##### МИНИОБРАНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«Тверской государственный технический университет»**

(ТвГТУ)

Факультет информационных технологий

Кафедра «Информационные системы»

Ветров А.Н.

#### МЕТОДИЧЕСКИЕ УКАЗАНИЯ

#### и задания к курсовым работам

по дисциплине «Теория вероятностей и математическая статистика»

vetrov\_48@mail.ru

Тверь, 2019

**Оглавление**

[Введение 3](#_Toc466762343)

[1. РАСЧЕТ ОСНОВНЫХ ПОКАЗАТЕЛЕЙ СТАТИСТИКИ 4](#_Toc466762344)

[2. Построение гистограммы 8](#_Toc466762345)

[3. Проверка соответствия закона распределения наблюдаемым данным 10](#_Toc466762346)

[4. Проверка гипотезы о равенстве средних величин при известной дисперсии 11](#_Toc466762347)

[5. Проверка гипотезы о равенстве дисперсий 15](#_Toc466762348)

[6. Проверка гипотезы о равенстве средних величин при неизвестной десперсии 17](#_Toc466762349)

[7. Однофакторный дисперсионный анализ 2](#_Toc466762350)2

[8. Двухфакторный дисперсионный анализ без повторений и с повторениямии 26](#_Toc466762351)

9. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ…………………………………………27

[Список литературы 31](#_Toc466762352)

Введение

Курсовая работа является завершающим этапом освоения дисциплины, на этом этапе студент самостоятельно решает поставленную задачу с применением полученные во время лекций и практик знания. Качество решения поставленной задачи наглядно показывает, насколько студентом освоена программа по предмету.

Теория вероятностей и статистика это самостоятельная дисциплина, которая основывается на высшей математике. Эта дисциплина находит широкое применение в экономике и инженерном деле.

Данная работа состоит из 8 разделов:

1. Расчет основных показателей статистики.
2. Построение гистограммы.
3. Проверка соответствия закона распределения наблюдаемым данным.
4. Проверка гипотезы о равенстве средних величин при известной дисперсии.
5. Проверка гипотезы о равенстве средних величин при неизвестной дисперсии.
6. Проверка гипотезы о равенстве дисперсий.
7. Однофакторный дисперсионный анализ.
8. Двухфакторный дисперсионный анализ без повторений и с повторениями.

Работа выполняется в табличном процессоре MS Excel с использованием статистических функций и пакета анализа.

Статистическая информация для выполнения заданий генерируется студентами самостоятельно с помощью инструмента анализа «*Генерация»* пакета ECXEL в соответствии с вариантом задания (№ по списку).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Вариант | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| mx | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 |
| σx | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |

mx – математическое ожидание

σx – стандартное отклонение.

Статистическую информацию для выполнения задания генерируем с помощью инструмента анализа: **Сервис→Анализ данных…→Генерация случайных чисел**. Для генерации используем исходные данные своего варианта, характеризующие генерируемый ряд случайных чисел (выборку).

*Пример*

Математическое ожидание ***m*x** =5;

Стандартное отклонение **σx** = 19.

В соответствии с исходными данными сгенерирован ряд из 31 значение случайной величины (табл.1), распределённой нормально. На рис.1 представлено окно «Генерация случайных чисел», в которое введены ***m*x**, **σx**.

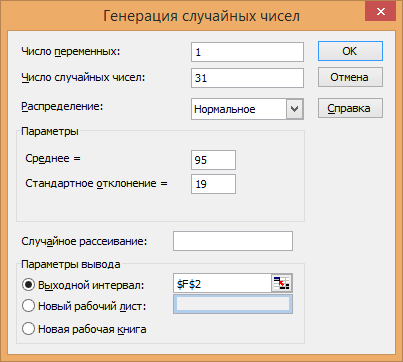


Рис.1. Диалоговое окно «Генерация случайных чисел»

# 1. РАСЧЕТ ОСНОВНЫХ ПОКАЗАТЕЛЕЙ СТАТИСТИКИ

Показатели описательной статистики можно разбить на несколько групп.

1. *Показатели положения* описывают положение данных на числовой оси. Примеры таких показателей: минимальный и максимальный элементы выборки (первый и последний члены вариационного ряда), верхний и нижний квартили (ограничивают зону, в которую попадают 50% центральных элементов выборки), средняя арифметическая, средняя гармоническая, медиана и другие характеристики.

2. *Показатели разброса* описывают степень разброса данных относительно своего центра. К этой группе относятся: дисперсия, стандартное отклонение, размах выборки (разность между максимальным и минимальным элементами), межквартильный размах (разность между верхней и нижней квартилью), эксцесс и т.п. Эти показатели определяют, насколько кучно основная масса данных группируется около центра.

3. *Показатели асимметрии* характеризуют симметрию распределения данных около своего центра. К ним можно отнести коэффициент асимметрии, положение медианы относительно среднего и т.п.

4. *Показатели, описывающие закон распределения*, дают представление о законе распределения данных. Сюда относятся таблицы частот, таблицы частостей, полигоны, кумуляты, гистограммы.

На практике чаще всего используются следующие показатели: средняя арифметическая, медиана, дисперсия, стандартное отклонение.

**Расчёт показателей описательной статистики**

1. Среднее.

Функция СРЗНАЧ рассчитывает значение средней арифметической величины по формуле



где *xi* – i-ое значение выборки, *n* – число наблюдаемых значений выборки.

2. Медианой (Me) называется значение признака, приходящееся на середину ранжированной (упорядоченной) совокупности. Используется функция МЕДИАНА.

3. Модой (Мо) называется чаще всего встречающаяся варианта или то значение признака, которое соответствует максимальной точке теоретической кривой распределения. Используется функция МОДА.

4. Выборочная дисперсия рассчитывается по выборочным данным. Для этого используется выражение



где  – среднее арифметическое выборки.

Для определения в Excel можно воспользоваться функцией ДИСП.

5. Выборочное стандартное отклонение оценивает разброс возможных значений случайной величины вокруг её среднего. Формула для расчета стандартного отклонения



Для определения в Excel можно воспользоваться функцией СТАНДОТКЛОН.

6. Стандартная (средняя) ошибка повторной собственно-случайной выборки определяется по формуле



где *sx* – выборочная дисперсия, *n* – число наблюдаемых значений выборки.

7. Эксцесс характеризует так называемую «крутость», т.е. островершинность или плосковершинность кривой распределения. За исходную принята кривая нормального распределения (*Ek* = 0), если *Ek* > О, распределение островершинное, если *Ek* < 0 – плосковершинное.

Для определения в Excel можно воспользоваться функцией ЭКСЦЕСС, которая рассчитывает значение эксцесса как для симметричных, так и для асимметричных распределений.

8. Симметричным является распределение, в котором частоты любых двух вариант, равноотстоящих в обе стороны от центра распределения, равны между собой. Для симметричных распределений средняя арифметическая, мода и медиана равны между собой. С учетом этого показатель асимметрии основан на соотношении показателей центра распределения: чем больше разница между х, Mo, Me, тем больше асимметрия ряда. При этом если Mo < Me, асимметрия правосторонняя, если Mo > Me – асимметрия левосторонняя.

Функция СКОС определяет величину асимметрии по выборочной совокупности. При этом если As > О – асимметрия правосторонняя (положительная), если As < О — асимметрия левосторонняя

9. Функции МИН и МАКС используются для определения минимального и максимального значений признака в выборке.

10. Интервалрассчитывают как разность между наибольшим (*хmах*) и наименьшим (*хmin*) значениями выборки, т.е.

*R* = *xmax – xmin*

и называется размах вариации.

11. Функция СЧЕТ используется для определения величины *n*.

12. Функции НАИБОЛЬШИЙ и НАИМЕНЬШИЙ определяют k-ое максимальное и минимальное значения в выборке.

13. Уровень надёжности.

Предельная ошибка выборки связана со средней ошибкой выборки соотношением



где *t* – коэффициент доверия, который определяется в зависимости от того, с какой доверительной вероятностью нужно гарантировать результаты выборочного обследования.

В Excel коэффициент доверия *t* рассчитывается через функцию СТЬЮДРАСПОБР, в которой в качестве аргументов задаются уровень значимости *α* и число степеней свободы *df*. Уровень значимости *α* связан с доверительной вероятностью *β* выражением

*α* = *1* – *β*.

Число степеней свободы *df* зависит от объема выборки *n* и связано с ним выражением *df = n – 1*.

Границы доверительного интервала для математического ожидания находятся из выражения



**Задание:**

1. Сгенерировать 100 значений нормально распределенной случайной величины с параметрами mx, σ**x**
2. Рассчитать значения показателей описательной статистики.
3. Рассчитанные значения свести в таблицу вида

|  |  |  |
| --- | --- | --- |
| № | Показатель | Значение |
| 1 | … | … |
| 2 | … | … |

# 2. Построение гистограммы

Пусть для изучения количественного (дискретного, непрерывного) признака **X**из генеральной совокупности извлечена выборка **х1, х2, …, хn(n**– объём выборки).

Наблюдающиеся значения признака называют вариантами, а последовательность значений признака, записанных в возрастающем порядке – вариационным рядом.

***Статистическим распределением выборки*** называют перечень вариант **хi**вариационного ряда и соответствующих им частот **nj** или относительных частот **ωj.** При этом сумма всех частот **nj** равна объёму выборки**n**, а сумма всех относительных частот равна единице.

Статистическое распределение выборки можно задать в виде последовательности интервалов и соответствующих им частот.

***Эмпирической функцией распределения (функцией распределения выборки)*** называют функцию **F\*(x),** определяющую для каждого значения **x** относительную частоту события **X<x**:

**F\*(x)=nx/n**,

где **nx**- число вариант, меньших x,

**n** – объём выборки.

**Свойства эмпирической функции распределения.**

*Свойство 1.*

Значение эмпирической функции распределения принадлежит отрезку **[0, 1].**

*Свойство 2.*

Эмпирическая функция распределения **–** неубывающая по своему аргументу функция.

*Свойство 3.*

Если **x1** – наименьшая варианта, а **xk** – наибольшая варианта, то **F\*(x)=**0при **x<x1 , F\*(x)=**1при **x>xk**.

При *дискретном* распределении признака **X*полигоном частот*** называют ломаную линию, отрезки которой соединяют точки:

(**x1, n1**), (**x1, n2**), …, (**xk, nk**),

где

**хi**– варианта

**ni**- частота, соответствующая варианте **хi.**

***Полигон относительных частот*** – ломаная линия, отрезки которой соединяют точки:

**(x1, n1**), (**x1, n2**), …, (**xk, ni**),

где

**хi**– варианта

**ni**- относительная частота, соответствующая варианте **хi.**

При *непрерывном* распределении признака **X**весь интервал, в котором заключены все наблюдаемые значения признака, разбивается на ряд частичных интервалов длины **h**которые называются интервалами группировки и находится сумма частот вариант, попавших в **j**-ый интервал - **nj**.

***Гистограмма частот*** – ступенчатая фигура, состоящая из прямоугольников, основаниями которой служат частичные интервалы длины **h,** а высоты равны **nj.**

Относительной частотой **ωj** попадания случайной величины в j – й интервал называется отношение **nj/n,** где **n** – общее число наблюдений.

***Гистограмма относительных частот*** – ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длины **h**, а высотами – относительные частоты **ωj.**

Для неравных интервалов используется понятие плотность частоты **nj/hj**. Относительная плотность определяется отношением **ωi/hi,** где **hi –** длина **j** - того интервала.

Для определения числа интервалов k можно использовать формулу



Эмпирическая функция распределения строится по наколенным частотам, так как это показано в таблице

|  |  |  |  |
| --- | --- | --- | --- |
| j(№ интервала) | Границы интервала | Частоты | Накопленные частоты |
| 1 | x1- x2 | n1 | s1 = n1 |
| 2 | x2 – x3 | n2 | s2= n1 + n2 |
|  | …. |  |  |
| k | xk - xk+1 | nk | sk= n1 + n2 +…+nk |

Таким образом F1 = 0, F2 = s1, F3 = s2, …, Fk+1 = sk.

***Ненормированной эмпирической функцией распределения*** называется ломанная, соединяющая точки (x1, F1), (x2, F2),…, (xk+1, Fk+1)

Если используются относительные частоты, то мы будем говорить о ***нормированной эмпирической функции распределения***

**Задание:**

1. Сгенерировать 100 значений нормально распределенной случайной величины с параметрами mx, σ**x**
2. Построить гистограмму
3. Построить нормированную эмпирическую функцию распределения

# Проверка соответствия закона распределения наблюдаемым данным

В данной работе проверяется соответствие опытных данных нормальному закону. В качестве исходных данных используются результаты 2-ого задания.

Проверяется гипотеза

H0: p(x) = N(0; 1)

Альтернативная гипотеза :

H1: p(x) ≠ N(0; 1)

Последовательность расчетов:

1.Определяеся число значений признака попадающих в j – ый интервал и среднее значение признака для каждого интервала.

2.Вычисляется среднее значение вариационного ряда**x**.

3.Вычисляется выборочная дисперсия и стандартное отклонение вариационного ряда.

4.Вычисляются значения функции плотности нормального распределения для каждого интервала по формуле pj = НОРМРАСП(), в качестве x используется среднее значение на интервале, параметр ИНТЕГРАЛЬНАЯ = 0.

5.Расчитываются теоретические частоты нормального распределения по формуле



где h – длина интервала,n – общее число наблюдаемых значений признака.

6.Расчитывается значение критерия **χ2** по формуле



Данные группируются в таблицу

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| j | Границы интервала | | Эмпирическая частота | Середи­на ин­тервала, | pj(xj,,σ) | Теоретическая частота |  |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  | n |  |  |  |  |
|  |  |  |  | sx |  |  | χ2расч. |
|  |  |  |  |  | α = 0,05 | df = | χ2крит. |

8.Вычисляется критическое значение c помощью функции ХИ2ОБРс параметрами ВЕРОЯТНОСТЬ= 0,05; СТЕПЕНИ СВОБОДЫ = k– 3, где k число интервалов.

На основе полученных расчетного и критического значений критерия χ2 сделайте вывод о соответствии наблюдаемых данных нормальному закону распределения.

*Пример*

1. Используя данные расчета проведенного в главе 2, определяем число значений признака попадающих в j-ый интервал и среднее значение признака  для каждого интервала.

2. Вычисляем среднее значение вариационного ряда .



3. Вычисляем выборочную дисперсию  и стандартное отклонение вариационного ряда.



4. Вычисляем значения функции плотности нормального распределения для каждого интервала по формуле *pj* = НОРМРАСП( ), в качестве *x* используется среднее значение на интервале, параметр ИНТЕГРАЛЬНАЯ = 0 (рис.2). Расчёт сведен в табл.1.

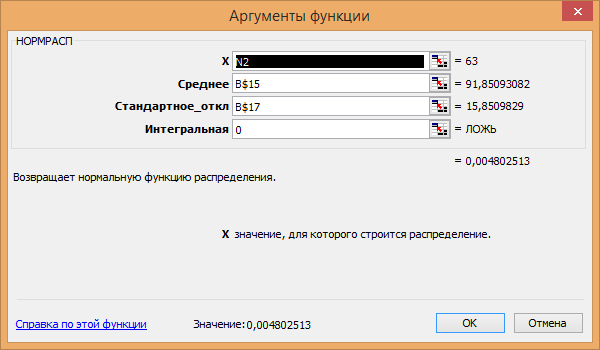


Рис.2. Диалоговое окно «Аргументы функции» для НОРМРАСП.

1. Рассчитываем теоретические частоты нормального распределения по формуле



где *h* – длина интервала, *n* – общее число наблюдаемых значений признака.

=0,004803x8x31=1,19

Остальные расчёты сведены в табл. 1.

6. Рассчитываем значение критерия *χ2* по формуле



7. Вычисляется критическое значение c помощью функции ХИ2ОБР с параметрами ВЕРОЯТНОСТЬ = 0,05; СТЕПЕНИ СВОБОДЫ = k – – 3, где k число интервалов.

Таблица1.

Результаты проверки соответствия закона распределения наблюдаемым данным

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| j | Границы  интервала | | Эмпирическая. частота | Середина  интервала | Плотность  *pj* (*xj,,σ*) | Теоретическая  частота |  |
| 1 | 59 | 67 | 3 | 63 | 0,004803 | 1,191023 | 2,747551 |
| 2 | 67 | 75 | 3 | 71 | 0,010595 | 2,62761 | 0,052776 |
| 3 | 75 | 83 | 1 | 79 | 0,018119 | 4,493419 | 2,715967 |
| 4 | 83 | 91 | 5 | 87 | 0,024017 | 5,956187 | 0,153503 |
| 5 | 91 | 99 | 9 | 95 | 0,024676 | 6,119769 | 1,355562 |
| 6 | 99 | 107 | 4 | 103 | 0,019653 | 4,873906 | 0,156694 |
| 7 | 107 | 115 | 4 | 111 | 0,012132 | 3,008809 | 0,326528 |
| 8 | 115 | 123 | 2 | 119 | 0,005805 | 1,439751 | 0,218009 |
|  |  |  |  |  |  |  | 7,72659 |
|  |  |  |  |  |  |  | 11,0705 |

8. Вывод о нормальности распределения данных.

Так как , то гипотеза о нормальности распределения СВ принимается.

# 4. Проверка гипотезы о равенстве средних величин при известной дисперсии

Иногда оказывается, что средний результат  из основной серии опытов отличается от среднего результата  другой серии опытов. Необходимо определить случайно или нет, это различие, т.е. можно ли считать, что результат эксперимента представляет собой выборка из двух независимых генеральных совокупностей с одинаковыми средними, или средние этих совокупностей не равны.

Формальная постановка этой задачи выглядит следующим образом – изучаются две случайные величины, распределённые **по нормальному** закону:

,

где *σ –* стандартное отклонение, *m* – математическое ожидание (среднее)*.*

Предполагается, что дисперсии  и  известны, а математические ожидания не известны.

Пусть имеются две серии наблюдений величины Χ и Υ.

Χ: *х1*, *х2*, …, *хn1*.

Υ: *y1*, *y2*, …, *yn2*.

Выдвигаем следующую гипотезу, что *mx= my*. На основании наблюдений необходимо подтвердить или опровергнуть эту гипотезу. Если подтвердится нулевая гипотеза, то можно говорить о том, что различия между средними величинами в двух выборках статистически незначимо, т.е. объясняется как случайной ошибкой.

Для проверки этой гипотезы используется Z-тест. Для этого рассчитывается z-критерий (z-статистика), который определяется следующим образом:



где – среднее арифметическое значение из серии *n* наблюдений.

Z-критерий распределён нормально с нулевым математическим ожиданием и единичной дисперсией.

Нулевая гипотеза:

*.*

Альтернативная гипотеза:



Последовательность проведения тестирования:

1. Вычисляем статистику *Z.*
2. Задаёмся уровнем значимости .
3. Определяем критические точки.
4. Сравниваем рассчитанное в п.1 значение *Z* со значением критических точек:

Если значение *Z-*статистики будет по абсолютной величине больше чем значение критической точки, то нулевая гипотеза отклоняется при данном уровне значимости . Это означает, что две совокупности, из которых сделана выборка, различны и, следовательно, средние значения и математические ожидания для этих выборок не равны. В противном случае принимается нулевая гипотеза о равенстве средних значений, и можно рассматривать эти две совокупности как одну общую с одним и тем же математическим значением.

**ЗАДАНИЕ.** Сгенерировать 2 нормально распределенные переменные. Первая переменная генерируется в соответствии с Вашим вариантом. При генерации второй переменной математическое ожидание увеличивается на 2, а стандартное отклонение на 0,5

Проверить гипотезу о равенстве средних аеличин

# 5. Проверка гипотезы о равенстве дисперсий

Для того чтобы определить на основе выборочных данных равны ли дисперсии или нет, мы рассмотрим процедуру проверки гипотезы о равенстве дисперсий двух независимых нормально распределённых случайных величин. Эта задача имеет также самостоятельное значение, поскольку дисперсия характеризует точность работы приборов или технологических процессов, обработки данных и т.п. Убедившись в равенстве двух дисперсий, мы тем самым убеждаемся, например, в том, что два прибора обеспечивают одинаковую точность.

В математической статистике доказывается, что если гипотеза о равенстве дисперсий двух случайных величин выполняется: *H0:* = , то величина  распределена в соответствии с законом распределения Фишера.

Это отношение *F* называют *дисперсионным отношением Фишера* и используют в качестве критерия проверки нулевой гипотезы.

Распределение Фишера характеризуется наличием степеней свободы, которые вычисляются по формулам:



Поскольку величина *F* – неотрицательная, то критическая область данной величины будет принадлежать интервалу (0;+∞).

Альтернативными являются гипотезы:

Н1: > при >

Н1: < при <

Если удвоенное значение р – уровня будет больше, чем уровень значимости (0,05) то нулевая гипотеза о равенстве дисперсий не отклоняется

**ЗАДАНИЕ.** Используя данные задания 4 проверить гипотезу о равенстве дисперсий

# 6. Проверка гипотезы о равенстве средних величин при неизвестной дисперсии

Для проверки гипотезы о равенстве средних (математических ожиданий) двух независимых нормальных распределений с неизвестными дисперсиями  и  используется t-тест

Относительно дисперсий  и  можно выдвинуть следующие два предположения:

1) Обе дисперсии неизвестны, но предполагается, что они равны между собой, т.е. =.

2) Обе дисперсии неизвестны и предполагается, что они не равны между собой, т.е. ≠.

* В случае, когда обе дисперсии неизвестны, но предполагается что они равны между собой, мы имеем дело с двумя оценками  и  одной и той же дисперсии =.

То в этом случае строится объединённая оценка:



где *S2* – это объединённая оценка дисперсии ==.

В математической статистике доказывается, что если нулевая гипотеза о равенстве математических ожиданий *H0: mx=my* выполняется, то величина *t* вычисляется по формуле:



где  и  – средние арифметические величины, *n1* – число наблюдений в первой выборке, *n2* – число наблюдений во второй выборке, *S* – выборочное стандартное отклонение.

Статистика *t* имеет распределение Стьюдента. Число степеней свободы определяется по формуле:



Эту *t*-статистику и используют в качестве критерия при проверке нулевой гипотезы о равенстве математических ожиданий. Схема проверки аналогична проверке при использовании *Z-*теста.

* В случае, когда дисперсии неизвестны и предполагается, что они не равны, используется аналог *Z-*теста с заменой дисперсий их оценками.

 - это распределение близко к распределению Стьюдента.

Число степеней свободы вычисляется по следующей формуле:



В данном случае *t*-статистику, используемую для проверки нулевой гипотезы о равенстве средних величин при различных неизвестных дисперсиях, называют критерием *Фишера-Беренса*.

**Задание:**

Требуется для вашего варианта проверить гипотезу *H0: mx=my*, предположив, что соответствующие генеральные совокупности имеют нормальное распределение

1. с одинаковыми дисперсиями;
2. с различными дисперсиями.

7**. Однофакторный дисперсионный анализ**

Ранее нами были рассмотрены процедуры для оценки значимости различий между средними значениями двух выборок. Однако часто необходимо сравнивать средние значения трёх и более числа выборок. В случае, когда необходимо сравнить средние значения большого числа выборок, используется метод дисперсионного анализа (ANOVA – Analysis of Variance), который устанавливает влияние отдельных факторов на изменчивость какого – либо признака, значения которого могут быть получены опытным путем в виде случайной величины Y. В зависимости от числа факторов, различают однофакторный и многофакторный дисперсионный анализ.

**Однофакторный дисперсионный анализ**

Величину Y называют результативным признаком, а конкретную реализацию фактора A – уровнем (группой) фактора A или способом обработки и обозначают через A(i) . Всего имеется c уровней фактора A. Обозначим их А(1),А(2),…,А(с) .

Задачу однофакторного дисперсионного анализа можно продемонстрировать на следующем примере.

*Пример*

Необходимо определить существует ли разница между прочностью парашютов, сотканных из синтетических волокон разных поставщиков. Результаты эксперимента (сила разрыва) приведены в таблице .

Таблица 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Поставщик 1** | **Поставщик 2** | **Поставщик 3** | **Поставщик 4** |
|  | 18,5 | 26,3 | 20,6 | 25,4 |
|  | 24,0 | 25,3 | 25,2 | 19,9 |
|  | 17,2 | 24,0 | 20,8 | 22,6 |
|  | 19,9 | 21,2 | 24,7 | 17,5 |
|  | 18,0 | 24,5 | 22,9 | 20,4 |
| среднее | 19,5 | 24,26 | 22,84 | 21,16 |
| ст.откл | 2,69 | 1,92 | 2,13 | 2,98 |

.

Пусть m1, m2,…, mс –математические ожидания результативного признака Υ при соответствующих уровнях фактора А. В данном примере результативный признак Υ- сила разрыва, уровни фактора А – группы поставщиков.

Если при изменении уровня фактора групповые математические ожидания не изменяются, т.е. выполняется условие равенства мат.ожиданий : H0: m1=m2=…=mсто считается, что результативный признак не зависит от фактора А. В противном случае такая зависимость имеется (H1: не все мат.ожидания равны).

Поскольку мат.ожидания не известны, необходимо подтвердить гипотезу об их равенстве на основе выборочных данных.

Эту гипотезу Н0: m1 = m2=…= mс можно подтвердить с помощью F – критерия Фишера, если выполняются следующие условия:

1. наблюдения должны быть случайными, независимы и проводиться в одинаковых условиях.
2. экспериментальные данные должны иметь нормальный закон распределения
3. их дисперсии должны быть одинаковыми.

Если эти условия выполняются, то можно приступать непосредственно к процедуре дисперсионного анализа, т.е. к проверке гипотезы о равенстве средних величин: Н0: m1 = m2=…= mс

Проверить эту гипотезу можно, изучая вариации отдельных значений признака. Общая изменчивость значений признака может быть вызвана как изменчивостью значений признака между различными группами (межгрупповая вариация), так и изменчивостью значений признака внутри группы (внутригрупповая вариация). Для измерения степени вариации используется показатель – сумма квадратов отклонений.

Общая (полная) вариация определяется полной суммой квадратов отклонений.



где  - общее среднее.

.

 - среднее значение в j –ой группе



Межгрупповая вариация, вызванная влиянием фактора A на X определяется по формуле

,

Внутригрупповая вариация определяется равенством



В общем случае выполняется равенство , т.е. полная вариация значений признаков определяется суммой межгрупповой и внутригрупповой вариации.

Для проверки гипотезы о равенстве средних величин используется *F-критерий* Фишера, статистика которого определяется отношением.



Статистика F-критерия подчиняется распределению Фишера с числом степеней свободы , где **n** – общее число наблюдений, **c** - число уровней фактора A.

Показатель MS определяется как сумма квадратов отклонения, приходящаяся на одну степень свободы.

,

,

где SSA – сумма квадратов отклонения, вызванная влиянием фактора A на X, а SSвн - сумма квадратов отклонения, вызванная влиянием остаточных факторов на Y.

Для проверки гипотезы определяется правосторонняя критическая область, т.е. вычисляется **Fкр**при уровне значимости ** (см. функцию Excel F.ОБР.ПХ) и проверяется попадание рассчитанного значения Fрасч – статистики в интервал (**Fкр;+∞**). Если попадает, то гипотеза отклоняется, в противном случае принимается.

Прежде чем использовать F – критерий Фишера необходимо установить на основе имеющихся выборочных данных, являются ли генеральные дисперсии результативного признака при различных условиях фактора одинаковыми или нет. Проверяется гипотеза

H0:σ1=σ2=σ3…=σ против гипотезы Н1: не все дисперсии одинаковы.

Для проверки равенства трёх или более дисперсий используется критерий Бартлетта *w*.

,

где *q* вычисляется по формуле:



где

*с* –число уровней фактора *А*;

*n1, …,nj, …, nс* -число наблюдений для *1,…, j,…, с*-ого уровня фактора *А*.

 - внутригрупповая дисперсия, соответствующая *j-ому* уровню фактора *А*.



- среднее арифметическое значение результирующего показателя (признака) при *j-ом* уровне фактора *А.*

,



При выполнении гипотезы о равенстве дисперсии **Н0: = =…=** критерий имеет распределение χ2 (хи – квадрат) с числом степеней свободы .

Для проверки гипотезы при заданном уровне значимости ** находится правосторонняя критическая точка **wкр***.*, которая определяет область отклонения - интервал **(wкр;+∞).** Если рассчитанное значение **w** попадает в эту область, то мы отклоняем гипотезу при уровне значимости *.* В противном случае гипотеза принимается.

Более мощным является модифицированный критерий Левенэ.

Вычисляются абсолютные величины разностей между наблюдениями и медианами в каждой группе (см. пример). Результат представлен в таблице 3.

таблица 3.

|  |  |  |  |
| --- | --- | --- | --- |
| **Поставщик 1** | **Поставщик 2** | **Поставщик 3** | **Поставщик 4** |
| 0,0 | 1,8 | 2,3 | 5,0 |
| 5,5 | 0,8 | 2,3 | 0,5 |
| 1,3 | 0,5 | 2,1 | 2,2 |
| 1,4 | 3,3 | 1,8 | 2,9 |
| 0,5 | 0,0 | 0,0 | 0,0 |

Выполняется однофакторный дисперсионный анализ полученных значений абсолютных разностей

таблица 4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Однофакторный дисперсионный анализ | | |  |  |  |  |
|  |  |  |  |  |  |  |
| ИТОГИ |  |  |  |  |  |  |
| *Группы* | *Счет* | *Сумма* | *Среднее* | *Дисперсия* |  |  |
| Столбец 1 | 5 | 8,7 | 1,74 | 4,753 |  |  |
| Столбец 2 | 5 | 6,4 | 1,28 | 1,707 |  |  |
| Столбец 3 | 5 | 8,5 | 1,7 | 0,945 |  |  |
| Столбец 4 | 5 | 10,6 | 2,12 | 4,007 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Дисперсионный анализ | |  |  |  |  |  |
| *Источник вариации* | *SS* | *df* | *MS* | *F* | *P-Значение* | *F критическое* |
| Между группами | 1,77 | 3 | 0,59 | 0,20679986 | 0,890188801 | 3,238866952 |
| Внутри групп | 45,648 | 16 | 2,853 |  |  |  |
|  |  |  |  |  |  |  |
| Итого | 47,418 | 19 |  |  |  |  |

Поскольку Fрасч = 0,2067 < 3,2388, нулевая гипотеза о равенстве дисперсий не отклоняется. Между дисперсиями внутри каждой группы существенной разницы нет, т. е. условие об однородности данных выполняется .

Для проведения однофакторного дисперсионного анализа существует инструмент в пакете анализа Excel, который так и называется « Однофакторный дисперсионный анализ».

Здесь задаются следующие параметры:

1. входной интервал (вводится вся таблица с исходными данными);
2. вид группирования (по столбцам/ по строкам);
3. метки;
4. поле **;

указать выходной интервал

Результаты анализа для примера (табл.5), приведенного выше, выведены в таблицу

Таблица 5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Однофакторный дисперсионный анализ** | | | |  |  |  |
|  |  |  |  |  |  |  |
| **ИТОГО** |  |  |  |  |  |  |
| ***Группы*** | ***Счет*** | ***Сумма*** | ***Среднее*** | ***Дисперсия*** |  |  |
| **Поставщик 1** | **5** | **97,6** | **19,52** | **7,237** |  |  |
| **Поставщик 2** | **5** | **121,3** | **24,26** | **3,683** |  |  |
| **Поставщик 3** | **5** | **114,2** | **22,84** | **4,553** |  |  |
| **Поставщик 4** | **5** | **105,8** | **21,16** | **8,903** |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| **ДИСПЕРСИОННЫЙ АНАЛИЗ** | |  |  |  |  |  |
| ***Источник вариации*** | ***SS*** | ***df*** | ***MS*** | ***F*** | ***P-значение*** | ***F крит.*** |
| **Между группами** | **63,2855** | **3** | **21,09516667** | **3,461628925** | **0,0413656** | **3,238866952** |
| **Внутри групп** | **97,504** | **16** | **6,094** |  |  |  |
|  |  |  |  |  |  |  |
| **Итого** | **160,7895** | **19** |  |  |  |  |

Поскольку Fрасч = 3,4616<Fкрит=3,2388, гипотезу о равенстве средних величин H0: m1=m2= m3 =m4 отклоняется в пользу гипотезы H1.

При обнаружении значительных различий между математическими ожиданиями необходимо определить, какие именно группы отличаются друг от друга. Для этого используется процедура множественного сравнения Тьюки – Крамера, описанная ниже.

1. Вычисляются разности , где i ≠ j – номера групп, между средними значениями **c(c – 1)** групп;
2. Вычисляется критический размах процедуры Тьюки – Крамера по формуле



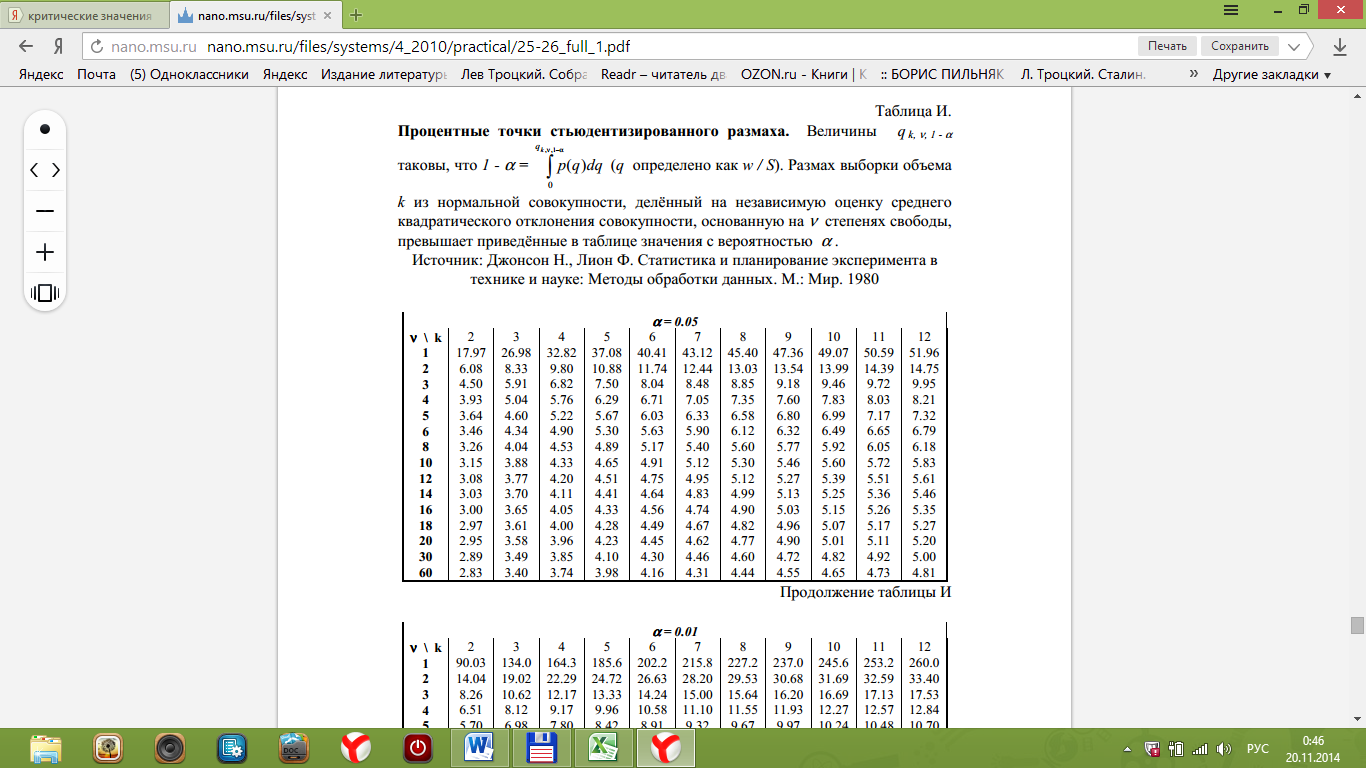
где Qu –верхнее критическое значение распределения стьюдентизированного размаха, имеющего c степеней свободы в числителе df1 = c и df2 = n степеней свободы в знаменателе, n-общее число наблюдений, n1 и n2 число наблюдений в i-ой и j-ой группах соответственно.

1. Каждая из c(c – 1)/2 пар разностей математических ожиданий сравнивается с рассчитанным критическим размахом.

Элементы пары считаются значительно различными, если модуль разности между ними  превышает критический размах. Результаты расчетов приведены ниже.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Absolute | Std. Error | Critical |  |  |  |
| **Comparison** | Difference | of Difference | Range | **Results** |  |  |
|  | 4,74 | 1,10399275 | 4,4712 | **Means are different** | | |
|  | 3,32 | 1,10399275 | 4,4712 | **Means are not different** | | |
|  | 1,64 | 1,10399275 | 4,4712 | **Means are not different** | | |
|  | 1,42 | 1,10399275 | 4,4712 | **Means are not different** | | |
|  | 3,1 | 1,10399275 | 4,4712 | **Means are not different** | | |
|  | 1,68 | 1,10399275 | 4,4712 | **Means are not different** | | |

Таблица стьюдентизованного распределения Qu κ=df1, ν=df2.



**ЗАДАНИЕ**

1. Сгенерировать 4 нормально распределенные переменные. Первые 3 переменные генерируется в соответствии с Вашим вариантом. При генерации четвертой переменной математическое ожидание увеличивается на 2, а стандартное отклонение не изменяется
2. Используя модифицированный критерий Левенэ проверить гипотезу о равенстве дисперсий.
3. . Используя инструмент анализа « Однофакторный дисперсионный анализ» проверить гипотезу о равенстве математических ожиданий.
4. При обнаружении значительных различий между математическими ожиданиями необходимо определить, какие именно группы отличаются друг от друга, используя процедуру множественного сравнения Тьюки – Крамера

# 8. Двухфакторный дисперсионный анализ без повторений и с повторениямии

Аналогично задаче однофакторного дисперсионного анализа можно рассмотреть задачу о действии на результативный признак *Y* двух факторов – *A* и *B*. Логика однофакторного и двухфакторного дисперсионного анализа во многом схожа и состоит в следующем.

Пусть  – математическое ожидание результативного признака *Y* при уровне *A(j)* (*j* = 1, 2, … *ca*);  – математическое ожидание результативного признака Y при уровне *B(l)* (*l* = 1, 2, … *cb*). Если при изменении уровня фактора *A* групповые математические ожидания не изменяются, то считаем, что результативный признак не зависит от фактора *А*, в противном случае такая зависимость имеется. Аналогично, если при изменении уровня фактора *B* сохраняется равенство групповых математических ожиданий, то считаем, что *Y* не зависит от фактора *B*. Но поскольку числовые значения математических ожиданий неизвестны, возникает задача проверки следующих гипотез:



Проверять эти гипотезы, так же как и в задаче однофакторного дисперсионного анализа, можно только при соблюдении следующих требований:

1) при различных сочетаниях уровней факторов *A* и *B* наблюдения независимы;

2) при каждом сочетании уровней факторов *A* и *B* результативный признак *Y* имеет нормальный закон распределения с постоянной для различных сочетаний генеральной дисперсией *σ2*.

Основой проведения двухфакторного дисперсионного анали­за служит комбинационная группировка по двум факторам с последующим разложением полной вариации результативного признака *SSОБЩ* на сумму вариаций фактора *A* и фактора *B*, вариации взаимодействия и случайной ошибки по формуле



где  – полная вариация, определяется как полная сумма квадратов отклонения от общего среднего:

.

Общее среднее значение равно:



*SSА* – вариация фактора *A*, вызванная влиянием на *Y* фактора *A*;



где  – среднее значение, соответствующее *j*-ому уровню фактора A



*SSВ* – вариация фактора В, вызванная влиянием на *Y* фактора *B*;



где  – среднее значение, соответствующее *l*-ому уровню фактора *B*



*SSВ* – вариация взаимодействия. Определяет эффект взаимодействия между факторами *A* и *B*.



где  – среднее значение, соответствующее *i*-ому уровню фактора *A* и *l*-ому уровню фактора *B*



*SSВ* – случайная ошибка. Показатель вариации, вызванной влиянием на *Y* остаточных факторов.

Если каждую сумму квадратов отклонений (вариацию) разделить на соответствующее число степеней свободы, то получится четыре типа дисперсии *МSА*, *МSВ*, *МSАВ*, *МSЕ*.





В двухфакторном дисперсионном анализе применяются три разных критерия.

1. Для проверки гипотезы об отсутствии эффекта фактора *A*  и альтернативной гипотезы *H1:* не все *mj* равны используется *F-критерий Фишера*:



В математической статистике доказывается, что если гипотеза  верна, то величина *FA* имеет F-распределение с числом степеней свободы *df1* = (*сa* – – 1) и *df2* = (*сА* – 1) (*сВ* – 1).

2. Для проверки гипотезы об отсутствии эффекта фактора B  и альтернативной гипотезы *H1:* не все *ml* равны используется *F-критерий Фишера*:



В математической статистике доказывается, что если гипотеза  верна, то величина *FВ* имеет F-распределение с числом степеней свободы *df1* = (*cb* – 1) и *df2* = (*ca* – 1) (*cb* – 1).

3. Для проверки гипотезы об отсутствии эффекта взаимодействия факторов *A* и *B* – : взаимодействие факторов *A* и *B* равно нулю и альтернативной гипотезы – : взаимодействие факторов *A* и *B* не равно нулю используется *F-критерий Фишера*:



Проверка выдвинутых гипотез осуществляется так же, как и при однофакторном дисперсионном анализе, и состоит в нахождении правосторонних критических интервалов  с после­дующим контролем попадания (или непопадания) в данный интер­вал расчетных значений *FA* (или *FB*) Если расчетное значение по­падает в критический интервал, то гипотеза  () отвергается, т.е. считается, что фактор *A* (*B*) влияет на результативный признак *Y*

Двухфакторный дисперсионный анализ может иметь две раз­новидности: без повторений и с повторениями. В первом случае каждому уровню факторов соответствует только одна выборка данных, во втором — определенным уровням факторов может со­ответствовать более одной выборки данных.

*Пример*

1. Сгенерированы исходные данные для анализа, представленные в табл.6.

Таблица 6.

Четыре нормально распределённые выборки по 2 на каждую переменную

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1-ая  переменная | 2-ая  переменная | 3-я  переменная | 4-ая  переменная |
| 1-ая | 96,56642 | 102,6006 | 117,1175 | 75,17628 |
| группа | 141,3495 | 85,01397 | 84,21796 | 108,8279 |
|  | 108,3845 | 112,0894 | 89,14946 | 77,90009 |
|  | 83,816 | 91,26086 | 92,55361 | 74,80965 |
|  | 141,8192 | 117,6365 | 95,43391 | 125,2423 |
|  | 87,72672 | 126,2366 | 101,8187 | 87,86416 |
|  | 86,15352 | 103,9551 | 86,88417 | 107,3838 |
|  | 73,1155 | 84,20715 | 80,73287 | 86,10379 |
|  | 97,04241 | 84,0885 | 65,27771 | 94,44711 |
|  | 89,59141 | 129,1113 | 73,73314 | 118,9077 |
| 2-ая | 84,05879 | 83,02714 | 111,2919 | 100,1878 |
| группа | 87,90936 | 82,06905 | 74,1734 | 119,4141 |
|  | 93,14024 | 75,46683 | 126,9855 | 71,48985 |
|  | 107,2834 | 80,61729 | 85,59402 | 78,89242 |
|  | 55,35531 | 73,1825 | 98,2151 | 107,6148 |
|  | 70,82559 | 76,22452 | 66,0066 | 106,3478 |
|  | 63,48997 | 83,27096 | 96,75027 | 91,89195 |
|  | 104,5162 | 70,56275 | 92,17939 | 88,00212 |
|  | 97,39804 | 107,8439 | 101,0735 | 94,61252 |
|  | 111,1621 | 103,8345 | 83,60792 | 127,2286 |

2. Проведения двухфакторного дисперсионного анализа с повторениями для выяснения на основе выборочных данных табл.6 факта влияния контролируемых факторов *A(группа)* и *B(переменная)* на результативный признак *Y*.

Таблица 7

Результаты анализа

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Двухфакторный дисперсионный анализ с повторениями | | | | | |  |
|  |  |  |  |  |  |  |
| ИТОГИ | 1-ая переменная | 2-ая переменная | 3-я переменная | 4-ая переменная | Итого |  |
| *1ая-гр* |  |  |  |  |  |  |
| Счет | 10 | 10 | 10 | 10 | 40 |  |
| Сумма | 1005,565 | 1036,2 | 886,919 | 956,6627 | 3885,347 |  |
| Среднее | 100,5565 | 103,62 | 88,6919 | 95,66627 | 97,13367 |  |
| Дисперсия | 553,5287 | 298,6593 | 211,0579 | 338,8565 | 356,1788 |  |
|  |  |  |  |  |  |  |
| *2ая-гр* |  |  |  |  |  |  |
| Счет | 10 | 10 | 10 | 10 | 40 |  |
| Сумма | 875,139 | 836,0994 | 935,8775 | 985,682 | 3632,798 |  |
| Среднее | 87,5139 | 83,60994 | 93,58775 | 98,5682 | 90,81995 |  |
| Дисперсия | 363,3798 | 156,4353 | 313,268 | 299,2436 | 294,796 |  |
|  |  |  |  |  |  |  |
| *Итого* |  |  |  |  |  |  |
| Счет | 20 | 20 | 20 | 20 |  |  |
| Сумма | 1880,704 | 1872,3 | 1822,797 | 1942,345 |  |  |
| Среднее | 94,03521 | 93,61498 | 91,13983 | 97,11723 |  |  |
| Дисперсия | 479,0908 | 320,9404 | 254,6727 | 304,474 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| Дисперсионный анализ | | |  |  |  |  |
| *Источник вариации* | *SS* | *df* | *MS* | *F* | *P-Значение* | *F критическое* |
| Выборка | 797,2629 | 1 | 797,2629 | 2,516584 | 0,117037 | 3,973897 |
| Столбцы | 360,9016 | 3 | 120,3005 | 0,379732 | 0,767892 | 2,731807 |
| Взаимодействие | 2217,254 | 3 | 739,0846 | 2,332942 | 0,0812 | 2,731807 |
| Внутри | 22809,86 | 72 | 316,8037 |  |  |  |
|  |  |  |  |  |  |  |
| Итого | 26185,28 | 79 |  |  |  |  |

Анализ полученных результатов и проверка гипотез:

1. Гипотеза  не отклоняется, так как  меньше критического значения (соответственно *p*-значение больше уровня значимости α = 0,05);

2. Гипотеза  не отклоняется, так как  меньше критического значения (соответственно *p*-значение больше уровня значимости α = 0,05);

3. Гипотеза  не отклоняется, так как  меньше критического значения (соответственно *p*-значение больше уровня значимости α = 0,05).

**ЗАДАНИЕ**

1. Сгенерировать 4 нормально распределенные переменные. Первые 3 переменные генерируется в соответствии с Вашим вариантом. При генерации четвертой переменной математическое ожидание увеличивается на 2, а стандартное отклонение не изменяется
2. Проверить гипотезу 
3. Проверить гипотезу 

**9. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ**

Задача регрессионного анализа состоит в применении статистических методов для описания взаимосвязей между случайными величинами (признаками) с помощью математических моделей и оценки параметров этих моделей на основе данных статистического наблюдения.

Признаки разделяют на два класса: зависимые (результирующие) и независимые (факторные, предикторные, объясняющие), то есть те, от которых зависят результирующие.

Для количественной оценки зависимости между результирующими и факторными признаками определяются структура математической модели (вид аппроксимирующей функции) и ее параметры (коэффициенты аппроксимирующей функции). Оценка параметров (коэффициентов) модели производится на основе данных статистического наблюдения по совокупности показателей исследуемой зависимости. При проведении эконометрического исследования предполагается, что значения результирующей переменной носят случайный характер, поскольку зависят не только от факторных переменных, но и от тех факторов, которые мы не учитываем явно.

Обозначим: Y - зависимая переменная, а x1, x2,…,xj,…,xm – это m независимых (факторных) переменных. Здесь j – номер факторной переменной. Формально эконометрическая модель записывается следующим образом

Y = f(x1, x2,…,xm) + ε. (1)

Здесь f(x1, x2,…,xm) - аппроксимирующая функция m независимых переменных (детерминированная компонента), ε – случайная компонента, отражающая влияние факторов, не учтенных в модели. Уравнение вида y = f(x1, x2,…,xm) называется уравнением регрессии. Частным видом эконометрической модели вида (1) является модель множественной линейной регрессии

Y = a0 + a1x1 + a2x2 + ... + amxm + ε

Коэффициенты a1, a2, …, am называются параметрами уравнения регрессии. Пусть имеется n статистических наблюдений и i – номер наблюдения. Данные статистического наблюдения можно представить в виде таблицы, состоящей из n строк вида (yi, x1i, x2i,…,xji,…,xmi). На основе имеющихся статистических наблюдений можно подобрать параметры a1, a2, …, am таким образом, чтобы уравнение множественной линейной регрессии наилучшим образом описывало бы наблюдаемые данные. Значения параметров a1, a2, …, am, определенные тем или иным способом называются оценками. Зная оценки параметров, модели можно использовать далее для прогнозирования значений результирующих показателей при тех или иных сочетания значений факторных показателей. Используя регрессионные модели можно проводить многовариантные расчеты по принципу «что будет, если…?». Эти модели находят широкое применение при принятии ответственных решений на практике.

**Простая линейная регрессия**

Простаялинейная регрессия используется для исследования зависимости двух переменных. Уравнение простой линейной регрессии можно записать в виде

yi = a0 + a1xi + εi(2)

где ε1,…εn- независимые одинаково распределенные случайные величины, определяющие действие различных неучтенных факторов на изменение результирующего показателя Y.

Для определения оценок параметров в уравнении (2) широко используется ***метод наименьших квадратов (МНК),*** суть которого заключается в следующем***.***

Определим величину ei следующим образом:

ei = yi – (a0 + a1xi).

Величина ei называется отклонением (остатком) наблюдаемого значения результирующей переменной yi в i – ом наблюдении от расчетного. Отклонение ei является оценкой случайной компоненты εi. По­строим оценку параметров (a0, a1) так, чтобы сумма их квадратов отклонений была минимальной:

 (3)

Сумму минимимизируем по (a0, a1),приравнивая нулю соответствующие производные.В результате получим систему уравнений линейных относительно a0 и a1*.* Ее решение  легко находится:

**** (4) и (5)

Тесноту связи изучаемых явлений оценивает линейный коэффициент парной корреляции ryx. Для линейной регрессии (-1≤ryx≤1)

ryx = a1sx/sy

sx =, sy =,

здесь sx и sy - стандартные (среднеквадратические) отклонения по x и y.

**Проверка качества подгонки регрессионной модели к наблюдаемым данным**

Качество построенной модели можно оценить с помощью коэффициента (индекса) детерминации:

R2 = =,

где  - выборочная дисперсия результирующего показателя y;  - остаточная дисперсия, объясняемая случайной компонентой, - дисперсия, объясняемая регрессией. Чем больше значение этого показателя ( а оно изменяется от 0 до 1), тем лучше уравнение регрессии объясняет рассеяние наблюдаемых значений результирующего показателя y относительно средней величины, тем меньшее влияние на это рассеяние оказывают случайные факторы.

Для оценки качества уравнения регрессии также используется показатель, связанный с индексом детерминации. Это показатель называется множественный коэффициент корреляции

R = √ R2

Для парной линейной регрессии R = ryx. Степень тесноты связи результирующей и факторных переменных можно оценить, используя шкалу Чеддока

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Значение R | 0,1 – 0,3 | 0,3 – 0,5 | 0,5 – 0,7 | 0,7 – 0,9 | 0,9 - 0,99 |
| Степень связи | слабая | умеренная | заметная | высокая | очень высокая |

При значении R более 0,7 величина R2 будет более 50%. Это означает, что более 50% вариации результирующей переменной объясняется уравнением регрессии.

**Дисперсионный анализ**

Для определения статистической значимости показателя тесноты связи и существенности связи между результирующей и факторными переменными проводится дисперсионный анализ. Задача состоит в исследовании дисперсии результирующего показателя.

Проверка гипотезы о существенности связи результирующей и факторных переменных в уравнения регрессии (статистической значимости множественного коэффициента корреляции) осуществляется с помощью F-критерия Фишера. Величина F-критерия связана с коэффициентом детерминации R2:

 .

Проверка существенности связи в уравнения регрессии с помощью F-критерия проводится при условии нормальности распределения ошибки регрессии.

Для проверки вычисляется F-статистика:



где :



Из таблиц распределения Фишера определяется **критическое** значение Fdf1,df2,α при заданном уровне значимости α и степенях свободы df1 = 1, df2 = n-2, где **уровень значимости** α **– вероятность совершения ошибки первого рода**.

Если Fpасч> Fdf1,df2,α , то полученное значение множественного коэффициента корреляции можно считать статистически значимым. В противном случае полагаем R = 0, что свидетельствует об отсутствии линейной зависимости между результирующей и факторными переменными в уравнения регрессии

В пакетах программ используется **другой способ проверки** гипотезы о существенности связи результирующей и факторных переменных в уравнения регрессии. Там автоматически рассчитывается *p-уровень* (pF),т.е. значение вероятности, соответствующее расчетному значению F-критерия.



f(F)

F



Fрасч

α

p – уровень

Если pF<α, то полученное значение множественного коэффициента корреляции можно считать статистически значимым. В противном случае полагаем R = 0, что свидетельствует об отсутствии линейной зависимости между результирующей и факторными переменными в уравнения регрессии. Чем меньше значение p-уровня, тем надежнее полученные оценки.

**Проверка статистической значимости коэффициентов уравнения регрессии**

При таком же предположении можно проверить гипотезы относительно каждого коэффициента с использованием t-статистики Стьюдента:

a0, a1 – коэффициенты уравнения регрессии,

r – коэффициент корреляции.

t-статистика для коэффициента уравнения регрессии a0 – ;

t-статистика для коэффициента уравнения регрессии a1 – ;

t-статистика для коэффициента корреляции r – .

Ma0, ma1, mr – стандартные ошибки.

; ; .

Для проверки значимости этих коэффициентов необходимо сравнить полученные расчетные значения ta0, ta1, tr с табличным значением распределения Стьюдента с df степенями свободы при уровне значимости α, т.е. с tdf,α (df = n-2).

Если расчетное значение по абсолютной величине больше табличного, то нулевая гипотеза H0

Н0: a0 =0,

Н0: a1 = 0,

Н0: r = 0.

отвергается и значение соответствующего коэффициента считается статистически значимым при данном уровне значимости α.

Другой способ проверки заключается в сравнении p – уровня критерия Стьюдента (ptj) с уровнем значимости α. Если ptj<α, то полученное значение проверяемого коэффициента уравнения регрессии можно считать статистически значимым.

Связь между F-критерием Фишера и t – статистикой Стьюдента выражается равенством:

 .

Таким образом, проверка гипотез о значимости коэффициентов регрессии и корреляции равна проверке гипотезы о существенности связи между переменными (результирующей и факторными) в уравнении регрессии.

Качество уравнения регрессии можно также оценить с помощью средней ошибки аппроксимации



**Множественная регрессия**

Обобщением линейной регрессионной модели с двумя переменными является многомерная регрессионная модель (или модель множественной регрессии). Уравнение линейной регрессии . В экономике широко используется степенная функция вида: . Эта функция используется для изучения спроса и потребления, для построения производственной функции, где y – выпуск, а x – факторы производства и др.

*Коэффициенты линейной модели уравнения регрессии* называются коэффициентами чистой регрессии. В случае полинома  коэффициенты характеризуют среднее изменение результата, при изменении соответствующего фактора на одну единицу и при неизменной величине остальных факторов.

В степенной функции коэффициенты чистой регрессии показывают, на сколько процентов изменится результат, при изменении соответствующего фактора на один процент и при фиксированном значении остальных факторов. Они играют роль коэффициентов эластичности.

Решение уравнения регрессии находится с помощью метода наименьших квадратов. Анализ полученного решения заключается в проверке полученного уравнения регрессии путем расчета коэффициента множественной детерминации:



и F – статистики:

.

Если известен коэффициент детерминации R2, то F – статистка может быть рассчитана следующим образом:



Рассчитанное значение сравнивается с табличным Fdf1,df2,α (), где m – число независимых переменных, n – число наблюдений. Либо для расчетного значения F – статистики определяется p – уровень, который сравнивается с уровнем значимости α, так как это было описано в предыдущем разделе.

Недостатком коэффициента детерминации является то, что он увеличивается при добавлении новых переменных, хотя это и не обязательно означает улучшения качества регрессионной модели. Поэтому лучше пользоваться скорректированным коэффициентом детерминации, который определяется по формуле:



Проверка значимости коэффициентов регрессии аналогична проверке коэффициентов парной регрессии и сводится к вычислению значения статистики

,

где - среднеквадратическая ошибка коэффициента регрессии aj и сравнении ее с критическим значением. Альтернативой является вычисление p – уровня критерия Стьюдента (ptj) для каждого коэффициента aj и сравнение его с уровнем значимости α так, как это было описано выше.

Уравнение регрессии может быть преобразовано к стандартизованному масштабу 

, где j – номер переменной.

Значения коэффициентов βj можно определить из уравнения:

,

где  - коэффициенты взаимной корреляции между xk  и xj.

Основное достоинство стандартизованного уравнения регрессии в том, что β - коэффициенты позволяют ранжировать факторы по степени их воздействия на результат. Чем больше значение β - коэффициента, тем больший вклад вносит соответствующая факторная переменная в значение результирующей.

Коэффициенты чистой регрессии bj связаны со стандартизованными коэффициентами βj соотношением



**Решение с помощью MS Exel**

С помощью инструмента анализа данных **Регрессия,** помимо результатов регрессионной статистики, дисперсионного анализа и доверительных интервалов, можно получить остатки и графики подбора линии регрессии, остатков и нормальной вероятности. Порядок действия следующий:

* + в главном меню выберите **Сервис/Анализ данных/Регрессия.** Щелкните по кнопке **OK**;
  + заполните диалоговое окно ввода данных и параметров вывода. Щелкните по кнопке **OK**;

Пример использования инструмента **Регрессия** приведен в файле Пример.xls

Исходные данные

|  |  |
| --- | --- |
| 596 | 913 |
| 417 | 1095 |
| 354 | 606 |
| 526 | 876 |
| 934 | 1314 |
| 412 | 593 |
| 525 | 754 |
| 367 | 528 |
| 364 | 520 |
| 336 | 539 |
| 409 | 540 |
| 452 | 682 |
| 367 | 537 |
| 328 | 589 |
| 460 | 626 |
| 380 | 521 |
| 439 | 626 |
| 344 | 521 |
| 401 | 658 |
| 514 | 746 |
|  |  |
| СрЗначY | СрЗначX |
| 446,25 | 689,2 |
| СтОтклY | СтОтклY |
| 136,1179 | 212,9822 |

Две последние строки содержат выборочные средние значения и выборочные стандартные отклонения sx, sy, рассчитанные с помощью функции СТАНДОТКЛ.

Исходные данные вводятся в окно ввода (рис.1).

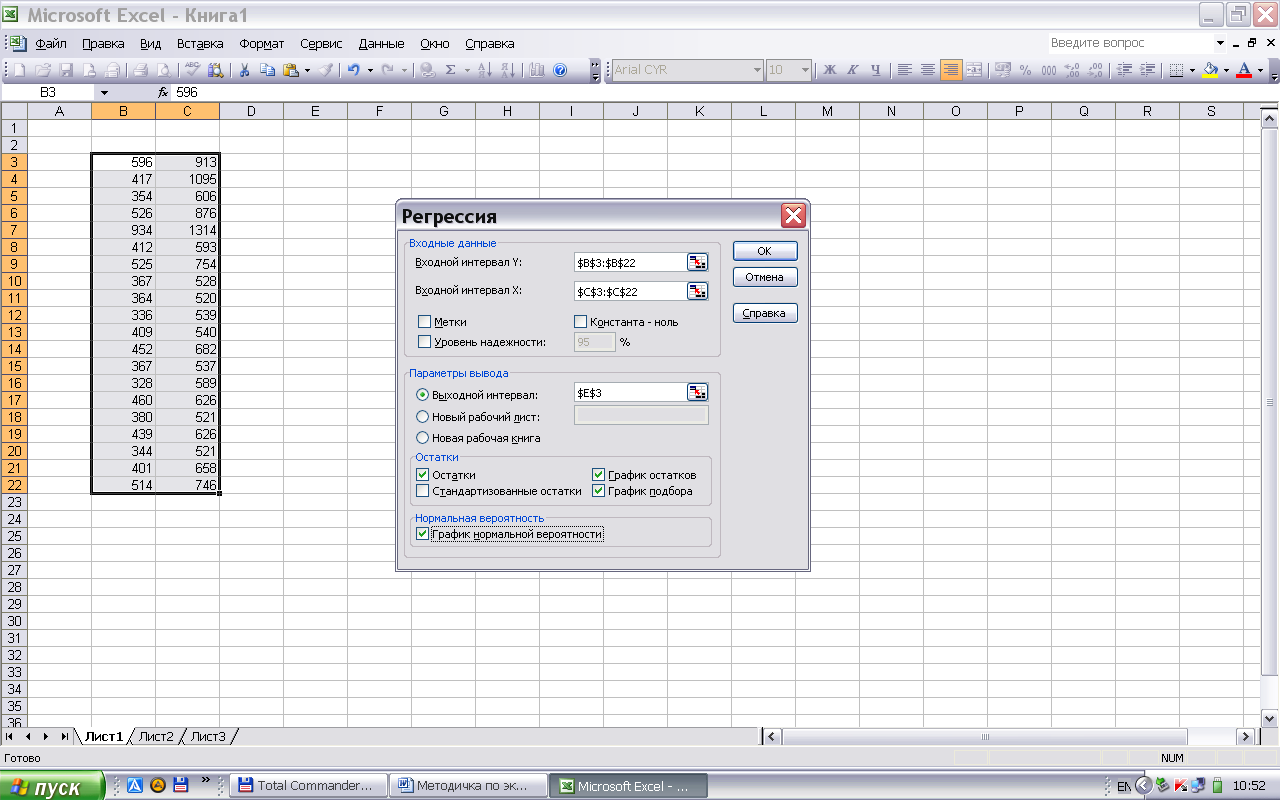


Рис 1. Окно Ввода

Результаты расчета приведены ниже. Здесь:

Множественный R - R

R-квадрат - R2

Нормированный R-квадрат - 

Значимость F - pF

Коэффициенты - значения коэффициентов (оценки)

Y-пересечение - a0

Переменная X1 – a1

P-Значение - ptj

Предсказанное Y - 

Остатки 

|  |  |
| --- | --- |
| ВЫВОД ИТОГОВ | |
|  |  |
| *Регрессионная статистика* | |
| Множественный R | 0,840977 |
| R-квадрат | 0,707243 |
| Нормированный R-квадрат | 0,690979 |
| Стандартная ошибка | 75,66752 |
| Наблюдения | 20 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Дисперсионный анализ | | |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Значимость F* |  |
| Регрессия | 1 | 248973,4 | 248973,4 | 43,48445 | 3,42E-06 |  |
| Остаток | 18 | 103060,3 | 5725,573 |  |  |  |
| Итого | 19 | 352033,8 |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* |
| Y-пересечение | 75,82389 | 58,66674 | 1,292451 | 0,212547 | -47,4303 | 199,0781 |
| Переменная X 1 | 0,537473 | 0,081506 | 6,594274 | 3,42E-06 | 0,366235 | 0,70871 |
|  |  |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
| ВЫВОД ОСТАТКА | |  |
|  |  |  |
| *Наблюдение* | *Предсказанное Y* | *Остатки* |
| 1 | 566,5364 | 29,46363 |
| 2 | 664,3564 | -247,356 |
| 3 | 401,5323 | -47,5323 |
| 4 | 546,6499 | -20,6499 |
| 5 | 782,0629 | 151,9371 |
| 6 | 394,5451 | 17,45486 |
| 7 | 481,0782 | 43,92178 |
| 8 | 359,6094 | 7,390582 |
| 9 | 355,3096 | 8,690363 |
| 10 | 365,5216 | -29,5216 |
| 11 | 366,0591 | 42,94091 |
| 12 | 442,3802 | 9,619803 |
| 13 | 364,4467 | 2,553329 |
| 14 | 392,3952 | -64,3952 |
| 15 | 412,2817 | 47,71827 |
| 16 | 355,8471 | 24,15289 |
| 17 | 412,2817 | 26,71827 |
| 18 | 355,8471 | -11,8471 |
| 19 | 429,4809 | -28,4809 |
| 20 | 476,7784 | 37,22156 |

График подбора содержит наблюдаемые и предсказанные значения, иллюстрирует размах отклонений рассчитанных значений от наблюдаемых для переменной Y.

**Переменная X 1 График подбора**

0

500

1000

0

500

1000

1500

**Переменная X 1**

**Y**

Y

Предсказанное Y

График нормального распределения используется для визуальной проверки выполнения условий Маркова-Гаусса



# Контрольное задание

Исходные данные для 20 наблюдений приведены в таблице.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X0 | X1 | X2 | X3 | X4 | X5 |
| 1 | 9,7 | 1,59 | ,26 | 2,05 | ,32 | ,14 |
| 2 | 8,4 | ,34 | ,28 | ,46 | ,59 | ,66 |
| 3 | 9,0 | 2,53 | ,31 | 2,46 | ,30 | ,31 |
| 4 | 9,9 | 4,63 | ,40 | 6,44 | ,43 | ,59 |
| 5 | 9,6 | 2,16 | ,26 | 2,16 | ,39 | ,16 |
| 6 | 8,6 | 2,16 | ,30 | 2,69 | ,32 | ,17 |
| 7 | 12,5 | ,68 | ,29 | ,73 | ,42 | ,23 |
| 8 | 7,6 | ,35 | ,26 | ,42 | ,21 | ,08 |
| 9 | 6,9 | ,52 | ,24 | ,49 | ,20 | ,08 |
| 10 | 13,5 | 3,42 | ,31 | 3,02 | 1,37 | ,73 |
| 11 | 9,7 | 1,78 | ,30 | 3,19 | ,73 | ,17 |
| 12 | 10,7 | 2,40 | ,32 | 3,30 | ,25 | ,14 |
| 13 | 12,1 | 9,36 | ,40 | 11,51 | ,39 | ,38 |
| 14 | 9,7 | 1,72 | ,28 | 2,26 | ,82 | ,17 |
| 15 | 7,0 | ,59 | ,29 | ,60 | ,13 | ,35 |
| 16 | 7,2 | ,28 | ,26 | ,30 | ,09 | ,15 |
| 17 | 8,2 | 1,64 | ,29 | 1,44 | ,20 | ,08 |
| 18 | 8,4 | ,09 | ,22 | ,05 | ,43 | ,20 |
| 19 | 13,1 | ,08 | ,25 | ,03 | ,73 | ,20 |
| 20 | 8,7 | 1,36 | ,26 | ,17 | ,99 | ,42 |

В соответствии со своим вариантом необходимо:

* + построить 3 уравнения линейной регрессии, последовательно увеличивая число факторных переменных от одной до 3;
  + определить оценки параметров в уравнениях регрессии;
  + определить качество полученных уравнений регрессии и их статистическую значимость;
  + оценить статистическую значимость параметров регрессии;
  + построить графики остатков для полученных регрессий;
  + выбрать лучшую модель;
  + для этой модели рассчитать нормированные коэффициенты **βj**

По каждому пункту сделать выводы.

**Варианты задания** определяются по списку следующим образом.

В качестве независимой переменной используйте X0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | X1 | X2 | X3 | X4 | X5 |
|  |  |  |  |  |  |
| 1 | + |  | + |  | + |
| 2 | + | + |  |  | + |
| 3 | + | + | + |  |  |
| 4 | + |  | + | + |  |
| 5 | + | + | + |  |  |
| 6 |  | + | + | + |  |
| 7 |  | + | + |  | + |
| 8 |  | + |  | + | + |
| 9 |  | + |  | + |  |

В качестве независимой переменной используйте X1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | X0 | X2 | X3 | X4 | X5 |
| 10 | + |  | + |  | + |
| 11 | + | + |  |  | + |
| 12 | + | + | + |  |  |
| 13 | + |  | + | + |  |
| 14 | + | + | + |  |  |
| 15 |  | + | + | + |  |
| 16 |  | + | + |  | + |
| 17 |  | + |  | + | + |
| 18 |  | + |  | + |  |

В качестве независимой переменной используйте X2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № | X0 | X1 | X3 | X4 | X5 |
| 19 | + |  | + |  | + |
| 20 | + | + |  |  | + |
| 21 | + | + | + |  |  |
| 22 | + |  | + | + |  |
| 24 | + | + | + |  |  |
| 25 |  | + | + | + |  |
| 26 |  | + | + |  | + |
| 27 |  | + |  | + | + |
| 28 |  | + |  | + |  |

**9. Простая линейная регрессия**

Простаялинейная регрессия используется для исследования зависимости двух переменных. Уравнение простой линейной регрессии можно записать в виде

yi = a0 + a1xi + εi(1)

где ε1,…εn- независимые одинаково распределенные случайные величины, определяющие действие различных неучтенных факторов на изменение результирующего показателя Y.

Для определения оценок параметров в уравнении (1) широко используется ***метод наименьших квадратов (МНК),*** суть которого заключается в следующем***.***

Определим величину ei следующим образом:

ei = yi – (a0 + a1xi).

Величина ei называется отклонением (остатком) наблюдаемого значения результирующей переменной yi в i – ом наблюдении от расчетного. Отклонение ei является оценкой случайной компоненты εi. По­строим оценку параметров (a0, a1) так, чтобы сумма их квадратов отклонений была минимальной:

 (2)

Сумму минимизируем по (a0, a1),приравнивая нулю соответствующие производные.В результате получим систему уравнений линейных относительно a0 и a1*.* Ее решение  легко находится:

**** (3) и (4)

Тесноту связи изучаемых явлений оценивает линейный коэффициент парной корреляции ryx. Для линейной регрессии

,

Степень тесноты связи результирующей и факторных переменных можно оценить, используя шкалу Чеддока

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Значение ryx | 0,1 – 0,3 | 0,3 – 0,5 | 0,5 – 0,7 | 0,7 – 0,9 | 0,9 - 0,99 |
| Степень связи | слабая | умеренная | заметная | высокая | очень высокая |

Для определения статистической значимости показателя тесноты связи и существенности связи между результирующей и факторными переменными проводится дисперсионный анализ. Задача состоит в исследовании дисперсии результирующего показателя.

Проверка гипотезы о существенности связи результирующей и факторных переменных в уравнения регрессии (статистической значимости множественного коэффициента корреляции) осуществляется с помощью F-критерия Фишера.

Для проверки вычисляется F-статистика:



где :



Здесь

 - наблюдаемые значения **y;**

 - значения **y**, вычисленные при соответствующих значениях **x;**

 - среднее значение **y.**

Из таблиц распределения Фишера определяется **критическое** значение Fdf1,df2,α при заданном уровне значимости α и степенях свободы df1 = 1, df2 = n-2, где уровень значимости α – вероятность совершения ошибки первого рода.

Если Fpасч> Fdf1,df2,α , то полученное значение множественного коэффициента корреляции можно считать статистически значимым. В противном случае полагаем ryx = 0, что свидетельствует об отсутствии линейной зависимости между результирующей и факторными переменными в уравнения регрессии

**Задание:**

Сгенерировать 45 значений случайной величины ε, считая, что случайная компонента ε распределена нормально с нулевым математическим ожиданием и единичной дисперсией. Для вычисления ε использовать генератор случайных чисел.

.

Таблица1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Номер*  *набл.* | *Значение*  *xi* | *Номер*  *набл* | *Значение*  *xi* | *Номер*  *набл* | *Значение*  *xi* |
| 1 | 6,5 | 16 | 9,3 | 31 | 10,4 |
| 2 | 10,3 | 17 | 5,7 | 32 | 10,2 |
| 3 | 7,7 | 18 | 12,9 | 33 | 18,0 |
| 4 | 15,8 | 19 | 5,1 | 34 | 13,8 |
| 5 | 7,4 | 20 | 3,8 | 35 | 6,0 |
| 6 | 14,3 | 21 | 17,1 | 36 | 11,9 |
| 7 | 15,4 | 22 | 8,2 | 37 | 9,4 |
| 8 | 21,1 | 23 | 8,1 | 38 | 13,7 |
| 9 | 22,1 | 24 | 11,7 | 39 | 12,0 |
| 10 | 12,0 | 25 | 13,0 | 40 | 11,6 |
| 11 | 9,5 | 26 | 15,3 | 41 | 9,1 |
| 12 | 8,1 | 27 | 13,5 | 42 | 6,6 |
| 13 | 8,4 | 28 | 10,5 | 43 | 7,6 |
| 14 | 15,3 | 29 | 7,3 | 44 | 9,9 |
| 15 | 4,3 | 30 | 13,8 | 45 | 14,7 |

Рассчитать 45 значений Y используя формулу

Y = 11,5 + (0,14N)x + 2ε (5)

где ε случайная компонента. Значения X возьмите из таблицы 1

Построить таблицу пар (Y,X), подставляя в уравнение (5) значения X из таблицы 1 и сгенерированные значения ε. N = Ваш номер по списку.

Используя данные таблицы (Y,X) рассчитать коэффициенты a0 и a1, а также коэффициент корреляции ryx и записать уравнение регрессии.

Сделать вывод о тесноте связи используя шкалу Чеддока.

Рассчитать F – статистику Фишера и сделать вывод о статистической значимости коэффициента корреляции.

# Список литературы

1. Гмурман, В.Е. Теория вероятностей и математическая статистика: учеб. пособие для вузов / В.Е. Гмурман. – 10-е изд., стер. – М.: Высш. шк., 2004. – 479 с.

2. Ветров, А.Н. Статистические программные системы: учебное  
пособие ⁄ А.Н. Ветров, А.Л. Борисов, Ю.Г. Козлова. – Тверь: Тверской  
государственный технический университет, 2014. – 152 с.

3. Ветров, А.Н. Методические указания и задания к курсовой работе по дисциплине Теория вероятностей и статистика / А.Н. Ветров. – Тверь: ТвГТУ, 2017. – 25 с.

5. Дисперсионный двухфакторный анализ [Электронный ресурс] // refdb.ru: сайт. – Режим доступа: https://refdb.ru/look/1210036.html

5. Шеффе, Г. Дисперсионный анализ / Г. Шеффе. – М.: Наука, 1980. 512 с.