МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)

Лабораторные работы
по математической статистике

Рекомендовано редакционно-издательским советом федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева» в качестве методических указаний для обучающихся по направлению прдготовки 23.03.01 Технология транспортных процессов

Составители: И.А. Тимбай
Е.П. Ростова

САМАРА
Издательство Самарского университета
2020

© Самарский университет, 2020

УДК 519.22(075)
ББК 22.172я7

Составители: ***И.А. Тимбай, Е.П. Ростова***

Рецензент д-р техн. наук, проф. С. А. И ш к о в

Лабораторные работы по математической статистике: методические указания / Сост.: *И.А.Тимбай, Е.П.Ростова;* Министерство науки и высшего образования, Самарский университет. - Самара: Издательство Самарского университета, 2020. 1CD-ROM (121Мб). - Загл. с титул. экрана. - Текст: электронный

Методические указания к лабораторным работам по математической статистике предназначены для студентов, обучающихся по направлению подготовки бакалавров 23.03.01 Технология транспортных процессов. Лабораторные работы выполняются на ЭВМ с применением современной компьютерно-математической системы MathCAD и офисной программы Excel. Электронные методические указания представляют собой веб-страницу (преобразованную в PDF-файл), содержащую текст, графические изображения, видео.

Методические указания могут быть рекомендованы студентам всех институтов и направлений подготовки/специальностей Самарского университета для самостоятельной работы и при выполнении соответствующих расчетно-графических работ и разделов курсовых работ.

Выполнены на кафедре высшей математики и на кафедре математических методов в экономике. Данные методические указания являются переработанной версией методических указаний, ранее изданных в соавторстве с Карпиловой О.М.

УДК 519.22(075)
ББК 22.172я7

Минимальные системные требования:
PC, процессор Pentium, 1 ГГц; оперативная память 2 Гб;
свободное место на винчестере 1 Гб; Microsoft Windows XP/7/8/10,
Linux, MacOS; разрешение экрана 1024x768 с глубиной цвета 16 бит;
DVD-ROM 2-х и выше; Adobe Reader.

© Самарский университет, 2020

Наверх

Введение

Цель лабораторных работ – дать студентам практику решения задач математической статистики с применением современной компьютерно-математической системы MathCAD и офисной программы Excel. Лабораторные работы по содержанию соответствуют программе курса математики для студентов, обучающихся по направлению подготовки бакалавров 23.03.01 Технология транспортных процессов. Данные методические указания могут быть использованы также студентами всех институтов и направлений/специальностей Самарского университета для самостоятельной работы и при выполнении соответствующих расчетно-графических работ и разделов курсовых работ.

При выполнении лабораторных работ студенты обучаются основным методам и приёмам решения задач математической статистики на компьютере, закрепляя знания, полученные на лекциях и практических занятиях. Лабораторные работы выполняются в математической системе MathCAD и офисной программы Excel. При этом данное электронное издание используется как файл справочной системы. Вначале студент знакомится с приведенными общими теоретическими положениями, затем выполняет задание по образцу, приобретая необходимые навыки решения математических задач, в конце выполняет индивидуальное задание.

Методические указания предназначены для следующих лабораторных работ:

№1. Статистический анализ одномерных данных.

№2. Статистический анализ двумерных данных.

В Приложении приведены некоторые приемы работы в системе MathCad.

Наверх

Статистический анализ одномерных данных

Лабораторная работа №1

1. Общие теоретические положения

Основная задача статистики

Основной задачей математической статистики является обработка и анализ результатов наблюдений для выявления статистических закономерностей. При этом результаты наблюдений рассматриваются как значения случайной величины, полученные в результате сбора данных об интересующем объекте или явлении.

Генеральная совокупность и выборка

Генеральной совокупностью называется весь набор однородных объектов, изучаемых относительно некоторого качественного или количественного признака. Число всех изучаемых объектов N называется объемом генеральной совокупности.

Выборка - это та часть генеральной совокупности, элементы которой подвергаются статистическому обследованию. Число n вошедших в выборку элементов называется объемом выборки.

Математическая статистика позволяет по выборке делать выводы о виде и параметрах закона распределения генеральной совокупности, а также оценивать степень точности и достоверности полученных выводов.

Вариационные ряды и их характеристики

Наблюдаемые значения случайной величины X называют *вариантами*.

Вариационный ряд – это запись результатов измерений какой-либо случайной величины в виде последовательности вариант.

Упорядоченный вариационный ряд – это последовательность вариант, записанная в порядке возрастания (или убывания).

Для обработки результатов статистических наблюдений их удобно оформлять в виде таблицы частот.

Статистическое распределение – таблица частот, в которой указаны значения случайной величины x_i и соответствующие частоты m_i, показывающие, сколько раз в выборке встретилось данное значение случайной величины.

Для получения интервальной таблицы частот (интервального вариационного ряда) весь диапазон измеренных значений случайной величины X делят на K равных интервалов (u_k, u_{k+1}) и подсчитывают количество m_k значений случайной величины, попавших на соответствующий интервал.

Иногда вместо таблицы частот составляют таблицу относительных частот $w_k = \frac{m_k}{n}$ (n - объём выборки), очевидно, $\sum_{k=1}^{K} w_k = 1$.

Количество интервалов можно определить по формуле Стургерса:

$$K = [1 + 1,44\ln(n)] + 1 ,$$

где n – объем выборки, а квадратные скобки означают целую часть числа.

Размах – это разность между наибольшим и наименьшим значениями вариант:

$$R = x_{max} - x_{min} .$$

Для графического представления вариационных рядов используют гистограмму.

Гистограмма – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат отрезки оси, равные интервалам (u_k, u_{k+1}), а высоты пропорциональны частотам m_k (или относительным частотам w_k).

Статистические оценки параметров генеральной совокупности

Одна из задач математической статистики – оценка параметров генеральной совокупности по данным выборки.

Статистические оценки бывают *точечные* (определяемые одним числом) и *интервальные* (определяемые двумя числами - концами интервала). Точечные оценки дают представление о величине соответствующего параметра, а интервальные характеризуют точность и достоверность оценки.

Для достоверности результатов точечная оценка должна быть несмещенной, состоятельной и эффективной. Этим условиям удовлетворяют следующие оценки:

выборочное среднее

$$X_{\text{В}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

– оценка для математического ожидания генеральной совокупности;

выборочная дисперсия ("*исправленная*")

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - X_{\text{В}})^2}{n-1}$$

– несмещенная оценка для дисперсии генеральной совокупности;

стандартное отклонение ("*исправленное*")

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - X_{\text{В}})^2}{n-1}}$$

– несмещенная оценка для дисперсии генеральной совокупности.

Следует отметить, что смещенной оценкой генеральной дисперсии служит выборочная дисперсия

$$D_{\text{В}} = \frac{\sum_{i=1}^{n} (x_i - X_{\text{В}})^2}{n} .$$

Если θ^*- статистическая оценка параметра θ, то говорят, что оценка вычислена с точностью δ, если $|\theta - \theta^*| < \delta$, то есть величина параметра θ попадает в интервал $(\theta^* - \delta\, ;\, \theta^* + \delta)$. Вероятность γ, с которой осуществляется это неравенство, называется *надежностью (доверительной вероятностью)* оценки, а значение $\alpha = 1 - \gamma$ – *уровнем значимости*.

Интервал $(\theta^* - \delta\, ;\, \theta^* + \delta)$, который покрывает неизвестный параметр с заданной надежностью γ, называется *доверительным интервалом*.

Доверительный интервал для оценки математического ожидания нормального распределения при неизвестной дисперсии имеет вид

$$\left(X_{\text{В}} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}\ ;\ X_{\text{В}} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}\right) ,$$

где $t_{1-\frac{\alpha}{2}} = t(1 - \frac{\alpha}{2}, n-1)$ - квантиль обратного распределения Стьюдента определяется по справочной таблице.

Доверительный интервал для оценки дисперсии нормального распределения при неизвестном математическом ожидании имеет вид

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\ ;\ \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}\right) \quad ,$$

где $\chi^2_{1-\frac{\alpha}{2}} = \chi^2(1-\frac{\alpha}{2}\ ,\ n-1)$, $\chi^2_{\frac{\alpha}{2}} = \chi^2(\frac{\alpha}{2}\ ,\ n-1)$ - квантили распределения χ^2 определяются по справочной таблице.

Проверка статистических гипотез

Статистической гипотезой называется любое предположение о виде или параметрах неизвестного распределения.

Выдвинутую статистическую гипотезу, которую подвергают проверке, называют *нулевой гипотезой* и обозначают H_0 . Каждая другая допустимая гипотеза, отличная от нулевой, называется *альтернативной* и обозначается H_1 .

В результате проверки гипотеза H_0 принимается или отвергается, причем возможны четыре случая:

а) H_0 верна, и её приняли; б) H_0 верна, но её отвергли;

в) H_0 неверна, но её приняли; г) H_0 неверна, и её отвергли.

Очевидно, что во втором и третьем случаях принятое решение было ошибочным.

Ошибка, совершенная в случае б), когда отвергается правильная гипотеза, называется *ошибкой первого рода*; ошибка, совершенная в случае в), когда принята неверная гипотеза, называется *ошибкой второго рода*.

Вероятность совершить ошибку первого рода обозначают α и называют *уровнем значимости* или α - *риском*. Величина α обычно выбирается 0,05 (5%) или 0,01 (1%).

Вероятность совершить ошибку второго рода обозначают β и называют β - *риском*.

Для проверки гипотезы о виде теоретического закона распределения обычно применяют критерии согласия. При этом выдвигается гипотеза H_0 о том, что исследуемая случайная величина X подчиняется определенному закону распределения, и вычисляют статистику Z меры расхождения теоретического (выбранного в гипотезе H_0) и эмпирического (полученного по экспериментальным данным) распределений. Закон распределения Z должен быть известен, и тогда для выбранного уровня значимости α устанавливают критическую точку $Z_{кр}$. Если вычисленная величина $Z \geq Z_{кр}$,то гипотезу H_0 отвергают, если $Z < Z_{кр}$, то гипотезу H_0 принимают.

На практике наиболее часто применяют критерий согласия χ^2 - Пирсона. В этом критерии в качестве статистики Z берут

$$\chi^2 = \sum_{k=1}^{K} \frac{(m_k - np_k)^2}{np_k} \ ,$$

которая при $n \to \infty$ имеет χ^2 - распределение с числом степеней свободы $r = K - l - 1$, где K – число интервалов эмпирического распределения (интервальной таблицы частот), l – число параметров теоретического распределения, p_k – теоретические вероятности попадания случайной величины на интервал $(u_k\ , u_{k+1})$, вычисляемые по формуле

$$p_k = \Phi\left(\frac{u_{k+1} - X_{\text{В}}}{s}\right) - \Phi\left(\frac{u_k - X_{\text{В}}}{s}\right) \quad ,$$

$\Phi(x) = \frac{1}{2\pi}\int_0^x e^{-\frac{t^2}{2}}\, dt$ – функция Лапласа.

После вычисления статистики, выбрав уровень значимости α (например α = 0,05), в таблице критических точек распределения χ^2 находят $\chi^2_{кр}$. Если $\chi^2 < \chi^2_{кр}$, то можно принять гипотезу о нормальном распределении, т. е. полученный теоретический закон хорошо аппроксимирует статистическое распределение. Если $\chi^2 \geq \chi^2_{кр}$, то гипотеза о выборе теоретического закона отвергается, т.е. полученный закон не согласуется с экспериментальными данными.

2. Содержание лабораторной работы

Дан протокол измерений случайной величины *X*. Необходимо:

1) получить интервальную таблицу частот;
2) построить гистограмму относительных частот;
3) получить точечные оценки для математического ожидания и дисперсии;
4) аппроксимировать гистограмму теоретическим нормальным законом распределения;
5) сделать выводы о согласовании теоретического и статистического законов распределений;
6) найти доверительные интервалы для математического ожидания и дисперсии.

Примечание. В учебных целях протокол измерений нормальной случайной величины X с математическим ожиданием a и средним квадратическим отклонением σ можно сгенерировать, используя формулу

$$x_i = (\sum_{j=1}^{12} r_j - 6)\,\sigma + a \quad ,$$

где r_j – случайная величина, распределенная равномерно.

Пример выполнения лабораторной работы в MathCad

Лабораторная работа №1

Статистический анализ одномерных данных

Приближенное разыгрывание нормальной случайной величины с параметрами a, σ

$n := 325$ - количество реализаций случайной величины

$a := 0.65$ - математическое ожидание

$\sigma := 3.9$ - среднее квадратическое отклонение

$i := 1..n$ ORIGIN := 1 - нижняя граница индексации массивов

$$X_i := \left[\left(\sum_{j=1}^{12} rnd(1)\right) - 6\right]\cdot\sigma + a$$

- формула моделирования (разыгрывания) возможных значений нормальной случайной величины

функция генерации случайных чисел *rnd* (x) при каждом обращении к ней возвращает случайное число с равномерным распределением на отрезке [0,x]

Простейшая статистическая обработка

Построение упорядоченного вариационного ряда

$X := sort(X)$ функция sort(X) возвращает вектор, содержащий элементы в порядке возрастания их значений

X =

	1
1	-8.989
2	-8.89
3	-8.066
.	.
.	.
323	9.119
324	9.355
325	10.421

Определение максимального и минимального значений выборки и размаха R

$Xmin := min(X)$ $Xmin = -8.989$

$Xmax := max(X)$ $Xmax = 10.421$

$R := Xmax - Xmin$ $R = 19.41$

Определение количества интервалов в гистограмме по правилу Стургерса

$K := \mathrm{ceil}(1 + 1.44 \cdot \ln(n))$ $\quad K = 10 \quad$ функция ceil(x) возвращает наименьшее целое число большее или равное x

Определение длины интервала Δx

$$\Delta x := \frac{Xmax - Xmin}{K} \qquad \Delta x = 1.941$$

Определение границ интервалов

$k := 1..K$ - номер интервала

$$u_1 := Xmin \qquad u_{k+1} := Xmin + k \cdot \Delta x$$

Определение частоты попадания выборочных значений в k-й интервал

$$m_k := \sum_{i=1}^{n} \left(u_k \le X_i < u_{k+1}\right) \qquad m_1 := \sum_{i=1}^{n} \left(X_i < u_2\right) \qquad m_K := \sum_{i=1}^{n} \left(X_i \ge u_K\right)$$

Определение относительной частоты попадания выборочных значений в k-й интервал

$$w_k := \frac{m_k}{n}$$

Определение высоты прямоугольника на каждом интервале

$$h_k := \frac{w_k}{\left(u_{k+1} - u_k\right)}$$

номер интервала	границы интервалов		частота	относительная частота	высота столбца гистограммы
$k =$	$u_k =$	$u_{k+1} =$	$m_k =$	$w_k =$	$h_k =$
1	-8.989	-7.048	7	0.022	0.011
2	-7.048	-5.107	13	0.04	0.021
3	-5.107	-3.166	27	0.083	0.043
4	-3.166	-1.225	43	0.132	0.068
5	-1.225	0.716	67	0.206	0.106
6	0.716	2.657	60	0.185	0.095
7	2.657	4.598	53	0.163	0.084
8	4.598	6.539	33	0.102	0.052
9	6.539	8.48	18	0.055	0.029
10	8.48	10.421	4	0.012	$6.341 \cdot 10^{-3}$

$$\sum m = 325 \qquad \sum w = 1$$

<u>Замечание.</u> Вектор частот можно вычислить с помощью встроенной функции MathCAD

$m1 := \mathrm{hist}(u, X)$ $\qquad m1 =$

	1
1	7
2	13
3	27
4	43
5	67
6	60
7	53
8	33
9	18
10	4

$$\sum m1 = 325$$

Построение гистограммы относительных частот

$$fs(x) := \sum_{j=1}^{K} h_j \cdot \left(u_j \le x < u_{j+1}\right)$$ - эмпирическая плотность распределения

Построение рисунка 1 (видео)

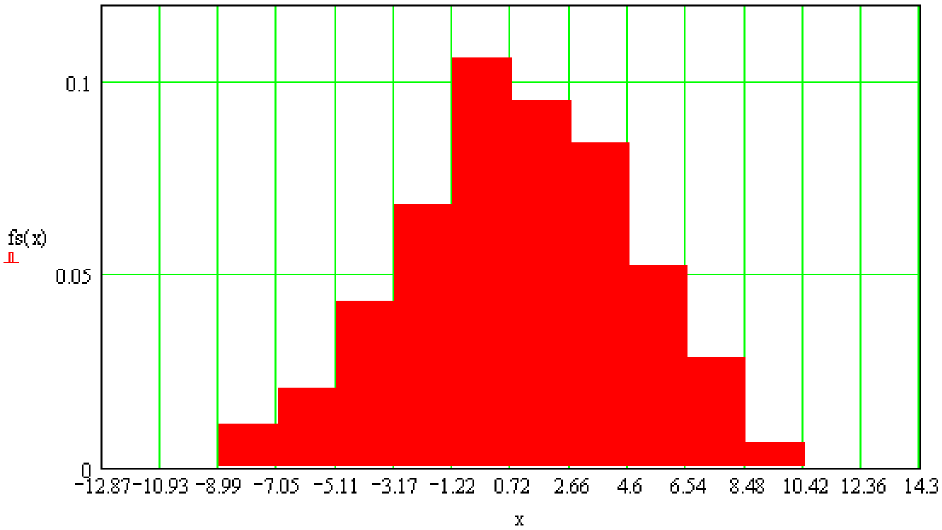


Рисунок-1. Эмпирическая плотность распределения

Заметим, что гистограмма относительных частот, являющаяся статистичесой оценкой плотности вероятностей генеральной совокупности, схожа с кривой плотности вероятностей нормального закона

Статистические оценки параметров распределения

Вычисление выборочного среднего

$$Xmean := \frac{\sum X}{n} \qquad Xmean = 0.822$$

Вычисление выборочного среднего с помощью встроенной функции MathCAD

$$mean(X) = 0.822$$

Вычисление выборочной дисперсии

$$D := \frac{\left[\overrightarrow{\sum (X - Xmean)^2}\right]}{n} \qquad D = 14.304$$

Вычисление выборочной дисперсии с помощью встроенной функции MathCAD

$$var(X) = 14.304$$

Вычисление выборочного среднего квадратического отклонения (стандартного)

$$\sigma := \sqrt{D} \qquad \sigma = 3.782$$

Вычисление выборочного среднего квадратического отклонения с помощью встроенной функции MathCAD

$$stdev(X) = 3.782$$

Вычисление исправленной дисперсии

$$s2 := \frac{n}{n-1}\cdot D \qquad s2 = 14.349$$

Вычисление исправленной дисперсии с помощью встроенной функции MathCAD

$$Var(X) = 14.349$$

Вычисление "исправленного" среднего квадратического отклонения

$$s := \sqrt{s2} \qquad s = 3.788$$

Вычисление "исправленного" среднего квадратического отклонения с помощью встроенной функции MathCAD

$$Stdev(X) = 3.788$$

Вычисление теоретической плотности вероятностей

Вычисление функции плотности распределения вероятностей нормального распределения

$$f(x) := \frac{1}{s\cdot\sqrt{2\cdot\pi}}\cdot e^{\frac{-(x-Xmean)^2}{2\cdot s^2}} \qquad f(Xmean) = 0.105$$

Вычисление функции плотности распределения вероятностей нормального распределения с помощью встроенной функции MathCAD

$f(x) := dnorm(x, Xmean, s)$ - функция dnorm(x,μ,σ) возвращает плотность распределения вероятностей нормального распределения (μ − среднее значение, $\sigma > 0$ − среднеквадратическое отклонение)

$f(Xmean) = 0.105$

Построение рисунка 2 (видео)

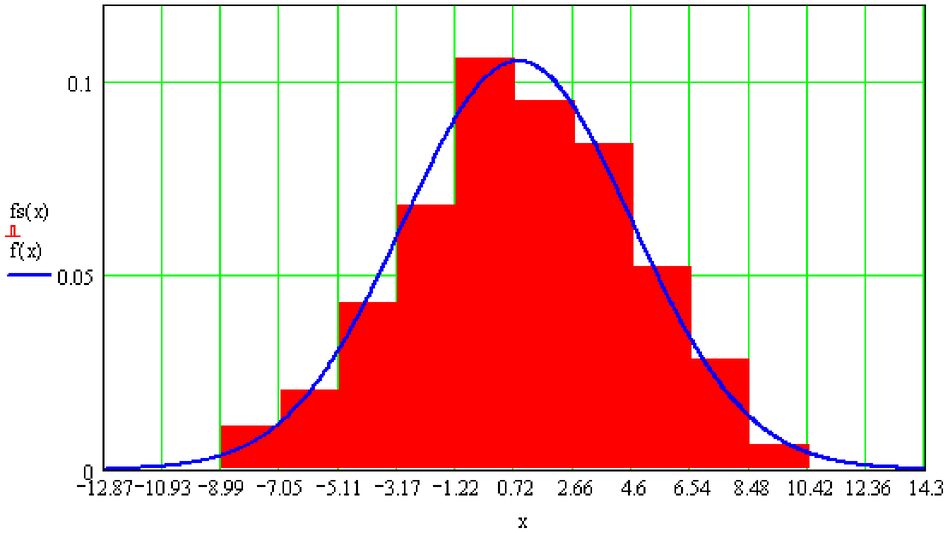


Рисунок-2. Эмпирическая и теоретическая плотности распределения

Проверка гипотезы о нормальном распределении случайной величины X, при использовании критерия Пирсона при уровне значимости α

$\alpha := 0.05$ - уровень значимости

$p_k := \text{pnorm}\left(u_{k+1}, \text{Xmean}, s\right) - \text{pnorm}\left(u_k, \text{Xmean}, s\right)$

Статистикой критерия Пирсона является величина

$$\chi 2 := \sum_{k=1}^{K} \frac{\left(m_k - p_k \cdot n\right)^2}{p_k \cdot n}$$

$\chi 2 = 3.992$

$\chi 2kp := \text{qchisq}(1 - \alpha, K - 2 - 1)$ функция qchisq(p,d) возвращает квантиль обратного Хи-квадрат распределения (d - определяет степени свободы, p - вероятность)

$\chi 2kp = 14.067$

Поскольку $\chi 2 < \chi 2kp$ - *гипотеза принимается*

Построение доверительного интервала для оценки математического ожидания

$\alpha := 0.05$ - уровень значимости

$\gamma := 1 - \alpha$ - заданная надежность (доверительная вероятность)

$t\gamma := \text{qt}\left(1 - \frac{\alpha}{2}, n - 1\right)$ функция qt(p,d) возвращает квантиль обратного распределения Стьюдента (d - определяет степени свободы, p - вероятность)

$t\gamma = 1.967$

Вычисление доверительных границ

$\theta x_1 := \text{Xmean} - t\gamma \cdot \frac{s}{\sqrt{n}}$ $\qquad$ $\theta x_2 := \text{Xmean} + t\gamma \cdot \frac{s}{\sqrt{n}}$

$\theta x_1 = 0.408$ $\qquad$ $\theta x_2 = 1.235$

Построение доверительного интервала для оценки дисперсии

Вычисление доверительных границ

$$\theta\sigma_1 := \frac{(n-1)\cdot s^2}{\text{qchisq}\left(1 - \frac{\alpha}{2}, n - 1\right)} \qquad \theta\sigma_2 := \frac{(n-1)\cdot s^2}{\text{qchisq}\left(\frac{\alpha}{2}, n - 1\right)}$$

$\theta\sigma_1 = 12.372$ $\qquad$ $\theta\sigma_2 = 16.842$

Пример выполнения лабораторной работы в Excel

Лабораторная работа № 1

Петрова П.П. *гр. 3205* *вариант 25*

Статистический анализ одномерных данных

n= 30 - объем выборки

i	Xi
1	-5.248
2	0.466
3	5.017
4	2.022
5	-0.497
6	5.073
7	5.286
8	-6.279
9	1.03
10	2.128
11	4.733
12	-4.436
13	1.578
14	1.063
15	-3.417
16	-3.734
17	-2.949
18	1.076
19	-1.078
20	-2.295
21	-7.381
22	2.826
23	0.101
24	3.221
25	-8.989
26	-3.266
27	-2.27
28	2.017
29	8.122
30	1.983

- варианты выборки

Простейшая статистическая обработка

Определение максимального и минимального значений выборки и размаха R

=МАКС(B11:B40)

- максимальное значение выборки Xmax= 8.122

=МИН(B11:B40)

- минимальное значение выборки Xmin= -8.989

- размах выборки R= 17.111

Определение количества интервалов в гистограмме по правилу Стургерса

=ОТБР(1+1.44*LN(B8))+1

K= 6

функция **ОТБР(ЧИСЛО)** усекает число до целого,
отбрасывая дробную часть числа, так что остается целое число

Определение длины интервала Δx

Δx= 2.851833

Определение границ интервалов

u1 = Xmin, ... , uk+1 = Xmin+k*Δx

Определение частоты попадания выборочных значений в k-й интервал

функция **ЧАСТОТА(массив_данных;массив_интервалов)** вычисляет
частоту появления значений в интервале значений и возвращает массив цифр
Примечание. Формулу необходимо ввести как формулу массива.
Выделите ячейки частот, начиная с ячейки, содержащей формулу.
Нажмите клавишу F2, а затем нажмите клавиши CTRL+SHIFT+ENTER.

Определение относительной частоты попадания выборочных значений в k-й интервал

wk=mk/n

Определение высоты прямоугольника на каждом интервале

hk=wk/Δx

номер интервала	границы интервалов		частота	относительная частота	высота столбца гистограммы	середина каждого интервала
k	**uk**	**uk+1**	**mk**	**wk**	**hk**	**b**
1	-8.989	-6.137	3	0.100	0.035	-7.563
2	-6.137	-3.285	4	0.133	0.047	-4.711
3	-3.285	-0.434	6	0.200	0.070	-1.859
4	-0.434	2.418	10	0.333	0.117	0.992
5	2.418	5.270	5	0.167	0.058	3.844
6	5.270	8.122	2	0.067	0.023	6.696
		проверка	30	1		

Построение рисунка 1 (видео)

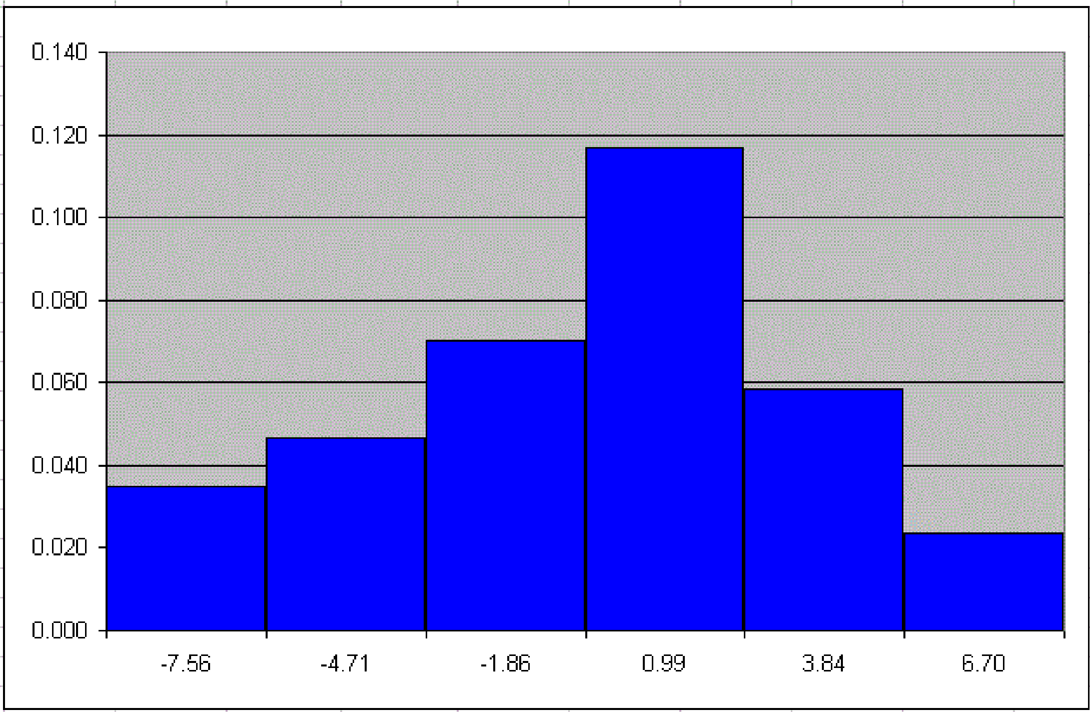


Рисунок 1. Эмпирическая плотность распределения

Заметим, что гистограмма относительных частот, являющаяся статистической оценкой плотности вероятностей генеральной совокупности, схожа с кривой плотности вероятностей нормального закона.

Статистические оценки параметров распределения

Вычисление выборочного среднего		Вычисление выборочного среднего с помощью встроенной функции Excel	
Xmean=	-0.137	Xmean=	-0.137 =СРЗНАЧ(B11:B40)
Вычисление выборочной дисперсии			
D=	16.215 =СУММПРОИЗВ(B11:B40;B11:B40)/B8-B$103^2		
Вычисление выборочного среднего квадратического отклонения			
sig=	4.027		
Вычисление исправленной дисперсии		Вычисление исправленной дисперсии с помощью встроенной функции Excel	
s2=	16.774	s2=	16.774 =ДИСП(B11:B40)
Вычисление исправленного среднего квадратического отклонения		Вычисление исправленного среднего квадратического отклонения с помощью встроенной функции Excel	
s=	4.096	s=	4.096 =СТАНДОТКЛОН(B11:B40)

Вычисление теоретической плотности вероятностей

Вычисление функции плотности распределения вероятностей нормального распределения с помощью встроенной функции Excel

НОРМРАСП(x;среднее;стандартное_отклонение;ЛОЖЬ) - функция возвращает функцию плотности нормального распределения

$$f(x) := \frac{e^{\frac{-(x-Xmean)^2}{2\cdot s^2}}}{s\cdot\sqrt{2\cdot\pi}}$$

середина каждого интервала	функция плотности нормального распределения
b	**f(b)**
-7.563	0.0188
-4.711	0.0522
-1.859	0.0892
0.992	0.0938
3.844	0.0607
6.696	0.0242

Построение рисунка 2 (видео)

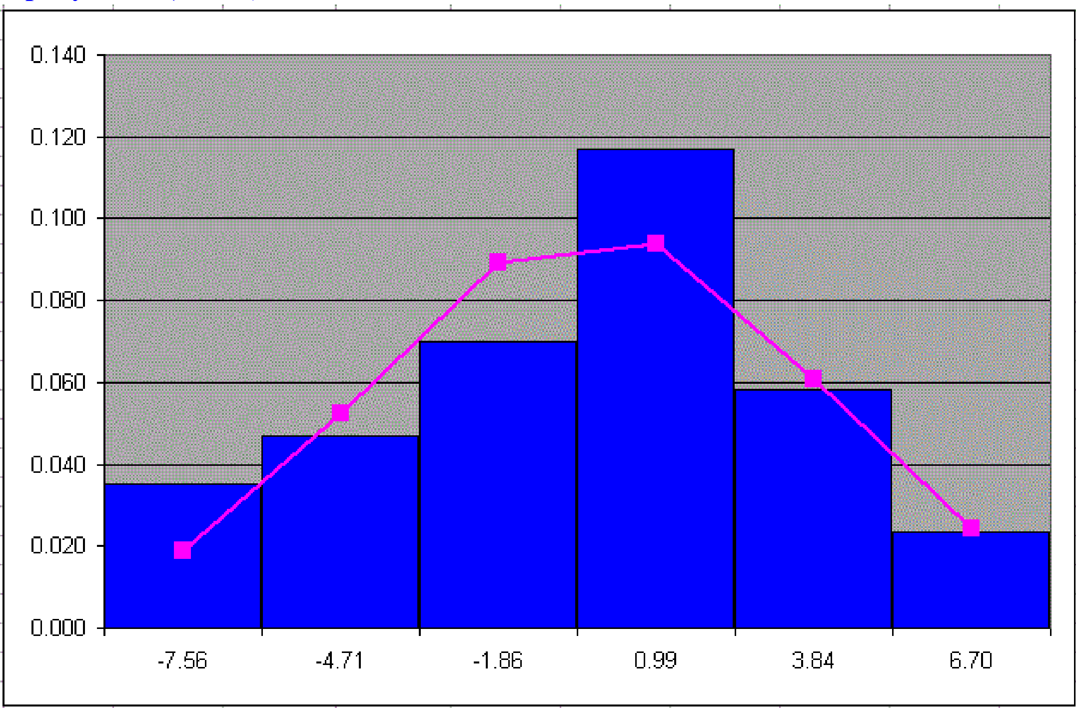


Рисунок 2. Эмпирическая и теоретическая плотности распределения

Проверка гипотезы о нормальном распределении случайной величины X, при использовании критерия Пирсона при уровне значимости α

α= 0.05 - уровень значимости

p(к)	p(к+1)	p	
0.01533	0.071441	0.056111	1.029851
0.071441	0.221	0.149558	0.052806
0.221	0.471102	0.250102	0.301102
0.471102	0.733627	0.262525	0.572945
0.733627	0.906606	0.172979	0.00691
0.906606	0.978124	0.071518	0.009874
			1.973489

функция НОРМРАСП(x;среднее;стандартное_откл;интегральная) возвращает нормальную функцию распределения для указанного среднего и стандартного отклонения.

1.973489 -Статистика ϰ2 Пирсона, вычисленная по выборке

$$\chi^2 := \sum_{k=1}^{K} \frac{(m_k - p_k \cdot n)^2}{p_k \cdot n}$$

ϰ2кр= 7.814725

функция ХИ2ОБР(вероятность;степени_свободы) возвращает значение, обратное к односторонней вероятности распределения ϰ2 (хи-квадрат).

Как видно вычисленная по выборке статистика удовлетворяет условию ϰ2<ϰ2кр, а значит, нет оснований отвергнуть гипотезу.

Гипотеза принимается

Построение доверительного интервала для оценки математического ожидания

α= 0.05 - уровень значимости

γ= 0.95 - заданная надежность (доверительная вероятность)

tγ= 2.045231

СТЬЮДРАСПОБР(вероятность;степени_свободы) - функция возвращает обратное распределения Стьюдента

Вычисление доверительных границ

θx1= -1.66588 θx2= 1.392747

Построение доверительного интервала для оценки дисперсии

α= 0.05 - уровень значимости

γ= 0.95 - заданная надежность (доверительная вероятность)

Вычисление доверительных границ

θσ1= 10.63898 θσ2= 30.31326

Задания к лабораторной работе №1

Количество реализаций случайной величины:	n = № группы + № варианта
Математическое ожидание:	a = № варианта - 10
Среднее квадратическое отклонение:	σ = (50 - № варианта) * 0.1

Наверх

Статистический анализ двумерных данных

Лабораторная работа №2

1. Общие теоретические положения

Двумерная случайная величина

Двумерной называют случайную величину (X,Y), возможные значения которой есть пары чисел (x,y). Составляющие X и Y, рассматриваемые одновременно, образуют *систему* двух случайных величин.

Двумерную величину геометрически можно истолковывать как случайную точку $M(X,Y)$ на плоскости xOy либо как случайный вектор OM.

Статистическая и корреляционная зависимость

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределение другой.

Если статистическая зависимость проявляется в том, что при изменении одной величины изменяется среднее значение другой, то зависимость называется *корреляционной*.

Выборочные оценки числовых характеристик распределения

Пусть имеется n наблюдений случайного вектора (X,Y). По данным наблюдениям вычисляются следующие выборочные оценки числовых характеристик распределения:

выборочное среднее случайной величины X

$$X_{\text{В}} = \frac{\sum_{i=1}^{n} x_i}{n} ;$$

выборочное среднее случайной величины Y

$$Y_{\text{В}} = \frac{\sum_{i=1}^{n} y_i}{n} ;$$

выборочная дисперсия случайной величины X

$$D_X = \frac{\sum_{i=1}^{n} (x_i - X_B)^2}{n} ;$$

выборочная дисперсия случайной величины Y

$$D_Y = \frac{\sum_{i=1}^{n} (y_i - Y_B)^2}{n} ;$$

выборочное среднее квадратическое отклонение случайной величины X

$$\sigma_X = \sqrt{D_X} \ ;$$

выборочное среднее квадратическое отклонение случайной величины Y

$$\sigma_Y = \sqrt{D_Y} \ ;$$

выборочный коэффициент корреляции

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - X_B)(y_i - Y_B)}{n \ \sigma_X \ \sigma_Y} \ .$$

Выборочный коэффициент корреляции является оценкой коэффициента корреляции генеральной совокупности, который характеризует степень линейной зависимости между случайными величинами X и Y, причем

$$-1 \le r_{XY} \ge 1 \ .$$

Если случайные величины X и Y независимы, то коэффициент корреляции $r_{XY} = 0$, если $r_{XY} = \pm 1$, то X и Y связаны линейной функциональной зависимостью.

<u>Проверка гипотезы о значимости выборочного коэффициента корреляции</u>

Для того чтобы при уровне значимости α проверить нулевую гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности нормальной двумерной случайной величины при конкурирующей гипотезе об отличии его от нуля, надо вычислить выборочное значение критерия

$$T = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $(n-2)$ найти критическую точку

$$t_{kp} = t(1 - \frac{\alpha}{2} \ , \ n-2) \ .$$

Если $|T| < t_{kp}$ - нет оснований отвергнуть нулевую гипотезу. Если $|T| \ge t_{kp}$ - нулевую гипотезу отвергают.

Если $|T| < t_{kp}$ - нет оснований отвергнуть нулевую гипотезу. Если $|T| \ge t_{kp}$ - нулевую гипотезу отвергают.

Когда основная гипотеза отвергается на определенном уровне значимости, это значит, что выборочный коэффициент корреляции значимо отличается от нуля и, следовательно, величины X и Y коррелированны. Когда основная гипотеза принимается, это значит, что значение выборочного коэффициента корреляции не сильно отличается от нуля и является случайным.

<u>Уравнения регрессии</u>

Если дано распределение системы двух случайных величин X и Y, то *регрессией* Y на X называется любая функция, приближенно представляющая статистическую зависимость Y от X.

Случайные величины X и Y называются *линейно коррелированными*, если линии регрессии являются прямыми. Эти “прямые регрессии” задаются следующими уравнениями:
регрессия Y относительно X

$$y = Y_B + r_{XY} \ \frac{\sigma_Y}{\sigma_X}(x - X_B) \ ;$$

регрессия X относительно Y

$$x = X_B + r_{XY}\frac{\sigma_X}{\sigma_Y}(y - Y_B) \quad .$$

2. Содержание лабораторной работы

Дан протокол измерений случайного вектора (X,Y) . Необходимо:
1) построить эмпирические распределения в системе координат XY;
2) получить выборочные оценки характеристик распределения;
3) проверить гипотезу о значимости выборочного коэффициента корреляции;
4) составить уравнения прямых регрессий и построить эти прямые на одном рисунке с эмпирическими распределениями.

Примечание. В учебных целях протокол измерений двумерной нормальной совокупности (X,Y) с математическими ожиданиями a_X, a_Y и средними квадратическими отклонениями σ_X, σ_Y можно сгенерировать, используя формулы

$$x_i = (\sum_{j=1}^{12} r_j - 6)\,\sigma_X + a_X \quad , \qquad y_i = (\sum_{j=1}^{12} r_j - 6)\,\sigma_Y + a_Y \quad ,$$

где r_j – случайная величина, распределенная равномерно.

Пример выполнения лабораторной работы в MathCad

Лабораторная работа №2

Статистический анализ двумерных данных

Моделирование случайного вектора (X,Y),
имеющего двумерный нормальный закон распределения

$n := 325$ - количество реализаций случайной величины

$aX := 0.65$ - математическое ожидание составляющей X

$aY := 3.8$ - математическое ожидание составляющей Y

$\sigma X := 3.9$ - среднее квадратическое отклонение составляющей X

$\sigma Y := 2.335$ - среднее квадратическое отклонение составляющей Y

$rXY := -0.37$ - коэффициент корреляции

$i := 1..n$ ORIGIN := 1 - нижняя граница индексации массивов

$$\xi_i := \left(\sum_{k=1}^{12} rnd(1)\right) - 6$$

функция генерации случайных чисел *rnd* (x) при каждом обращении к ней возвращает случайное число с равномерным распределением на отрезке [0,x]

$$\eta_i := \left(\sum_{k=1}^{12} rnd(1)\right) - 6$$

$$X_i := \sigma X\cdot\xi_i + aX \qquad Y_i := \sigma Y\cdot\left(rXY\cdot\xi_i + \sqrt{1 - rXY^2}\cdot\eta_i\right) + aY$$

$$X = \begin{array}{|c|c|}\hline & 1 \\ \hline 1 & -5.248 \\ \hline 2 & 0.466 \\ \hline 3 & 5.017 \\ \hline 4 & 2.022 \\ \hline \end{array} \qquad Y = \begin{array}{|c|c|}\hline & 1 \\ \hline 1 & 7.464 \\ \hline 2 & 2.968 \\ \hline 3 & 0.046 \\ \hline 4 & 6.786 \\ \hline \end{array}$$

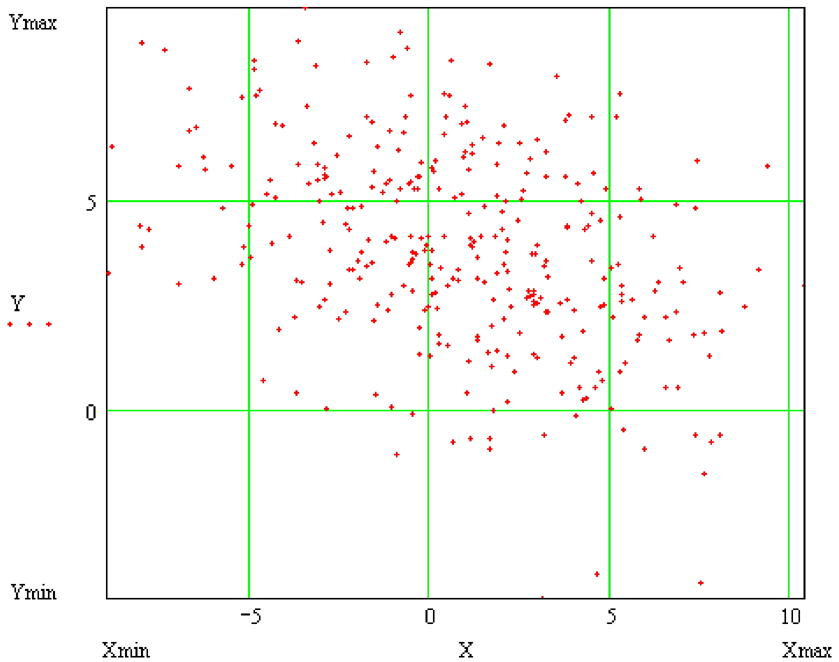
$$\vdots \qquad\qquad \vdots$$

$$\begin{array}{|c|c|}\hline 323 & 1.704 \\ \hline 324 & -1.715 \\ \hline 325 & 6.924 \\ \hline \end{array} \qquad \begin{array}{|c|c|}\hline 323 & 1.046 \\ \hline 324 & 4.062 \\ \hline 325 & 3.398 \\ \hline \end{array}$$

$Xmax := \max(X)$ $\quad Xmax = 10.421$ $\quad Xmin := \min(X)$ $\quad Xmin = -8.989$

$Ymax := \max(Y)$ $\quad Ymax = 9.633$ $\quad Ymin := \min(Y)$ $\quad Ymin = -4.459$

Построение рисунка (видео)



Нахождение выборочных оценок числовых характеристик распределения

Вычисление выборочного среднего

$$Xmean := \frac{\sum X}{n} \qquad Xmean = 0.822 \qquad Ymean := \frac{\sum Y}{n} \qquad Ymean = 3.854$$

Вычисление выборочной дисперсии

$$DX := \frac{\sum_{i=1}^{n} \left(X_i - Xmean\right)^2}{n} \qquad DX = 14.304 \qquad DY := \frac{\sum_{i=1}^{n} \left(Y_i - Ymean\right)^2}{n} \qquad DY = 5.643$$

Вычисление выборочного среднего квадратического отклонения (стандартного)

$$\sigma x := \sqrt{DX} \qquad \sigma x = 3.782 \qquad \sigma y := \sqrt{DY} \qquad \sigma y = 2.376$$

Вычисление выборочного коэффициента корреляции

$$rxy := \frac{\frac{1}{n}\cdot\sum_{i=1}^{n}\left[\left(X_i - Xmean\right)\cdot\left(Y_i - Ymean\right)\right]}{\sigma x\cdot\sigma y} \qquad rxy = -0.392$$

Вычисление выборочного коэффициента корреляции с помощью встроенной функции MathCAD

$rxy := corr(X,Y)$ функция corr(X,Y) возвращает скаляр - коэффициент корреляции

$rxy = -0.392$

Проверка гипотезы о значимости выборочного коэффициента корреляции

Проверка нулевой гипотезы H_0: о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: о неравенстве нулю генерального коэффициента корреляции

$\alpha := 0.05$ - уровень значимости

$$T := \frac{rxy\cdot\sqrt{n-2}}{\sqrt{1-rxy^2}}$$ - выборочное значение критерия $T = -7.647$

$$tkp := qt\left(1-\frac{\alpha}{2}, n-2\right)$$

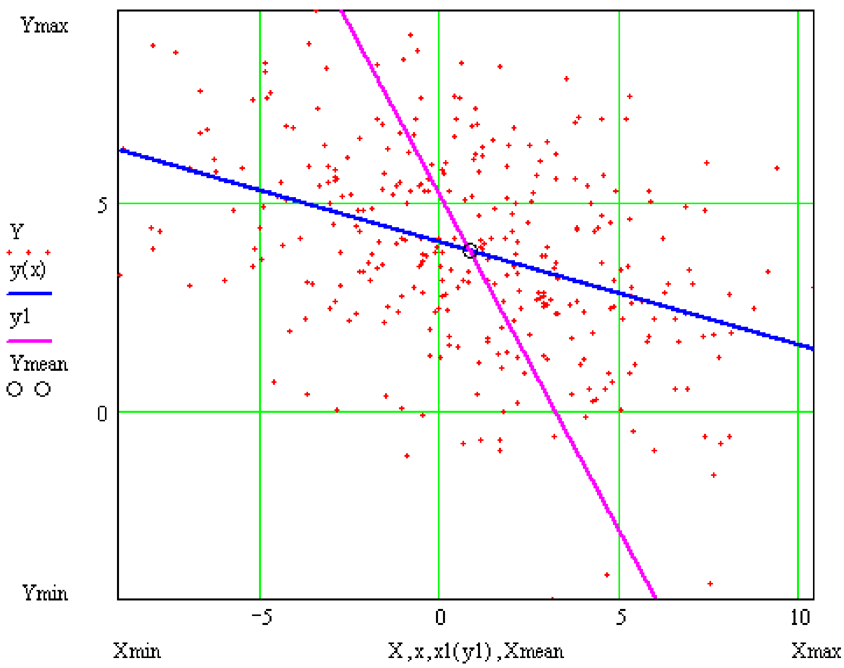
функция qt(p,d) возвращает квантиль обратного распределения Стьюдента (d - определяет степени свободы, p - вероятность)

$tkp = 1.967$ - критическая точка

Поскольку |T| > tkp - нулевую гипотезу отвергаем, т.е.выборочный коэффициент корреляции значимо отличается от нуля, следовательно величины X и Y коррелированы.

Нахождение эмпирических уравнений регрессии Y на X и X на Y

$$y(x) := Ymean + rxy\cdot\frac{\sigma y}{\sigma x}\cdot(x - Xmean) \qquad x1(y1) := Xmean + rxy\cdot\frac{\sigma x}{\sigma y}\cdot(y1 - Ymean)$$



Задания к лабораторной работе №2

Количество реализаций случайной величины:	n = № группы + № варианта
Математическое ожидание составляющей X:	aX = № варианта - 10
Математическое ожидание составляющей Y:	aY = № варианта - 5
Среднее квадратическое отклонение составляющей X:	σX = (50 - № варианта) * 0.1
Среднее квадратическое отклонение составляющей Y:	σY = (30 - № варианта) * 0.1
Коэффициент корреляции	rXY = - 0.5 + № варианта / 25

Наверх

Библиографический список

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика. -М.: Высшая школа, 2002.

2. *Дьяконов В.П., Абраменкова И.В.* MathCAD 7.0 в математике, физике и в Internet. -М.: Нолидж, 1999.

3. Теория вероятностей и математическая статистика: метод. указания к курсовой работе / сост.: О.М. Карпилова, О.П. Шевченко. - Самара: Самар. гос. аэрокосм. ун-т. 2004.

Наверх

Приложение. Некоторые приёмы работы в системе MathCad

Документ программы MathCad называется рабочим листом. Он содержит объекты: *формулы и текстовые блоки*. В ходе расчетов формулы обрабатываются последовательно, слева направо и сверху вниз, а текстовые блоки игнорируются.

Ввод информации осуществляется в месте расположения курсора. Программа MathCad использует три вида курсоров. Если ни один объект не выбран, используется *крестообразный курсор*, определяющий место создания следующего объекта. При вводе формул используется *уголковый курсор*, указывающий текущий элемент выражения. При вводе данных в текстовый блок применяется *текстовый курсор* в виде вертикальной черты.

Ввод текста

Текст, помещенный в рабочий лист, содержит комментарии и описания и предназначен для ознакомления, а не для использования в расчетах. Программа MathCad определяет назначение текущего блока автоматически при первом нажатии клавиши ПРОБЕЛ. Если введенный текст не может быть интерпретирован как формула, блок преобразуется в текстовый и последующие данные рассматриваются как текст. Создать текстовый блок без использования автоматических средств позволяет команда Insert - Text Region (Вставка - Текстовый блок). Иногда требуется встроить формулу внутрь текстового блока. Для этого служит команда Insert - Math Region (Вставка - Формула).

Ввод текста (видео)

Ввод формул

Формулы – основные объекты рабочего листа. Новый объект по умолчанию является формулой. Чтобы начать ввод формулы, надо установить крестообразный курсор в нужное место и начать ввод букв, цифр, знаков операций. Для управления порядком операций используют скобки, которые можно вводить вручную. Уголковый курсор позволяет автоматизировать такие действия. Чтобы выделить элементы формулы, которые в рамках операции должны рассматриваться как единое целое, используют клавишу ПРОБЕЛ . При каждом ее нажатии уголковый курсор “расширяется”, охватывая элементы формулы, примыкающие к данному. После ввода знака операции элементы в пределах уголкового курсора автоматически заключаются в скобки.

Элементы формул можно вводить с клавиатуры или с помощью панелей управления. Панели управления открываются с помощью меню Viev (Вид).

Введенное выражение обычно вычисляют или присваивают переменной. Знак присваивания изображается как “:=”, а вводится путем нажатия сочетания клавиш SHIFT + : или кнопки соответствующей панели управления. Знак вычисления вводится символом “ = ”.

Ввод формул (видео)

Построение графиков

Чтобы построить двумерный график в координатных осях X - Y , надо дать команду Format -Graph - X-Y Plot (Вставка - График - Декартовы координаты). В области размещения графика находятся заполнители для указания отображаемых выражений и диапазона изменения величин. Граничные значения по осям выбираются автоматически, но их можно задать и вручную.

В одной графической области можно построить несколько графиков. Для этого надо у соответствующей оси перечислить несколько выражений через запятую.

Разные кривые изображаются разным цветом, а для форматирования графика надо дважды щелкнуть по области графика. Для управления отображением построенных линий служит вкладка Traces (Линии) в открывающемся диалоговом окне. Текущий формат каждой линии приведен в списке, а под списком расположены элементы управления, позволяющие изменять формат. Поле Legend Label (Описание) задает описание линии, которое отображается в графической области только при сбросе флажка Hide Legend(Скрыть описание). Список Symbol (Символ) позволяет выбрать маркеры для отдельных точек, список Line (Тип линии) задает тип линии, список Color (Цвет) – цвет. Список Type (Тип) определяет способ связи отдельных точек, а список Width (Толщина) – толщину линии.

Построение графика (видео)

Организация гипертекстовой ссылки

Гипертекстовая ссылка – это выделенное слово или фраза, активизация которой вызывает переход к какому-либо объекту – например к новому документу или файлу. Гипертекстовая ссылка организуется следующим образом. Вначале текст ссылки выделяется, а затем нажимается кнопка инструментальной панели Insert - Hyperlink В появившемся простом окне нужно указать полное (с путем) имя файла, который будет загружаться и отображаться в момент активизации фрагмента – гиперссылки.

Наверх

Редактор Т.К. Кретинина
Компьютерная верстка И.И. Спиридонава

Подписано для тиражирования 30.09.2020
Объём издания 121 Мб.
Тираж 10 дисков

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)
443086, САМАРА, МОСКОВСКОЕ ШОССЕ, 34.

Изд-во Самарского университета.
443086, Самара, Московское шоссе, 34.

Наверх