

Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Локтионова Оксана Геннадьевна
Должность: проректор по учебной работе
Дата подписания: 10.11.2023 03:15:07
Уникальный программный ключ:
0b817ca911e6668abb13a5d426d730e5f1c11eabbf73e943df4a4851fd356d089

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра биомедицинской инженерии



ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Методические указания к лабораторным работам по дисциплине
«Математическая статистика»

Курск 2017

УДК 004.93:61

Составители: О.В. Шаталова, К.Д.А. Кассим.

Рецензент

Доктор технических наук, профессор Р.А. Томакова

Основы математической статистики: методические указания к лабораторным работам / Юго-Зап. гос. ун-т; сост.: О.В. Шаталова, К.Д.А. Кассим. Курск, 2017. 124 с.

Предназначено для студентов специальности 30.05.03 «Медицинская кибернетика» по дисциплине «Математическая статистика». Может быть использована аспирантами, обучающимися по направлениям 05.11.13 – Системный анализ, управление и обработка информации и 05.11.17 – Приборы, системы и изделия медицинского назначения.

Текст печатается в авторской редакции

Подписано в печать . Формат 60×84 1/16. Бумага офсетная.
Усл. печ. л. 7,21. Уч.-изд. л. 6,53. Тираж 100 экз. Заказ .
Юго-Западный государственный университет.
305040, г. Курск, ул. 50 лет Октября, 94.

Семестр 3

Лабораторная работа №1 «Моделирование случайных чисел с заданным законом распределения»

Целью работы является 1) практическое ознакомление с алгоритмами моделирования случайных чисел с заданным законом распределения; 2) изучение основных способов статистической оценки характеристик случайных чисел.

Краткие теоретические сведения

Дискретные случайные величины

Слова "случайная величина" в обыденном смысле употребляют тогда, когда хотят подчеркнуть, что неизвестно, каким будет конкретное значение этой величины. Причем иногда за этими словами скрывается просто незнание, какова эта величина.

Математик употребляет эти же слова "случайная величина", вкладывая в них определенное содержание.

«Действительно, - говорит он, - мы не знаем, какое значение примет эта величина в данном конкретном случае, но мы знаем, какие значения она может принимать, и знаем, каковы вероятности тех или иных значений. На основании этих данных мы не можем точно предсказать результат одного испытания, связанного с этой случайной величиной, но можем весьма надежно предсказать совокупность результатов большого числа испытаний. Чем больше испытаний, тем точнее будут наши предсказания».

Итак, чтобы задать случайную величину, надо указать, какие значения она может принимать, и каковы вероятности этих значений.

Случайная величина X называется дискретной, если она может принимать дискретное множество значений x_1, x_2, \dots, x_n .

Формально случайная дискретная величина X определяется таблицей

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}. \quad (1.1)$$

где x_1, x_2, \dots, x_n - возможные значения величины X ;

p_1, p_2, \dots, p_n - соответствующие вероятности.

Точнее говоря, вероятность $P\{X = x_i\}$ того, что случайная величина X примет значение x_i , равна:

$$P\{X = x_i\} = p_i. \quad (1.2)$$

Таблица (1) называется распределением случайной дискретной величины.

Числа x_1, x_2, \dots, x_n могут быть вообще говоря, любыми. Однако вероятности p_1, p_2, \dots, p_n должны удовлетворять двум условиям:

$$p_i > 0 \quad (1.3)$$

и

$$p_1 + p_2 + \dots + p_n = 1. \quad (1.4)$$

Последнее условие означает, что X обязана в каждом случае принять одно из значений x_1, x_2, \dots, x_n .

Кроме распределения случайной величины, которая является исчерпывающей характеристикой, вводятся числовые характеристики, основными среди которых являются математическое ожидание и дисперсия.

Получение случайных величин на ЭВМ

Сама постановка вопроса "получение случайных чисел на ЭВМ" иногда вызывает недоумение: ведь все, что делает компьютер, должно быть заранее запрограммировано; откуда же может появиться случайность?

Специалисты считают, что в этом вопросе есть определенные трудности, но они относятся скорее к философии, так что мы на них останавливаться не будем. Отметим лишь, что случайные

величины, о которых шла речь в предыдущем разделе это идеальные математические понятия.

Вопрос о том, можно ли с их помощью описать какое-либо явление природы, решается опытным путем. Такое описание всегда является приближенным. Более того, случайная величина, которая вполне удовлетворительно описывает какую-то физическую величину в одном классе явлений, может оказаться плохой характеристикой этой же величины при исследовании других явлений. Точно так же дорога, которую на карте страны можно считать прямой (идеальной математической прямой "без ширины"), становится полосой с изгибами на крупномасштабном плане населенного пункта.

Обычно различают три способа получения случайных величин:

- из заранее составленных таблиц случайных чисел;
- физические генераторы случайных чисел;
- с помощью формул (генераторов или датчиков) псевдослучайных чисел.

Поскольку "качество" используемых в имитационном моделировании случайных чисел проверяется с помощью специальных тестов, можно не интересоваться тем, как эти числа получены: лишь бы они удовлетворяли принятой системе тестов.

Числа, получаемые по какой-либо формуле и имитирующие значения случайной величины X , называются псевдослучайными числами. Под словом "имитирующие" подразумевается, что эти числа удовлетворяют ряду тестов так, как если бы они были значениями этой случайной величины.

Основой или «сырьем» для моделирования случайных величин с заданным законом распределения являются так называемые базовые случайные числа. Совокупность $\{R_i\}, i = 1, 2, \dots$ независимых равномерно распределенных на отрезке $[0, 1]$ случайных величин называется последовательностью базовых случайных чисел.

Мы называем эти числа псевдослучайными потому, что фактически они остаются полностью детерминированными в том смысле, что если каждое обращение к соответствующей формуле (точнее, к алгоритму) начинается с одними и теми же исходными данными (константами и начальными значениями), то на выходе получаются одинаковые последовательности чисел R .

В настоящее время почти все стандартные библиотечные программы вычисления равномерных случайных чисел основаны на конгруэнтных методах, разработанных Лемером.

Основная формула мультипликативного конгруэнтного метода Лемера имеет вид:

$$R_{i+1} = aR_i \pmod{m}, \quad (1.5)$$

где a и m – неотрицательные целые числа.

Согласно этому выражению, нужно взять случайное число R_i , умножить его на постоянный коэффициент a и взять модуль полученного числа m (т.е. разделить на aR_i и остаток считать как R_{i+1}). Поэтому для вычисления (или генерирования) последовательности R_i нам необходимы начальные значения R_0 , множитель a и модуль m . Выбираются a , R_0 и m так, чтобы обеспечить максимальную длину (или, как говорят, период) неповторяющейся последовательности R_i и минимальную корреляцию между генерируемыми числами.

На рисунке 1.1 показан фрагмент среды MathCad, на котором проиллюстрирована математическая реализация этого метода.

Переменной A присваивается значение $a = 5^{13} = 1220703125$, переменной m – значение $m = 2^{31} + 1 = 2147483649$. Функция $\text{mod}(x_1, x_2)$ вычисляет остаток от целочисленного деления первого аргумента во второй. Получаем последовательность $\{X\}$ псевдослучайных чисел, равномерно распределенных от 0 до m . Делим каждый член этой последовательности на m , получаем базовую последовательность $\{R_i\}$ – числа равномерно распределенные от 0 до 1.

Методы генерации псевдослучайных чисел с заданным законом распределения

Базовые случайные числа позволяют генерировать новые случайные последовательности, подчиняющиеся любому закону распределения.

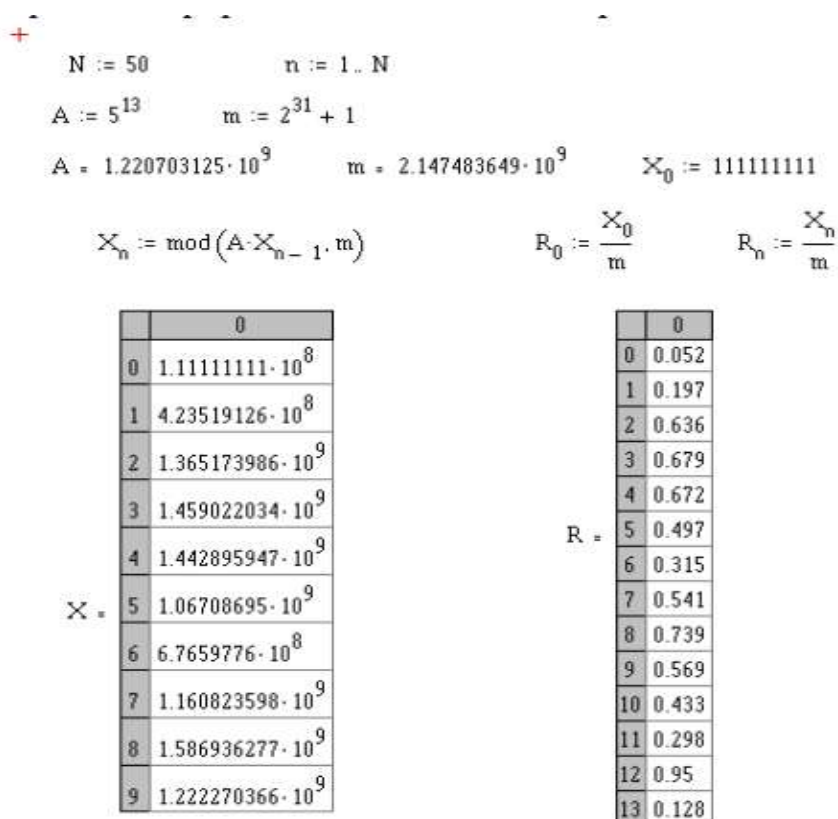


Рисунок 1.1 – Моделирование базовой последовательности мультипликативным конгруэнтным методом. Фрагмент среды MathCad

Существует два основных пути преобразования базовых случайных чисел $\{R_i\}$, в случайные числа $\{y_i\}$, распределенные по заданному закону распределения.

Один из них, который называется методом инверсии, состоит в реализации определенных арифметических операций над базовым числом R_i , чтобы получить y_i .

Второй метод основывается на моделировании условий соответствующей предельной теоремы теории вероятностей. Кроме указанных двух основных подходов можно также выделить эвристические способы генерирования случайных чисел.

*Метод инверсии**Моделирование случайной величины, равномерной на (a, b)*

Предположим, что нам необходимо составить программу для моделирования входного потока заявок распределенного по равномерному закону в интервале (a, b).

Уравнение метода инверсии (1.6) для рассматриваемого случая выглядит так:

$$\int_a^y \frac{dy}{b-a} = R, \quad (1.6)$$

где R – равномерно распределенное случайное число на (0; 1), т.е. базовое число. Это интегральное уравнение решается легко и ответ ясен:

$$\frac{y-a}{b-a} = R. \quad (1.7)$$

Отсюда мы имеем явное выражение для y :

$$y = a + R(b-a), \quad (1.8)$$

где R – как обычно, базовое случайное число.

Моделирование экспоненциальной случайной величины

Как известно, случайная величина x , распределенная по экспоненциальному закону описывается следующей плотностью распределения:

$$p(x) = \lambda e^{-\lambda x} \quad (1.9)$$

На рисунке 1.2 построены графики экспоненциальных плотностей распределения при различных параметрах λ .

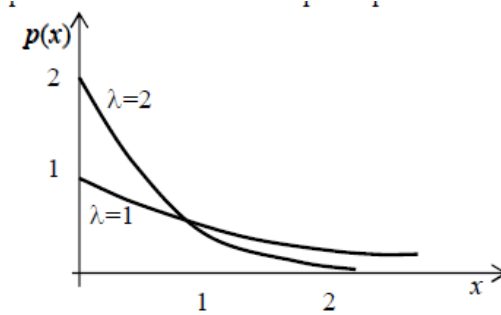


Рисунок 1.2 – Экспоненциальная плотность вероятностей $p(x) = \lambda e^{-\lambda x}$ с разными значениями параметра λ

Экспоненциальному распределению, как правило, подчиняется случайный интервал времени τ между поступлениями заявок в систему массового обслуживания. Поэтому весьма важно уметь моделировать потоки заявок разной интенсивности λ .

Напомним, что математическое $M[\tau]$ ожидание экспоненциально распределенной случайной величины τ равно:

$$M[\tau] = 1 / \lambda,$$

$$\text{а дисперсия: } D[\tau] = 1 / \lambda^2.$$

Чтобы найти алгоритм имитации экспоненциально распределенных чисел τ , применим метод инверсии:

$$\int_0^{\tau} \lambda e^{-\lambda x} = R \quad (1.10)$$

$$1 - e^{-\lambda \tau} = R, \quad (1.11)$$

откуда

$$\tau = -\frac{1}{\lambda} \ln(1 - R), \quad (1.12)$$

но, поскольку случайная величина $(1 - R)$ распределена точно так же, как R , и находится в том же интервале $(0, 1)$, то (1.12) можно заменить на более удобную формулу:

$$\tau = -\frac{1}{\lambda} \ln R, \quad (1.13)$$

что дает искомый ответ.

Моделирование нормальной случайной величины на основе центральной предельной теоремы

Нормальное (или гауссово) распределение (рисунок 1.3) - это, несомненно, один из наиболее важных и часто используемых в имитационном моделировании видов непрерывных распределений.

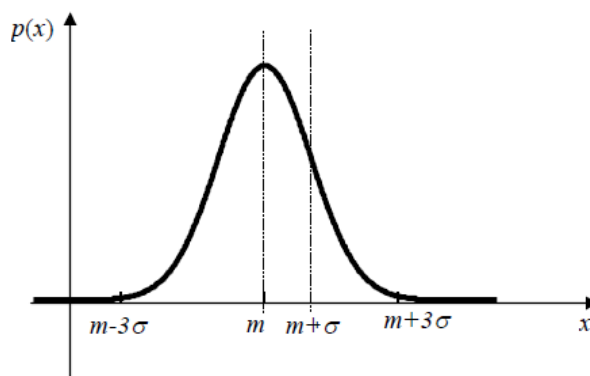


Рисунок 1.3 – Нормальная (гауссовская) плотность вероятностей

Плотность вероятности нормально распределенной случайной величины записывается так:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad (1.14)$$

где m и σ - параметры нормального распределения $m = M_x$ - математическое ожидание; σ - среднеквадратическое отклонение.

Интегральная функция распределения нормальной случайной величины равна

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(x-m)^2}{2\sigma^2}} dx. \quad (1.15)$$

Поэтому алгоритмы моделирования нормальных случайных чисел базируются на предельных теоремах теории вероятностей. Центральная предельная теорема говорит о том, что сумма n одинаково распределенных независимых случайных величин x со средним M_x и дисперсией D_x стремится к нормально распределенной величине с параметрами nM_x и nD_x при бесконечном увеличении n . Следствием теоремы является, в частности, и то, что для получения нормальной выборки, можно воспользоваться базовыми случайными числами R . Идея алгоритма состоит в следующем. Определим новую случайную величину s в виде суммы базовых чисел R_i , ($i=1, 2, 3, \dots, n$):

$$s = R_1 + R_2 + \dots + R_n. \quad (1.16)$$

Тогда, согласно утверждению центральной предельной теоремы, случайная величина s является асимптотически нормальной величиной с математическим ожиданием M_s и дисперсией D_s равными соответственно:

$$M_s = n / 2, \quad (1.17)$$

и

$$D_s = n / 12. \quad (1.18)$$

Для практического использования формула (1.16) неудобна (поясните почему), поэтому введем вспомогательную случайную величину z равную

$$z = \frac{(s - n/2)}{\sqrt{n/12}} \quad (1.19)$$

Из (1.19) следует, что z – случайная величина, распределенная по нормальному закону с нулевым средним и единичной дисперсией. Тогда для любого нормального распределения со средним μ и дисперсией σ^2 случайное отклонение y , соответствующее указанным выше n случайным числам, получается из формулы

$$\frac{(y - \mu)}{\sigma} = z = \frac{\left(s - \frac{n}{2}\right)}{\sqrt{n/12}} \quad (1.20)$$

Следовательно,

$$y = \mu + \frac{\sigma(s + n/2)}{\sqrt{n/12}} = \mu + \frac{\sigma}{\sqrt{n/12}} \left(\sum_{i=1}^n R_i + n/2 \right). \quad (1.21)$$

Согласно той же предельной теореме, нормальность достигается быстро даже при сравнительно небольших значениях n . В практических задач n обычно принимается равным 12. При этом последняя формула упрощается и принимает вид:

$$y = \mu + \sigma \left(\sum_{i=1}^{12} R_i + 6 \right). \quad (1.22)$$

Формула (1.22) и дает алгоритм моделирования нормальных случайных чисел с требуемыми параметрами μ и σ .

Описанный метод считается малоэффективным, так как требует генерации нескольких случайных базовых чисел R для получения одного нормального выборочного значения y .

Оценка статистических характеристик случайных величин

При решении многих прикладных задач необходимые вероятностные характеристики соответствующих случайных величин неизвестны исследователю и должны определяться по экспериментальным данным. Такое статистическое описание результатов наблюдений, построение и проверка различных математических моделей, использующих понятие вероятности, составляют основное содержание математической статистики. Фундаментальными понятиями статистической теории являются понятия генеральной совокупности и выборки.

Генеральная совокупность - совокупность всех мыслимых (возможных) результатов наблюдений над случайной величиной, которые в принципе могут быть проведены при данных условиях.

Содержательный смысл этого понятия состоит в том, что предполагается существование некоторых вполне определенных свойств, неслучайных закономерностей, присущих данной совокупности. Эти свойства и должны быть определены исследователем. Фактически эти свойства являются объективным отображением вероятностных свойств изучаемого объекта, которые могут быть охарактеризованы с помощью соответствующих законов распределения вероятностей или связанных с ними числовых параметров. Как правило, считается, что указанные свойства не изменяются во времени.

Выборка - это конечный набор x_1, x_2, \dots, x_N значений случайной величины, полученный в результате наблюдений. Число элементов N выборки называется ее объемом или размером.

Заметим, что выборка может иметь и совпадающие значения x_i случайной величины X . Интуитивно понятно, что чем больше объем выборки, тем более точно она должна отражать статистические свойства случайной величины. Определение. Выборка называется репрезентативной (представительной), если она достаточно полно характеризует свойства генеральной совокупности.

Для обеспечения репрезентативности выборки чаще всего используют метод случайного выбора элементов. Предполагается, что при таком выборе каждая возможная выборка фиксированного объема имеет одну и ту же вероятность выбора, а последовательные наблюдения взаимно независимы.

Оцениванием в статистике называется указание приближенного значения интересующего нас параметра (или функции от некоторых параметров) на основе наблюдаемых (экспериментальных) данных, представленных в виде выборки ограниченного объема.

Оценка - это правило вычисления приближенного значения параметра (или функции от некоторых параметров) по наблюдаемым данным.

При многократном извлечении выборок одного и того же объема и последующем нахождении множества оценок одного и того же параметра получаются различные числовые значения этих

оценок, изменяющиеся от одной выборки к другой случайным образом.

Иными словами, любая оценка произвольного параметра есть случайная величина. В этом состоит принципиальное отличие оценки от самого параметра.

Элементарные статистические процедуры

В случае гауссовского распределения для истинного математического ожидания m_x существует его оценка \tilde{m}_x , вычисляемая по выборке объема n случайной величины $X = (x_1, x_2, \dots, x_n)$:

$$\tilde{m}_x = \frac{1}{n} \sum_{i=1}^n x_i. \quad (1.23)$$

Для истинной дисперсии D_x (характеристика рассеивания случайной величины око ее математического ожидания) ее оценка \tilde{D}_x при известном математическом ожидании m_x вычисляется так:

$$\tilde{D}_x = \tilde{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_x)^2, \quad (1.24)$$

где σ_x является среднеквадратическим отклонением.

В случае неизвестного математического ожидания дисперсию \tilde{D}_x нужно вычислять по формуле:

$$\tilde{D}_x = \tilde{\sigma}_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) \quad (1.25)$$

Приведенные оценки являются несмещенными и асимптотически эффективными.

Не будем забывать о том, что оценки сами являются случайными величинами, а значит, обладают некоторым разбросом, который оценивается дисперсией. Дисперсии $D \{ \}$ вышеуказанных оценок соответственно таковы:

дисперсия оценки среднего:

$$D\{\tilde{m}_x\} = \sigma_x^2 / n; \quad (1.26)$$

дисперсия оценки \tilde{D}_x в случае известного математического ожидания:

$$D\{\tilde{D}_x\} = 2\sigma_x^4 / n; \quad (1.27)$$

в случае, если не известно математическое ожидание:

$$D\{\tilde{D}_x\} = 2\sigma_x^4 / (n - 1). \quad (1.28)$$

После вычисления точечных оценок обычно переходят к построению вариационного ряда, диаграммы накопленных частот и гистограммы выборки.

Пусть имеется набор (выборка) экспериментальных данных x_1, x_2, \dots, x_n . Вариационный ряд (или ряд распределения) z_1, z_2, \dots, z_n получают из исходных данных путем расположения x_m ($m = 1, 2, \dots, n$) в порядке возрастания от x_{\min} до x_{\max} так, чтобы $x_{\min} = z_1 \leq z_2 \leq \dots \leq z_n = x_{\max}$.

Диаграмма накопленных частот $P_n(x)$ является эмпирическим аналогом интегрального закона распределения $P(x)$ и ее строят в соответствии с формулой

$$P_n(x) = \sum_{j=1}^{\mu_n(x)} \frac{1}{n} \quad (1.29)$$

где $\mu_n(x)$ - число элементов в выборке, для которых значение $x_j < x$.

Практически это делается так. На оси абсцисс указывают значения наблюдений x_{\min} (или z_1). Значение по оси ординат равно нулю левее точки x_{\min} ; в точке x_{\min} и далее во всех других точках x_m диаграмма имеет скачок, равный $1/n$. Если существует λ совпадающих значений x_m , то в этом месте на диаграмме

происходит скачок, равный λ/n . Ясно, что для величин $x > x_{\max}$ значение диаграммы накопленных частот равно 1. Отметим, что если $n \rightarrow \infty$, то $P_n(x) \rightarrow P(x)$.

Пример. Пусть имеется выборка объема 5:

$$x_1 = 5; x_2 = 2; x_3 = 4; x_4 = 5; x_5 = 7.$$

Вариационный ряд для данной выборки будет таким:

$$z_1 = 2; z_2 = 4; z_3 = 5; z_4 = 5; z_5 = 7.$$

Соответствующая диаграмма накопительных частот представлена на рисунке 1.4.

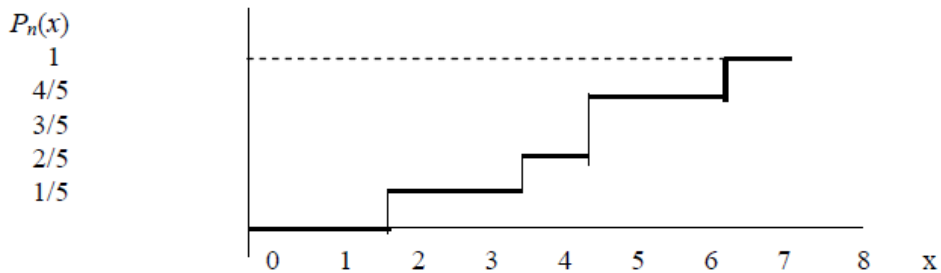


Рисунок 1.4 – Диаграмма накопленных частот

Гистограмма $f_n(x)$ является эмпирическим аналогом функции плотности распределения $f(x)$. Последовательность построения гистограммы такова.

По оценочной формуле находят предварительное количество квантов (интервалов) K , на которое нужно разбить на ось Ox :

$$K=1+3.2\lg n;$$

найденное значение K округляется до ближайшего целого числа.

Формула для K является эмпирической, что означает примерное значение. Ее величина связана с той целью, чтобы в один квант попало хотя бы одно выборочное значение x_i . Интересно, что зависимость количества интервалов K от объема выборки n равна (таблица 1.1)

Таблица 1.1 - Зависимость количества интервалов K от объема выборки n

n	100	200	300	500	1000
K	7	8	9	10	11

Далее определяют длину каждого кванта (интервала):

$$\Delta x = (x_{\max} - x_{\min}) / K,$$

которую для удобства построений можно несколько округлить в ту или иную сторону.

Середину области изменения выборки (центр распределения)

$$(x_{\max} + x_{\min}) / 2$$

принимают за центр некоторого интервала, после чего находят границы и окончательное количество указанных интервалов так, чтобы в совокупности они перекрывали всю область от x_{\min} и x_{\max} .

Далее подсчитывают количество наблюдений n_m , попавшее в каждый квант: n_m равно числу членов вариационного ряда, для которых справедливо неравенство

$$x_m \leq z_1 < x_m + \Delta x.$$

Здесь x_m и $x_m + \Delta x$ - границы m -го интервала. Отметим, что при использовании этой формулы значения z_1 , попавшие на границу между $(m-1)$ и m -м интервалами, относят к m -му интервалу.

Далее подсчитывают относительное количество (относительную частоту) наблюдений n_m / n , попавших в данный квант.

Наконец, строят гистограмму, представляющую собой ступенчатую кривую, значение которой на m -й интервале $(x_m, x_m + \Delta)$ ($m=1, 2, \dots, K$) постоянно и равно n_m / n , или с учетом

условия $\int_{-\infty}^{\infty} p_n(z) dz = 1$, равно $(n_m / n) \cdot \Delta x$.

Практическая часть

1. Смоделируйте базовую последовательность объемом $N=1000$ мультипликативным конгруэнтным методом.

2. Напишите одну комплексную программу моделирования выборки случайных чисел, оценки математического ожидания и дисперсии для всех ниже перечисленных распределений:

а) равномерное на интервале (a, b) ;

б) экспоненциальное с параметром λ ;

с) нормальное с параметрами μ и σ , используя метод суммирования или какой-либо один из эвристических методов.

3. Самостоятельно задав параметры распределений, промоделируйте выборки всех вышеуказанных распределений. Объем каждой выборки принять $N=1000$.

4. Вычислите оценки математического ожидания и дисперсии каждой из полученных в п. 2 последовательностей случайных чисел для следующих объемов выборки $N_1=10$, $N_2=20$, $N_3=50$, $N_4=100$ и $N_5=1000$. Сравните полученные оценки с заданными в пп. 2 параметрами. Постройте графики зависимостей оценок от объема выборки. Оцените относительные погрешности для какой-либо одной выборки.

5. Для всех выборок разных распределений, рассчитайте и постройте:

- диаграммы накопленных частот;
- гистограммы распределений.

6. Сравните гистограммы с графиками теоретических распределений. Для сравнения постройте также вышеуказанные гистограммы на одном графике с функциями распределений.

Варианты лабораторной работы. Параметры распределений

№ вар	a	b	λ	μ	σ
1	3	11	1	0	1
2	10	16	5	0	2
3	1	6	1	0	3
4	4	7	2	0	4
5	9	17	4	0	5
6	1	11	1	0	6
7	8	13	4	1	1

8	3	9	1	1	2
9	2	7	1	1	3
10	9	14	5	1	4
11	6	11	3	1	5
12	3	7	2	1	6
13	7	12	3	2	1
14	5	12	3	2	2
15	7	13	4	2	3
16	3	12	1	2	4
17	2	12	2	2	5
18	4	13	2	10	6
19	4	8	5	10	1
20	10	14	4	10	2
21	9	16	1	10	3
22	0	7	3	10	4
23	6	6	5	10	5
24	10	11	3	10	6
25	6	11	4	5	1
26	9	18	2	5	2
27	4	7	1	5	3
28	2	6	2	5	4
29	4	8	5	5	5
30	9	13	4	5	6

Пример построения гистограммы и диаграммы накопленных частот в среде MathCad 5.0 для базовой последовательности R . Оценки математического ожидания и дисперсии приведены ниже на рисунках.

Вычисляем количество квантов (интервалов), на которые разбиваем ось ox

$$K := \text{Floor}(1 + 3.2 \cdot \log(N))$$

$$1 + 3.2 \cdot \log(N) = 10.6$$

$$K = 10$$

$$k := 1..K$$

границы области построения гистограммы

$$\max(R) = 0.999$$

$$\min(R) = 8.945 \cdot 10^{-4}$$

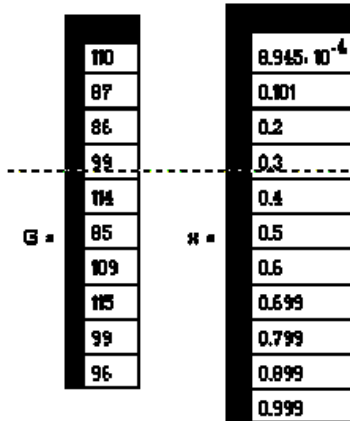
ширина одного интервала.

$$\Delta_k := \frac{\max(R) - \min(R)}{K}$$

массив границ интервалов

$$k_k := k \cdot \Delta_k + \min(R) \quad k_0 := \min(R)$$

$$G := \text{hst}(x, R)$$



Функция построения гистограммы. Вычисляет сколько значений массива R попадает в интервалы, определенные массивом x .

Возвращает массив, размерность которого на единицу меньше размерности массива x , поскольку количество интервалов на единицу меньше количества точек, ограничивающих эти интервалы

+

Рисунок 1.5 – Количество квантов, на которые разбивается ось Ox

Теоретическое значение закона распределения (плотности вероятности)

$$p(x) := \begin{cases} 0, & x < 0, \\ 1, & 0 < x < 1, \\ 0, & x > 1, \end{cases} \quad \text{ин} := -0.2, -0.19, \dots, 1.2$$

Теоретическое значение интегральной функции распределения

$$F(x) := \begin{cases} 0, & x < 0, \\ x, & 0 < x < 1, \\ 1, & x > 1, \end{cases}$$

График, на котором построена теоретическая функция распределения, совмещенная с гистограммой. Гистограмма нормирована, относительная частота попадания в интервал делится на ширину интервала, тогда высота столбика может быть соотнесена со значением теоретической функции распределения в пределах каждого интервала.

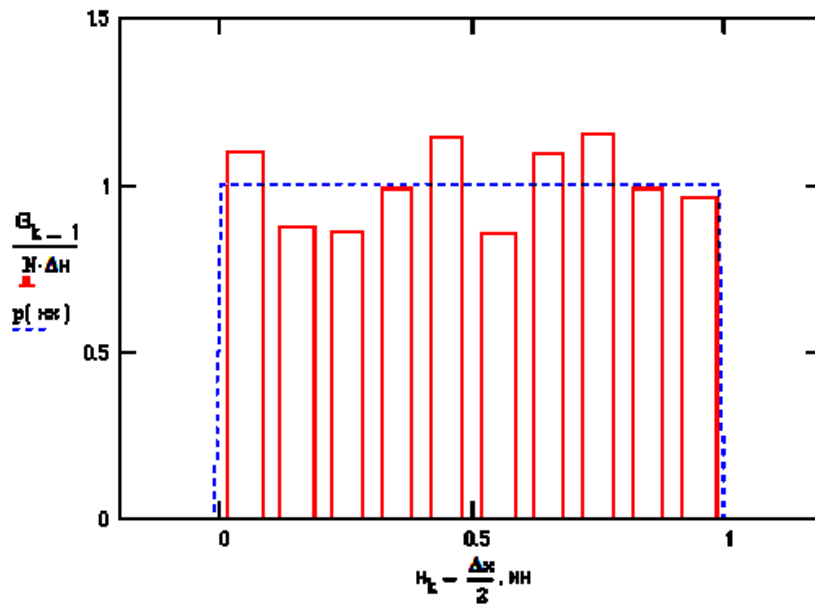


Рисунок 1.6 – Теоретическая функция распределения

$$G_k := \sum_{i=1}^k G_{i-1}$$

Вычисление массива значений для диаграммы накопленных частот

График, на котором диаграмма накопленных частот совмещена с интегральной функцией распределения

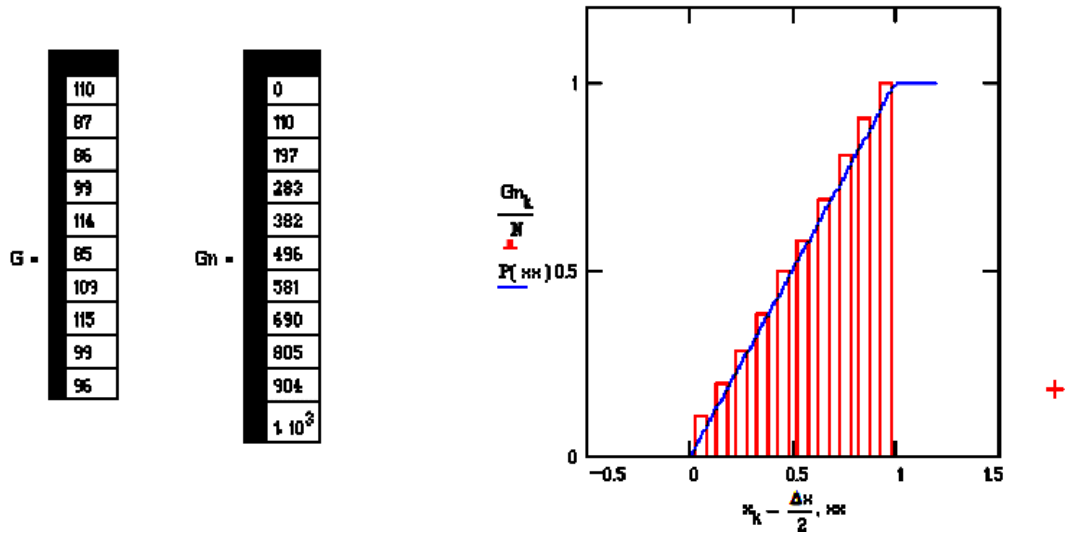


Рисунок 1.7 – Вычисление массива значений для диаграммы накопленных частот

оценка математического ожидания в зависимости от объема выборки

$$M(N, R) := \frac{1}{N} \sum_{i=1}^N R_i \quad M(10, R) = 0.528 \quad M(N, R) = 0.507 \quad MSR := M(N, R)$$

несмещенная оценка дисперсии в зависимости от объема выборки

$$D(N, R) := \frac{1}{N-1} \sum_{i=1}^N (R_i - MR)^2 \quad m := 2..N$$

графики данных зависимостей

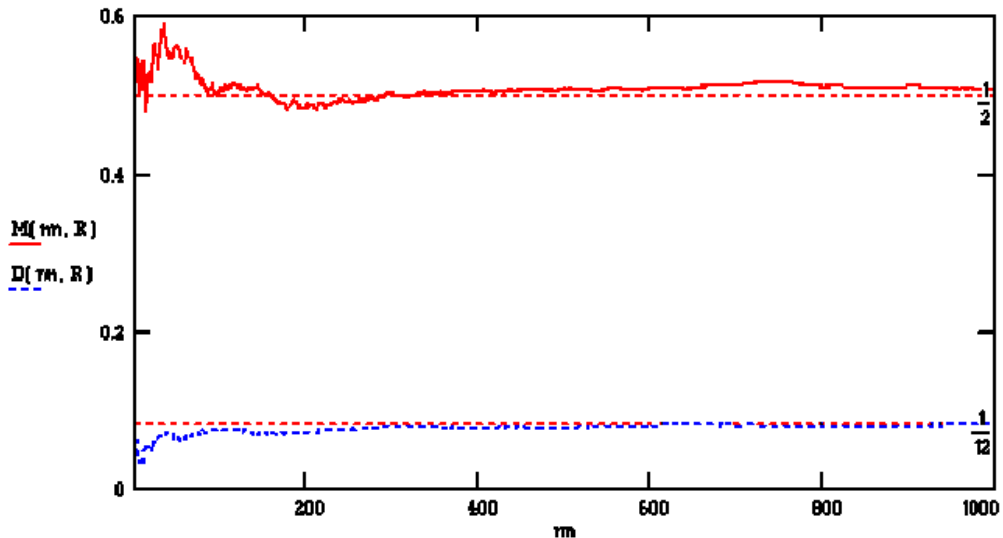


Рисунок 1.8 – Оценка математического ожидания в зависимости от объема выборки

Контрольные вопросы

1. Что такое распределение случайной дискретной величины?
2. Что такое дискретная случайная величина? Дайте развернутый ответ.
3. Каков алгоритм получения случайных величин на ЭВМ?
4. Раскройте понятие «методы генерации псевдослучайных чисел с заданным законом распределения».
5. Что такое метод инверсии?
6. Что значит «моделирование случайной величины равномерной на (a, b)»?
7. Что значит «моделирование экспоненциальной случайной величины»?
8. Что значит «моделирование нормальной случайной величины на основе центральной предельной теоремы»?

9. Что такое оценка статистических характеристик случайных величин?
10. Что такое элементарные статистические процедуры?

Лабораторная работа №2 **«Элементарные задачи математической статистики»**

Цель работы: Рассмотреть возможности MathCAD для решения элементарных задач математической статистики. Научиться использовать возможности MathCAD для ввода и вывода файловых данных. Познакомиться с расчетом основных выборочных характеристик в среде MathCAD. Научиться представлять графически выборку случайных величин в виде гистограмм и полигонов.

Краткие теоретические сведения

В большинстве статистических расчетов приходится иметь дело либо со случайными данными, полученными в ходе какого-либо эксперимента (которые выводятся из файла или печатаются непосредственно в документе), либо с результатами генерации случайных чисел.

Случайной выборкой называется случайный вектор, элементы которого независимы и одинаково распределены. Обычно под *выборкой* подразумевают результаты независимых измерений, которые проводятся в одинаковых условиях.

Ввод и вывод файлов данных

Важный компонент ввода-вывода — это ввод-вывод во внешние файлы. Ввод внешних данных в документы Mathcad применяется чаще вывода, поскольку Mathcad имеет гораздо лучшие возможности представления результатов расчетов, чем многие пользовательские программы. Для общения с внешними файлами данных в Mathcad имеется несколько разных способов.

Самый простой из них — использовать имеющееся семейство встроенных функций.

- READPRN (“file”) - чтение данных в матрицу из текстового файла;
- WRITEPRN (“file”) - запись данных в текстовый файл;
- APPENDPRN (“file”) - дозапись данных в существующий текстовый файл;
- file — путь к файлу.

Встроенная функция APPENDPRN может применяться и для создания нового файла. Иными словами, если файла с заданным именем не существовало, то он, после применения, будет создан и наполнен теми данными, которые Вами определены в документе.

Для удобства можно использовать функцию CWD - указания полигона, где необходимо создать файл или где находится считываемый файл.

Можно задавать как полный путь к файлу, например, C:\Мои документы, так и относительный, имея в виду, что он будет отсчитываться от папки, в которой находится файл с документом Mathcad. В качестве имени файла можно использовать русские буквы.

Пример 1: Запись данных в файл "da.ta.txt"

$x := \text{rpost}(N, a, \sigma)$ создание выборки случайных величин распределенных по нормальном закону;

CWD:= "D:\tmp\" устанавливается текущий рабочий каталог.

WRITEPRN(data.txt) :=x запись в файл «data»созданной ранее выборки x.

Моделирование выборок из стандартных распределений

Mathcad обладает богатой библиотекой встроенных функций, предназначенных для генерации выборок из генеральных совокупностей с наиболее распространенными стандартными распределениями.

Вставку рассмотренных ранее статистических функций в программы удобно осуществлять с помощью диалогового окна *Insert Function* (Вставка функции).

Для этого необходимо выполнить следующие действия:

1. Установить курсор на место вставки функции в документе.

2. Вызвать диалоговое окно *Insert Function* нажатием кнопки $f(x)$ на стандартной панели инструментов или командой меню *Insert/Function* (Вставка/Функция), или нажатием клавиш <Ctrl>+<E>.

3. Выбрать в списке *Function Category* (Категория функции) выберите одну из категорий статистических функций. Категория *Probability Density* (Плотность вероятности) содержит встроенные функции для плотности вероятности, Категория *Probability*

Distribution (Функция распределения) — для вставки функций или квантилей распределения, Категория *Random Numbers* (Случайные числа) — для вставки функции генерации случайных чисел.

4. Выбрать в списке *Function Name* (Имя функции) функцию, соответствующую требующемуся закону распределения. При выборе элемента списка в текстовом поле в нижней части окна будет появляться информация о назначении выбранной функции и ее параметрах.

5. Вставить выбранную функцию в документ нажатием кнопки "Ок".

Функции Mathcad для расчета численных характеристик

В Mathcad имеется ряд встроенных функций для расчетов числовых статистических характеристик рядов случайных данных.

- $\text{mean}(x)$ — выборочное среднее значение, оценка математического ожидания выборки;
- $\text{median}(x)$ — выборочная медиана (*median*) — значение аргумента, которое делит гистограмму плотности вероятностей на две равные части;
- $\text{var}(x)$ — выборочная дисперсия выборки (*variance*);
- $\text{stdev}(x)$ — среднеквадратичное (или "стандартное") отклонение выборки (*standard deviation*);
- $\text{max}(x)$, $\text{min}(x)$ — максимальное и минимальное значения выборки;
- $\text{mode}(x)$ — наиболее часто встречающееся значение выборки.

Построение гистограмм

Гистограммой называется график, аппроксимирующий по случайным данным плотность их распределения. При построении гистограммы область значений случайной величины (a , b) разбивается на некоторое количество *bin* сегментов, а затем подсчитывается процент попадания данных в каждый сегмент. Для построения гистограмм в Mathcad имеется несколько встроенных функций.

Гистограмма с произвольными сегментами разбиения

- $\text{hist}(\text{intvls}, x)$ - вектор частоты попадания данных в интервалы гистограммы;
- intvls - вектор, элементы которого задают сегменты построения гистограммы в порядке возрастания $a < \text{int } v_i < b$;
- x - вектор случайных данных.

Если вектор intvls имеет bin элементов, то и результат hist имеет столько же элементов.

1. Для построения гистограмм созданную случайную величину предварительно необходимо упорядочить. Для этого в Mathcad имеется встроенная функция.

$\text{sort}(x)$ - сортировка выборки в порядке возрастания;

Для того, чтобы построить гистограмму, нужно сначала сгруппировать выборочные данные, записанные в массиве x , и сохранить граничные очки интервалов группировки в векторе intvls , размерность которого равна числу интервалов.

2. Сформировать вектор intvls границ интервалов.

3. Определить процент попадания данных в каждый сегмент.

4. Построить гистограмму.

Пример 2. Построение гистограммы

$N:=1000$

$x:=\text{binom}(N, \Delta, 0,5)$

$\text{bin}:=30$; кол-во равных сегментов, на кот. разбивается весь диапазон

; определение границы интервала построения гистограммы

$\text{lower}:=\text{floor}(\text{min}(x))$: наибольшее целое число $\leq \text{min}(x)$

$\text{upper}:=\text{ceil}(\text{max}(x))$; наименьшее целое число $\geq \text{max}(x)$

$h := \frac{\text{upper} - \text{lower}}{\text{bin}}$; размер сегмента

$j:=0 \dots \text{bin}$; счетчик сегментов

$\text{int } j := \text{lower} + h \cdot j$

$f := \frac{1}{N \cdot h} \cdot \text{hist}(\text{int}, x)$; массив начальных точек каждого

сегмента

$int:=int+0.5h$; от левой границы каждого сегмента к его центру;
нормирование гистограммы для удобства отображения на одном графике вместе с плотностью распределения

В векторе *int* можно задать произвольные границы сегментов разбиения так, чтобы они имели разную ширину.

Недостаток упрощенной формы функции *hist* состоит в том, что необходимо дополнительно определять вектор сегментов построения гистограммы. От этого недостатка свободна функция *histogram*.

Гистограмма с разбиением на равные сегменты

histogram (bin, x) — матрица гистограммы размера $bin*2$, состоящая из столбца сегментов разбиения и столбца частоты попадания в них данных;

- *bin* — количество сегментов построения гистограммы;
- *x* — вектор случайных данных.

Пример 3. Построение гистограммы (упрощенный вариант)

```
N:=100; созданы выборки сл. величины
x:=exp(N, 1)
bin:=30; кол-во равных сегментов, на кот. разбивается
весь диапазон
f := histogram(bin, x)
```

Аргументы у обеих процедур *hist()* и *histogram()* одинаковы: первый определяет интервалы для создания гистограммы, а второй - это выборка, на основе которой строится гистограмма. Первый аргумент может быть либо вектором конечных точек интервалов для группировки данных выборки, либо целым числом, задающим число интервалов. В последнем случае весь диапазон значений в выборке разбивается на равные интервалы.

Создание графика гистограммы


Для того чтобы создать график в виде гистограммы необходимо:

1. Построить двумерный график, по оси x откладываются границы интегралов, по оси y частота(процент) попадания значения сл. величины в заданные интервалы.

2. Перейти в диалоговом окне *Formatting Currently Selected Graph* (Форматирование) выбранного графика (например, двойным щелчком мыши) в раздел *Traces* (Графики). Установить в поле *Type* (Тип) элемент списка *bar* (столбцы) или *solidbar* (гистограмма). Тип *solidbar* специально предназначен для гистограмм.

3. Нажать кнопку ОК.

Процедура *histogram()* инициализирует объект (матрицу), содержащий срединные точки интервалов гистограммы (первый столбец) и столбец частот, попадания в заданные интервалы.

Для построения таких графиков по оси x откладывается столбец срединные точки интервалов (столбец матрицы с нулевым индексом), а по оси y - столбец с частотами распределения данных по интервалам гистограммы (столбец с первым индексом). Индекс  столбца вводится с помощью соответствующей пиктограммы п панели *Matrix* или комбинации клавиш <Ctrl>+<6>.

Полигон частот

Иная форма графического представления группированных данных - полигон частот. *Полигон частот* - это ломанная линия, соединяющая точки с координатами (\bar{x}_i, h_i) , т.е. с абсциссами, равными серединам интервалов группировки, и ординатами, равными соответствующим частотам. Если соединить центры элементарных сегментов гистограммы ломанной линией, то получится график полигона.

Задание 1. Создайте выборку из 100 случайных величин с нормальным распределением, среднее значение $m=0,1*k$ (k - номер варианта) и стандартное отклонение $\sigma =0,5$. Запишите данную выборку в файл с произвольным именем. Рассчитайте с помощью встроенных функций *MathCad* числовые статистические характеристики созданной выборки. Постройте гистограмму двумя способами и полигон частот

Порядок выполнения работы:

1. С помощью встроенной функции из категории Random Numbers (Случайные числа) получить заданную выборку.
2. Записать полученную величину в файл с произвольным названием. В отчет вставьте фрагмент этого документа.
3. Упорядочить значения в выборке случайной величины по возрастанию.
4. С помощью стандартных функций Mathcad, получить числовые характеристики: \min и \max значения выборки, выборочное среднее, выборочную дисперсию, среднеквадратическое отклонение, выборочную медиану.
5. Используя функцию дозаписи, добавить в созданный ранее файл числовые характеристики выборки. Фрагмент вновь созданного файла привести в отчете.
6. Считать полученный файл.
7. Выполните расчет гистограммы с помощью функцию $\text{hist}(\text{int}, x)$ Отобразить на графиках гистограмму и плотность распределения на одном и полигон частот на другом.
8. Выполните расчет гистограммы, используя функцию $\text{histogram}(\text{int}, x)$.
9. Выведите на экран результаты процедур $\text{hist}()$ и $\text{histogram}()$. Сравните их.
10. Понаблюдайте, как изменится внешний вид гистограммы, если изменить количество интервалов разбиения выборки. Сделать выводы.

В отчете представить все необходимые фрагменты, сделанные в Mathcad, и требуемые выводы.

Контрольные вопросы

1. Что такое случайная выборка?
2. Что такое выборка?
3. Как происходит ввод и вывод данных в MathCad?
4. Как в MathCad произвести моделирование выборок из стандартных распределений?
5. Что такое функция MathCad для расчета численных характеристик?
6. Как в MathCad построить гистограммы?

7. Что такое гистограмма с произвольным сегментом разбиения?
8. Что такое гистограмма с разбиением на равные сегменты?
9. Приведите алгоритм создания графика гистограммы.
10. Что такое полигон частот?

Лабораторная работа №3

«Числовые характеристики дискретных случайных величин»

Цель: изучить способы вычисления числовых характеристик случайных величин с использованием пакета MathCad.

Задание: решить представленную задачу.

Краткие теоретические сведения

Математическое ожидание

Математическим ожиданием дискретной случайной величины называется сумма произведений всех ее возможных значений на вероятности этих значений.

Если случайная величина принимает значения с разной вероятностью, математическое ожидание вычисляется по формуле

$$M(X) = \sum_{i=0}^n x_i \cdot p_i$$

Пример 1. Найти математическое ожидание дискретной случайной величины, закон распределения которой задан таблицей:

X	1	2	3	4	5
P	0,15	0,25	0,3	0,2	0,1

Зададим векторы

$$x := (1\ 2\ 3\ 4\ 5)^T \quad p := (0.15\ 0.25\ 0.3\ 0.2\ 0.1)^T$$

Найдем математическое ожидание

$$M := \sum_{i=0}^{\text{last}(x)} x_i \cdot p_i$$

$$M = 2.85$$

Если случайная величина принимает ряд значений с равной вероятностью, то математическое ожидание определяется как

среднее арифметическое значение некоторого количественного признака выборки.

В MathCad среднее значение выборки можно подсчитать с помощью функции $mean(x)$.

Пример 2. При измерении величины силы тока были получены следующие значения: 0,45; 0,49; 0,44; 0,42; 0,48; 0,41; 0,44; 0,56; 0,47; 0,45; 0,52; 0,43. Вычислить выборочное среднее

$$X := (0.45 \ 0.49 \ 0.44 \ 0.42 \ 0.48 \ 0.41 \ 0.56 \ 0.47 \ 0.45 \ 0.52 \ 0.43)$$

$$mean(X) = 0.463$$

При обработке экспериментальных данных среднее значение выборки считается равным значению параметра. Это утверждение верно только в том случае, если выборка является генеральной, т.е. содержит все возможные значения измеряемой величины. В реальной ситуации с генеральными совокупностями работать невозможно, а всегда приходится делать из них некоторые небольшие выборки. В зависимости от условий отбора и объема выборки она может передавать особенности генеральной совокупности с различной точностью. При этом такие характеристики, как среднее значение и дисперсия, приобретают случайный характер. Исследование особенностей поведения такого рода величин – очень сложная и важная статистическая задача.

Дисперсия и среднеквадратичное отклонение

В статистике дисперсией называется среднее арифметическое квадратов отклонений случайной величины от ее среднего значения:

$$D = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}.$$

В общем случае дисперсия является характеристикой степени рассеяния значений выборки по сравнению с ее средней величиной.

В MathCad простая выборочная дисперсия вычисляется с помощью функции $var(x)$. Кроме того, существует и функция

$\text{Var}(x)$, которая определяет исправленную дисперсию, которая на практике используется для несмещенной оценки генеральной дисперсии при малом объеме выборки:

$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}.$$

На практике используют не саму дисперсию, а квадратный корень из нее, который называется среднеквадратичным отклонением. В MathCad существуют две функции для вычисления этого параметра: $\text{stdev}(x)$ – выборочное стандартное отклонение и $\text{Stdev}(x)$ – исправленное среднеквадратичное отклонение.

Пример 3. Подбрасывается игральный кубик. Случайная величина X – количество выпавших очков. Найти дисперсию и среднеквадратичное отклонение случайной величины X .

$$X := (1\ 2\ 3\ 4\ 5\ 6)^T$$

$$\text{var}(X) = 2.917 \quad \text{stdev}(X) = 1.708$$

Аналогичные результаты получаются и при использовании формул:

$$D := \frac{1}{\text{length}(X)} \sum_{i=0}^{\text{last}(X)} (X_i - \text{mean}(X))^2$$

$$\sigma := \sqrt{\frac{1}{\text{length}(X)} \cdot \sum_{i=0}^{\text{last}(X)} (X_i - \text{mean}(X))^2},$$

$$D = 2.917 \quad \sigma = 1.708.$$

Мода и медиана

Модой в статистике называют варианту, которая встречается в выборке наиболее часто. В MathCad подсчитать моду выборки можно с помощью встроенной функции $\text{mode}(x)$. В случае, если все варианты встречаются в выборке с одинаковой частотой, система

выдаст сообщение: No value occurs more then any others (ни одна величина не встречается чаще, чем все остальные).

Медианой называется варианта, которая делит вариационный ряд (рассортированную выборку) на две части, равные по количеству вариант. То есть если количество элементов выборки нечетное и равно $2k+1$, то медианой будет являться $(k+1)$ -й элемент. В случае четного количества вариант медиана определяется как среднее арифметическое между k -м и $(k+1)$ элементами выборки. В MathCad медиана вычисляется с помощью встроенной функции $\text{median}(x)$.

Пример 4. Вычисление моды и медианы

$$X:=(1 \ 1 \ 0 \ 8 \ 3 \ 7)$$

$$\text{mode}(X)=1 \qquad \text{median}(X)=4$$

Статистические функции работают не только с векторами - столбцами, но и с векторами - строками.

Размах варьирования

Важная характеристика рассеяния вариационного ряда - размах варьирования может быть просто вычислена в MathCad с помощью двух специальных матричных функций: $\text{max}(x)$ - находит максимальное значение в выборке, $\text{min}(x)$ - функция находит минимальную величину в выборке. Используя описанные функции, размах варьирования можно задать как

$$R=\text{max}(x)-\text{min}(x).$$

Пример 5. Вычисление размаха варьирования. Для задания вектора выборки воспользуемся генератором случайных чисел, распределенных по показательному закону:

$$X:=\text{rexp}(1000,4)$$

$$\text{Max}(X)=2.892 \quad \text{min}(X)=1.58 \times 10^{-5}$$

$$R=\text{max}(X)-\text{min}(X)=2.892.$$

Наибольший общий делитель и наименьшее общее кратное

Для решения некоторых задач в статистике бывает необходимым определить, на какое максимальное целое число делятся без остатка все величины в выборке. В MathCad очень просто вычислить такое число. Для этого необходимо воспользоваться встроенной функцией $\text{gcd}(x)$ (от англ. Greatest common divisor - наибольший общий делитель).

Схожей с описанной является задача поиска наименьшего числа, которое делится без остатка на все значения элементов выборки. В MathCad ее можно решить с помощью встроенной функции $\text{lcm}(x)$ (сокращение от Leastcommonmultiple – наименьшее общее кратное).

Пример 6. Найти наибольший общий делитель и наименьшее общее кратное.

$$X := (2\ 4\ 8\ 16\ 32\ 64)$$

$$\text{gcd}(X) = 2 \quad \text{lcm}(X) = 64$$

Задача. Для заданных случайных величин найти числовые характеристики (математическое ожидание, дисперсию, среднее квадратическое отклонение, моду, медиану), размах варьирования, а также наибольший делитель и наименьшее общее кратное элементов массива X .

№ варианта						
1	X	-1	0	2	4	7
	p	0,1	0,3	0,3	0,2	0,1
2	X	3	5	6	8	10
	p	0,2	0,2	0,3	0,2	0,1
3	X	-5	-4	-3	-1	1
	p	0,3	0,3	0,2	0,1	0,1
4	X	-1	2	3	4	6
	p	0,3	0,2	0,2	0,2	0,1
5	X	4	5	7	8	10
	p	0,1	0,4	0,3	0,1	0,1
6	X	-2	-1	0	2	5
	p	0,2	0,2	0,3	0,2	0,1
7	X	-4	-1	0	1	2
	p	0,1	0,1	0,2	0,5	0,1

8	X	-3	-2	-1	2	3
	p	0,2	0,4	0,1	0,1	0,2
9	X	6	7	9	10	12
	p	0,2	0,3	0,1	0,2	0,2
10	X	0	2	4	5	6
	p	0,4	0,2	0,2	0,1	0,1
11	X	-3	-2	0	1	2
	p	0,3	0,2	0,2	0,2	0,1
12	X	1	2	4	7	11
	p	0,1	0,1	0,3	0,3	0,2
13	X	-1	2	5	7	10
	p	0,2	0,2	0,3	0,1	0,2
14	X	4	7	9	11	13
	p	0,3	0,1	0,2	0,2	0,2

Отчет о выполненной работе должен содержать:

1. Тему и цель работы.
2. Индивидуальное задание согласно варианту.
3. Решение предложенных задач.

Контрольные вопросы

1. Дайте определение основных числовых характеристик случайной величины (математическое ожидание, дисперсия, среднее квадратическое отклонение, мода, медиана). Каков их вероятностный смысл?
2. Что называется размахом выборки?

Лабораторная работа №4 «Вычисление числовых характеристик выборки»

Цель: изучить возможности Mathcad по вычислению числовых характеристик выборки

Задание: решить представленную задачу.

Краткие теоретические сведения

Моменты

В теории вероятностей и математической статистике, помимо математического ожидания и дисперсии, используются и другие числовые характеристики случайных величин. В первую очередь это начальные и центральные моменты.

Начальным моментом k -го порядка случайной величины X называется математическое ожидание k -й степени случайной величины X , т.е. $\alpha_k = M(x^k)$.

Центральным моментом k -го порядка случайной величины ξ называется величина μ_k , определяемая формулой

$$\mu_k = M[(x - M)^k].$$

Заметим, что математическое ожидание случайной величины – начальный момент первого порядка $\alpha_1 = M$, а дисперсия – центральный момент второго порядка:

$$\mu_2 = M[(x - M)^2] = D(x).$$

Существуют формулы, позволяющие выразить центральные моменты случайной величины через ее начальные моменты. Одна из таких формул приведена выше:

$$D = M(x - M)^2 = \mu_2 - \alpha_1^2.$$

В дальнейшем будет использована формула

$$\mu_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3.$$

Асимметрия

В теории вероятностей и математической статистике в качестве меры асимметрии распределения служит коэффициент асимметрии, который определяется формулой:

$$\beta = \frac{\mu_3}{\sigma^3}$$

где μ_3 - центральный момент третьего порядка;

$$\sigma = \sqrt{D} = \sqrt{\mu_2} \text{ - среднеквадратичное отклонение}$$

Коэффициент асимметрии – безразмерная величина, а по его знаку можно судить о характере асимметрии. Для симметричной СВ $\mu_3 = 0$.

Указание. Для того чтобы определить точность коэффициента асимметрии, выделенное выражение для него щелкните в строке FloatingPoint в меню Symbolics и укажите в окне диалога число десятичных знаков в выводе.

Эксцесс

Нормальное распределение наиболее часто используется в теории вероятностей и математической статистке, и поэтому график плотности вероятностей нормального распределения стал своего рода эталоном, с которым сравнивают другие распределения. Одним из параметров, определяющих отличие сравниваемого распределения от нормального, является эксцесс.

Эксцесс γ случайной величины ξ определяется равенством

$$\gamma = \frac{\mu_4}{(D)^2} - 3.$$

По известным нам свойствам математического ожидания и определению центрального момента получим формулу для μ_4

$$\mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4.$$

У нормального распределения, естественно, $\gamma = 0$. Если $\gamma > 0$, то это означает, что график плотности вероятностей $p(x)$ сильнее «заострен», чем у нормального распределения, если же $\gamma < 0$, то «заостренность» графика $p(x)$ меньше, чем у нормального распределения.

Порядок выполнения работы

1. Изучить теоретический материал.
2. Вычислить для выборки заданной случайным образом выборочные моменты 3 и 4-го порядков выборочный эксцесс E , коэффициент асимметрии.
3. Оформить отчет.

Пример выполнения задания приведен на рисунке 4.1.

Рисунок 4.1 – Пример выполнения задания в MathCad

Контрольные вопросы

1. Раскройте понятие «моменты».
2. Что такое центральный момент?
3. Что такое начальный момент?
4. Что такое асимметрия?
5. Что такое коэффициент асимметрии?
6. Что такое эксцесс?
7. Что такое математическое ожидание?

Лабораторная работа №5 «Применения MATHCAD для решения задач теории вероятности»

Цель:

1. Закрепить знания об основных законах распределения случайной величины.
2. Научить применять MathCad для построения функций распределения и плотности распределения случайной величины.
3. Закрепить навыки расчета основных числовых характеристик случайных величин.
4. Научиться применять MathCad для расчета основных числовых характеристик случайных величин.

Функции и инструменты MathCad

Прежде чем приступать к решению задач теории вероятностей в MathCad, познакомимся с инструментами, которые предоставляет пакет для их решения.

Напомним, что дискретная случайная величина x , принимающая значения $x_1 < x_2 < \dots < x_i < \dots$ с вероятностями $p_1, p_2, \dots, p_i, \dots$, может быть задана распределением – таблицей вида (X_i - значение случайной величины, p – вероятность появления именно этого значения).

X	x_1	x_2	...	x_i	...	x_n
p	p_1	p_2	...	p_i	...	p_n

Для дальнейшей работы ряд распределения необходимо сделать вариационным.

Вариационный ряд – это дискретный ряд распределения, у которого значения X_i располагаются в порядке возрастания.

Такие таблицы в среде MathCad записывается в виде матрицы размерности $2 \times n$.

Указания. Распределение случайной величины в MathCad записывается в виде матрицы A

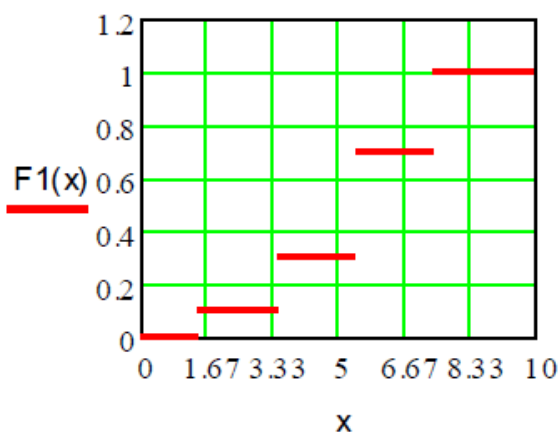




Рисунок 5.1 – Функция распределения

Функцию распределения, заданную разными выражениями на разных интервалах изменения аргументов, можно определить следующим образом:

1. Введите имя функции переменной x и знак присваивания (знак присваивания «:=»).

2. Используйте панель программных элементов (Programming) (кнопка ). С помощью кнопки AddLine добавьте необходимое число строк для задания функции распределения (рисунок 5.1).

3. Введите в помеченной позиции нуль, щелкните по кнопке if и введите неравенство, определяющие первый интервал изменения аргумента (символ ∞ можно ввести щелчком по соответствующей кнопке в панели  (Calculus)).

4. Затем перейдите во вторую строку определения функции, введите $A_{2,1}$ - имя переменной, содержащей значение p_1 , или число 0.2 – значение.

5. Введите неравенство, определяющие второй интервал изменения аргумента (знак можно ввести щелчком по соответствующей кнопке в панели отношений (Boolean)); выделите, нажимая клавишу <SPACE>, вторую строку определения функции, щелкните по кнопке AddLine и введите, действуя, как описано выше определение функции на следующем интервале.

В результате у Вас должно получиться следующее:

$$F(x) := \begin{cases} 0 & \text{if } -\infty < x < A_{1,1} \\ A_{2,1} & \text{if } A_{1,1} \leq x < A_{1,2} \\ A_{2,1} + A_{2,2} & \text{if } \dots \\ \dots & \dots \dots \\ \dots & \dots \dots \\ 1 & \text{if } A_{1,5} \leq x < \infty \end{cases}$$

Рисунок 5.2 – Функция распределения случайной величины

На приведенном рисунке 5.2 функция распределения определена с использованием имен переменных.

Замечание. Следует помнить, что MathCad не совсем корректно строит графики ступенчатых функций, соединяя отрезками прямых значения функции в точке скачка. Более точный график функции распределения представляет собой отрезки, параллельные оси абсцисс, с «выколотым» правым концом.

Случайные величины. Функции распределения

Каждая случайная величина полностью определяется своей функцией распределения.

Функция распределения любой случайной величины обладает следующими свойствами:

- $F(x)$ определена на всей числовой прямой R ;
- $F(x)$ не убывает, т.е. если $x_1 \leq x_2$, то $F(x_1) \leq F(x_2)$;
- $F(-\infty) = 0$;
- $F(+\infty) = 1$;
- $P(a < x < b) = F(b) - F(a)$.

Важно помнить, что функция распределения является «паспортом» случайной величины: она содержит всю информацию об этой случайной величине, и поэтому изучение случайной величины заключается в исследовании ее функции распределения, которую часто называют просто распределением.

Для проведения вычислений со случайными величинами (непрерывными и дискретными) в MathCad есть богатая библиотека встроенных функций наиболее распространенных стандартных распределений. Каждое распределение представлено в библиотеке

тремя функциями – плотностью вероятностей, функцией распределения и функцией, обратной к функции распределения. Имена всех встроенных функций, определяющих плотности вероятностей, начинаются с буквы d, определяющих функции распределения – с буквы p.

Например, для работы с нормальным распределением предназначены функции $dnorm(x, h, s)$, $p(norm(x, h, s))$ и $qnorm(x, h, s)$.

Наиболее распространенные распределения дискретных случайных величин

Познакомимся с дискретными случайными величинами, которые чаще всего используются при решении практических задач. Эти случайные величины имеют биномиальные, геометрические и пуассоновские распределения.

Биномиальное распределение (схема Бернулли)

Пусть проводится серия из n независимых испытаний, каждое из которых заканчивается либо «успехом», либо «неуспехом». Пусть в каждом испытании (опыте) вероятность успеха p , а вероятность неудачи – $q=1-p$. С таким испытанием можно связать случайную величину x , равную числу успехов в серии из n испытаний. Эта величина принимает целые значения от 0 до n .

Ее распределение называется биномиальным и определяется формулой Бернулли

$$p_k = P(\xi = k) = C_n^k p^k q^{n-k}$$

где $0 < p < 1$, $q = 1 - p$, $k = 0, 1, \dots, n$, $C_n^k = \frac{n!}{k!(n-k)!}$

В MathCad для вычисления плотности вероятности и функции распределения случайной величины, имеющей биномиальное распределение, предназначены функции $dbinom(k, n, p)$ и $pbinom(k, n, p)$, значения которых – соответственно $p(k)$ и $F(k)$.

Геометрическое распределение

Со схемой испытаний Бернулли можно связать еще одну случайную величину: h – число испытаний до первого успеха. Эта величина принимает бесконечное множество значений от 0 до $+\infty$, и ее распределение определяется формулой

$$p_k = P(\eta = k) = p^k q$$

где $0 < p < 1$, $q = 1 - p$, $k = 0, 1, \dots, n$.

В MathCad для вычисления плотности вероятности и функции распределения случайной величины, имеющей геометрическое распределение, предназначены функции $dgeom(k, p)$ и $pgeom(k, p)$, значения которых – соответственно $p(k)$ и $F(k)$.

Пуассоновское распределение

Пуассоновское распределение имеет случайная величина m , принимающая значения $k = 0, 1, 2, \dots$ с вероятностями

$$p_k = P(\mu = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots,$$

где $\lambda > 0$ - параметр пуассоновского распределения.

В MathCad для вычисления вероятности и функции распределения случайной величины, имеющей пуассоновское распределение, предназначены функции $dpois(k, \lambda)$ и $ppois(k, \lambda)$, значения которых – соответственно $p(k)$ и $F(k)$.

Наиболее распространенные частные распределения непрерывных случайных величин

Равномерное распределение

Непрерывная случайная величина ξ , принимающая значение на отрезке $[a, b]$, распределена равномерно на $[a, b]$, если плотность распределения $p(x)$ и функция распределения случайной величины ξ имеют соответственно вид

$$p(x) = \begin{cases} 0, & x \notin [a, b] \\ \frac{1}{b-a}, & x \in [a, b] \end{cases} \quad F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & x > b \end{cases}$$

В MathCad значения в точке x плотности распределения и функции распределения случайной величины имеющее равномерное распределение на отрезке $[a, b]$, вычисляются встроенными функциями соответственно $\text{dunif}(x, a, b)$ и $\text{punif}(x, a, b)$.

Экспоненциальное (показательное) распределение

Непрерывная случайная величина ξ имеет показательное распределение с параметром $\lambda > 0$, если плотность распределения имеет вид

$$p(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases} \quad F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \lambda e^{-\lambda x}, & x > 0 \end{cases}$$

В MathCad значения в точке x плотности распределения и функции распределения случайной величины, имеющей экспоненциальное распределение с параметром λ , вычисляются встроенными функциями соответственно $\text{dexpr}(x, \lambda)$ и $\text{rexpr}(x, \lambda)$.

Нормальное распределение

Это распределение играет исключительно важную роль в теории вероятностей и математической статистике. Случайная величина ξ нормально распределена с параметрами m и σ , ($\sigma > 0$), если ее плотность распределения имеет вид

$$p_{\xi}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right).$$

В MathCad значения в точке x плотности распределения и функции распределения нормальной случайной величины с параметрами a , σ вычисляются встроенными функциями соответственно $dnorm(x, a, s)$ и $pnorm(x, a, s)$.

Распределение Стьюдента

Пусть случайная величина ξ имеет стандартное нормальное распределение, а случайная величина $\chi_n^2 - \chi^2$ распределение с n степенями свободы. Если ξ и χ_n^2 независимы, то про случайную величину $\tau_n = \frac{\xi}{\sqrt{\chi_n^2/n}}$ говорят, что она имеет распределение

Стьюдента с числом степеней свободы n . Доказано, что плотность вероятности этой величины вычисляется по формуле

$$p_{\tau n}(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}, x \in \mathbb{R}$$

При больших n распределение Стьюдента практически не отличается от $N(0, 1)$.

В MathCad значения в точке x плотности распределения и функции Стьюдента с n степенями свободы вычисляются встроенными функциями соответственно $dt(x, n)$ и $pt(x, n)$.

Числовые характеристики случайных величин

Каждая случайная величина полностью определяется своей функцией распределения. В то же время при решении практических задач достаточно знать несколько числовых параметров, которые позволяют представить основные особенности случайной величины в сжатой форме.

К таким величинам относятся, в первую очередь, математическое ожидание и дисперсия.

Математическое ожидание случайной величины

Математическое ожидание – число, вокруг которого сосредоточены значения случайной величины.

Если ξ - дискретная случайная величина с распределением,

ξ	x_1	x_2	...	x_n
p	p_1	p_2	...	p_n

то ее математическим ожиданием – оно обозначается M – называется величина

$$M[x] = \sum_{i=1}^n p_i x_i.$$

Математическое ожидание непрерывной случайной величины с плотностью вероятностей $p(x)$ вычисляется по формуле

$$M[x] = \int_{-\infty}^{\infty} xp(x)dx.$$

Дисперсия случайной величины

Дисперсия случайной величины характеризует меру разброса значений случайной величины около ее математического ожидания. Если случайная величина ξ имеет математическое ожидание M , то дисперсией случайной величины ξ называется величина $D = M[(X - M)^2]$. Легко показать, что $D = M[x^2] - (M[x])^2$. Эта универсальная формула одинаково хорошо применима как для дискретных случайных величин, так и для непрерывных. Величина M^2 вычисляется по формулам:

$$M[x^2] = \sum_{i=1}^n p_i x_i^2 \quad M[x^2] = \int_{-\infty}^{\infty} x^2 p(x)dx,$$

для дискретных и непрерывных случайных величин соответственно.

Еще одним параметром для определения меры разброса значений случайной величины является среднеквадратическое отклонение σ , связанное с дисперсией соотношением $\sigma = \sqrt{D}$.

Моменты

В теории вероятностей и математической статистике, помимо математического ожидания и дисперсии, используются и другие числовые характеристики случайных величин. В первую очередь это начальные и центральные моменты.

Начальным моментом k -го порядка случайной величины X называется математическое ожидание k -й степени случайной величины X , т.е. $\alpha_k = M(x^k)$.

Центральным моментом k -го порядка случайной величины ξ называется величина μ_k , определяемая формулой

$$\mu_k = M[(x - M)^k].$$

Заметим, что математическое ожидание случайной величины – начальный момент первого порядка $\alpha_1 = M$, а дисперсия – центральный момент второго порядка:

$$\mu_2 = M[(x - M)^2] = D(x).$$

Существуют формулы, позволяющие выразить центральные моменты случайной величины через ее начальные моменты. Одна из таких формул приведена выше:

$$D = M(x - M)^2 = \mu_2 - \alpha_1^2.$$

В дальнейшем будет использована формула

$$\mu_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3.$$

Асимметрия

В теории вероятностей и математической статистике в качестве меры асимметрии распределения служит коэффициент асимметрии, который определяется формулой:

$$\beta = \frac{\mu_3}{\sigma^3}$$

где μ_3 - центральный момент третьего порядка;

$$\sigma = \sqrt{D} = \sqrt{\mu_2} \text{ - среднеквадратичное отклонение}$$

Коэффициент асимметрии – безразмерная величина, а по его знаку можно судить о характере асимметрии. Для симметричной СВ $\mu_3 = 0$.

Указание. Для того чтобы определить точность коэффициента асимметрии, выделенное выражение для него щелкните в строке FloatingPoint в меню Symbolics и укажите в окне диалога число десятичных знаков в выводе.

Эксцесс

Нормальное распределение наиболее часто используется в теории вероятностей и математической статистке, и поэтому график плотности вероятностей нормального распределения стал своего рода эталоном, с которым сравнивают другие распределения. Одним из параметров, определяющих отличие сравниваемого распределения от нормального, является эксцесс.

Эксцесс γ случайной величины ξ определяется равенством

$$\gamma = \frac{\mu_4}{(D)^2} - 3.$$

По известным нам свойствам математического ожидания и определению центрального момента получим формулу для μ_4

$$\mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4.$$

У нормального распределения, естественно, $\gamma = 0$. Если $\gamma > 0$, то это означает, что график плотности вероятностей $p(x)$ сильнее «заострен», чем у нормального распределения, если же $\gamma < 0$, то «заостренность» графика $p(x)$ меньше, чем у нормального распределения.

Порядок выполнения работы

Задание 1.

Постройте с помощью MathCAD график функции распределения для случайной величины:

X	1	0	7	4	-2
P	0.1	0.5	0.1	0.1	0.2

Порядок выполнения задания

1. Задайте случайную величину A в виде матрицы.
2. Определите функцию распределения случайной величины $F(x)$.
3. Определите функцию распределения $G(x)$ той же случайной величины с использованием конкретных значений переменных.
4. Постройте графики функций распределений $F(x)$ и $G(x)$. Отредактируйте и сравните графики.
В отчете представить два графика $F(x)$ и $G(x)$, сделать выводы о их сходстве и объясните, почему так произошло.

Задание 2.

Порядок выполнения задания

1. Задайте случайные величины, имеющие биномиальное распределение. Параметры распределения задайте самостоятельно. Постройте графики распределения и функции распределения

случайной величины. (Подберите удобные параметры отображения графиков). Удобно строить оба графика в одной системе координат.

2. Проверьте для них $\sum_{k=0}^n p_k = 1$.

3. Вычислите вероятность попадания значений случайной величины в выбранный интервал.

4. Найдите значение k , для которого величина $P(\xi = k)$ максимальна (медиану). Исследуйте (понаблюдайте) зависимость этой вероятности от параметров распределения.

Указание. Для того, чтобы определить по графику распределения наиболее вероятное значение случайной величины, щелкните в меню Format (Формат) в пункте Graph (График) по строке Trace (Следование), установите перекрестье маркера на точке максимума распределения и выведите в рабочий документ вероятность значения, указанного в окне X-Value (Величина X).

5. Измените значения параметров распределения и повторите вычисления. Сравните полученные результаты. Выводы привести в отчете.

Повторить п. 1-5 для других распределений (геометрическое и пуассоновское).

В отчете представьте по одному варианту для каждого распределения: параметры распределения, графики вероятности и функции распределения, значение наиболее вероятного значения СВ и значение вероятности попадания СВ в указанный диапазон. В отчет вставлять фрагменты из MathCAD.

Задание 3.

Порядок выполнения задания

1. Введите параметры равномерного распределения. (Можно использовать любые).

2. Определите плотность вероятности и функцию распределения случайной величины.

3. Постройте графики.

4. Поэкспериментируйте с параметрами различных распределений.

5. Сделайте выводы о зависимости выходных данных от входных параметров.

Повторить п. 1-5 для других распределений (экспоненциального, нормального распределения и распределения Стьюдента).

В отчете представьте по одному варианту для каждого распределения: параметры распределения, графики вероятности и функции распределения и выводы о зависимости выходных данных от входных параметров. В отчет вставлять фрагменты из MathCad.

Задание 4.

Порядок выполнения задания

1. Вычислите математическое ожидание и дисперсии для случайных величин, имеющих дискретные распределения (биномиальные, геометрические и пуассоновские распределения).

В качестве исходных параметров возьмите:

$n=10$ – число испытаний;

$p=0.1 \cdot h$ – вероятность наступления события;

$\lambda = 4 + k$ - параметр пуассоновского распределения;

$N=1000$ – число испытаний стремится к ∞ ;

k – номер варианта студента по журналу;

2. Сверите полученные значения со справочными данными.

В отчете привести все расчетные формулы и результаты сравнения для каждого распределения в виде фрагментов из MathCad.

Задание 5.

Порядок выполнения задания

1. Вычислите математические ожидания и дисперсии для случайных величин, имеющих непрерывные распределения (равномерное, экспоненциальное, нормальное, распределение Стьюдента). В качестве исходных параметров возьмите (при желании можно использовать любые другие) следующие параметры:

$a=1*h$, $b=3*h$, $n=5$, $\lambda =0.1*h$, $\sigma =0.05*h$ – необходимые параметры распределений;

h – порядковый номер студента по журналу.

2. Сверите полученные значения со справочными данными.

В отчете привести все расчетные формулы и результаты для каждого распределения в виде фрагментов из MathCad.

Задание 6 (Дополнительное).

Вычислите коэффициент асимметрии случайной величины X с равномерным распределением (или любого другого).

Порядок выполнения задания

1. Определите значения параметров распределения случайной величины.

2. Вычислите коэффициент асимметрии.

3. Сверьте полученное значение со справочными данными для данного распределения.

4. Постройте график плотности вероятности.

5. Сделайте выводы, опираясь на значение коэффициента асимметрии и форму распределения.

Задание 7. (Дополнительное)

Вычислите эксцесс случайной величины ξ с равномерным распределением (или любым другим).

Порядок выполнения задания

1. Определите значения параметров распределения случайной величины.

2. Вычислите коэффициент асимметрии.

3. Сверьте полученное значение со справочными данными для данного распределения.

4. Постройте график плотности вероятности.

5. Сделайте выводы, опираясь на значение коэффициента асимметрии и форму распределения.

Справочный материал приведен ниже.

Законы распределения случайных величин и их характеристики

Вид закона распределения		Аналитическое выражение	Основн. числовые хар-ки
Плотность распределения	Функция распределения		
Дискретные случайные величины			
<p>Биноминальный (Бернулли)</p>		$p_k = C_n^k p^k q^{n-k}$ <p>где $0 < p < 1, q = 1 - p,$ $k = 0, 1, \dots, n,$ $C_n^k = \frac{n!}{k!(n-k)!}$</p>	$M(X) = np,$ $D(X) = npq,$ $A = \frac{q-p}{\sqrt{npq}},$ $E = \frac{1-6pq}{npq}$
<p>Геометрический</p>		$p_k = p^k q$ <p>где $0 < p < 1, q = 1 - p,$ $k = 0, 1, \dots, n$</p>	$M(X) = \frac{1-p}{p},$ $D(X) = \frac{q}{p^2},$ $A = \frac{2-p}{\sqrt{1-p}},$ $E = 6 + \frac{p^2}{1-p}$
<p>Пуассоновский</p>		$p_k = \frac{\lambda^k}{k!} e^{-\lambda},$ <p>$k = 0, 1, 2, \dots,$ где $\lambda > 0.$</p>	$M(X) = \lambda,$ $D(X) = \lambda,$ $A = \lambda^{-1/2},$ $E = \lambda^{-1}$
Непрерывные случайные величины			
<p>Равномерный</p>	$p(x) = \begin{cases} 0, & x \notin [a, b] \\ \frac{1}{b-a}, & x \in [a, b] \end{cases}$ $F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & x > b \end{cases}$	$M(X) = \frac{a+b}{2},$ $D(X) = \frac{(b-a)^2}{12},$ $A = 0,$ $E = -\frac{6}{5}.$	

Рисунок 5.2 – Законы распределения случайных величин

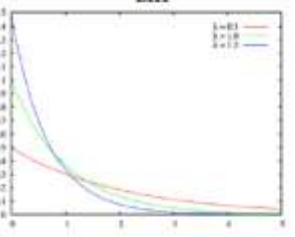
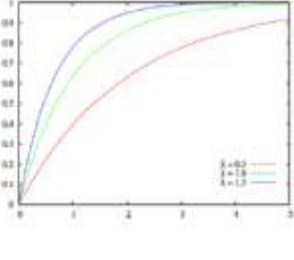
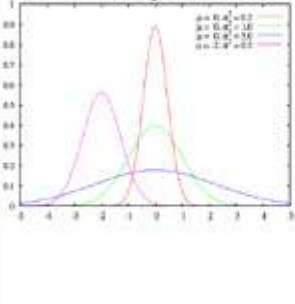
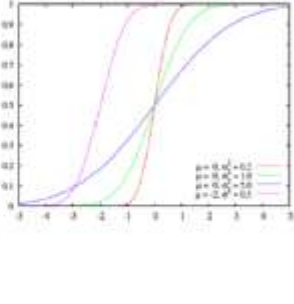
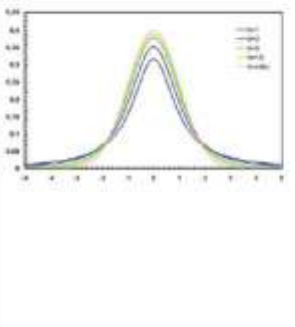
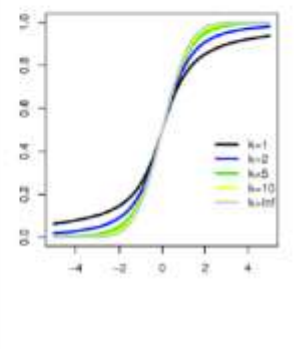
<p style="text-align: center;">Экспоненциальный</p> 		$p(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$ $F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \lambda e^{-\lambda x}, & x > 0 \end{cases}$ $\lambda > 0.$	$M(X) = \lambda^{-1},$ $D(X) = \lambda^{-2},$ $A = 2, E = 6.$
<p style="text-align: center;">Нормальный</p> 		$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$ $F(x) = \Phi(x), \text{ где } \Phi(x) - \text{ функция Лапласа:}$ $\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^x \exp\left(-\frac{(z)^2}{2\sigma^2}\right) dz$ <p style="text-align: center;">где $z = x - m$</p>	$M(X) = m,$ $D(X) = \sigma^2,$ $A = 0, E = 0.$
<p style="text-align: center;">Стюдента</p> 		$p(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$ $x \in R$ <p style="text-align: center;">где $\Gamma\left(\frac{n}{2}\right)$ - гамма-функция Эйлера</p>	$M(X) = 0,$ если $n > 1$ $D(X) = \frac{n}{n-2},$ если $n > 2,$ $A = 0,$ если $n > 3,$ $E = \frac{6}{n-4},$ если $n > 4.$

Рисунок 5.3 – Законы распределения случайных величин

4 семестр

Лабораторная работа №1. Интервальное оценивание

1.1 Понятие доверительного интервала

Пусть имеется выборка $X = (X_1, \dots, X_n)$ из распределения F_θ с неизвестным параметром $\theta \in \Theta \subseteq \mathbb{R}$. Задача интервального оценивания заключается в том, чтобы найти интервал, который покрывает оцениваемый параметр с заданной наперед вероятностью.

Интервал (θ^-, θ^+) называется доверительным интервалом для параметра θ уровня доверия $1 - \varepsilon$, если для любого $\theta \in \Theta$

$$P_\theta(\theta^- < \theta < \theta^+) \geq 1 - \varepsilon \quad (1.1)$$

Если в соотношении (1.1) вероятность в точности равна $1 - \varepsilon$ (или стремится к $1 - \varepsilon$), то интервал называют точным (или асимптотически точным) доверительным интервалом уровня доверия $1 - \varepsilon$.

1.2 Построение точных доверительных интервалов для параметров нормального распределения

Пусть $X = (X_1, \dots, X_n)$ - выборка объема n из нормального распределения N_{a, σ^2} . Рассмотрим возможные задачи интервального оценивания:

1. Пусть известно σ^2 , а $a \in \mathbb{R}$ - неизвестный параметр. Требуется построить точный доверительный интервал для параметра a .

$$P_{a, \sigma^2} \left(\bar{X} - \frac{\tau_{1-\varepsilon/2} \cdot \sigma}{\sqrt{n}} < a < \bar{X} + \frac{\tau_{1-\varepsilon/2} \cdot \sigma}{\sqrt{n}} \right) = 1 - \varepsilon, \quad (1.2)$$

где $\tau_{1-\varepsilon/2}$ - квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения (см. Приложение Б, таблица 1).

2. Пусть известен параметр a , требуется построить доверительный интервал для σ^2 .

$$P_{a,\sigma^2} \left(\frac{n \cdot s_1^2}{g_2} < \sigma^2 < \frac{n \cdot s_1^2}{g_1} \right) = 1 - \varepsilon, \quad (1.3)$$

где s_1^2 - выборочная дисперсия, g_1 и g_2 - квантили распределения "хи- квадрат" с n степенями свободы ($\chi_{\alpha,n}^2$) уровня $\alpha = \varepsilon/2$ и $\alpha = 1 - \varepsilon/2$ соответственно (см. Приложение Б, таблица 3).

3. Доверительный интервал для σ^2 при неизвестном a :

$$P_{a,\sigma^2} \left(\frac{(n-1) \cdot s_0^2}{g_2} < \sigma^2 < \frac{(n-1) \cdot s_0^2}{g_1} \right) = 1 - \varepsilon, \quad (1.4)$$

где s_0^2 - несмещенная выборочная дисперсия.

4. Доверительный интервал для a при неизвестном σ^2 :

$$P_{a,\sigma^2} \left(\bar{X} - \frac{t_{1-\varepsilon/2,n-1} \cdot s_0}{\sqrt{n}} < a < \bar{X} + \frac{t_{1-\varepsilon/2,n-1} \cdot s_0}{\sqrt{n}} \right) = 1 - \varepsilon, \quad (1.5)$$

где $t_{1-\varepsilon/2,n-1}$ - квантиль распределения Стьюдента с $n-1$ степенью свободы уровня $1-\varepsilon/2$ (см. Приложение Б, таблица 2).

Приведем пример построения доверительного интервала.

Пусть имеется выборка объема N из нормального распределения, для которого известен параметр $\sigma^2 = 4$, а параметр a неизвестен. Требуется построить точный доверительный интервал для параметра a . На рисунке 1.1 приведен текст программы,

реализующей метод построения доверительного интервала в среде Mathcad.

Ввод исходных данных

x - вектор, представляющий выборку из нормального распределения

dx - известное среднеквадратическое отклонение

$q=1-\epsilon$ - уровень доверия

$x :=$



C:\1..w1.txt

вектор, представляющий выборку считывается из файла

\vec{x}	0	1	2	3	4	5	6	7	8	9	
	0	0.432	-0.501	2.332	0.494	3.503	1.233	-1.313	1.969	2.207	1.188

вывод на экран элементов выборки (их можно просмотреть, щелкнув по вектору x и воспользовавшись линейкой прокрутки)

$dx := 2$

$q := 0.9$

ввод значений среднеквадратического отклонения и уровня доверия

$N := \text{length}(x)$

$N = 700$

нахождение объема выборки

$$Mx = \frac{\sum_{i=0}^{N-1} x_i}{N}$$

$Mx = 1.97$

вычисление среднего выборочного

$$p := 1 - \frac{1 - q}{2}$$

$v := \text{qnorm}(p, 0, 1)$

$v = 1.645$

нахождение квантиля стандартного нормального распределения уровня $1-\epsilon/2$

$$\text{upper} := Mx + \frac{v \cdot dx}{\sqrt{N}}$$

$\text{upper} = 2.064$

$$\text{lower} := Mx - \frac{v \cdot dx}{\sqrt{N}}$$

$\text{lower} = 1.346$

нахождение границ доверительного интервала

$\text{length} := \text{upper} - \text{lower}$

$\text{length} = 0.249$

вычисление длины доверительного интервала

Рисунок 1.1. Построение доверительного интервала для параметра а нормального распределения при известном σ

1.3 Задание к лабораторной работе

Даны две выборки одной случайной величины с нормальным распределением N_{a, σ^2} объема n_1 и n_2 соответственно.

Для вариантов с нечетным номером:

1. Для обеих выборок построить точный доверительный интервал уровня доверия Q_0 для параметра a , считая:

а) σ неизвестным,

б) σ известным и равным σ_0 .

2. В одной системе координат построить графики зависимости длины доверительного интервала от уровня доверия Q для всех четырех случаев (объем выборки равен n_1 , σ неизвестно; объем выборки равен n_1 , σ известно; объем выборки равен n_2 , σ неизвестно; объем выборки равен n_2 , σ известно). При этом Q придать минимум 50 разных значений через равные промежутки.

Проанализировать взаимное расположение полученных графиков и объяснить его.

Для вариантов с четным номером:

1. Для обеих выборок построить точный доверительный интервал

уровня доверия Q_0 для параметра σ^2 , считая:

а) a неизвестным,

б) a известным и равным a_0 .

2. В одной системе координат построить графики зависимости длины доверительного интервала от уровня доверия Q для всех четырех случаев (объем выборки равен n_1 , a неизвестно; объем выборки равен n_1 , a известно; объем выборки равен n_2 , a неизвестно; объем выборки равен n_2 , a известно). При этом Q придать минимум 50 разных значений через равные промежутки.

Проанализировать взаимное расположение полученных графиков и объяснить его.

Указания: Выборки необходимо считать с двух текстовых файлов - "di-V.txt" "di-V-1.txt" (V- номер вашего варианта).

Для написания программы можно пользоваться следующими встроенными функциями: функцией `length(x)` для определения объема выборки, функциями для нахождения значений квантилей распределений

`qnorm(p,a,a)` (возвращает значение квантиля нормального распределения N, a, σ^2 уровня p), `qchisq(p,d)` (возвращает значение квантиля распределения "хи-квадрат" с d степенями свободы уровня p), `qt(p,d)` (возвращает значение квантиля распределения Стьюдента с d степенями свободы уровня p). Остальные статистические функции должны быть запрограммированы.

Варианты заданий

1. $\sigma_0 = 2; q_0 = 0,9$
2. $\sigma_0 = 0; q_0 = 0,8$
3. $\sigma_0 = 3; q_0 = 0,7$
4. $\sigma_0 = 2; q_0 = 0,5$
5. $\sigma_0 = 1; q_0 = 0,6$
6. $\sigma_0 = 3; q_0 = 0,9$
7. $\sigma_0 = 0,5; q_0 = 0,8$
8. $\sigma_0 = -1; q_0 = 0,8$
9. $\sigma_0 = 1,5; q_0 = 0,7$
10. $\sigma_0 = 0,5; q_0 = 0,8$
11. $\sigma_0 = 1; q_0 = 0,5$
12. $\sigma_0 = -5; q_0 = 0,6$
13. $\sigma_0 = 1,2; q_0 = 0,7$
14. $\sigma_0 = 4; q_0 = 0,8$
15. $\sigma_0 = 2,5; q_0 = 0,75$
16. $\sigma_0 = 10; q_0 = 0,6$
17. $\sigma_0 = 3,2; q_0 = 0,9$
18. $\sigma_0 = 0; q_0 = 0,75$
19. $\sigma_0 = 3; q_0 = 0,75$
20. $\sigma_0 = 3; q_0 = 0,5$

Лабораторная работа №2. Проверка гипотезы о виде распределения с помощью критерия согласия Смирнова

2.1 Статистическая гипотеза и статистический критерий. Критерий согласия.

Пусть в результате некоторого эксперимента получена выборка

$X = (X_1, X_2, \dots, X_n)$ из некоторого распределения F .

Статистической гипотезой H называется любое утверждение о виде неизвестного распределения, или о параметрах известного распределения наблюдаемой в эксперименте случайной величины.

Гипотеза H называется простой, если она однозначно определяет распределение выборки: $H = \{F = F_1\}$, иначе H называют сложной, например, $H = \{F \in \{F\}\}$, где $\{F\}$ - некоторое семейство распределений: $\{F\} = \{F(x, \theta), \theta \in \Theta\}$.

Если выдвигаются всего две гипотезы, то одну из них принято называть основной (H_0), а другую альтернативной (конкурирующей) (H_1).

Правило, согласно которому проверяемая гипотеза H_0 принимается или отвергается, называется статистическим критерием. Дадим формальное определение критерия.

Статистическим критерием для проверки гипотез H_1, \dots, H_k называется любое измеримое отображение $\rho: R^n \rightarrow \{H_1, \dots, H_k\}$.

Если $\rho(X) = H_i$, то мы принимаем гипотезу H_i (или считаем $\theta = \theta_i$ в параметрическом случае).

Качество критерия характеризуется набором вероятностей ошибочных решений.

Будем говорить, что произошла ошибка i -го рода, если гипотеза H_i отвергнута, когда она верна. Вероятностью ошибки i -го рода критерия ρ называется

$$q_i(\delta) = P_{H_i}(\rho(X) \neq H_i).$$

Если удастся выбрать критерий P так, что все числа $q_i(p)$ малы, то мы будем объявлять, что верна гипотеза H_k , если $\rho(X) = H_k$. При этом мы будем ошибаться примерно в доле случаев q_i , если верна гипотеза H_i .

Уровнем значимости статистического критерия называют вероятность ошибочно отвергнуть основную проверяемую гипотезу, когда она верна (вероятность ошибки первого рода).

Рассмотрим случай, когда о распределении наблюдений X имеется две гипотезы: $H_0 = \{F = F_1\}$ при альтернативе $H_1 = \{F \neq F_1\}$. В этом случае любой критерий $\rho(X): R^n \rightarrow \{H_0, H_1\}$ принимает не более двух значений, то есть область R^n делится на две части

$$R^n = S \cup (R^n \setminus S)$$

так, что

$$\rho(X) = \begin{cases} H_0, & X \in R^n \setminus S, \\ H_1, & X \in S. \end{cases}$$

Область S , в которой принимается альтернативная гипотеза, называется критической областью критерия ρ .

Обозначим $q = q(\rho)$ - уровень значимости критерия ρ , тогда

$$q = q_1(\rho) = P_{H_0}(\rho(X) \neq H_0) = P_{H_0}(\rho(X) = H_1) = P_{H_0}(X \in S).$$

По своему смыслу критическая область должна строиться так, чтобы событие $X \in S$ было маловероятным. В конкретных задачах ρ выбирают обычно равной 0,1; 0,05; 0,01 и так далее.

Критериями согласия называют критерии для проверки простой гипотезы H_0 при сложной альтернативе $H_1 = \{H_0 \text{ неверна}\}$.

Критерии согласия принимают или отвергают основную гипотезу исходя из величины некоторой статистики $T(X)$, характеризующей отклонение эмпирических данных от соответствующих (гипотезе H_0) гипотетических значений, распределение которой в случае справедливости H_0 можно было бы определить.

Предположим, такая статистика и ее распределение при

гипотезе H_0 найдены. Пусть T - множество всевозможных значений статистики T . Определим для фиксированного заранее достаточно малого числа $\alpha > 0$ подмножество $\tilde{T} \subset T$, так чтобы вероятность осуществления события $T(X) \in \tilde{T}$ в случае справедливости гипотезы H_0 удовлетворяла условию

$$P(T(X) \in \tilde{T} | H_0) \leq \alpha$$

Тогда правило проверки гипотезы H_0 можно сформулировать следующим образом: Если для данной выборки X значение статистики $T(X) \in \tilde{T}$, то в предположении справедливости гипотезы H_0 произошло маловероятное событие и гипотеза должна быть отвергнута как противоречащая статистическим данным. В противном случае нет основания отказываться от принятия гипотезы H_0 . **На языке критериев это можно записать так:**

$$\rho(X) = \begin{cases} H_0, & T(X) \in T \setminus \tilde{T} \\ H_1, & T(X) \in \tilde{T} \end{cases}$$

2.2 Критерий Смирнова для проверки гипотезы о виде распределения

Имеется выборка $X = (X_1, X_2, \dots, X_n)$ из некоторого распределения F . Проверяется простая гипотеза $H_0 = \{F = F_1\}$ против сложной альтернативы $H_1 = \{F \neq F_1\}$. В том случае, когда распределение F_1 имеет непрерывную функцию распределения F_1 , для проверки гипотезы можно воспользоваться критерием Смирнова.

Пусть $F_n^*(x)$ - эмпирическая функция распределения. Рассмотрим статистику

$$T_n = \sqrt{n} \sup_x (F_n^*(x)) = \sqrt{n} \max_{X_i} (F_n^*(X_i) - F_1(X_i))$$

Теорема 4.1 В случае справедливости гипотезы H_0 при $n \rightarrow +\infty$ статистика T_n имеет распределение Смирнова:

$$\lim_{n \rightarrow \infty} P(T_n < x) = S(x),$$

здесь $S(x) \equiv 1 - e^{-2x^2}$ - функция Смирнова.

Зададим некоторый уровень значимости q и найдем пороговое значение статистики C_q из условия $S(C_q) = 1 - q$. Построим критерий Смирнова для проверки гипотезы о виде распределения:

$$\rho(X) = \begin{cases} H_0, & T_n(X) \leq C_q, \\ H_1, & T_n(X) > C_q. \end{cases}$$

Приведем пример применения этого критерия. Пусть в результате эксперимента получена выборка объема $N=200$ из распределения некоторой случайной величины. Проверим гипотезу о том, что эта случайная величина имеет стандартное нормальное распределение. Текст программы, реализующий проверку данной гипотезы с помощью критерия Смирнова, приведен на рисунке 2.1.

`n := 200`

Задается объем выборки

`x :=`

Выборка считывается из файла

`C:\asmirn.txt`

$$F(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-\frac{t^2}{2}} dt$$

Задается теоретическая функция распределения.

$$I(t,y) := t < y \quad Fn(y) := \frac{1}{n} \cdot \sum_{i=0}^{n-1} I(x_i, y)$$

Определяется эмпирическая функция распределения

$$MAX(x,F) := \begin{cases} c \leftarrow Fn(x_0) - F(x_0) \\ \text{for } i \in 1..n-1 \\ c \leftarrow Fn(x_i) - F(x_i) \text{ if } Fn(x_i) - F(x_i) > c \\ c \end{cases}$$

$$T(x) := \sqrt{n} \cdot MAX(x,F)$$

$$T(x) = 0.507$$

Строится статистика и вычисляется ее значение

$$C(q) := \sqrt{\frac{-\ln(q)}{2}} \quad q = 0.05 \quad C(q) = 1.224$$

Находится пороговое значение $C(q)$, где q - уровень значимости. В данном случае $T(x) < C(q)$, поэтому гипотеза о распределении случайной величины по нормальному закону $N_{0,1}$ принимается

Рисунок 2.1. Проверка гипотезы о распределении случайной величины по нормальному закону $N_{0,1}$ с помощью критерия Смирнова.

2.3 Задание к лабораторной работе

В файле smirn-V.txt (V - номер вашего варианта) задана выборка из некоторого распределения. Задав некоторый уровень значимости α , спомощью критерия Смирнова проверить следующие гипотезы:

Варианты заданий

1. а) $H_0 = \{F = N_{1,2}\}$ против $H_1 = \{F \neq N_{1,2}\}$;
 б) $H_0 = \{F = E_3\}$ против $H_1 = \{F \neq E_3\}$;
 в) $H_0 = \{F = F_{3,5}\}$ против $H_1 = \{F \neq F_{3,5}\}$.
2. а) $H_0 = \{F = U_{2,4}\}$ против $H_1 = \{F \neq U_{2,4}\}$;
 б) $H_0 = \{F = K_{0,1}\}$ против $H_1 = \{F \neq K_{0,1}\}$;
 в) $H_0 = \{F = \Gamma_{2,2}\}$ против $H_1 = \{F \neq \Gamma_{2,2}\}$.
3. а) $H_0 = \{F = \beta_{3,2}\}$ против $H_1 = \{F \neq \beta_{3,2}\}$;
 б) $H_0 = \{F = K_{1,1}\}$ против $H_1 = \{F \neq K_{1,1}\}$;
 в) $H_0 = \{F = E_4\}$ против $H_1 = \{F \neq E_4\}$.
4. а) $H_0 = \{F = \beta_{3,4}\}$ против $H_1 = \{F \neq \beta_{3,4}\}$;
 б) $H_0 = \{F = F_{5,5}\}$ против $H_1 = \{F \neq F_{5,5}\}$;
 в) $H_0 = \{F = U_{-1,1}\}$ против $H_1 = \{F \neq U_{-1,1}\}$.
5. а) $H_0 = \{F = K_{2,4}\}$ против $H_1 = \{F \neq K_{2,4}\}$;
 б) $H_0 = \{F = N_{2,4}\}$ против $H_1 = \{F \neq N_{2,4}\}$;
 в) $H_0 = \{F = U_{0,1}\}$ против $H_1 = \{F \neq U_{0,1}\}$.
6. а) $H_0 = \{F = \beta_{5,4}\}$ против $H_1 = \{F \neq \beta_{5,4}\}$;
 б) $H_0 = \{F = F_{2,4}\}$ против $H_1 = \{F \neq F_{2,4}\}$;
 в) $H_0 = \{F = E_3\}$ против $H_2 = \{F \neq E_3\}$.
7. а) $H_0 = \{F = U_{3,6}\}$ против $H_1 = \{F \neq U_{3,6}\}$;
 б) $H_0 = \{F = K_{1,1}\}$ против $H_1 = \{F \neq K_{1,1}\}$;

8. в) $H_0 = \{F = \Gamma_{2,3}\}$ против $H_1 = \{F \neq \Gamma_{2,3}\}$.
 а) $H_0 = \{F = N_{-1,3}\}$ против $H_1 = \{F \neq N_{-1,3}\}$;
 б) $H_0 = \{F = E_{1,5}\}$ против $H_1 = \{F \neq E_{1,5}\}$;
 в) $H_0 = \{F = \beta_{3,5}\}$ против $H_1 = \{F \neq \beta_{3,5}\}$.
9. а) $H_0 = \{F = F_{2,3}\}$ против $H_1 = \{F \neq F_{2,3}\}$;
 б) $H_0 = \{F = K_{0,2}\}$ против $H_1 = \{F \neq K_{0,2}\}$;
 в) $H_0 = \{F = \Gamma_{2,2}\}$ против $H_1 = \{F \neq \Gamma_{2,2}\}$.
10. а) $H_0 = \{F = \beta_{3,4}\}$ против $H_1 = \{F \neq \beta_{3,4}\}$;
 б) $H_0 = \{F = E_5\}$ против $H_1 = \{F \neq E_5\}$;
 в) $H_0 = \{F = U_{0,5}\}$ против $H_1 = \{F \neq U_{0,5}\}$.
11. а) $H_0 = \{F = \Gamma_{4,4}\}$ против $H_1 = \{F \neq \Gamma_{4,4}\}$;
 б) $H_0 = \{F = F_{2,4}\}$ против $H_1 = \{F \neq F_{2,4}\}$;
 в) $H_0 = \{F = E_3\}$ против $H_1 = \{F \neq E_3\}$.
12. а) $H_0 = \{F = N_{5,1}\}$ против $H_1 = \{F \neq N_{5,1}\}$;
 б) $H_0 = \{F = E_{2,5}\}$ против $H_1 = \{F \neq E_{2,5}\}$;
 в) $H_0 = \{F = U_{3,5}\}$ против $H_1 = \{F \neq U_{3,5}\}$.
13. а) $H_0 = \{F = \beta_{3,2}\}$ против $H_1 = \{F \neq \beta_{3,2}\}$;
 б) $H_0 = \{F = K_{0,1}\}$ против $H_1 = \{F \neq K_{0,1}\}$;
 в) $H_0 = \{F = E_2\}$ против $H_1 = \{F \neq E_2\}$.
14. а) $H_0 = \{F = U_{0,2}\}$ против $H_1 = \{F \neq U_{0,2}\}$;
 б) $H_0 = \{F = \Gamma_{3,3}\}$ против $H_1 = \{F \neq \Gamma_{3,3}\}$;
 в) $H_0 = \{F = K_{2,1}\}$ против $H_1 = \{F \neq K_{2,1}\}$.
15. а) $H_0 = \{F = F_{5,3}\}$ против $H_1 = \{F \neq F_{5,3}\}$;
 б) $H_0 = \{F = E_2\}$ против $H_1 = \{F \neq E_2\}$;
 в) $H_0 = \{F = \Gamma_{2,4}\}$ против $H_1 = \{F \neq \Gamma_{2,4}\}$.
16. а) $H_0 = \{F = N_{0,4}\}$ против $H_1 = \{F \neq N_{0,4}\}$;
 б) $H_0 = \{F = E_{3,5}\}$ против $H_1 = \{F \neq E_{3,5}\}$;

- в) $H_0 = \{F = F_{4,5}\}$ против $H_1 = \{F \neq F_{4,5}\}$.
17. а) $H_0 = \{F = \Gamma_{4,2}\}$ против $H_1 = \{F \neq \Gamma_{4,2}\}$;
- б) $H_0 = \{F = E_5\}$ против $H_1 = \{F \neq E_5\}$;
- в) $H_0 = \{F = U_{-2,2}\}$ против $H_1 = \{F \neq U_{-2,2}\}$.
18. а) $H_0 = \{F = \beta_{3,6}\}$ против $H_1 = \{F \neq \beta_{3,6}\}$;
- б) $H_0 = \{F = F_{2,2}\}$ против $H_1 = \{F \neq F_{2,2}\}$;
- в) $H_0 = \{F = E_{4,5}\}$ против $H_1 = \{F \neq E_{4,5}\}$.
19. а) $H_0 = \{F = F_{3,4}\}$ против $H_1 = \{F \neq F_{3,4}\}$;
- б) $H_0 = \{F = N_{5,5}\}$ против $H_1 = \{F \neq N_{5,5}\}$;
- в) $H_0 = \{F = U_{-1,0}\}$ против $H_1 = \{F \neq U_{-1,0}\}$.
20. а) $H_0 = \{F = K_{2,3}\}$ против $H_1 = \{F \neq K_{2,3}\}$;
- б) $H_0 = \{F = E_{1,5}\}$ против $H_1 = \{F \neq E_{1,5}\}$;
- в) $H_0 = \{F = U_{0,6}\}$ против $H_1 = \{F \neq U_{0,6}\}$.

Лабораторная работа №3. Проверка параметрической гипотезы о виде распределения с помощью критерия согласия χ^2 Пирсона

3.1 Критерий согласия Пирсона

Пусть имеется выборка $X = (X_1, X_2, \dots, X_n)$ из некоторого распределения F . Будем проверять гипотезу о принадлежности, наблюдаемой в опыте случайной величины некоторому семейству распределений. То есть будем проверять сложную гипотезу $H_0 = \{F \in \{F_\theta\}\}$ против сложной альтернативы $H_1 = \{F \notin \{F_\theta\}\}$.

Чтобы воспользоваться критерием Пирсона, выборочные данные предварительно группируют. Разделим область выборочных данных на интервалы $\Delta_1, \Delta_2, \dots, \Delta_k$. Обозначим за $v_j (j=1, 2, \dots, k)$ число элементов выборки, попавших в интервал $\Delta_j (v_1 + \dots + v_k = n)$. Эмпирические вероятности попадания элементов выборки в Δ_j обозначим q_j :

$$q_j = \frac{v_j}{n}, j = 1, \dots, k.$$

За $p_j(\theta)$ обозначим теоретические вероятности попадания значения случайной величины в интервал группировки Δ_j в случае, если выполняется гипотеза H_0 .

Составим статистику, характеризующую отклонение выборочных данных (т.е. вероятностей q_j) от соответствующих гипотетических значений (p_j):

$$\chi_n^2 = \chi_n^2(\theta) = n \sum_{j=1}^k \frac{(q_j - p_j(\theta))^2}{p_j(\theta)}.$$

Непосредственно использовать эту статистику для построения критерия нельзя - для начала надо исключить неопределенность, связанную с неизвестным параметром θ . Для этого поступают

следующим образом - заменяют θ некоторой оценкой $\tilde{\theta}_n = \tilde{\theta}_n(X)$ найденной по выборке X .

$$\tilde{\chi}_n^2 = \tilde{\chi}_n^2(\tilde{\theta}_n) = n \sum_{j=1}^n \frac{(q_j - p_j(\tilde{\theta}_n))^2}{p_j(\tilde{\theta}_n)}.$$

Такую статистику уже можно однозначно вычислить для каждой заданной реализации выборки X .

Теорема 3.1 Если верна гипотеза H_0 и k - размерность векторного параметра θ , то при фиксированном k и при $n \rightarrow \infty$ то

$$\tilde{\chi}_n^2 \rightarrow \xi,$$

где случайная величина ξ имеет распределение χ^2 с $k-1-1$ степенями свободы.

Приведем схему использования критерия согласия χ^2 :

1. По заданной выборке X найти оценку векторного параметра θ .
2. Найти теоретические вероятности попадания значений случайной величины в интервалы группировки Δ_j .
3. Вычислить значение статистики $\tilde{\chi}_n^2$ (см. Приложение Б, таблица 3).
4. По заданному уровню значимости q найти пороговое значение статистики C_q из условия $\chi_{k-1-1}^2(C_q) = 1 - q$.
5. Если $\tilde{\chi}_n^2 < C_q$ то гипотезу H_0 принимаем, в противном случае отклоняем.

Замечание: Малочисленные частоты ($v_j < 5$), следует объединить, объединив соответствующие интервалы группировки. Теоретические частоты следует вычислять уже после объединения интервалов. При определении числа степеней свободы в этом случае вместо k следует взять число интервалов, получившихся после объединения.

Приведем пример использования критерия χ^2 для проверки

параметрической гипотезы о виде распределения.

Пусть имеется выборка объема $n = 500$ из некоторого распределения F . Требуется проверить гипотезу $H_0 = \{F = \Gamma_{1,\lambda}\}$. На рис. 3.1 приведен текст программы для проверки этой гипотезы с помощью критерия χ^2 . Обратите внимание на то, что так как один из параметров гамма-распределения уже задан, то по выборке оценивается только параметр λ .

N := 500

Задается объем выборки

x :=



C:\.lpears2.txt

Вектор, задающий выборку считывается из файла

$x^T =$		0	1	2	3	4	5	6	7	8	9
	0	8.95	2.992	3.925	4.055	5.859	5.455	1.031	1.435	0.4	1.99

i := 0..N-1

lower := floor(min(x))

lower = 0

определяются границы области выборочных данных

upper := floor(max(x)) + 1

upper = 11

k := upper - lower

k = 11

Задается число интервалов группировки данных

i := 0..k-1

 $v_{1,0} := \text{lower} + i$

Формируется таблица частот.

 $v_{1,1} := \text{lower} + (i + 1)$

Для i-го столбца

 $I(i, y) := y < v_{1,1} \wedge y \geq v_{1,0}$

В 1-й строке записано начало i-го интервала группировки
 во 2-й строке - конец этого интервала,
 в 3-й строке - количество попаданий значений случайной
 величины в i-й интервал группировки

$$v_{1,2} := \sum_{j=0}^{N-1} I(i, x_j)$$

$v^T =$		0	1	2	3	4	5	6	7	8	9	10
	0	0	1	2	3	4	5	6	7	8	9	10
	1	1	2	3	4	5	6	7	8	9	10	11
	2	41	141	105	78	55	32	29	10	4	4	1

$x :=$	0	1	2	3	4	5	6	7
	1	2	3	4	5	6	7	10
	42	124	131	86	61	34	14	7

 $x := x^T$

Формируется новая таблица частот, в которой объединены интервалы группировки, количества попаданий значений случайной величины в которые незначительны

L := rows(x)

L = 8

$$Mx := \frac{1}{N} \cdot \sum_{i=0}^{N-1} x_i$$

 $\lambda := Mx$ $\lambda = 3.08$ Методом моментов находится оценка параметра λ

i := 0..L-1

$$q_i := \frac{x_{i,2}}{N}$$

Находятся эмпирические частоты попадания случайной величины в i-й интервал группировки

$$q^T = (0.084 \ 0.248 \ 0.262 \ 0.172 \ 0.122 \ 0.068 \ 0.028 \ 0.014)$$

 $\alpha := 1$

$$p_i = \int_{q_{i,0}}^{q_{i,1}} \frac{\alpha^\lambda}{\Gamma(\lambda)} \cdot t^{\lambda-1} \cdot e^{-t} dt$$

Находятся теоретические частоты попадания значений случайной величины в i-й интервал группировки

$$p^T = (0.072 \ 0.233 \ 0.253 \ 0.19 \ 0.119 \ 0.067 \ 0.035 \ 0.029)$$

$$h := N \cdot \sum_{i=0}^{L-1} \frac{(q_i - p_i)^2}{p_i} \quad h = 7.233$$

$$h_{кр} = \chi^2_{0.95}(L-2-1) \quad h_{кр} = 11.07$$

В данном случае $h < h_{кр}$, поэтому гипотеза о том, что случайная величина имеет гамма-распределение с параметрами $\alpha=1$, $\lambda=2.921$ принимается

Рисунок 3.1. Проверка гипотезы о том, что случайная величина имеет гамма-распределение $\Gamma_{1,\lambda}$

3.2 Задание к лабораторной работе

Варианты заданий

1. а) В ходе испытания 400 ламп накаливания была получена выборка, элементы которой - длительности их горения (в часах). Данные приведены в файле pearson-1a.txt. С помощью критерия Пирсона проверить гипотезу о распределении времени горения ламп по показательному закону с уровнем значимости $\alpha = 0.05$.

б) В течение 3 месяцев (90 дней) в супермаркете вели статистику о количестве проданных за день буханок белого хлеба. Полученная выборка приведена в файле pearson-1b.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества проданных в день буханок по закону Пуассона с уровнем значимости $\alpha = 0.05$.

2. а) В тепличном хозяйстве проводился контроль урожайности томатов некоторого сорта. Было измерено, сколько килограммов томатов было собрано за сезон с каждого из 100 выбранных кустов. Полученные данные приведены в файле pearson-2a.txt. С помощью критерия Пирсона проверить гипотезу о равномерном распределении урожайности томатов с уровнем значимости $\alpha = 0.02$.

б) В результате испытаний 300 дискет на количество циклов перезаписывания, которые дискета выдерживает без выхода из строя, была получена выборка случайной величины, значение которой - количество циклов, выдержанных дискетой до поломки. Выборка приведена в файле pearson-2b.txt с помощью критерия Пирсона проверить гипотезу о распределении этой случайной величины по закону Пуассона с уровнем значимости $\alpha = 0.02$.

3. а) В некоторой местности в течение 300 суток регистрировалась среднесуточная температура воздуха (в градусах Цельсия). В результате была получена выборка, приведенная в файле pearson-3a.txt. С помощью критерия Пирсона проверить гипотезу о равномерном распределении температуры воздуха с уровнем значимости $\alpha = 0.03$.

б) Для проверки качества работы заводского оборудования было подсчитано количество нестандартных деталей, изготовленных каждым из 200 станков за неделю. Полученная выборка приведена в файле pearson-3b.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества нестандартных изделий по закону Пуассона с уровнем значимости

$q = 0.05$.

4. а) В ходе школьного медицинского осмотра был измерен рост 100 первоклассников. Полученные данные приведены в файле pearson-4a.txt. С помощью критерия согласия Пирсона проверить гипотезу о распределении роста учеников по нормальному закону с уровнем значимости $q = 0.05$.

б) Опыт, состоящий в одновременном подбрасывании 5 игральных костей, повторили 150 раз. Событие Ав единичном испытании состоит в том, что на кости выпало не более двух очков. Данные о том, сколько раз при каждом из 150 подбрасываний повторилось событие А приведены в файле pearson-4b.txt. С помощью критерия Пирсона проверить гипотезу о распределении исследуемой случайной величины по биномиальному закону с уровнем значимости $q = 0.05$.

5. а) В результате испытания 300 батареек на длительность их работы (в часах) была получена выборка, приведенная в файле pearson-5a.txt. С помощью критерия Пирсона проверить гипотезу о распределении длительности работы батареек по показательному закону с уровнем значимости $q = 0.03$.

б) Опыт состоял в том, что человек одновременно вытягивал из 3-х карточных колод по 1 карте и складывал их обратно. Опыт было повторен 200раз и каждый раз подсчитывалось сколько вынималось карт пиковой масти. Данные приведены в файле pearson-5b.txt. С помощью критерия Пирсона проверить гипотезу о распределении исследуемой случайной величины по биномиальному закону с уровнем значимости $q = 0.04$.

6. а) В результате взвешивания 500 бильярдных шаров была получена выборка, представленная в файле pearson-6a.txt(вес задан в граммах). С помощью критерия Пирсона проверить гипотезу о равномерном распределении веса шаров с уровнем значимости $q = 0.01$.

б) После поступления на склад 200 коробок стеклянных елочных игрушек кладовщик проверил, сколько игрушек в каждой коробке было повреждено в результате транспортировки. Результаты его проверки приведены в файле pearson-6b.txt. С помощью критерия Пирсона проверить гипотезу распределении количества поврежденных игрушек в 1 коробке по закону Пуассона

с уровнем значимости $\alpha = 0.03$.

7. а) По 100 магазинам города был проведен мониторинг цен на красную икру. В файле pearson-7a.txt приведены цены за 100 граммов икры в каждом из магазинов (в рублях). С помощью критерия Пирсона проверить гипотезу о нормальном распределении цен на икру с уровнем значимости $\alpha = 0.01$.

б) Среди студентов третьего курса (250 человек) был проведен тест по теории вероятностей, состоявший из 20 вопросов. Случайная величина ξ - количество вопросов, на которые каждый студент ответил верно. В итоге была получена выборка из распределения случайной величины ξ , которая приведена в файле pearson-7b.txt. С помощью критерия согласия Пирсона проверить гипотезу о распределении случайной величины ξ по биномиальному закону с уровнем значимости $\alpha = 0.02$ (параметр биномиального закона оценить по выборке).

8. а) Работники рыбного хозяйства выясняли пригодность некоторого озера для разведения карпов, запустив в это озеро 200 меченых мальков и отследив продолжительность их жизни. Длительность жизни карпов (в месяцах) приведена в файле pearson-8a.txt. С помощью критерия Пирсона проверить гипотезу о распределении продолжительность жизни карпов по показательному закону с уровнем значимости $\alpha = 0.05$.

б) Опыт, состоящий в одновременном подбрасывании 4 монет, повторили 200 раз. Данные о том, сколько "гербов" выпало при каждом из повторений опыта, приведены в файле pearson-8b.txt. С помощью критерия Пирсона проверить гипотезу о распределении числа одновременно выпадающих "гербов" по биномиальному закону с уровнем значимости $\alpha = 0.03$.

9. а) Производителем компьютерной техники проводился эксперимент по выявлению длительности безотказной работы жестких дисков. Было исследовано 300 жестких дисков, результаты эксперимента представлены в файле pearson-9a.txt. С помощью критерия Пирсона проверить гипотезу о распределении длительности безотказной работы дисков по показательному закону с уровнем значимости $\alpha = 0.02$.

б) В обувном магазине решили выявить наиболее ходовой размер женской обуви. Для этого продавцы фиксировали размер каждой проданной пары обуви. Данные приведены в файле pearson-

9b.txt. С помощью критерия Пирсона проверить гипотезу о распределении предпочтений покупателей относительно размера обуви по закону Пуассона с уровнем значимости $q = 0.04$.

10. а) Была собрана статистика о продолжительности жизни жителей Красноярска. Данные о продолжительности жизни 1000 человек приведены в файле pearson-10a.txt. С помощью критерия Пирсона проверить гипотезу о распределении продолжительности жизни красноярцев по нормальному закону с уровнем значимости $q = 0.2$.

б) Среди семян ржи имеется некоторое количество семян сорняков. В 300 пакетах семян по 500 штук посчитали количество семян сорняков. Данные приведены в файле pearson-10b.txt. С помощью критерия Пирсона проверить гипотезу о распределении предпочтений покупателей относительно размера обуви по биномиальному закону с уровнем значимости $q = 0.05$ (параметр роценить по выборке).

11. а) В одном из родильных домов была собрана статистика о весе новорожденных мальчиков, родившихся в течение года. Данные о весе 500 детей приведены в файле pearson-11a.txt. С помощью критерия Пирсона проверить гипотезу о равномерном распределении веса новорожденных мальчиков с уровнем значимости $q = 0.01$.

б) Оператор сотовой связи ведет учет количества смс, отправленных за день ее абонентами. Данные о количестве смс, отправленных 600 абонентами за 1 день приведены в файле pearson-11b.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества отправленных смс по закону Пуассона с уровнем значимости $q = 0.01$.

12. а) Имеются данные о количестве электроэнергии, потребленной за 1 месяц жителями 500 квартир. Данные приведены в файле pearson- 12a.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества потребленной энергии по равномерному закону с уровнем значимости $q = 0.05$

б) Любительница бразильских сериалов стала записывать, сколько серий продолжался каждый из просмотренных ею фильмов. Данные о количестве серий в 120 сериалах приведены в файле pearson-12b.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества серий в сериалах по

биномиальному закону с уровнем значимости $\alpha = 0.05$.

13. а) На молочной ферме в течении недели регистрировали удои

коров. В файле pearson-13a.txt приведены средние удои за неделю 120 коров. С помощью критерия Пирсона проверить гипотезу о нормальном распределении удоев коров на молочной ферме с уровнем значимости $\alpha = 0.04$.

б) Студентами 1 курса (400) человек была написана контрольная по русскому языку. Результаты контрольной (по 100 бальной шкале) приведены в файле pearson-13b.txt. С помощью критерия Пирсона проверить гипотезу о распределении баллов студентов по биномиальному закону с уровнем значимости $\alpha = 0.02$.

14. а) На некотором месторождении было взято 100 проб руды на содержание железа. Данные о процентном содержании железа в каждой пробе приведены в файле pearson-14a.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества потребленной энергии по равномерному закону с уровнем значимости $\alpha = 0.05$.

б) Данные о количестве заказных писем, отправленных через некоторое почтовое отделение за 1 день в течение последних 3 месяцев (90 дней) приведены в файле pearson-14b.txt. С помощью критерия Пирсона проверить гипотезу о распределении числа отправляемых в день писем по закону Пуассона с уровнем значимости $\alpha = 0.05$.

15. а) На мелькомбинате было переработано 800 тонн зерна, данные

о количестве муки, получившейся из каждой тонны зерна, приведены в файле pearson-15a.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества полученного при переработке зерна муки по равномерному закону с уровнем значимости $\alpha = 0.02$.

б) На автозаправочной станции 5 суток регистрировали количество автомобилей, заправившихся на бензоколонке в течение каждого часа. Данные приведены в файле pearson-15b.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества заправляющихся на бензоколонке автомобилей в течение часа по биномиальному закону с уровнем значимости

$$q = 0.05$$

16. а) Таксист в течение 3 месяцев (90 дней) измерял расход бензина на своей машине (в литрах на 100 километров). Данные его измерений приведены в файле pearson-16a.txt. С помощью критерия Пирсона проверить гипотезу о распределении количества потребляемого топлива по нормальному закону с уровнем значимости $q = 0.05$.

б) В библиотеке проверили 400 книг на наличие недостающих страниц. В файле pearson-16b.txt приведены данные о том, сколько страниц не доставало в каждой книге. С помощью критерия Пирсона проверить гипотезу о распределении числа недостающих в книгах страниц по закону Пуассона с уровнем значимости $q = 0.01$.

17. а) Интернет-провайдер ведет статистику интернет-трафика для каждого абонента. В файле pearson-17a.txt приведен объем входящего трафика (в мегабайтах) для каждого из 500 абонентов сети за 1 неделю. С помощью критерия Пирсона проверить гипотезу о распределении количества потребленной энергии по показательному закону с уровнем значимости $q = 0.05$.

б) В файле pearson-17b.txt приведены результаты социологического опроса о том, сколько книг за год прочитал каждый из опрошенных (всего 1000 человек). С помощью критерия Пирсона проверить гипотезу о распределении числа недостающих в книгах страниц по закону Пуассона с уровнем значимости $q = 0.02$.

18. а) Оператор сотовой связи ведет учет времени разговоров своих абонентов по исходящей связи. Данные о количестве минут, которые каждый из 500 абонентов проговорил по исходящей связи за 1 сутки, приведены в файле pearson-18a.txt. С помощью критерия Пирсона проверить гипотезу о длительности разговоров абонентов по исходящей связи по показательному закону с уровнем значимости $q = 0.03$.

б) В файле pearson-18b.txt. приведены данные о том, сколько ограблений совершалось ежедневно за последние полгода (180 дней) в городе N. С помощью критерия Пирсона проверить гипотезу о распределении числа ограблений, совершаемых за одни сутки по биномиальному закону с уровнем значимости $q = 0.04$.

19. а) В файле pearson-19a.txt приведены данные о курсе некоторой валюты (в рублях) за последние полгода (180 дней). С помощью критерия Пирсона проверить гипотезу о распределении

стоимости этой валюты по нормальному закону с уровнем значимости $\alpha = 0.05$.

б) На консультации перед экзаменом преподаватель решил выяснить степень готовности студентов. Он спросил каждого о том, сколько вопросов студент уже подготовил (всего на экзамен вынесено 40 вопросов). Данные о готовности 120 студентов приведены в файле pearson-19a.txt. С помощью критерия Пирсона проверить гипотезу о распределении числа подготовленных студентами вопросов по биномиальному закону с уровнем значимости $\alpha = 0.05$.

20. а) В файле pearson-20a.txt приведена стоимость 1 килограмма говядины в 150 магазинах города. С помощью критерия Пирсона проверить гипотезу о распределении стоимости говядины по нормальному закону с уровнем значимости $\alpha = 0.04$

б) Среди студентов университета (300 человек) был проведен опрос о количестве фильмов, просмотренных ими в кинотеатрах. Данные опроса приведены в файле pearson-20b.txt. С помощью критерия Пирсона проверить гипотезу о распределении числа просмотренных в кинотеатрах фильмов по закону Пуассона с уровнем значимости $\alpha = 0.03$.

Лабораторная работа №4. Проверка гипотезы однородности

Одной из важных задач прикладной статистики является задача проверки однородности статистического материала.

Пусть имеются две независимые выборки

$$X = (X_1, X_2, \dots, X_n) \quad \text{и} \quad Y = (Y_1, Y_2, \dots, Y_m),$$

описывающие один и тот же процесс, явление и т.д., но полученные в разное время или в разных условиях. Требуется установить, являются ли они выборками из одного и того же распределения.

Пусть X - выборка из распределения F , а Y - выборка из распределения G . Требуется проверить гипотезу однородности $H_1 = \{F = G\}$: против альтернативы $H_2 = \{H_1 \text{ неверна}\}$.

4.1 Критерий однородности Колмогорова-Смирнова

Этот критерий применяется для непрерывных случайных величин и основан на статистике

$$T_{nm} = \sqrt{\frac{nm}{n+m}} \sup_{t \in \mathbb{R}} |F_n^*(t) - G_m^*(t)|,$$

где $F_n^*(t)$ и $G_m^*(t)$ - эмпирические функции распределения, построенные по выборкам X и Y .

Теорема 6.1 Если гипотеза H_1 верна, то $T_{nm} \rightarrow \xi$ при $n, m \rightarrow 0$,

где случайная величина ξ имеет распределение Колмогорова с функцией распределения $K(t)$.

По заданному уровню значимости q найдем C_q из условия $K(C_q) = 1 - q$. Построим критерий согласия Колмогорова-Смирнова:

$$\rho(X, Y) = \begin{cases} H_1, & T_{n,m} < C_q, \\ H_2, & T_{n,m} \geq C_q. \end{cases}$$

Таким образом, для проверки гипотезы однородности по критерию Колмогорова-Смирнова необходимо следовать

k наблюдений проводились над одной случайной величиной. Другими словами, если p_{ij} - вероятность появления i -го исхода в испытаниях j -й серии, то гипотеза однородности означает утверждение:

$$(p_{ij}, \dots, p_{sj}) = (p_1, \dots, p_s), j = 1, 2, \dots, k,$$

где $p = (p_1, \dots, p_s)$ - некоторый неизвестный вектор вероятностей ($p_1 + \dots + p_s = 1$).

Следуя принципу χ^2 , в качестве меры отклонения опытных данных от их гипотетических значений следовало бы выбрать статистику

$$\chi_n^2(p) = \sum_{i=1}^s \sum_{j=1}^k \frac{(v_{ij} - n_j p_i)^2}{n_j p_i}.$$

Но так как p_i неизвестны, то их нужно предварительно оценить. Оцениваем эти вероятности методом максимального правдоподобия. Получаем следующие оценки:

$$\hat{p}_i = \frac{\mu_i}{n}, \quad \mu_i = \sum_{j=1}^k v_{ij}, \quad i = 1, \dots, s, \quad n = n_1 + n_2 + \dots + n_k \quad (4.1)$$

Таким образом получена следующая статистика критерия:

$$\chi_n^2(p) = n \left(\sum_{i=1}^s \sum_{j=1}^k \frac{v_{ij}^2}{n_j \mu_i} - 1 \right). \quad (4.2)$$

Теорема 6.2 При $n \rightarrow \infty$ $\chi_n^2(p) \rightarrow \xi$ где случайная величина ξ имеет распределение χ^2 с $(s-1)(k-1)$ степенями свободы.

Запишем алгоритм проверки гипотезы однородности с помощью критерия χ^2 :

1. По выборкам X^1, \dots, X^k строим вектор наблюдаемых значений Y .

2. Для каждого исхода Y_1, \dots, Y_s вычисляем число его реализаций в j -й серии.

3. Получаем оценки вероятностей P_1, \dots, P_s по формуле (4.1).

4. Вычисляем значение статистики $\chi_n^2(p)$ по формуле (4.2).

5. По заданному уровню значимости q найдем C_q из условия $\chi_{(s-1)(k-1)}^2(C_q) = 1 - q$ (см. Приложение Б, таблица 3).

6. Гипотезу однородности принимаем, если $\chi_n^2(p) < C_q$ и отклоняем в противном случае.

На рисунке 4.1 приведен текст программы, реализующей проверку гипотезы однородности двух выборок по критерию χ^2 в среде Mathcad.

4.3 Задание к лабораторной работе

а) Было проверено 2 партии теннисных мячей, произведенных на одном заводе. Первая партия состоит из N_1 штук, вторая - из N_2 штук. Каждый мяч был взвешен, веса мячей из первой партии приведены в файле homo-V-1.txt, второй партии - hom-V-2.txt (V - это номер вашего варианта). Проверить гипотезу однородности двух партий теннисных мячей с уровнем значимости q .

б) В файлах homog-V-1.txt, homog-V-2.txt, homog-V-3.txt

(V - это номер вашего варианта) находятся 3 независимые выборки, описывающих работу 3-х смен на заводе, изготавливающих одинаковые детали на одном и том же оборудовании. Элементы выборок - это количества бракованных деталей, произведенных каждым рабочим смены. В первой смене работало N_1 рабочих, во второй - N_2 , в третьей - N_3 . Проверить гипотезу однородности для этих выборок с уровнем значимости q .

Варианты заданий

1. а) $N_1 = 500, N_2 = 400, q = 0.02$; б)

$N_1 = 200, N_2 = 180, N_3 = 190, q = 0.05$.

2. а) $N_1 = 300, N_2 = 250, q = 0.01$; б)

$N_1 = 200, N_2 = 220, N_3 = 195, q = 0.04$.

3. a) $N_1 = 350, N_2 = 250, q = 0.04$; б) $N_1 = 190, N_2 = 210, N_3 = 195, q = 0.03$.
4. a) $N_1 = 220, N_2 = 250, q = 0.05$; б) $N_1 = 185, N_2 = 205, N_3 = 195, q = 0.02$.
5. a) $N_1 = 250, N_2 = 240, q = 0.03$; б) $N_1 = 210, N_2 = 205, N_3 = 200, q = 0.02$.
6. a) $N_1 = 350, N_2 = 400, q = 0.01$; б) $N_1 = 185, N_2 = 190, N_3 = 188, q = 0.05$.
7. a) $N_1 = 400, N_2 = 420, q = 0.04$; б) $N_1 = 220, N_2 = 215, N_3 = 218, q = 0.03$.
8. a) $N_1 = 300, N_2 = 310, q = 0.02$; б) $N_1 = 202, N_2 = 205, N_3 = 200, q = 0.04$.
9. a) $N_1 = 420, N_2 = 430, q = 0.05$; б) $N_1 = 198, N_2 = 202, N_3 = 200, q = 0.03$.
10. a) $N_1 = 500, N_2 = 490, q = 0.01$; б) $N_1 = 199, N_2 = 201, N_3 = 220, q = 0.05$.
11. a) $N_1 = 480, N_2 = 485, q = 0.02$; б) $N_1 = 181, N_2 = 183, N_3 = 187, q = 0.04$.
12. a) $N_1 = 450, N_2 = 420, q = 0.04$; б) $N_1 = 202, N_2 = 218, N_3 = 200, q = 0.02$.
13. a) $N_1 = 380, N_2 = 390, q = 0.03$; б) $N_1 = 150, N_2 = 200, N_3 = 210, q = 0.05$.
14. a) $N_1 = 390, N_2 = 360, q = 0.01$; б) $N_1 = 180, N_2 = 182, N_3 = 185, q = 0.03$.
15. a) $N_1 = 400, N_2 = 490, q = 0.02$; б) $N_1 = 179, N_2 = 180, N_3 = 185, q = 0.04$.
16. a) $N_1 = 350, N_2 = 500, q = 0.05$; б) $N_1 = 184, N_2 = 220, N_3 = 202, q = 0.02$.
17. a) $N_1 = 470, N_2 = 490, q = 0.01$; б) $N_1 = 199, N_2 = 213, N_3 = 200, q = 0.05$.

18. а) $N_1 = 360, N_2 = 300, q = 0.04$; б) $N_1 = 211, N_2 = 201, N_3 = 203, q = 0.03$.
19. а) $N_1 = 380, N_2 = 310, q = 0.02$; б) $N_1 = 214, N_2 = 200, N_3 = 198, q = 0.01$.
20. а) $N_1 = 410, N_2 = 390, q = 0.03$; б) $N_1 = 200, N_2 = 220, N_3 = 204, q = 0.05$.

$N1 := 200$ $x1 := \text{rpois}(N1, 6)$ Генерируем две выборки $x1$ и $x2$ из распределения Пуассона
 $N2 := 180$ $x2 := \text{rpois}(N2, 6)$ P_0 объема 200 и 180 соответственно
 $i := 0$ $y_i := x1_i$

```
form1(x1, N1, r, z) :=
  for i ∈ 0..r
    y_i ← z_i
    j ← r + 1
    flag ← 0
    for i ∈ 0..N1 - 1
      for k ∈ 0..j - 1
        flag ← 1 if y_k = x1_i
      if flag = 0
        y_j ← x1_i
        j ← j + 1
    flag ← 0
  y
```

Функция формирования вектора y , компоненты которого - различные наблюдаемые значения в обеих сериях опытов

$z := \text{form1}(x1, N1, 0, y)$ $r := \text{length}(z)$ Формируем вектор y , применив функцию form1(...)
 $y := \text{form1}(x2, N2, r - 1, z)$ вначале к вектору $x1$, а затем к вектору $x2$

$y^T =$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	6	3	7	8	10	11	5	4	9	1	2	12	13	14

$s := \text{length}(y)$ $s = 15$ Длина вектора y - число различных наблюдаемых значений

```
form2(x1, x2, N1, N2, y, s) :=
  for i ∈ 0..s - 1
    v_{i,0} ← 0
    v_{i,1} ← 0
  for i ∈ 0..s - 1
    for j ∈ 0..N1 - 1
      v_{i,0} ← v_{i,0} + 1 if y_i = x1_j
  for i ∈ 0..s - 1
    for j ∈ 0..N2 - 1
      v_{i,1} ← v_{i,1} + 1 if y_i = x2_j
  v
```

Функция для формирования матрицы v , каждый элемент которой v_{ij} - число реализаций исхода y_i в j -й серии опытов

$v := \text{form2}(x1, x2, N1, N2, y, s)$ Формируем матрицу v

$v^T =$

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	35	23	17	21	10	4	24	28	16	1	15	3	2	0
1	0	25	22	18	15	14	6	34	23	10	3	8	1	0	1

$i := 0..s - 1$ $\mu_i := v_{i,0} + v_{i,1}$ $P_i := \frac{\mu_i}{N1 + N2}$ Формируем вектор p , каждый элемент которого p_i - оценка вероятности исхода y_i , найденная методом максимального правдоподобия

$P^T =$

	0	1	2	3	4	5	6	7
0	$2.632 \cdot 10^{-3}$	0.158	0.118	0.092	0.095	0.063	0.026	0.153

$\chi = (N1 + N2) \sum_{i=0}^{s-1} \left[\frac{(v_{i,0})^2}{N1 \cdot \mu_i} + \frac{(v_{i,1})^2}{N2 \cdot \mu_i} \right] - 1$ $\chi = 14.901$ Вычисляем значение статистики $\chi^2 = \chi_{N1+N2}(P)$ и сравниваем ее значение с $\chi_{кр}$. Гипотеза однородности принимается с уровнем значимости 0.05.

$\chi_{кр} := \text{qchisq}(0.95, s - 1)$ $\chi_{кр} = 23.685$

Рисунок 4.1. Проверка гипотезы однородности с помощью критерия χ^2

Лабораторная работа №5. Проверка гипотезы случайности

5.1 Построение критерия для проверки гипотезы случайности


В различных статистических задачах исходные данные $X = (X_1, \dots, X_n)$ рассматривают как случайную выборку из некоторого распределения F , т.е. считают компоненты X_i вектора данных X независимыми и одинаково распределенными случайными величинами. Однако, иногда такое предположение нуждается в проверке.

Математически задачу можно сформулировать так: проверить гипотезу $H_0 = \{F_x(x) = F(x_1) \cdot \dots \cdot F(x_n), x = (x_1, \dots, x_n)\}$, где $F(x)$ - некоторая функция распределения, против альтернативной гипотезы $H_1 = \{H_0 \text{ неверна}\}$.

Критерий для проверки этой гипотезы строится исходя из следующих соображений: если гипотеза случайности действительно имеет место, то компоненты вектора X "равноправны" и поэтому данные не должны быть ни в каком смысле упорядочены. Следовательно, критерий проверки гипотезы H_0 можно построить на основании статистик, измеряющих степень беспорядка исходных данных.

Одной из таких статистик является число инверсий в выборке. Говорят, что компонента X_i образует μ_i инверсий, если в вариационном ряду, построенном по выборке X левее X_i , стоит μ_i элементов выборки с большими номерами.

На рисунке 5.1 приведен текст программы, вычисляющей количество инверсий для любого элемента выборки по заданному вариационному ряду

$y :=$  CA..rand1.txt

Считываем из файла вариационный ряд. 1-я строка в y^T - это значение элементов выборки, а 2-я строка - это номера соответствующих элементов в исходной выборке $X=(X_0, \dots, X_{N-1})$

$$y^T =$$

	0	1	2	3	4	5	6	7	8	9
0	0.036	0.049	0.063	0.074	0.076	0.095	0.1	0.113	0.161	0.165
1	22	52	113	132	78	54	99	82	5	144

$N := \text{length}(y^{(0)})$ $N = 150$

Определяем объем исходной выборки

```

inv(k) :=
  num ← 0
  for i ∈ 0..N-1
    m ← i if  $y_{i,1} = k$ 
    for i ∈ 0..m-1
      num ← num + 1 if  $y_{i,1} > y_{m,1}$ 
  num

```

Функция, вычисляющая количество инверсий, образованных k-м элементом исходной выборки

$\text{inv}(54) = 3$

Элемент исходной выборки X_{54} образовал 3 инверсии

Рисунок 5.1. Вычисление количества инверсий, образованных элементом X_k

Общее число инверсий для выборки X можно найти по формуле:

$$T_n(X) = \mu_1 + \dots + \mu_{n-1}. \quad (5.1)$$

Нормируем статистику T_n следующим образом:

$$T_n^*(x) = \left(T_n(X) - \frac{n(n-1)}{4} \right) \cdot \frac{6}{n^{3/2}}. \quad (5.2)$$

Теорема 5.1 При $n \rightarrow \infty$ $T_n^* \rightarrow \xi$, где случайная величина ξ имеет стандартное нормальное распределение $N_{0,1}$.

По заданному уровню значимости q найдем C_q из условия $\Phi(-C_q) = q/2$ ($\Phi(x)$ — функция Лапласа). Построим критерий проверки гипотезы случайности:

$$\rho(X) = \begin{cases} H_0, & \text{если } T_n^* \leq C_q, \\ H_1, & \text{в противном случае.} \end{cases}$$

Таким образом, получаем следующий алгоритм проверки гипотезы случайности:

1. По заданной выборке X составляем вариационный ряд.
2. Считаем значение статистик T_n и T_n^* по формулам (5.1), (5.2).
3. Для заданного уровня значимости q определяем C_q из условия $\Phi(-C_q) = q/2$ (см. Приложение Б, таблица 1).
4. Если $T_n^* \leq C_q$, то гипотезу случайности принимаем, в противном случае отклоняем.

5.2 Задание к лабораторной работе

В файле `rand-V.txt` находится некоторая последовательность чисел. Можно ли считать эту последовательность случайной выборкой из некоторого распределения с уровнем значимости $q = 0.0V$ (V - номер вашего варианта)?

Лабораторная работа №6. Проверка гипотезы о независимости, вычисление коэффициента корреляции, построение уравнения линейной регрессии

6.1 Проверка гипотезы независимости с помощью критерия χ^2

Предположим, что в некотором эксперименте наблюдается случайная величина $\psi = (\xi, \eta)$ с неизвестной функцией распределения $F_\psi(x, y)$, и есть основание предполагать, что компоненты ξ и η независимы. В этом случае надо проверить гипотезу независимости $H_0 = \{F_\psi(x, y) = F_\xi(x) \cdot F_\eta(y)\}$, где $F_\xi(x)$ и $F_\eta(y)$ - некоторые одномерные функции распределения, против альтернативной гипотезы $H_1 = \{H_0 \text{ неверна}\}$.

Итак, пусть имеется выборка $(X, Y) = ((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ из распределения случайной величины $\psi = (\xi, \eta)$. Простой критерий согласия для проверки гипотезы H_0 для этой выборки можно построить, основываясь на методике χ^2 .

Как известно, эту методику применяют для дискретных моделей с конечным числом исходов, поэтому условимся считать, что случайная величина ξ принимает конечное число различных значений, которые обозначим u_1, u_2, \dots, u_s , а вторая компонента η - k значений v_1, v_2, \dots, v_k .

Если исходная модель имеет другую структуру, то предварительно группируют возможные значения случайных величин отдельно по первой и второй компонентам: множество значений ξ разбивается на интервалов $\Delta_1, \Delta_2, \dots, \Delta_s$, множество значений η на k интервалов $\nabla_1, \nabla_2, \dots, \nabla_k$, а само множество значений $\psi = (\xi, \eta)$ на $N = sk$ прямоугольников $\Delta_i \times \nabla_j$

Обозначим через v_{ij} число наблюдений пары (u_i, v_j) (или число элементов выборки, принадлежащих прямоугольнику $\Delta_i \times \nabla_j$, если данные группируются), так что
$$\sum_{i=1}^s \sum_{j=1}^k v_{ij} = n$$
. Результаты наблюдений

удобно

расположить в виде таблицы сопряженности двух признаков:

	η_j				
ξ_i	v_1	v_2	...	v_k	Сумма
u_1	v_{11}	$v_{12} \dots$		v_{1k}	$v_{1\bullet}$
u_2	v_{21}	$v_{22} \dots$		v_{2k}	$v_{2\bullet}$
...
u_s	v_{s1}	$v_{s2} \dots$		v_{sk}	$v_{s\bullet}$
Сумма	$v_{\bullet 1}$	$v_{\bullet 2} \dots$		$v_{\bullet k}$	n

Далее вычисляем значение статистики


$$\hat{X}_n^2 = n \left(\sum_{i,j} \frac{v_{ij}^2}{v_{i\bullet} v_{\bullet j}} - 1 \right).$$

Теорема 6.1 В случае справедливости гипотезы H_0 при $n \rightarrow \infty$ $\hat{X}_n^2 \rightarrow \xi$, где случайная величина ξ имеет распределение χ^2 с $(s-1)(k-1)$ степенями свободы.

Построим критерий согласия для проверки гипотезы независимости:

$$\rho(X, Y) = \begin{cases} H_0, & \text{если } \hat{\chi}_T^2 < \chi_{1-q, (s-1)(k-1)}^2, \\ H_1, & \text{если } \hat{\chi}_T^2 \geq \chi_{1-q, (s-1)(k-1)}^2 \end{cases}$$

На рисунке 6.1 приведен текст программы, реализующей проверку гипотезы независимости по критерию χ^2 в среде Mathcad.

A :=  C:\independent.tbl Считываем из файла матрицу A, 1-й столбец которой - значения случ. величины X, а 2-й - соответствующие значения случ. величины Y

x := A (a) $x^T =$

	0	1	2	3	4	5	6	7	8
0	0.342	0.33	0.339	0.371	3.333	3.315	0.1	0.249	2.406

y := A (d) $y^T =$

	0	1	2	3	4	5	6	7	8
0	2.835	0.958	0.724	1.209	3.032	1.073	2.323	5.851	2.279

N := rows(A) N = 200 Определяем объем выборки (X,Y)

U := $\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 7 \\ 1 & 2 & 3 & 4 & 5 & 7 & 14 \\ 77 & 61 & 23 & 12 & 9 & 12 & 6 \end{pmatrix}^T$ $l_X := rows(U)$ $l_X = 7$ Составляем таблицы частот для каждого из признаков:
U - таблица частот для выборки X,
V - таблица частот для выборки Y

V := $\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 & 12 \\ 66 & 51 & 35 & 22 & 12 & 7 & 7 \end{pmatrix}^T$ $l_Y := rows(V)$ $l_Y = 7$ l_X и l_Y - число интервалов группировки в соответствующих таблицах

$$I(i,j,x,y) := x \geq U_{i,0} \wedge x < U_{i,1} \wedge y \geq V_{j,0} \wedge y < V_{j,1}$$

$$i := 0..l_X - 1 \quad j := 0..l_Y - 1$$

$$v_{i,j} := \sum_{k=0}^{N-1} I(i,j,x_k,y_k)$$

$$v = \begin{pmatrix} 28 & 17 & 14 & 9 & 3 & 3 & 3 \\ 22 & 15 & 11 & 4 & 4 & 4 & 1 \\ 6 & 8 & 4 & 3 & 1 & 0 & 1 \\ 4 & 3 & 3 & 2 & 0 & 0 & 0 \\ 2 & 3 & 1 & 1 & 2 & 0 & 0 \\ 4 & 5 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 & 2 \end{pmatrix}$$

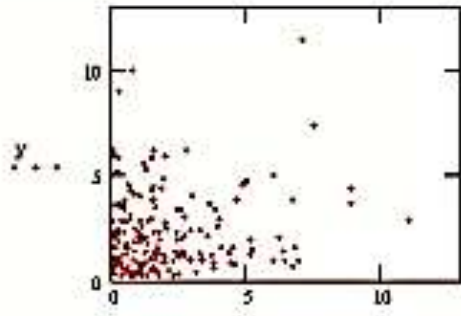
Составляем таблицу сопряженности 2-х признаков

$$\chi := N \cdot \left[\sum_{i=0}^{l_X-1} \sum_{j=0}^{l_Y-1} \frac{(v_{i,j})^2}{U_{i,2} \cdot V_{j,2}} - 1 \right] \quad \chi = 40.315$$

Вычисляем значение статистики

q := 0.05 $\chi_{кр} := qchi2q[1 - q, (l_X - 1) \cdot (l_Y - 1)]$ $\chi_{кр} = 50.998$

Т.к. $\chi < \chi_{кр}$, то гипотезу о независимости случайных величин X и Y принимаем с уровнем значимости $q=0.05$



Для наглядности элементы выборки (X,Y) нанесем на плоскость

Рисунок 6.1. Проверка гипотезы независимости с помощью критерия χ^2 .

6.2 Выборочный коэффициент корреляции. Проверка гипотезы о значимости выборочного коэффициента корреляции

Как известно из курса теории вероятностей, коэффициент корреляции

$$r = \frac{M(\xi\eta) - M(\xi)M(\eta)}{\sigma(\xi)\sigma(\eta)}$$

характеризует наличие (или отсутствие) линейной зависимости между двумя случайными величинами ξ и η .

При $r \neq 0$ случайные величины ξ и η называются коррелированными, а при $r = 0$ - некоррелированными.

Необходимо помнить, что в общем случае некоррелированность случайных величин еще не означает их независимости.

Коэффициент корреляции удовлетворяет неравенству

$$-1 \leq r \leq 1,$$

и если $r = \pm 1$, то ξ и η связаны линейной функциональной зависимостью.

Пусть в результате эксперимента получена выборка $(X, Y) = ((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ из распределения случайной величины (ξ, η) .

Исходя из определения коэффициента корреляции и точечных оценок для математического ожидания и среднеквадратического отклонения, дадим определение выборочного коэффициента корреляции r_B :

$$r_B = \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X} \cdot \bar{Y}}{s_0(X) s_0(Y)},$$

где \bar{X} и \bar{Y} - выборочные средние для выборок X и Y , а $s_0(X)$ и $s_0(Y)$ - соответствующие несмещенные выборочные средние

квадратические отклонения.

r_B является точечной оценкой коэффициента корреляции случайных величин ξ и η .

Разумеется, из того, что найденный по выборке $r_B \neq 0$ не следует, что коэффициент корреляции генеральной совокупности $r \neq 0$. В связи с этим проверяют гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности (гипотезу о значимости выборочного коэффициента корреляции): $H_0 = \{r = 0\}$ против альтернативной гипотезы $H_1 = \{r \neq 0\}$.

Если гипотеза H_0 будет принята, что это будет означать, что выборочный коэффициент корреляции незначим, а величины ξ и η некоррелированы, если же будет принята H_1 , то что выборочный коэффициент корреляции значим, а величины ξ и η коррелированы.

Предположим, что двумерная генеральная совокупность (ξ, η) распределена нормально. Тогда, если при проверки гипотезы о значимости выборочного коэффициента корреляции мы получим, что ξ и η некоррелированы, то мы имеем право сделать вывод, что они независимы.

Вычисляем значение следующей статистики:

$$\hat{T}_n = \frac{r_B \sqrt{n-2}}{\sqrt{1-r_B^2}}$$

Теорема 6.2 *В случае справедливости гипотезы H_0 при $n \rightarrow \infty$ $\hat{T}_n \rightarrow \xi$, где случайная величина ξ имеет распределение Стьюдента с $n-2$ степенями свободы.*

Построим критерий для проверки гипотезы о значимости выборочного коэффициента корреляции:

$$\rho(X, Y) = \begin{cases} H_0, & \text{если } \hat{T}_n < t_{1-q/2, n-2} \\ H_1, & \text{если } \hat{T}_n \geq t_{1-q/2, n-2} \end{cases},$$

где $t_{1-q/2, n-2}$ - квантиль распределения Стьюдента уровня $1 - q/2$ (см. Приложение Б, таблица 2).

6.3 Линейная регрессия

Пусть наблюдаемая случайная величина η зависит от случайной величины ξ . Обозначим через $f(x)$, функцию задающую зависимость среднего значения η от значений ξ

$$M(\eta/\xi = x) = f(x).$$

Уравнение $y = f(x)$ называется уравнением регрессии.

Проведем экспериментов, в результате которых случайная величина ξ примет последовательно значения X_1, X_2, \dots, X_n , и получим соответствующие значения случайной величины η : Y_1, Y_2, \dots, Y_n . Обозначим разницу между Y_i и ее математическим ожиданием

$$\alpha_i = Y_i - M(\eta/\xi = X_i) = Y_i - f(X_i).$$

Обычно предполагают, что α_i - независимы и распределены нормально с параметрами $0, \sigma^2$.

Требуется по значениям X_1, \dots, X_n и Y_1, \dots, Y_n оценить как можно точнее функцию $f(x)$. Сначала заранее определяют вид функции $f(x)$. Будем предполагать, что $f(x)$ - линейная функция

$$f(x) = ax + b.$$

Оценки неизвестных параметров a и b находят с помощью метода максимального правдоподобия или метода наименьших квадратов, суть которого мы рассмотрим несколько позже.

Эти оценки выглядят следующим образом:

$$a = \frac{\sigma(\eta)}{\sigma(\xi)} r, b = M(\eta) - rM(\xi) \frac{\sigma(\eta)}{\sigma(\xi)}.$$

Прямая

$$y = M(\eta) + r \frac{\sigma(\eta)}{\sigma(\xi)} (x - M(\xi))$$

называется *прямой среднеквадратической регрессии η на ξ* .

Величина $\Delta = \sigma^2(\eta)(1 - r^2)$ называется остаточной дисперсией η на ξ . Она определяет величину ошибки приближенного равенства $\eta \approx a\xi + b$. Если $r = \pm 1$, то ошибка равна нулю, а величины η и ξ связаны линейной функциональной зависимостью.

Теперь, заменяя $M(\xi)$, $M(\eta)$, $\sigma(\xi)$, $\sigma(\eta)$ и r на их точечные оценки, получаем уравнение выборочной прямой среднеквадратической регрессии η на ξ :

$$y = \bar{Y} + r_B \frac{s_0(Y)}{s_0(X)} (x - \bar{X})$$

Аналогично получается уравнение выборочной прямой среднеквадратической регрессии ξ на η :

$$x = \bar{X} + r_B \frac{s_0(X)}{s_0(Y)} (y - \bar{Y})$$

6.4 Задание к лабораторной работе

а) В файле ind-V.txt (V - это номер вашего варианта) в виде матрицы задана выборка (X, Y) из двумерного распределения. Первый столбец матрицы - значения X, второй столбец - соответствующие значения Y. Проверить гипотезу о независимости случайных величин, представленных выборками X и Y с уровнем значимости q .

б) В файле cor-V.txt (V - это номер вашего варианта) находятся выборка (X, Y) из двумерного нормального распределения случайной величины (ξ, η). Первый столбец матрицы - значения X, второй столбец - соответствующие значения Y. Найти выборочный коэффициент корреляции. С уровнем значимости q проверить гипотезу о значимости выборочного

коэффициента корреляции. Являются ли величины ξ и η независимыми?

На плоскости Охунанести элементы выборки (X, Y) и построить прямую среднеквадратической регрессии η на ξ , определить остаточную дисперсию η на ξ . Сделать вывод о правомерности описания зависимости $\eta(\xi)$ линейной функцией.

Варианты заданий

1. а) $q = 0.02$; б) $q = 0.05$.
2. а) $q = 0.01$; б) $q = 0.04$.
3. а) $q = 0.04$; б) $q = 0.03$.
4. а) $q = 0.05$; б) $q = 0.02$.
5. а) $q = 0.03$; б) $q = 0.02$.
6. а) $q = 0.01$; б) $q = 0.05$.
7. а) $q = 0.04$; б) $q = 0.03$.
8. а) $q = 0.02$; б) $q = 0.04$.
9. а) $q = 0.05$; б) $q = 0.03$.
10. а) $q = 0.01$; б) $q = 0.05$.
11. а) $q = 0.02$; б) $q = 0.04$.
12. а) $q = 0.04$; б) $q = 0.02$.
13. а) $q = 0.03$; б) $q = 0.05$.
14. а) $q = 0.01$; б) $q = 0.03$.
15. а) $q = 0.02$; б) $q = 0.04$.
16. а) $q = 0.05$; б) $q = 0.02$.
17. а) $q = 0.01$; б) $q = 0.05$.
18. а) $q = 0.04$; б) $q = 0.03$.
19. а) $q = 0.02$; б) $q = 0.02$.
20. а) $q = 0.03$; б) $q = 0.02$.

Лабораторная работа №7. Дисперсионный анализ

Дисперсионный анализ - это статистический метод, предназначенный для выявления влияния отдельных факторов на результат эксперимента. Суть метода заключается в том, что общая вариация результирующего показателя расчленяется на части, соответствующие совместному и отдельному влиянию различных качественных факторов, и остаточную вариацию, аккумулирующую влияние неучтенных факторов. Статистическое изучение этих частей позволяет делать выводы о том, действительно ли тот или иной качественный фактор оказывает влияние на результирующий показатель.

Дисперсионный анализ основан на следующих допущениях:

1) наблюдения результирующего фактора ξ - это нормально распределенная случайная величина с центром распределения $M\xi = \phi(b_1, \dots, b_m)$, где b_1, \dots, b_m - это независимых управляющих качественных факторов;

2) дисперсия единичного наблюдения, обусловленная случайными ошибками, постоянна во всех опытах и не зависит от b_1, \dots, b_m .

По числу факторов, влияние которых исследуется, различают однофакторный и многофакторный дисперсионный анализ.

7.1 Однофакторный дисперсионный анализ

Как следует из названия, данным методом исследуется влияние на результирующий признак одного качественного показателя.

Пусть в результате эксперимента получено r групп выборочных значений результирующего признака

$X_{ij} (j = 1, \dots, n_i, i = 1, \dots, r)$, соответствующих значениям качественного фактора; n_i - это количество

наблюдений для i -го значения качественного фактора $\left(\sum_{i=1}^r n_i = n \right)$.

Пусть $a_i (i = 1, \dots, r)$ - групповые средние, а $a = \frac{1}{r} \sum_{i=1}^r a_i$ - общее

(генеральное) среднее.

Будем проверять гипотезу $H_0 = \{a_1 = \dots = a_r = a\}$ о том, что качественный фактор не влияет на результирующий признак против альтернативной гипотезы $H_1 = \{H_0 \text{ неверна}\}$.

Определим общее и групповые выборочные средние (соответственно \bar{X} и \bar{X}_i):

$$\bar{X} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ji}, \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ji}$$

Как известно, выборочные групповые средние являются несмещенными и состоятельными оценками средних a_i .

Представим полную сумму квадратов отклонений результирующего признака от общего среднего в виду двух сумм квадратов отклонений:

$$Q = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ji} - \bar{X})^2 = \sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ji} - \bar{X}_i)^2 = Q_1 + Q_2$$

Сумма Q_1 представляет собой сумму квадратов отклонений групповых средних значений от общего среднего значения ("сумма квадратов между группами"), т.е. вариацию, обусловленную качественным фактором, а сумма Q_2 является суммой квадратов отклонения каждой величины от соответствующего группового среднего значения ("сумма квадратов внутри групп"), т.е. остаточную вариацию, обусловленную случайными отклонениями от групповых средних.

Теорема 7.1 В случае справедливости гипотезы H_0 величина

$$F = \frac{Q_1 / (r - 1)}{Q_2 / (n - r)}$$

имеет распределение Фишера с $r - 1, n - r$ степенями свободы.

Отсюда, для проверки гипотезы H_0 при уровне значимости α получаем следующий критерий:

$$\rho(X) = \begin{cases} H_0, & \text{если } F \leq F_{1-q, r-1, n-r}, \\ H_1, & \text{в противном случае} \end{cases}$$

На практике для вычисления сумм Q_1, Q_2, Q бывает удобнее пользоваться формулами

$$Q_1 = \sum_{i=1}^r \frac{\left(\sum_{j=1}^{n_i} X_{ji} \right)^2}{n_i} - \frac{\left(\sum_{i=1}^r \sum_{j=1}^{n_i} X_{ji} \right)^2}{n},$$

$$Q_2 = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ji}^2 - \sum_{i=1}^r \frac{\left(\sum_{j=1}^{n_i} X_{ji} \right)^2}{n_i}, \quad Q = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ji}^2 - \frac{\left(\sum_{i=1}^r \sum_{j=1}^{n_i} X_{ji} \right)^2}{n}$$

Приведем пример. Предположим, на экспертную оценку отправлено 15 видов товара. Каждого вида товара опрашивалось по 20 образцов. Оценив каждый образец, эксперт должен был дать среднюю оценку каждому виду товара. Экспертиза проводилась двумя экспертами. Необходимо выяснить, насколько субъективной была эта экспертиза. Экспертные оценки приведены в следующей таблице:

Средние оценки экспертов по каждому виду товара	Виды товара														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1-й эксперт	4	3,2	4,6	3,8	3,4	3,2	3,5	4,6	3,7	4,1	5	3,3	4	5	4,9
2-й эксперт	3,9	3,9	4,1	4,3	1,9	3,2	2,3	5	4,9	2,7	4,1	5	3,6	4	2,5

На рисунке 7.1 приведен текст программы в среде Mathcad, проверяющей гипотезу о том, что личность эксперта не влияет на

оценку товаров. По результатам статистического анализа эта гипотеза была принята.

$n_1 := 15$ $n_2 := 15$ $n := n_1 + n_2$ Вводим количество наблюдений для каждого из значений качественного фактора и вычисляем общее число наблюдений

$x := (4 \ 3.2 \ 4.6 \ 3.8 \ 3.4 \ 3.2 \ 3.5 \ 4.6 \ 3.7 \ 4.1 \ 5 \ 3.3 \ 4 \ 5 \ 4.9)^T$ Вводим значения результирующего признака при каждом значении качественного фактора

$y := (3.9 \ 3.9 \ 4.1 \ 4.3 \ 1.9 \ 3.2 \ 2.3 \ 5 \ 4.9 \ 2.7 \ 4.1 \ 5 \ 3.6 \ 4 \ 2.5)^T$

$$Q_1 := \frac{\left(\sum_{i=0}^{n_1-1} x_i\right)^2}{n_1} + \frac{\left(\sum_{i=0}^{n_2-1} y_i\right)^2}{n_2} - \frac{\left[\sum_{i=0}^{n_1-1} x_i + \sum_{i=0}^{n_2-1} y_i\right]^2}{n}$$

$Q_1 = 0.8$ Вычисляем "сумму квадратов между группами" - Q_1 , "сумму квадратов внутри групп" - Q_2 и "общую сумму квадратов" - Q

$$Q_2 := \sum_{i=0}^{n_1-1} (x_i)^2 + \sum_{i=0}^{n_2-1} (y_i)^2 - \left[\frac{\left(\sum_{i=0}^{n_1-1} x_i\right)^2}{n_1} + \frac{\left(\sum_{i=0}^{n_2-1} y_i\right)^2}{n_2} \right]$$

$Q_2 = 19.613$

$$Q := \sum_{i=0}^{n_1-1} (x_i)^2 + \sum_{i=0}^{n_2-1} (y_i)^2 - \frac{\left(\sum_{i=0}^{n_1-1} x_i + \sum_{i=0}^{n_2-1} y_i\right)^2}{n}$$

$Q = 20.414$ $Q_1 + Q_2 = Q = 1$

$R_1 := Q_1$ $R_2 := \frac{Q_2}{n-3}$ $R := \frac{Q}{n-1}$ Находим среднее квадратов - R_1, R_2, R_3

$R_1 = 0.8$ $R_2 = 0.726$ $R = 0.704$

$F := \frac{R_1}{R_2}$ $F = 1.102$ Вычисляем дисперсионное отношение F

$F_{кр} := qF(0.95, 1, n-2)$ $F_{кр} = 4.196$ Т. к. $F < F_{кр}$, то делаем вывод о том, личность эксперта существенно не влияет на оценку товаров

Рисунок 7.1. Проверка гипотезы об отсутствии влияния одного качественного фактора на результирующий показатель

7.2 Двухфакторный дисперсионный анализ

В данном случае исследуется наличие или отсутствие влияния на результирующий признак двух качественных показателей.

Пусть рассматривается два фактора - А и В. Фактор А может принимать r значений ($A = \{A_1, \dots, A_r\}$) а фактор В - s значений ($B = \{B_1, \dots, B_s\}$).

В результате эксперимента получены выборочные значения

результатирующего признака X_{jik} , $j=1, \dots, n_{ik}$, $i=1, \dots, r$, $k=1, \dots, s$; n_{ik} - это количество наблюдений при i -м значении качественного фактора A_i и k -м значении качественного фактора B_k .

В $\left(\sum_{i=1}^r \sum_{k=1}^s n_{ik} = n \right)$.

По указанной выборке будем проверять справедливость следующих гипотез:

H_A - о том, что качественный фактор A не влияет на результирующий признак,

H_B - о том, что фактор B не влияет на результирующий признак,

H_{AB} - о том, что взаимодействие факторов A и B не влияет на результирующий признак.

Для этого вводим общее и групповое выборочные средние:

$$\bar{X} = \frac{1}{n} \sum_{k=1}^s \sum_{i=1}^r \sum_{j=1}^{n_{ik}} X_{jik}, \quad \bar{X}_{ik} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} X_{jik}.$$

Вычисляем значения Q, Q_1, Q_2, Q_3, Q_4 по формулам:

$$Q = \sum_{k=1}^s \sum_{i=1}^r \sum_{j=1}^{n_{ik}} (X_{jik} - \bar{X})^2, \quad Q_1 = \sum_{k=1}^s \sum_{i=1}^r n_{ik} \left(\frac{1}{s} \sum_{k=1}^s \bar{X}_{ik} - \bar{X} \right)^2,$$

$$Q_2 = \sum_{k=1}^s \sum_{i=1}^r n_{ik} \left(\frac{1}{r} \sum_{i=1}^r \bar{X}_{ik} - \bar{X} \right)^2, \quad (7.1)$$

$$Q_3 = \sum_{k=1}^s \sum_{i=1}^r n_{ik} \left(\bar{X}_{ik} - \frac{1}{s} \sum_{k=1}^s \bar{X}_{ik} - \frac{1}{r} \sum_{i=1}^r \bar{X}_{ik} + \bar{X} \right)^2,$$

$$Q_4 = \sum_{k=1}^s \sum_{i=1}^r \sum_{j=1}^{n_{ik}} (X_{jik} - \bar{X}_{ik})^2.$$

Проверку гипотез проводим по следующему критерию:

а) Если $\frac{Q_1 / (r-1)}{Q_4 / (n-rs)} \geq F_{1-q, r-1, n-rs}$, то гипотеза H_A отвергается;

б) Если $\frac{Q_2 / (s-1)}{Q_4 / (n-rs)} \geq F_{1-q, s-1, n-rs}$, то гипотеза H_B отвергается;

в) Если $\frac{Q_3 / ((r-1)(s-1))}{Q_4 / (n-rs)} \geq F_{1-q, (r-1)(s-1), n-rs}$, то гипотеза H_{AB} отвергается (см. Приложение Б, таблица 4).

7.3 Задание к лабораторной работе

Исследовать влияние на результирующий показатель: а) одного качественного фактора, б) двух качественных факторов. Уровень значимости для проверки гипотез взять равным $q = 0.05$.

Указание: Под буквой б) данные необходимо считывать из текстового файла. Данные в файле располагаются в виде матрицы следующим образом: в 1-м столбце данные соответствуют значениям факторов (A_1, B_1) , во 2-м столбце - значениям (A_2, B_1) , в 3-м - (A_1, B_2) , в 4-м - (A_2, B_2) .

Соответственно, чтобы можно было воспользоваться формулами (7.1), необходимо правильно считать данные из файла, как это сделано, например, в программе, текст которой приведен на рисунке 7.2.

```

I :=
  C:\..\disp-1.txt
X_{0,0} := x^{(0)}   X_{0,1} := x^{(2)}   X_{1,0} := x^{(1)}   X_{1,1} := x^{(3)}
  
```

Рисунок 7.2. Пример считывания данных из файла

Если вы считали данные подобным образом, то в дальнейшем, чтобы обратиться к элементу X_{j_ik} , пользуйтесь записью $(X_{i,k})_j$.

7.4 Варианты заданий

1. а) Исследовать влияние посещения секций и кружков во внеклассное время на успеваемость школьников. Качественный фактор - количество часов, проводимых школьниками на дополнительных занятиях. Результирующий признак - средние баллы учеников по совокупности предметов за год. Согласно значениям качественного фактора ученики были поделены на 3 группы (по 16, 12 и 10 человек соответственно).

Данные о средних баллах школьников приведены в таблице:

б) Исследовать влияние на скорость прорастания семян томатов следующих факторов: температуры воздуха (фактор А) и влажности воздуха (фактор В). Значения фактора $A:A=A_1$ - температура воздуха ниже 22° , $A=A_2$ - температура воздуха выше 22° . Значения фактора $B:B=B_1$ - влажность воздуха ниже 90 процентов, $B=B_2$ - влажность воздуха выше 90 процентов. Время прорастания семян (в часах) при различных значениях факторов приведено в файле disp-1.txt.

Количество часов Т, проводимое на дополнительных занятиях	Средний балл за год															
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}	X _{16i}
T=0	3,74	3,35	4,48	5,00	3,93	3,88	3,40	4,10	4,21	3,81	4,64	4,58	3,99	3,72	2,97	3,70
0<T≤3	4,36	4,06	3,52	4,44	3,22	4,27	4,42	3,89	3,68	3,95	4,17	4,29				
T>3	2,55	4,03	3,48	3,46	3,74	3,35	3,87	3,26	4,65	2,34						

2. а) Исследовать влияние поведения цены за барель нефти на курс акций некоторого предприятия. Результирующий признак - цена акции предприятия. Согласно значениям качественного фактора (поведения цены на нефть) данные были поделены на 3 группы (по 15, 14 и 13 значений соответственно). Данные о курсе акций предприятия приведены в следующей таблице:

Поведение цены на нефть	Курс акции														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
цена растет	106,12	97,22	99,55	100,07	99,13	98,98	98,13	102,12	99,55	105,00	100,18	101,60	96,35	99,35	100,60
цена стабильна	97,96	102,12	102,07	103,75	98,46	100,92	98,84	99,50	103,10	103,64	96,82	100,98	95,11	98,55	
цена падает	100,58	97,99	101,65	101,45	100,41	99,83	99,04	100,81	96,69	100,46	100,73	101,11	101,22		

б) Исследовать влияние на успеваемость студентов по физике следующих факторов: посещаемости лекционных занятий (фактор А) и активности при работе на практических занятиях (фактор В). Значения фактора $A:A=A_1$ - посещаемость ниже 60 процентов, $A=A_2$ - посещаемость выше 60 процентов. Значения фактора $B:B=B_1$ - низкая активность, $B=B_2$ - высокая активность. Средние баллы студентов по физике за 4 семестра при различных значениях факторов приведены в файле disp-2.txt.

3. а) Исследовать влияние на урожайность огурцов уровня влажности воздуха в теплице. Результирующий признак -

количество килограммов огурцов, собранных с одного куста за сезон. Согласно значениям качественного фактора данные были поделены на 3 группы (по 16, 14 и 15 значений соответственно). Данные об урожайности огурцов приведены в таблице

Влажность воздуха	Урожай с одного куста огурцов															
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}	X _{16i}
низкая	19,84	18,64	19,90	22,47	20,72	21,45	24,50	21,95	17,92	22,05	20,75	20,52	20,04	19,12	20,27	23,99
средняя	18,20	19,96	23,91	18,51	23,05	17,24	20,69	19,34	17,64	17,81	20,57	24,69	16,08	17,03		
высокая	20,05	20,74	21,82	20,39	18,33	22,59	18,53	21,82	19,74	23,85	20,90	18,71	17,68	18,22	21,25	

б) Исследовать влияние на размер выпускаемой на заводе детали следующих факторов: станка, на котором производится деталь (фактор А) и рабочего, изготавливающего деталь (фактор В). Значения фактора $A : A = A_1$ - 1-й станок, $A = A_2$ - 2-й станок. Значения фактора $B : B = B_1$ - 1-й рабочий, $B = B_2$ - 2-й рабочий. Размеры получаемых деталей (в миллиметрах) при различных значениях факторов приведены в файле disp-3.txt.

Сезон	Количество потребленной энергии														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
зима	255,95	295,30	324,45	292,69	278,12	291,27	290,55	305,73	330,64	269,81,	329,45	283,01	298,33	314,13	327,66
лето	319,14	274,65	325,24	288,73	309,48	326,89	320,48	314,63	293,57	314,86	275,53	329,28	287,35	304,11	
межсезонье	301,13	331,65	330,68	318,07	335,80	338,58	332,63	307,37	343,26	308,51	352,80	315,01	334,53	280,54	364,32

4. а) Исследовать влияние фактора сезонности на среднее за сезон количество потребляемой энергии семьей из 3-х человек. Согласно значениям качественного фактора данные были поделены на 3 группы (по 15, 14 и 15 значений соответственно). Данные о количестве потребленной энергии (в кВт) приведены в таблице

б) Исследовать влияние на всхожесть семян следующих факторов: свежести семян (фактор А) и освещенности теплицы (фактор В). Значения фактора $A : A = A_1$ - свежие семена (прошлого урожая), $A = A_2$ - не свежие семена. Значения фактора $B : B = B_1$ - хорошая освещенность, $B = B_2$ - плохая освещенность. Доли взошедших семян из каждой пачки приведены в файле disp-4.txt.

5 а) Исследовать влияние уровня кислотности почвы на урожайность свеклы. Согласно значениям качественного фактора

данные были поделены на 3 группы (по 13, 14 и 15 значений соответственно). Данные о среднем урожае свеклы с 1 Га земли для каждого хозяйства приведены в таблице

Вид почвы	Средняя урожайность свеклы в 1 Га (в Цт.)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
кислая	546,10	594,95	570,20	605,79	565,66	601,77	592,77	598,43	576,24	600,90	579,93	603,81	572,39		
щелочная	592,60	615,74	589,76	583,63	584,38	619,75	581,59	577,47	611,24	586,72	609,74	611,95	609,35	602,63	607,07
нормальная	629,24	621,44	611,20	612,20	617,45	630,55	650,25	616,24	625,16	625,03	649,04	641,62	573,71	626,92	

б) Исследовать влияние на заболеваемость гриппом и ОРЗ следующих факторов: количества времени, выделяемых человеком на сон (фактор А) и регулярность занятий спортом (фактор В). Значения фактора А: $A = A_1$ - сон менее 8 часов в сутки, $A = A_2$ - сон более 8 часов. Значения фактора В: $V = V_1$ - занятие спортом не реже 1 раза в неделю, $V = V_2$ - реже 1 раза в неделю. Исследования проводились в нескольких городах. Данные о том, сколько раз в среднем болеет за год житель каждого города приведены в файле disp-5.txt.

б а) Исследовать влияние времени суток на количество вызовов скорой помощи. Анализ проводился в 15 городах с примерно одинаковой численностью населения в течение года. Данные о среднем количестве вызовов бригад скорой помощи в каждом городе приведены в таблице

Время суток	Среднее число вызовов скорой помощи за указанный период														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
8.00-16.00	62,98	89,72	86,03	95,21	91,80	88,17	73,01	81,72	85,98	81,28	91,88	91,29	70,46	49,99	85,75
16.00-24.00	94,45	95,84	91,08	82,14	97,79	100,08	96,74	103,86	105,55	112,58	81,53	125,17	110,20	99,24	95,50
0.00-8.00	132,64	133,18	132,43	141,26	135,21	127,13	133,41	134,68	146,26	133,27	133,86	145,24	132,56	155,29	137,19

б) Исследовать влияние на количество крупных ДТП (с участием более 2-х машин или с наличием пострадавших) следующий факторов: плотности транспортного потока (фактор А) и наличие гололеда (фактор В). Значения фактора А: $A = A_1$ - плотный транспортный поток, $A = A_2$ - свободное движение.

Значения фактора $V:V = V_1$ - наличие гололеда, $V = V_2$ - отсутствие гололеда. Исследования проводились в нескольких городах. Данные о среднем количестве крупных ДТП в час приведены в файле disp-6.txt.

7 а) Исследовать влияние прослушиваемой водителем музыки на скорость движения автомобиля по незагруженной транспортом трассе. Согласно значениям качественного фактора данные были поделены на 3 группы (по 10, 14 и 15 значений соответственно). Данные о средней скорости водителей приведены в таблице

Вид музыки	Средняя скорость автомобиля на трассе														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
классическая	74,75	82,83	88,62	82,04	84,69	93,39	90,19	92,05	76,44	76,95					
популярная	100,69	92,44	106,97	98,18	93,56	92,77	94,83	105,58	97,55	100,90	112,62	92,94	100,02	111,08	108,93
рок	104,81	105,32	107,65	109,92	107,38	98,16	88,85	119,11	114,72	93,44	106,44	121,97	101,80	84,69	90,89

б) Исследовать влияние на посещаемость человеком кинотеатров следующих факторов: возраста (фактор А) и активности человека как читателя художественной литературы (фактор В). Значения фактора $A:A = A_1$ - возраст от 18 до 30 лет, $A = A_2$ - от 31 года. Значения фактора $V:V = V_1$ - количество прочитанных за год книг 0-3, $V = V_2$ - более 3-х. Опрос проводился в нескольких кинотеатрах. Данные о том, сколько раз в год в среднем человек посещает кинотеатры, приведены в файле disp-7.txt.

8 а) Исследовать влияние уровня дохода семьи из 4-х человек на количество потребляемого за неделю хлеба. Согласно значениям качественного фактора данные были поделены на 3 группы (по 10, 15 и 11 значений соответственно). Данные о среднем потреблении хлеба (в кг) за неделю приведены в таблице

Уровень дохода семьи	Среднее количество потребляемого за неделю хлеба (в кг.)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
низкий	2,79	3,44	2,72	2,55	2,57	3,55	2,49	2,37	3,31	2,63					
средний	3,06	3,02	2,81	2,67	3,05	2,98	2,43	3,30	3,26	3,42	3,31	3,21	2,78	2,86	3,24
высокий	3,30	3,08	2,79	2,82	2,97	3,33	3,88	2,93	3,18	3,18	3,75				

б) Исследовать влияние на объем входящего интернет-трафика следующих факторов: отношения пользователя к компьютерным играм (фактор А) и наличия в домашней сети пользователя бесплатных ресурсов (фактор В). Значения фактора А: $A = A_1$ - пользователь увлекается компьютерными играми, $A = A_2$ - не увлекается. Значения фактора В: $V = V_1$ - большой объем бесплатных ресурсов в домашней сети, $V = V_2$ - бесплатных ресурсов не имеется, либо их мало. Данные о среднем входящем интернет-трафике (в Мб.) приведены в файле disp-8.txt.

9 а) Исследовать влияние возраста покупателя на количество покупаемых в магазине глазированных сырков. Согласно значениям качественного фактора данные были поделены на 3 группы (по 12, 15 и 11 значений соответственно). Данные о среднем количестве покупаемых сырков за одно посещение магазина приведены в таблице

Возраст покупателя	Среднее количество покупаемых за 1 раз глазированных сырков														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
до 25	1,26	1,37	0,99	1,14	1,08	1,21	0,76	0,65	2,52	0,86	0,36	1,38			
25-45	3,40	3,07	2,50	2,54	2,56	3,14	2,80	3,65	2,31	3,19	3,22	3,88	2,68	3,28	3,77
старше 45	0,59	2,15	0,98	1,73	0,97	0,88	1,71	0,54	1,44	1,71	2,23				

б) Исследовать влияние на склонность к сердечно-сосудистым заболеваниям у людей старше 50 лет следующих факторов: пола пациента (фактор А) и курения (фактор В). Значения фактора А: $A = A_1$ - пациент - женщина, $A = A_2$ - пациент - мужчина. Значения фактора В: $V = V_1$ - пациент курит, $V = V_2$ - пациент не курит. Данные о среднем количестве инфарктов на 100 человек старше 50 лет приведены в файле disp-9.txt.

10 а) Исследовать влияние климата на количество потребляемых мясных продуктов. Анализ проводился в нескольких регионах. Согласно значениям качественного фактора данные были поделены на 3 группы (по 13, 15 и 9 значений соответственно). Данные о среднем количестве потребляемых в неделю мясных продуктов (в кг.) жителями каждого региона приведены в таблице

Климатические условия	Среднее количество мяса, употребляемого жителем региона за неделю (в кг.)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
северные регионы	0,46	0,52	0,86	0,91	0,54	0,64	0,67	0,74	0,69	0,61	0,65	0,53	0,76		
средняя полоса	0,41	0,33	0,41	0,28	0,39	0,36	0,40	0,67	0,39	0,33	0,56	0,61	0,18	0,59	0,23

южные регионы	0,09	0,22	0,27	0,40	0,27	0,06	0,30	0,36	0,42						
---------------	------	------	------	------	------	------	------	------	------	--	--	--	--	--	--

б) Исследовать влияние на спрос на фотоуслуги следующих факторов: дня недели(фактор А) и времени года (фактор В). Значения фактора А: $A = A_1$ - рабочий день, $A = A_2$ - выходной. Значения фактора В: $B = B_1$ - теплое время года, $B = B_2$ - холодное время года. Данные о среднем количестве отпечатываемых за день фотографий по нескольким фотолабораториям приведены в файле disp-10.txt.

11 а) Исследовать влияние возраста на продолжительность разговоров по мобильному телефону. Согласно значениям качественного фактора данные были поделены на 3 группы (по 15, 15 и 14 значений соответственно). Данные о средней продолжительности телефонных разговоров в день (в минутах) приведены в таблице

Возраст абонента	Средняя продолжительность телефонных разговоров в день (в мин.)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
16-25 лет	4,91	4,64	4,67	5,05	5,10	4,82	5,33	5,12	5,26	5,10	4,75	5,20	4,77	4,99	4,94
25-45 лет	3,78	4,14	4,02	4,10	4,21	3,99	4,11	4,30	3,67	3,90	4,04	3,95	4,04	4,13	4,05
45-60 лет	1,17	0,81	0,94	1,23	1,00	1,04	0,91	1,23	1,33	0,65	1,38	1,13	1,51	1,03	

б) Исследовать влияние на количество потребляемой человеком минеральной воды следующих факторов: образа жизни (фактор А) и пола (фактор В). Значения фактора А: $A = A_1$ - занимается спортом, $A = A_2$ - не занимается спортом. Значения фактора В: $B = B_1$ - мужчина, $B = B_2$ - женщина. Данные о среднем количестве выпиваемой за неделю минеральной воды (в литрах) приведены в файле disp-11.txt.

12 а) Исследовать объективность судей, оценивавших спортсменов на некотором соревновании. Перед судьями выступало 15 спортсменов, совершивших по 10 подходов. Данные о средних баллах спортсменов (по 6-ти бальной шкале) приведены в таблице

эксперт	Средние баллы спортсменов														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
1-й	3,9	4,5	4,0	4,7	3,8	4,4	4,7	4,0	3,8	4,3	4,5	4,5	4,5	4,1	4,0
2-й	4,8	3,9	4,4	4,7	4,5	3,9	4,1	4,7	4,8	4,6	4,2	3,8	4,7	4,8	3,6
3-й	5,2	3,9	5,1	4,1	4,7	4,0	4,6	4,9	4,5	4,5	5,1	4,9	4,4	4,4	4,9

б) Исследовать влияние на заболеваемость ОРЗ следующих факторов: как человек провел летний отпуск (фактор А) и принимает ли он витаминные комплексы (фактор В). Значения фактора $A:A=A_1$ - совершил длительный выезд на природу/отдыхал на курорте, $A=A_2$ - провел отпуск дома. Значения фактора $B:B=B_1$ - принимает витаминные комплексы, $B=B_2$ - не принимает. Данные о среднем количестве перенесенных за год ОРЗ приведены в файле disp-12.txt.

13 а) Исследовать влияние "возраста" йогурта на содержание в нем молочнокислых бактерий. Согласно значениям качественного фактора данные были поделены на 3 группы (по 15, 13 и 13 значений соответственно). Данные о среднем содержании молочнокислых бактерий в 1 г. йогурта (в 10^6 КОЕ) приведены в таблице

Возраст йогурта	Содержание молочнокислых бактерий (в 10^6 КОЕ)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
1/3 от срока годности	22,04	21,99	22,08	21,92	22,10	21,92	21,98	22,21	21,99	22,11	21,94	22,09	22,02	22,05	22,00
2/3 от срока годности	16,09	15,94	15,89	16,14	16,16	15,93	16,09	15,95	15,95	16,09	15,89	16,06	15,94		
конец срока годности	9,98	10,10	9,93	10,10	9,68	9,88	10,02	10,08	10,09	10,12	10,05	9,96	10,31		

б) Исследовать влияние на срок службы стиральных машин следующих факторов: жесткости воды (фактор А) и используемое число оборотов центрифуги (фактор В). Значения фактора $A:A=A_1$ - жесткая вода, $A=A_2$ - рН нейтральная или мягкая вода. Значения фактора $B:B=B_1$ - 1000 оборотов, $B=B_2$ - 800 оборотов. Данные о средней продолжительности работы стиральных машин некоторой марки (в часах) приведены в файле disp-13.txt.

14. а) Исследовать влияние вида применяемых удобрений на содержание калия в фасоли. Согласно значениям качественного фактора данные были поделены на 3 группы (по 12, 11 и 13 значений соответственно). Данные о среднем калия в фасоли (в мг. на 100 г.) приведены в таблице

Вид удобрения	Среднее содержание калия в фасоли (в мг. на 100 г.)												
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}
1	1062,	1059,	1059,	1061,	1060,	1060,	1058,	1059,	1059,	1060,	1060,	1059,	
	1	3	7	3	3	6	3	7	1	8	8	0	
2	1062,	1061,	1059,	1060,	1060,	1061,	1059,	1060,	1058,	1061,	1057,		
	9	9	2	7	1	1	7	1	4	5	9		

б)

3	1059,	1059,	1059,	1058,	1059,	1058,	1061,	1058,	1059,	1059,	1061,	1060,	1059,
	7	2	0	8	8	5	4	5	3	7	3	2	6

Исследовать влияние на количество детей у женщин следующих факторов: места проживания (фактор А) и наличия высшего образования (фактор В). Значения фактора А: $A = A_1$ - сельская местность, $A = A_2$ - город. Значения фактора В: $B = B_1$ - есть высшее образование, $B = B_2$ - высшего образования нет. Данные были собраны по 30 сельским пунктам и 30 городам. Данные о среднем количестве детей у женщин 40 лет приведены в файле disp-14.txt.

14 а) Исследовать влияние года обучения студента в ВУЗе на посещаемость им лекционных занятий. Согласно значениям качественного фактора данные были поделены на 3 группы - (по 15, 14 и 13 значений соответственно). Данные посещаемости лекционных занятий студентами каждой специальности (в процентах) приведены в таблице

Год обучения	Средняя посещаемость лекционных занятий (в %)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
1-й	76,32	67,52	70,51	75,23	68,72	66,11	72,25	66,93	72,51	66,23	75,77	73,84	67,83	68,68	66,66
2-й – 3-й	63,06	61,53	58,18	61,09	61,24	63,21	62,58	55,92	56,45	58,92	66,67	63,11	63,02	62,15	
4-й – 5-й	51,82	49,86	48,87	51,81	54,51	58,69	48,14	52,42	53,48	53,23	57,99	54,53	55,78		

б) Исследовать влияние на курс акций некоторого предприятия следующих факторов: поведения курса акций предприятия А (фактор А) и поведения курса акций предприятия Б (фактор В). Значения фактора А: $A = A_1$ - курс растет, $A = A_2$ - курс падает. Значения фактора В: $B = B_1$ - курс растет, $B = B_2$ - курс падает. Данные о цене акции изучаемого предприятия приведены в файле disp-15.txt.

15 а) Исследовать влияние сорта яблок на содержание в них железа. Согласно значениям качественного фактора данные были поделены на 3 группы - (по 15, 13 и 15 значений соответственно). Данные о содержании железа в яблоках (в мг. на 100 г.) приведены в таблице

Сорт яблок	Содержание железа в разных сортах яблок (в мг. на 100 г.)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
сорт «А»	2,51	2,51	2,50	2,50	2,50	2,50	2,50	2,49	2,53	2,50	2,49	2,51	2,50	2,49	2,49
сорт «Б»	2,46	2,45	2,44	2,44	2,44	2,45	2,45	2,46	2,44	2,45	2,45	2,47	2,44		

1-й	7,0	6,4	7,4	6,3	6,0	7,7	5,4	6,5	6,4	6,7	7,6	7,7	6,2	5,8	6,4
2-й	5,9	7,7	7,1	7,5	8,1	7,0	7,6	8,5	5,4	6,5	7,2	6,8	7,2	7,7	7,2
3-й	8,5	4,1	7,3	6,6	7,6	8,1	6,7	8,1	5,1	9,2	8,5	7,1	7,1	5,4	5,7

б) Исследовать влияние на формирование цены на товар "С" следующих факторов: изменение спроса на товар "А" (фактор А) и изменение спроса на товар "В" (фактор В). Значения фактора $A:A = A_1$ - спрос растет, $A = A_2$ - спрос падает. Значения фактора $B:B = B_1$ - спрос растет, $B = B_2$ - спрос падает. Данные о средней цене на товар "С" по городу (в руб.) приведены в файле disp-18.txt.

18 а) Исследовать влияние сорта томатов на содержание в нем витамина С. Согласно значениям качественного фактора данные были поделены на 3 группы (по 15, 14 и 13 значений соответственно). Данные о среднем содержании витамина С в томатах (в мг. на 100 г.) приведены в таблице

Сорт томатов	Содержание витамина С в томатах (в мг. на 100 г.)														
	X _{1i}	X _{2i}	X _{3i}	X _{4i}	X _{5i}	X _{6i}	X _{7i}	X _{8i}	X _{9i}	X _{10i}	X _{11i}	X _{12i}	X _{13i}	X _{14i}	X _{15i}
сорт «А»	34,93	34,81	35,54	34,04	35,34	34,87	34,99	35,30	34,38	34,74	35,03	34,82	35,21	34,14	35,56
сорт «В»	34,06	35,42	34,73	35,72	34,60	35,60	35,35	35,51	34,90	35,58	35,00	35,66	34,79	34,76	
сорт «С»	35,34	35,06	34,31	34,98	35,17	35,73	35,13	34,73	35,10	35,41	35,26	34,31	35,14		

б) Исследовать влияние на склонность к инфекционным заболеваниям у детей следующих факторов: пола ребенка (фактор А) и употребления витаминных комплексов (фактор В). Значения фактора $A:A = A_1$ - девочка, $A = A_2$ - мальчик. Значения фактора $B:B = B_1$ - ребенок принимает витаминные комплексы, $B = B_2$ - не принимает. Данные о среднем количестве перенесенных ребенком инфекционных заболеваний за год приведены в файле disp-19.txt.

19 а) Исследовать влияние поведения атмосферного давления на количество вызовов скорой помощи. Анализ проводился в 15 городах с примерно одинаковой численностью населения в течение года. Данные о среднем количестве вызовов бригад скорой помощи (в день) в каждом городе приведены в таблице

	Среднее число вызовов скорой помощи в день
--	--

атм.давление	X_{1i}	X_{2i}	X_{3i}	X_{4i}	X_{5i}	X_{6i}	X_{7i}	X_{8i}	X_{9i}	X_{10i}	X_{11i}	X_{12i}	X_{13i}	X_{14i}	X_{15i}
падает	97,2	97,9	95,5	91,1	98,9	100,0	98,4	101,9	102,8	106,3	90,8	112,6	105,1	99,6	97,8
стабильно	86,5	76,7	88,7	84,9	80,8	82,5	79,2	82,5	90,9	84,3	89,6	89,5	82,3	78,0	79,5
растет	102,2	94,6	98,5	95,8	95,7	100,5	101,6	97,6	106,4	96,1	106,3	98,0	95,6	97,7	102,7

б) Исследовать влияние на продолжительность исходящих звонков с сотового телефона следующих факторов: пола абонента(фактор А) и количество отправляемых абонентом смс в день (фактор В). Значения фактора А: $A = A_1$ - женский пол, $A = A_2$ - мужской пол. Значения фактора В: $B = B_1$ - абонент отправляет до 5 смс в день, $B = B_2$ - больше 5 смс в день. Данные о средней продолжительности одного разговора абонента по исходящей связи приведены в файле disp-20.txt.

Лабораторная работа №8. Метод наименьших квадратов. Построение конкретных нелинейных моделей

8.1 Нелинейная регрессия

В лабораторной работе №8 мы имели дело с частным случаем регрессии - линейной регрессией. Теперь рассмотрим общий случай и общую постановку задачи регрессионного анализа.

Пусть имеется выборка $(X, Y) = ((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ из распределения случайной величины $\Psi = (\xi, \eta)$. И пусть известно, что случайные величины ξ и η зависимы. Важное прикладное значение имеет задача о представлении одной из этих величин как функции от другой.

Проведение регрессионного анализа можно разделить на три этапа: выбор формы зависимости (типа уравнения), вычисление параметров выбранного уравнения, оценка достоверности полученного уравнения.

Выбор вида уравнения регрессии производится на основании опыта предыдущих исследований, наблюдений расположения точек (X_i, Y_i) на плоскости и т.д.

Обозначим через $f(x, \theta)$ функцию задающую зависимость среднего значения η от значений ξ (здесь $\theta = (\theta_1, \dots, \theta_k)$ - вектор параметров):

$$M(\eta / \xi = x) = f(x, \theta).$$

Уравнение $y = f(x, \theta)$ называется уравнением регрессии.

Для определения неизвестных параметров $\theta_1, \dots, \theta_k$ можно использовать метод наименьших квадратов.

Суть этого метода состоит в том, что наилучшим считается такое положение линии регрессии, при котором сумма квадратов отклонений значений $f(X_i, \theta)$ от соответствующих Y_i минимальна. Метод состоит в минимизации функции

$$Q(\theta) = \sum_{i=1}^n (y_i - f(x_i, \theta))^2$$

Приведем пример построения нелинейной регрессии с использованием метода наименьших квадратов.

Пусть при проведении эксперимента получены следующие значения величин x и y :

x	6	6.1	6.3	6.5	6.7	7	7.5	8	8.2	8.5
y	4.5	4	3.5	3	2.5	2	1.5	1	0.7	0.5

Считая справедливой зависимость $y(x, D) = D_0 e^{D_1 x}$, находим неизвестные параметры D_0 и D_1 помощью метода наименьших квадратов. В результате получаем следующее уравнение регрессии:

$$y = 500.1 e^{-0.79x}$$

Текст программы, реализующей построение уравнения регрессии приведен на рисунке 8.1. В данной программе для минимизации функции $Q(D)$ используется встроенная функция `Minerr()`. Однако минимизацию можно провести известным методом исследования функции нескольких переменных на экстремум с помощью дифференциального исчисления.

$$X := (6 \ 6.1 \ 6.3 \ 6.5 \ 6.7 \ 7 \ 7.5 \ 8 \ 8.2 \ 8.5)^T$$

Вводятся элементы выборки (X,Y)

$$Y := (4.5 \ 4 \ 3.5 \ 3 \ 2.5 \ 2 \ 1.5 \ 1 \ 0.7 \ 0.5)^T$$

$$Q(D) := \sum_{i=0}^9 \left(Y_i - D_0 e^{D_1 X_i} \right)^2$$

Задается минимизируемая функция Q(D)

$$D := \begin{pmatrix} 100 \\ -1 \end{pmatrix}$$

Given

$$Q(D) = 0$$

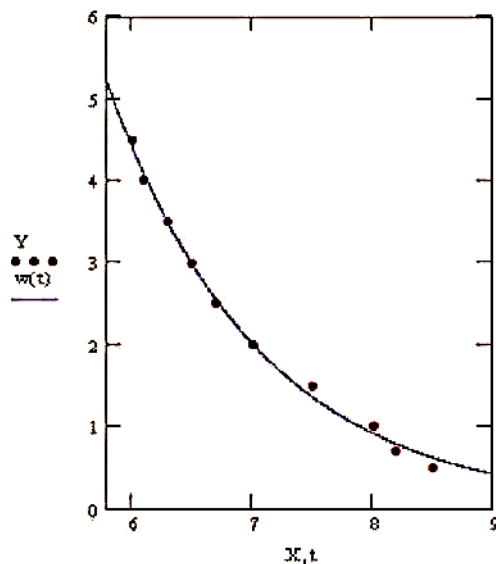
$$D := \text{Minim}(D)$$

$$D = \begin{pmatrix} 500.102 \\ -0.788 \end{pmatrix}$$

Находится вектор параметров D, при котором Q(D) достигает минимального значения

$$w(t) := D_0 e^{D_1 t}$$

Полученное уравнение регрессии



На графике отображены исходные данные и линия регрессии, соответствующая полученному уравнению

Рисунок 8.1 - Построение уравнения регрессии с помощью метода наименьших квадратов

8.2 Задание к лабораторной работе

В файле regrV.txt (V - это номер вашего варианта) в виде матрицы задана выборка (X, Y). Первый столбец матрицы - значения X, второй столбец - соответствующие значения Y.

1. С помощью метода наименьших квадратов построить уравнения регрессии, считая справедливыми следующие формы зависимости у от х:

$$\text{а) } y = a \sin (bx), \quad \text{б) } y = \log_a bx, \quad \text{в) } y = a_0 + a_1 x + a_2 x^2.$$

Поиск минимума функции $Q(D)$ проводить, исследуя эту функцию на экстремум с помощью частных производных.

2. На одном графике изобразить исходные данные и полученные линии регрессии. Сделать вывод о том, какая из функций наилучшим образом представляет зависимость y от x .

Некоторые параметрические семейства распределений

1. Равномерное распределение $U_{a,b}$. Функция плотности распределения и моменты распределения:

$$u_{a,b}(t) = \begin{cases} \frac{1}{b-a}, & t \in [a, b], \\ 0, & t \notin [a, b] \end{cases}, \quad MX = \frac{b+a}{2}, \quad DX = \frac{(b-a)^2}{12}.$$

2. Показательное распределение E_λ . Функция плотности распределения и моменты распределения:

$$e_\lambda(t) = \begin{cases} \lambda e^{-\lambda t}, & t > 0, \\ 0, & t \leq 0, \end{cases} \quad MX = \frac{1}{\lambda}, \quad DX = \frac{1}{\lambda^2}.$$

3. Гамма-распределение $\Gamma_{\alpha,\beta}$. Функция плотности распределения:

$$\gamma_{\alpha,\beta}(x) = \begin{cases} \frac{1}{\beta^{\alpha+1} \Gamma(\alpha+1)} x^\alpha e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & x \leq 0 \end{cases}, \quad \alpha > -1, \beta > 0.$$

Моменты распределения: $MX = (\alpha+1)\beta, DX = \beta^2(\alpha+1)$.

4. Распределение Пуассона $\Pi_\lambda (\lambda > 0)$:

$$P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!} \quad (m = 0, 1, 2, \dots), \quad MX = \lambda, \quad DX = \lambda$$

5. Геометрическое распределение G_p :

$$P(X = k) = (1-p)^k p, \quad p \in (0, 1), \quad k = 0, 1, 2, \dots$$

$$MX = \frac{1-p}{p}, DX = \frac{1-p}{p^2}.$$

6. Биномиальное распределение B_p^n :

$$P(X = k) = C_k^n p^k (1-p)^{n-k} (0 \leq k \leq n),$$

$$MX = np, DX = np(1-p).$$

7. Нормальное распределение N_{a, σ^2} . Функция плотности распределения и моменты распределения:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, MX = a, DX = \sigma^2$$

8. Бета-распределение $\beta_{m,n}$. Функция плотности распределения:

$$\beta_{m,n}(x) = \begin{cases} \frac{\Gamma(m+n)}{\Gamma(m) \cdot \Gamma(n)} x^{m-1} (1-x)^{n-1}, & x \in (0,1) \\ 0, & x \notin (0,1) \end{cases}$$

где $m > 0, n > 0$.

$$MX = \frac{m}{m+n}, DX = \frac{mn}{(m+n)^2(m+n+1)}.$$

Моменты распределения:

9. Логарифмически нормальное (логнормальное) распределение. Функция плотности распределения:

$$l(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - a)^2}{2\sigma^2}}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

Моменты распределения: $MX = e^{-\frac{\sigma^2}{2} + a}, DX = e^{\sigma^2 + 2a} (e^{\sigma^2} - 1).$

10. Распределение χ_n^2 Функция плотности распределения:

$$h_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{при } x > 0, \\ 0 & \text{при } x \leq 0. \end{cases}, \quad n = 1, 2, \dots$$

Моменты распределения: $MX = n$, $DX = 2n$.

11. Распределение Стьюдента T_k . Функция плотности распределения:

$$t_k(x) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \frac{1}{\left(1 + x^2/k\right)^{(k+1)/2}}, k = 1, 2, \dots$$

$$MX = 0; DX = k \frac{\Gamma(3/2)\Gamma(k/2-1)}{\sqrt{\pi}\Gamma(k/2)}, k > 2.$$

12. Распределение Фишера $F_{k,m}$. Функция плотности распределения:

$$f_{k,m}(x) = \left(\frac{k}{m}\right)^{k/2} \frac{\Gamma((k+m)/2)}{\Gamma(k/2)\Gamma(m/2)} \frac{x^{k/2-1}}{\left(1 + kx/m\right)^{(k+m)/2}}, x > 0; k, m > 0.$$

$$MX = \frac{m}{m-2}, m > 2; DX = \frac{2m^2(k+m-2)}{k(m-2)^2(m-4)^2}, m > 4$$

13. Распределение Коши $K_{m,n}$. Функция плотности распределения:

$$k_{m,n}(x) = \frac{1}{\pi} \frac{m}{m + (x-n)^2}, m > 0, -\infty < n < \infty.$$

MX и DX не существуют.

Таблица 1 - Квантили стандартного нормального распределения T_ε

ε	T_ε	ε	T_ε
0.010	2.3263	0.250	0.6745
0.025	1.9600	0.300	0.5244
0.050	1.6449	0.350	0.3853
0.100	1.2816	0.400	0.2533
0.150	1.0364	0.450	0.1257
0.200	0.8416	0.500	0.0000

Таблица 2 - Квантили распределения Стьюдента $t_{p,k}$

k	p= 0.750	p= 0.900	p= 0.990	p = 0.999
1	1.000	3.078	31.821	318
2	0.816	1.886	6.965	22.3
3	0.765	1.638	4.541	102
4	0.741	1.533	3.747	7.173
5	0.727	1.476	3.365	5.893
6	0.718	1.440	3.143	5.208
7	0.711	1.415	2.998	4.785
8	0.706	1.397	2.896	4.501
9	0.703	1.383	2.821	4.297
10	0.700	1.372	2.764	4.144
11	0.697	1.363	2.718	4.025
12	0.695	1.356	2.681	3.930
13	0.694	1.350	2.650	3.852
14	0.692	1.345	2.624	3.787
15	0.691	1.341	2.602	3.733
20	0.687	1.325	2.528	3.552
30	0.683	1.310	2.457	3.385
40	0.681	1.303	2.423	3.307
60	0.679	1.296	2.390	3.232
80	0.677	1.289	2.358	3.160
∞	0.674	1.282	2.326	3.090

Таблица 3 - Квантили распределения $\chi^2_{\alpha,k}$

k	$\alpha = 0.010$	$\alpha = 0.025$	$\alpha = 0.05$	$\alpha = 0.10$
1	0.00016	0.00098	0.00393	0.01580
2	0.0201	0.05060	0.1030	0.2110
3	0.1150	0.2160	0.3520	0.5840
4	0.297	0.484	0.711	1.106
5	0.554	0.831	1.150	1.161
6	0.872	1.240	1.640	2.200
7	1.240	1.690	2.170	2.830
8	1.650	2.180	2.730	3.490
9	2.090	2.700	3.330	4.170
10	2.560	3.250	3.940	4.870
11	3.050	3.820	4.570	5.580
12	3.570	4.400	5.230	6.300
13	4.110	5.010	5.890	7.040
14	4.660	5.630	6.570	7.790
15	5.230	6.260	7.260	8.550
16	5.81	6.91	7.96	9.31
17	6.41	7.56	8.67	10.1
18	7.01	8.23	9.39	10.9
19	7.63	8.91	10.1	11.7
20	8.26	9.59	10.9	12.4
21	8.90	10.3	11.6	13.2
22	9.54	11.0	12.3	14.0
23	10.2	11.7	13.1	14.8
24	10.9	12.4	13.8	15.7
25	11.5	13.1	14.6	16.5
30	15.0	16.8	16.5	20.6
35	18.5	20.6	22.5	24.8
40	22.2	24.4	26.5	29.1
45	25.9	28.4	30.6	33.4
50	29.7	32.4	34.8	37.7
75	49.5	52.9	56.1	59.8
100	70.1	74.2	77.9	82.4

Таблица 4 - Квантили распределения Фишера F_{p,k_1,k_2}

k2	$k_1 = 1$	$k_1 = 2$	$k_1 = 1$	$k_1 = 2$	$k_1 = 1$	$k_1 = 2$
1	161.4	199.5	4052	4999.5	405300	500000
2	18.51	19.00	98.50	99.00	998.5	999
3	10.13	9.55	34.12	30.82	167.0	148.5
4	7.71	6.94	21.20	18.00	74.14	61.25
5	6.61	5.79	16.26	13.27	47.18	37.12
6	5.99	5.14	13.75	10.92	35.51	27.00
7	5.59	4.74	12.25	9.55	29.25	21.69
8	5.32	4.46	11.26	8.65	25.42	18.49
9	5.12	4.26	10.56	8.02	22.86	16.39
10	4.96	4.10	10.04	7.56	21.04	14.91
11	4.84	3.98	9.65	7.21	19.69	13.81
12	4.75	3.89	9.33	6.93	18.64	12.97
13	4.67	3.81	9.07	6.70	17.81	12.31
14	4.60	3.74	8.86	6.54	17.14	11.78
15	4.54	3.68	8.68	6.36	16.59	11.34
16	4.49	3.63	8.53	6.23	16.12	10.97
17	4.45	3.59	8.40	6.11	15.72	10.66
18	4.41	3.55	8.29	6.01	15.38	10.39
19	4.38	3.52	8.18	5.93	15.08	10.16
20	4.35	3.49	8.10	5.85	14.82	9.95
25	4.24	3.39	7.77	5.57	13.88	9.22
30	4.17	3.32	7.56	5.39	13.29	8.77
40	4.08	3.23	7.31	5.18	12.61	8.25
60	4.00	3.15	7.089	4.98	11.97	7.76
120	3.92	3.07	6.85	4.79	11.38	7.32
то	3.84	3.00	6.63	4.61	10.83	6.91

Таблица 5 - Значения функции распределения Колмогорова $K(t)$

t	$K(t)$
1.36	0.9505
1.40	0.9603
1.45	0.9702
1.52	0.9803
1.63	0.9902

Литература

1. Васильев, А.В. Mathcad 13 на примерах/ А.В. Васильев. - СПб.: БХВ - Петербург, 2006. - 528с.
2. Кирьянов, Д. В. Самоучитель Mathcad 11/ Д.В. Кирьянов. - СПб.: БХВ-Петербург, 2003. - 560 с.
3. Новикова, Н.М. Компьютерный практикум по теории вероятности в среде МАТНСАД: учебно-методическое пособие для вузов/ Н.М. Новикова. – Воронеж: Издательско-полиграфический центр ВГУ, 2008. – 23 с.
4. Плис, А.И. Mathcad 2000. Математический практикум для экономистов и инженеров: Учеб. Пособие/ А.И. Плис, Н.А. Сливина. - М.: Финансы и статистика, 2000. -656с.: ил.