right)$, связанной с погрешностями параметров модели, и погрешностью собственно модели $\varepsilon^2 = S^2$.

6.3. Оценка дисперсии коэффициентов регрессии и доверительных интервалов

Оценка дисперсии коэффициента β_j находится по формуле: $s_j^2 = S^2\left[(X^TX)^{-1}\right]_{jj}$, где $\left[(X^TX)^{-1}\right]_{jj}$ соответствующий диагональный элемент матрицы $(X^TX)^{-1}$.

Доверительные интервал для σ^2 находится с использованием статистики $\chi^2 = (n-k-1)S^2/\sigma^2$, которая при нормальном распределении ε_i имеет распределение хи-квадрат с $(n-k-1)$ степенью свободы.

Для проверки значимости коэффициентов уравнения регрессии используем статистику $t_j = \dfrac{\beta_j^*}{\sqrt{S^2\left[(X^TX)^{-1}\right]_{jj}}}$, которая при истинности гипотезы $H_0: \beta_j = 0$, имеет распределение Стьюдента с $(n-k-1)$ степенью свободы. Если для заданного уровня значимости α значение $|t_j|$ больше критического $t_{крит} = t_{1-\alpha/2}(n-k-1)$, то нулевая гипотеза отвергается и коэффициент признается значимым. В противном случае коэффициент признается незначимым, и соответствующее слагаемое исключается из модели.

В пакете Excel рассчитывается также уровень значимости α статистики $|t_j|$, т.е. вероятность $P\left(x > \left|t_j\right|\right)$. Степень значимости параметров распределения качественно определяется по уровню значимости: не значимые ($\alpha \geq 0{,}100$), слабо значимые ($0{,}100 > \alpha \geq 0{,}050$), статистически значимые ($0{,}050 > \alpha \geq 0{,}010$), сильно значимые ($0{,}010 > \alpha \geq 0{,}001$), высоко значимые ($0{,}001 > \alpha$).

Для уровня значимости α доверительный интервал рассчитывается по формуле $\beta_j^* \pm t_\alpha\sqrt{S^2\left[(X^TX)^{-1}\right]_{jj}}$, где t_α – квантиль распределение Стьюдента с $n-k-1$ степенью свободы.

6.4. Пример построения уравнения регрессии

Имеется выборка значений совместно наблюдаемых величин X и Y:

X	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
Y	2,96	0,61	4,63	2,44	2,23	4,89	4,98	3,89	6,74	8,07
X	5,5	6	6,5	7	7,5	8	8,5	9	9,5	10
Y	8,34	9,56	9,30	12,35	11,46	11,09	7,91	8,16	6,54	7,88

Требуется подобрать подходящую модель регрессии, характеризующую зависимость Y от X, если известно, что ошибка $\sigma^2 = 1,3$.

Нанесем точки (X, Y) на координатную плоскость – построим корреляционное поле, соответствующее нашей выборке (рис. 1)

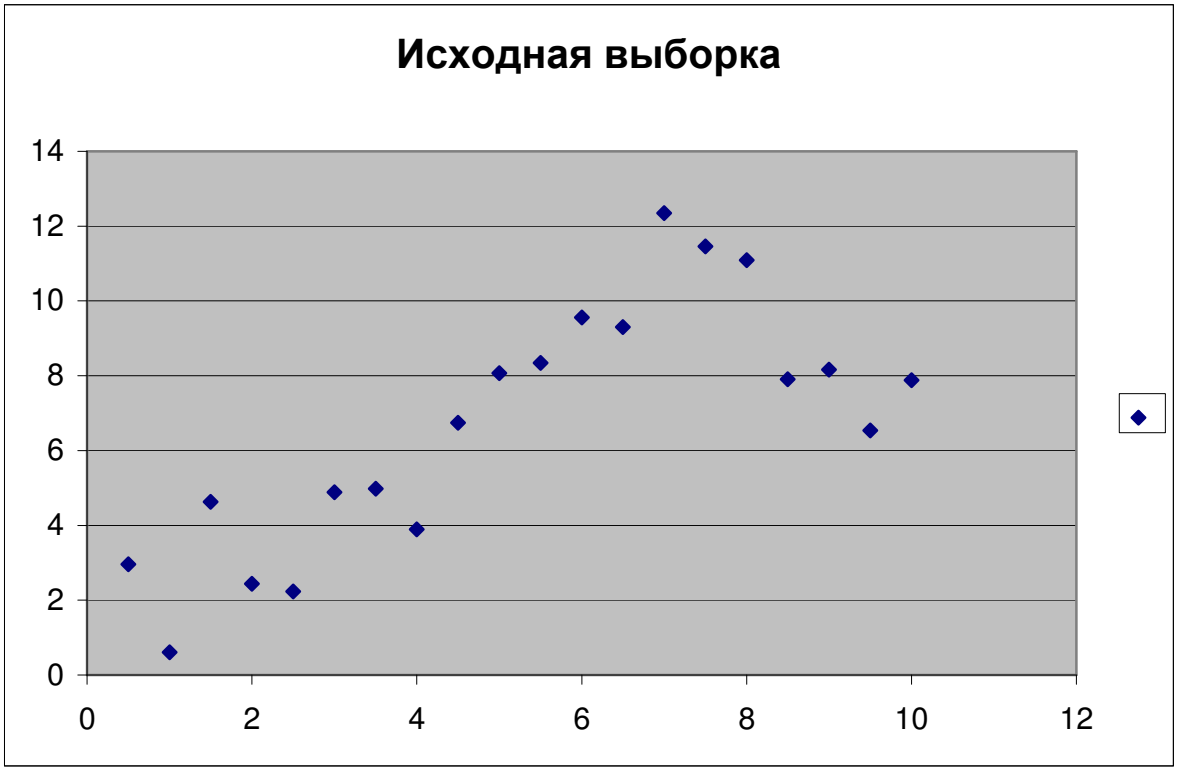


Рис. 1. Исходные данные

Видим, что существует зависимость, между значениями X и Y, причем зависимость явно нелинейная. Попробуем аппроксимировать эту зависимость для начала полиномами различных порядков. Возьмем в качестве уравнения регрессии квадратное уравнение:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Чтобы воспользоваться МНК для оценки коэффициентов, проведем линеаризацию модели, положив $x_1 = x$, $x_2 = x^2$, получим

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Тогда оценку вектора параметров, согласно МНК, найдем как

$$B^* = \left(X^T X\right)^{-1} X^T Y$$

Здесь X - матрица, первый столбец которой содержит единицы, а второй и последующий значения x_1 и x_2.

Для облегчения подбора модели можно воспользоваться встроенными функциями пакета EXCEL (для выбранной модели все равно потом потребуется провести все вычисления вручную, чтобы построить доверительные интервалы). В пакете анализа необходимо выбрать функцию “регрессия”, задать столбец значений Y и матрицу, соответствующую X (единичный столбец в этом случае задавать не надо). Если выбрать вывод остатков, то помимо регрессионной статистики, будут выведены и предсказанные значения Y, т.е.

$$y^* = \beta_0^* + \beta_1^* x + \beta_2^* x^2$$

Для нашей модели регрессионная статистика, полученная пакетом Fxcel будет иметь следующий вид:

Таблица 1.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,852379622
R-квадрат	0,72655102
Нормированный R-квадрат	0,694380552
Стандартная ошибка	1,820336831
Наблюдения	20

Таблица 2.

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	149,6702188	74,83510939	22,58750495	1,63336E-05
Остаток	17	56,32303623	3,313119778		
Итого	19	205,993255	10,841750	0,305589	

Таблица 3.

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-0,963028418	1,354297293	-0,711090853	0,486670901
Переменная X1	2,604940094	0,594036759	4,385149663	0,000403854
Переменная X2	-0,167559332	0,054953956	-3,049085893	0,007253372

(продолжение таблицы 3)

Нижние 95%	*Верхние 95%*
-3,82010793	1,894090386
1,351577249	3,858001699
-0,283477711	-0,051610008

Здесь в первой таблице:

1. Множественный R – корень квадратный из коэффициента детерминации $\sqrt{R^2}$;
2. R-квадрат – коэффициент детерминации $R^2 = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$;
3. Нормированный R-квадрат – это скорректированная величина коэффициента детерминации, вычисляемая по формуле $\hat{R}^2 = 1-(1-R^2)\dfrac{n-1}{n-k-1}$;
4. Стандартная ошибка – значение $S = \sqrt{S^2}$, где $S^2 = \dfrac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ - оценка дисперсии предсказания σ^2;
5. Наблюдения – объем выборки n.

Во второй таблице:

- df – степени свободы v;
- SS – сумма квадратов разностей:

5.3. между модельными значениями и средним

$$SS_{регрессия} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = k \cdot S_R^2;$$

5.4. остатки $SS_{остаток} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S^2 \cdot (n-k-1)$;

5.5. между исходными данными и средним

$$SS_{итого} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1) \cdot S_y^2;$$

- $MS = SS / df$;

- F – статистика $F_{Excel} = \dfrac{\dfrac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\dfrac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \dfrac{S_R^2}{S^2}$.

Если рассчитать $F = \dfrac{S^2}{S_y^2}$, то получим $F = \dfrac{3{,}313119}{10{,}841750} = 0{,}305589$.

- Значимость F – значение вероятности $F(x,k,n-k-1)$ при $x = \dfrac{S_R^2}{S^2}$, т.е. это уровень значимости принятия нулевой гипотезы H_0.

В третьей таблице:

- Коэффициенты – значения оценок коэффициентов $\beta_0{}^*, \beta_1{}^*, \beta_2{}^*$;
- Стандартная ошибка – значения оценок среднеквадратичных отклонений коэффициентов $s_j = \sqrt{S^2 \cdot \left[(X^T X)^{-1}\right]_{jj}}$;
- t-статистика – наблюдаемые значения статистик критерия проверки значимости коэффициентов соответственно $t_j = \dfrac{\beta_j^*}{\sqrt{S^2\left[(X^T X)^{-1}\right]_{jj}}}$;
- P-значения – достигнутые значения уровня значимости $P\left(x > \left|t_j\right|\right)$.
- Нижние и верхние границы 95%-го доверительного интервала $\beta_j^* \pm t_{0,05}(v)\sqrt{S^2\left[(X^T X)^{-1}\right]_{jj}}$.

Соответствующий график предсказанных значений в сравнении с исходными данными имеет вид (рис. 2):

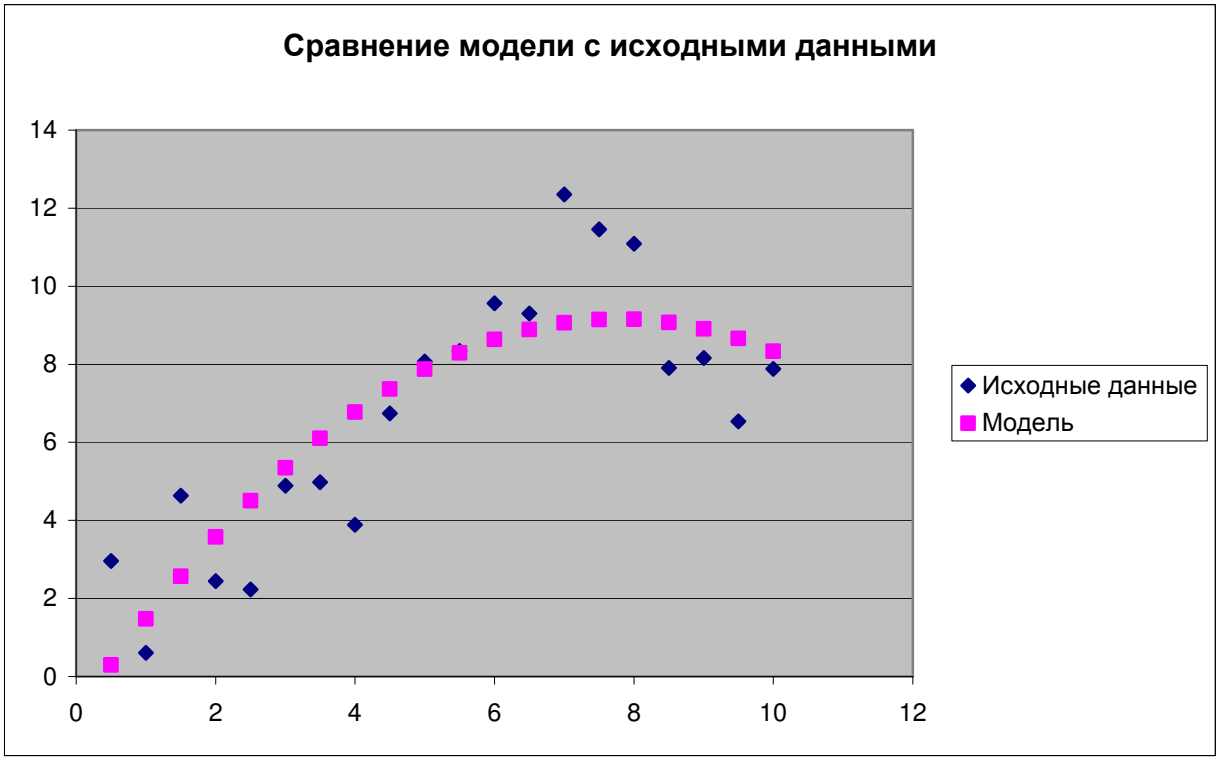


Рис. 2. Сравнение модели $y = \beta_0 + \beta_1 x + \beta_2 x^2$ с исходными данными

Отметим, что полученная оценка значения $\sigma = \sqrt{1,3} = 1,14$ велика: $S = 1,82$. Что касается коэффициентов модели, то, кроме β_0, все они значимо отличаются от нуля (достигнутый уровень значимости достаточно мал, поэтому можно отвергнуть гипотезу о равенстве коэффициентов нулю).

Попробуем улучшить модель, увеличим порядок полинома, пусть

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Проводим линеаризацию, полагая $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, и оцениваем коэффициенты новой модели.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,925507816
R-квадрат	0,856564717
Нормированный R-квадрат	0,829670602
Стандартная ошибка	1,358958106
Наблюдения	20

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	3,176148907	1,484433212	2,13963746	0,048142193
Переменная X 1	-1,619155418	1,194561911	-1,355438679	0,194104204
Переменная X 2	0,814063749	0,261005991	3,118946605	0,006612096

Переменная X 3	-0,062325275	0,016365816	-3,808259635	0,001545564

График показан на рис. 3

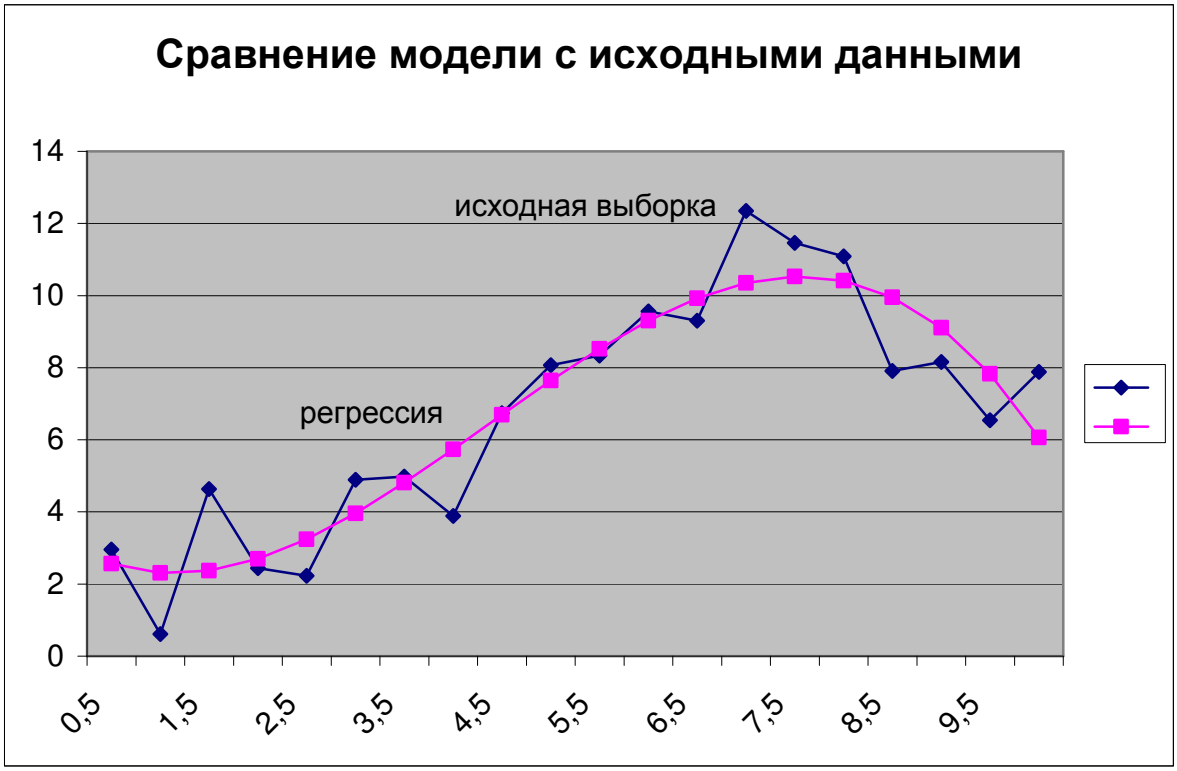


Рис. 3. Сравнение модели $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ с исходными данными

Заметим, что коэффициент детерминации увеличился, а оценка σ (стандартная ошибка) уменьшилась, что говорит о лучшем качестве модели по сравнению с предыдущей. Причем значение этой оценки близко к значениям $\sigma = 1,3$, указанному в задании. Из коэффициентов можно считать, что β_1 не значимо отличается от нуля (достигнутый уровень значимости $\alpha = 0,194$, говорит о том, что при истинности гипотезы $H_0 : \beta_1 = 0$, такое или большее значение t-статистики критерия могло наблюдаться с вероятностью 0,194). Поэтому, можно положить $\beta_1 = 0$ и модель соответственно примет вид:

$$y = \beta_0 + \beta_2 x^2 + \beta_3 x^3$$

Результаты регрессионного анализа для этой модели:

Регрессионная статистика	
Множественный R	0,916566765
R-квадрат	0,840094635
Нормированный R-квадрат	0,82128224

Стандартная ошибка	1,39201886
Наблюдения	20

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	1,356030259	0,648133062	2,092209668	0,051733578
Переменная X 1	0,469599546	0,060941974	7,70568321	6,05722E-07
Переменная X 2	-0,041727702	0,006223514	-6,70484584	3,6971E-06

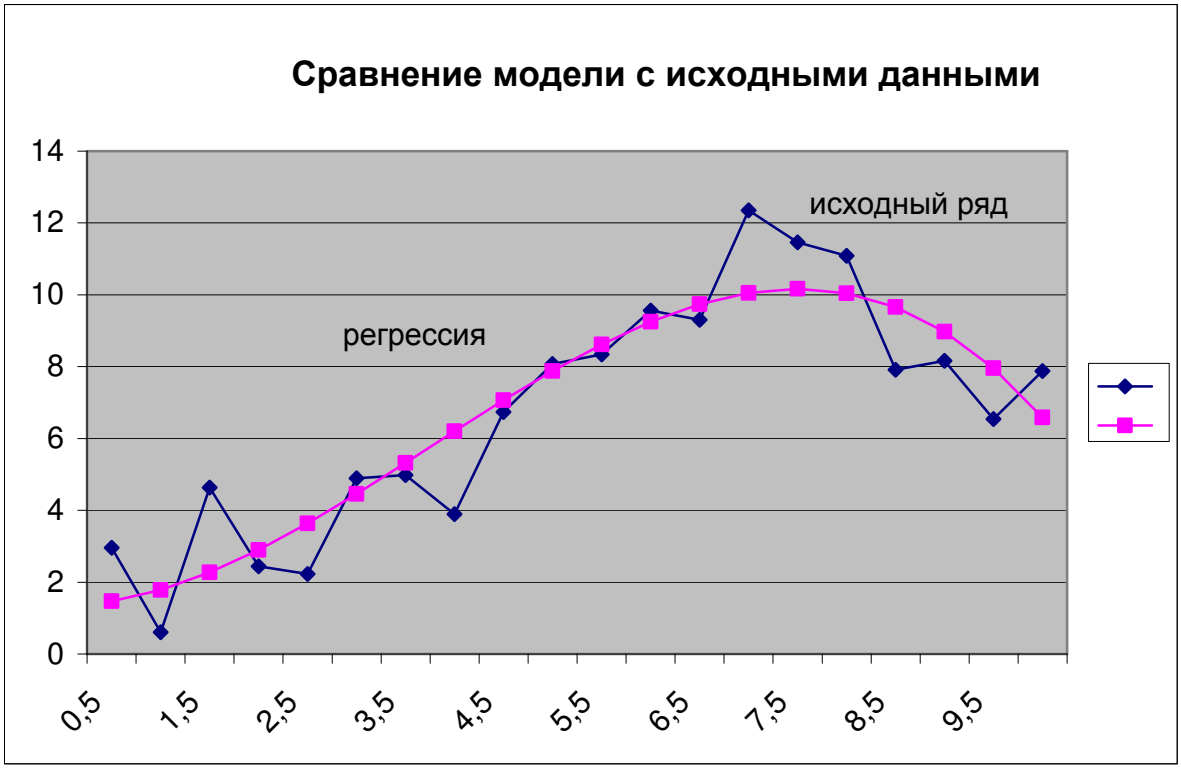


Рис. 4. Сравнение модели $y = \beta_0 + \beta_2 x^2 + \beta_3 x^3$ с исходными данными

И хотя, параметры модели немного ухудшились (сравните R-квадрат и стандартную ошибку!), тем не менее все коэффициенты получились значимыми, поэтому данная модель, предпочтительнее предыдущей.

Можно ли повысить еще качество модели? В классе полиномов это сделать не удастся. Повышение порядка полинома (можно проверить!), уже больше не понижает стандартную ошибку. Следовательно, улучшение нужно искать, используя иные классы функций. Например, можно предположить, что существует периодическая зависимость значений Y от X, тогда надо добавить в модель гармонические составляющие вида $\beta_1 cos(\omega x) + \beta_2 sin(\omega x)$. Частоты этих составляющих ω придется подбирать отдельно. Можно предположить, анализируя зависимость, что у нас присутствует периодическая составляющая с

частотой порядка 1 (одно колебание за весь интервал изменений X). Построим модель вида

$$y = \beta_0 + \beta_1 x + \beta_2 cos(2\pi x/10) + \beta_3 sin(2\pi x/10).$$

Получим оценки для этой модели:

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,934315451
R-квадрат	0,872945362
Нормированный R-квадрат	0,849122618
Стандартная ошибка	1,279008211
Наблюдения	20

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	5,080728355	0,88616922	5,733361352	3,08053E-05
Переменная X 1	0,308759559	0,159762044	1,932621493	0,0711836
Переменная X 2	-1,299656278	0,412270758	-3,152433814	0,006163641
Переменная X 3	-3,032825609	0,646493647	-4,691191664	0,000245237

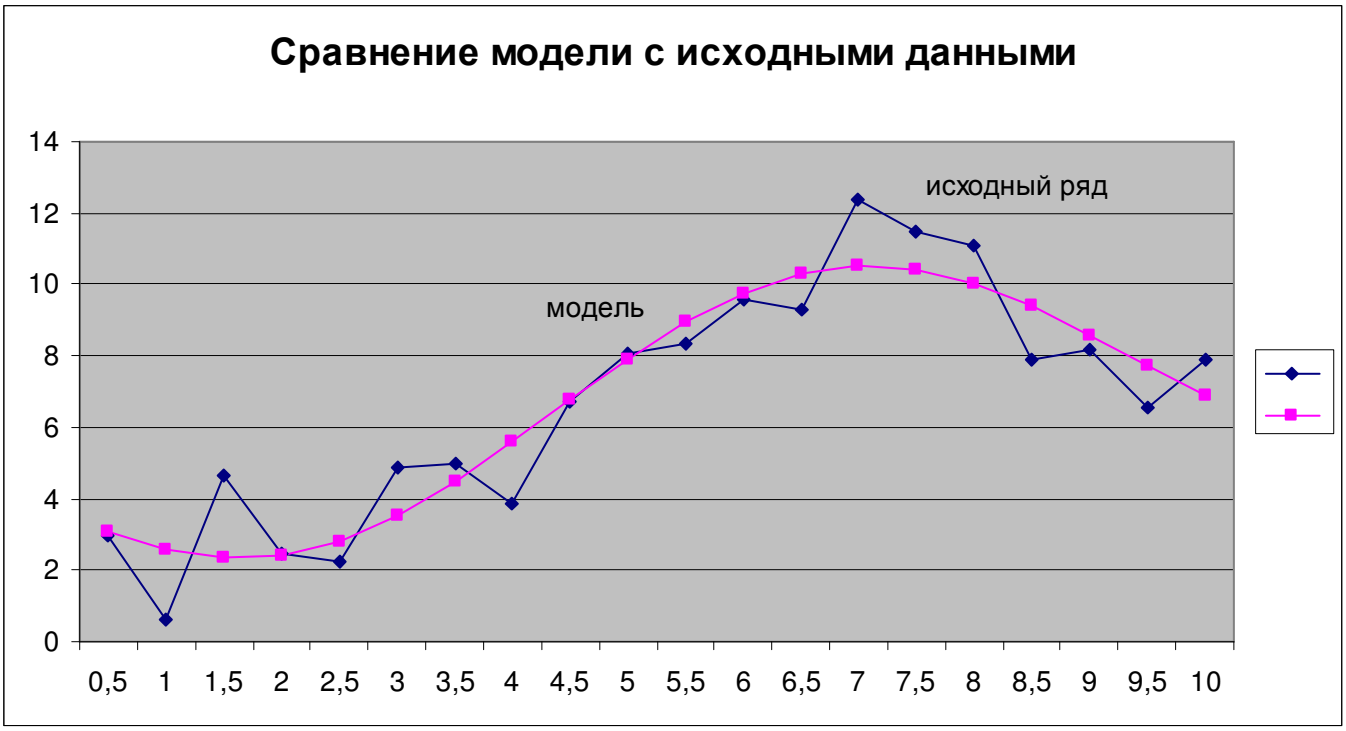


Рис. 5. Сравнение модели $y = \beta_0 + \beta_1 x + \beta_2 cos(2\pi x/10) + \beta_3 sin(2\pi x/10)$ с исходными данными

Заметим, что мы получили значение стандартной ошибки меньше заданой величины с. Это говорит о том, что дальше улучшать модель бессмысленно. Если мы продолжим, то будем по сути аппроксимировать случайные ошибки, а не реальную существующую зависимость Y от X. Качество модели выше, чем у модели вида $y = \beta_0 + \beta_2 x^2 + \beta_3 x^3$. Однако, данная модель содержит на один коэффициент больше, и кроме того содержит параметр, значение которого по сути определяется вручную. Поскольку модели принципиально различные, то предпочесть следует ту, которая более соответствует физике явления (если она известна). Мы остановимся на последней модели $y = \beta_0 + \beta_1 x + \beta_2 \mathrm{Cos}(2\pi x/10) + \beta_3 \mathrm{Cos}(2\pi x/10)$.

Найдем доверительный интервал для σ^2 соответствующий уровню 0,95. Находим квантили распределения хи-квадрат уровней 0,025 и 0,975 соответственно для числа степеней свободы $\nu = n - k - 1 = 20 - 3 - 1 = 16$ (k –число коэффициентов модели, не считая β_0): $\tau_{0,025} = 6{,}91$, $\tau_{0,975} = 28{,}85$. Тогда

$$\frac{(n-k-1)s^2}{\tau_{0,975}} < \sigma^2 < \frac{(n-k-1)s^2}{\tau_{0,025}} \Rightarrow 0{,}91 < \sigma^2 < 3{,}79 \Rightarrow 0{,}95 < \sigma < 1{,}95.$$

Доверительные интервалы для коэффициентов уравнения регрессии можно найти в Итоговой статистике.

Доверительный интервал для условного среднего $\tilde{y} = M(Y \mid X = x)$ для тех же значений X, что приведены в выборке, найдем по формуле $\left[\left(X^T B^* \right)_j \pm t_\alpha s \sqrt{\left(X (X^T X)^{-1} X^T \right)_{jj}} \right]$ (см. рис 6), где t_α квантиль распределение Стьюдента с $n-k-1$ степенью свободы (доверительный уровень возьмем 0,67, тогда $\alpha = 0{,}33$ и $t_\alpha = 1{,}0047$).

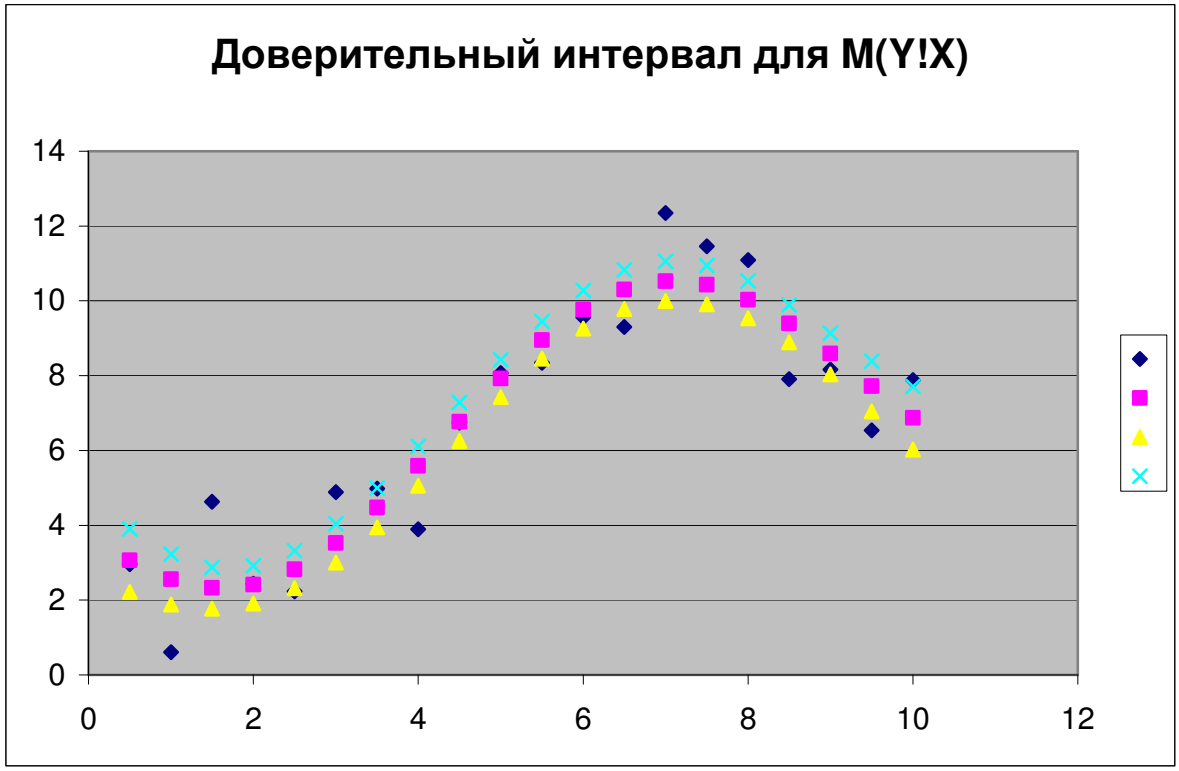


Рис. 6. Доверительный интервал для M(Y | X)

Доверительный интервал для значений y рассчитываем по формуле $y_j = \left[\left(X^T B^* \right)_j \pm t_\alpha s \left(1 + \sqrt{\left(X (X^T X)^{-1} X^T \right)_{jj}} \right) \right]$. Доверительный интервал для значений y, полученный по этим формулам, отображен на рис 7.

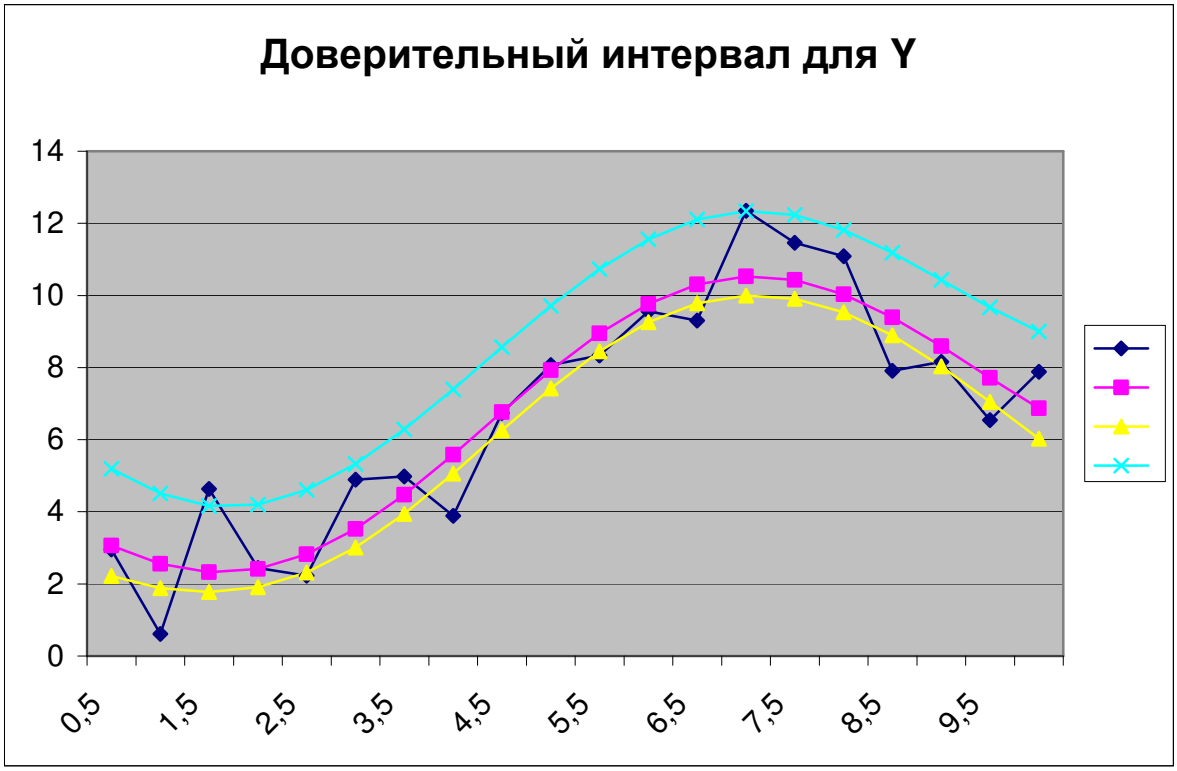


Рис. 7. Доверительный интервал для Y

Литература

1. Википедия. – Режим доступа: https://ru.wikipedia.org
2. Кобзарь А.И. Прикладная математическая статистика. – М.: ФИЗМАТЛИТ, 2012. – 813с.
3. Свешников А.А. Прикладные методы теории вероятностей. — Санкт-Петербург: Лань, 2012. – 480 с. [Электронный ресурс]. – Режим доступа: http://e.lanbook.com/books/element.php?pl1_cid=25&pl1_id=3184
4. Туганбаев А.А., Крупин В.Г. Теория вероятностей и математическая статистика — Санкт-Петербург: Лань, 2011. – 320 с. [Электронный ресурс]. – Режим доступа: http://e.lanbook.com/books/element.php?pl1_cid=25&pl1_id=652
5. Белов А.А., Баллод Б.А., Елизарова Н.Н. Теория вероятностей и математическая статистика: учебник для вузов. – Ростов н/Д: Феникс, 2008. – 318 с. (3 экз. в библиотеке ТУСУР)