**Лабораторный практикум  
по курсу «Компьютерный анализ статистических данных»**

**Преподаватель: Коробецкая Анастасия Александровна**

***kornast@yandex.ru***

Условием получения зачета является сдача всех трех лабораторных работ. По каждой работе выполняется ряд расчетов и делаются выводы.

Исходные данные для расчетов находятся в файле «КАСД Л.р. Варианты.xlsx», для каждой лабораторной – на своем листе.

Вариант = последняя цифра в номере зачетки (студенческого билета). Если цифра 0, то вариант 10.

К сдаче предоставляется одна книга Excel, содержащая лабораторные работы на разных листах. Выводы по каждой работе можно сдавать как письменно в том же файле, так и устно.

Имя файла обязательно должно содержать предмет, ФИО и группу, например, «КАСД 931-С Иванов И.И.xlsx».

Результаты для проверки и любые вопросы по работе можно присылать на электронную почту.

**Содержание**

[Подготовка Excel к работе 3](#_Toc434454021)

[Лабораторная работа №1 Описательная статистика 5](#_Toc434454022)

[Теоретические сведения 5](#_Toc434454023)

[Задание 16](#_Toc434454024)

[Пример 17](#_Toc434454025)

[Подготовка исходных данных 17](#_Toc434454026)

[Указания к выполнению работы 17](#_Toc434454027)

[Лабораторная работа №2 Гистограмма распределения 26](#_Toc434454028)

[Задание 34](#_Toc434454029)

[Пример 34](#_Toc434454030)

[Подготовка исходных данных 35](#_Toc434454031)

[Указания к выполнению работы 35](#_Toc434454032)

[Лабораторная работа №3 Статистические тесты. Исключение случайных выбросов 51](#_Toc434454033)

[Теоретические сведения 51](#_Toc434454034)

[Задание 58](#_Toc434454035)

[Пример 59](#_Toc434454036)

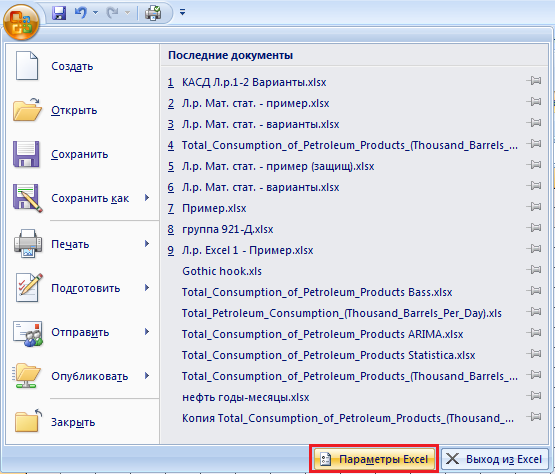
[Подготовка исходных данных 59](#_Toc434454037)

[Указания к выполнению работы 59](#_Toc434454038)

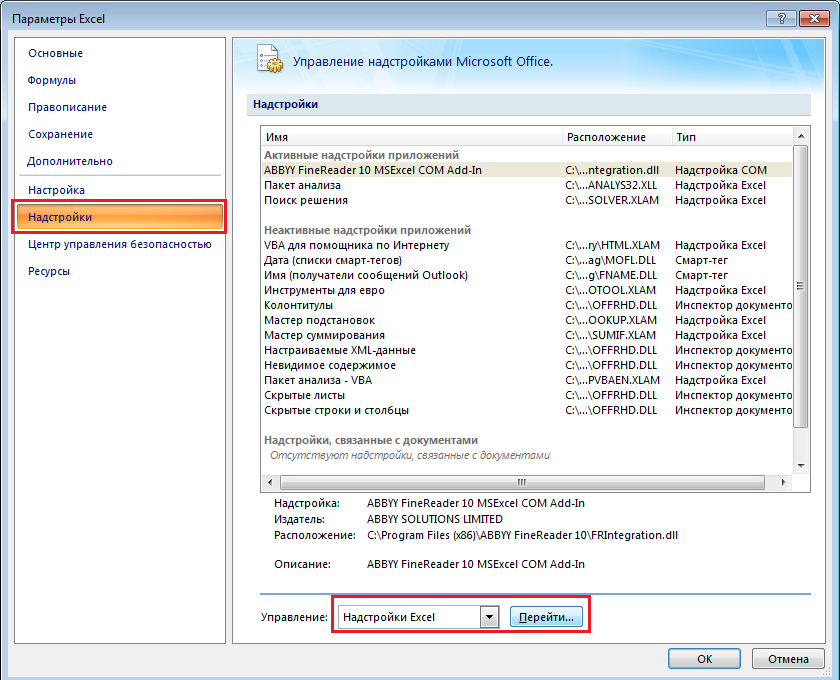
# Подготовка Excel к работе

Работа выполняется в MS Excel 2007 или выше. Для выполнения работ потребуется пакет «Анализ данных». По умолчанию он отключен.

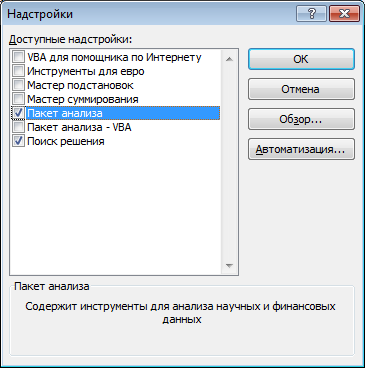
1. В меню Office нажать кнопку «Параметры Excel».



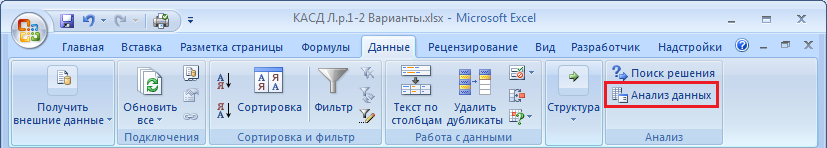
1. В окне параметров Excel выбрать «Надстройки», в выпадающем списке внизу выбрать «Надстройки Excel» и нажать кнопку «Перейти...»



3. В окне «Надстройки» поставить галочку «Пакет анализа» и нажать «OK».



4. В результате на Ленте на вкладке «Данные» должна появиться кнопка «Анализ данных».



# Лабораторная работа №1 Описательная статистика

## Теоретические сведения

### Основы статистики

Раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей, называется **математической статистикой**. Данный раздел математики имеет дело с массовыми явлениями и тесно связан с *теорией вероятностей*, так как базируется на ее математическом аппарате.

Методы статистики используются в различных науках и дисциплинах – физике, технике, биологии, медицине, экономике, социологии, психологии, юриспруденции и даже в лингвистике. С ее помощью можно изучать поведение потребителей на рынке, надежность технических систем и вероятность ошибок в сочинениях школьников.

**Статистика** – более широкое понятие, чем математическая статистика. Математическая статистика не изучает вопросы сбора исходных данных, проведения опросов и т.п., а только методы их обработки. Методы математической статистики не зависят от содержательного смысла изучаемого явления, но выводы по ее результатам – зависят.

*Цель* статистического исследования – исследование соотношений между статистическим данными (**описательная статистика**) и использование результатов исследований для прогнозирования и принятия решений (**аналитическая статистика**).

*Статистические данные* представляют собой данные, полученные в результате обследования большого число объектов или явлений.

**Статистическая совокупность** – множество обследуемых объектов (людей, семей, предприятий, автомобилей, животных, растений, ...).

По охвату статистической совокупности исследование может быть сплошное или не сплошное. При сплошном исследовании группа формируется из всех единиц изучаемого явления (**генеральная совокупность**), а при не сплошном – только часть этих единиц (**выборка**).

Пример

Исследуется средний рост студентов в России. Генеральная совокупность – рост абсолютно всех студентов всех вузов в конкретный момент времени, например на 1 сентября 2014г. Выборка – только для части студентов. Примеры выборок:

а) студенты только одного вуза

б) студенты Москвы, Самары и Томска города;

в) случайно выбранная 1000 студентов по всей стране.

Цель изучения выборочной совокупности (выборки) – получение информации не о ней самой, а о генеральной совокупности. Выборка – это неполная информация обо всех возможных представителях генеральной совокупности. В идеале, лучше исследовать генеральную совокупность, но на практике это не всегда удается (дорого, долго, физически невозможно).

Пример

Данные экзит-поллов на политических выборах – практический пример обследования выборочной совокупности (только часть избирателей опрашивается). Но это выборка позволяет предварительно судить о результатах голосования всех избирателей.

Поэтому обычно стремятся сделать так, чтобы выборка наилучшим образом представляла генеральную совокупность, т.е. была *репрезентативной* (*представительной*). Репрезентативная выборка – та, которая достаточно точно отражает свойства генеральной совокупности.

Пример

Для исследования среднего роста жителей города, было выбрано случайное здание и измерен рост находящихся в нем людей. Получили средний рост 2,14м. Причина – выбранное здание оказалось спортивным залом, в котором занимались баскетболисты. Такая выборка была *нерепрезентативной*.

Получить выборку можно разными способами. Самый простой – случайный выбор представителей генеральной совокупности.

При исследовании нас интересует не каждый объект целиком, а только какие-то его свойства (**признаки**).

Признаки могут быть:

1. *Количественные* (числовые) и *качественные* (нечисловые).

Пример

У объекта «человек» рассматриваются признаки: год рождения, пол, возраст, вес, рост, цвет глаз.

Количественные: год рождения, возраст, вес, рост.

Качественные: пол, цвет глаз.

1. *Моментные* и *интервальные*.

Пример

У объекта «фирма» рассматриваются признаки: общее число сотрудников; число сотрудников, принятых на работу; число уволенных сотрудников.

Моментные: общее число сотрудников (указать, на какое число).

Интервальные: число сотрудников, принятых на работу; число уволенных сотрудников (указать, за какой период времени – за месяц, за год, за день).

1. *Абсолютные* (в натуральных или денежных единицах) и *относительные* (в процентах или долях).

Пример

У объекта «семья» рассматриваются признаки: ежемесячный доход, количество членов семьи, доля затрат на продукты.

Абсолютные: наличие ежемесячный доход (руб.), количество членов семьи (чел.).

Относительные: доля затрат на продукты (% от общей суммы доходов).

1. *Случайные* и *неслучайные* (детерминированные), зависит от совокупности.

Пример

У объекта «студент» рассматриваются признаки: наличие зачетной книжки, курс, средний балл.

Случайные: курс, средний балл.

Неслучайные: наличие зачетной книжки.

Но если объект – «студент 1-го курса», то признак «курс» тоже будет неслучайным.

Статистика имеет дело со случайными количественными признаками, поэтому все данные в выборке должны быть числовыми. Обозначают выборки заглавными латинскими буквами *X*, *Y*, *Z*. Каждое значение в выборке называют *наблюдением* и обозначают *xi*, *yi*, *zi*, где *i* – номер наблюдения.

Пример

Исследуется *X* возраст студентов вуза (очной формы обучения) и *Y* средний балл. Всего опросили *n* = 100 студентов.

Исходная выборка имеет вид:

|  |  |  |  |
| --- | --- | --- | --- |
| *i* | *X* | *Все оценки* | *Y* |
| 1 | 18 | 4453 | 4 |
| 2 | 22 | 34433345444344555434444 | 3,9 |
| 3 | - | 5555555 | 5 |
| 4 | 17 | - | - |
| ... |  |  |  |
| 100 | 20 | 55443444444454 | 4,1 |

*x*1 = 18, *x*2 = 22, ... *y*1 = 4,0, *y*2 = 3,9, ...

Из нее получили сводную таблицу:

|  |  |  |
| --- | --- | --- |
| *X* | Количество студентов | Y |
| 18 и менее | 19 | 4,4 |
| 19 | 17 | 4,2 |
| 20 | 25 | 4,0 |
| 21 | 14 | 3,8 |
| 22 | 23 | 3,7 |
| 23 и более | 2 | 4,3 |
| *Всего* | *100* |  |

Можно заметить, что в среднем, в этом вузе с возрастом оценки ухудшаются.

Качественные признаки тоже «переводят» в числа.

Примеры

Пол: 1 - муж, 2 - жен.

Наличие машины: 0 - нет, 1 - есть.

Качество жилищных условий: в баллах от 0 до 10.

### Сведения из теории вероятностей

**Случайное событие** – это любое событие, о котором заранее неизвестно, произойдет оно или нет, или утверждение, о котором заранее неизвестно, истинно оно или ложно.

Примеры

Тело без опоры падает на поверхность земли (*неслучайное* событие).

При броске игральной кости на ней выпадет число 6 (*случайное*).

Завтра курс валюты снизится (*случайное*).

Тело человека состоит из клеток (*неслучайное*).

Прием лекарства поможет выздороветь (*случайное*).

При запросе к сайту он будет доступен (*случайное*).

Случайное событие определяется некоторой **вероятностью**. Вероятность события – интуитивно понятная характеристика, определяющая шанс на то, что событие произойдет.

*Формальное определение вероятности p* – отношение числа положительных исходов *m* (когда событие произошло) к общему числу исходов *n*.



Вероятность никогда не бывает отрицательной или больше 1.

**Случайная величина (СВ)** – число, значение которого заранее неизвестно. Для случайных величин задается вероятность появления конкретного значения или диапазона значений.

Случайные величины могут быть дискретными и непрерывными.

**Дискретная СВ** – множество отдельных значений (конечное или бесконечное). Каждому значению СВ можно сопоставить целое число. Сами значения СВ не обязательно целые.

**Непрерывная СВ** может принимать любое значение из некоторого диапазона. Количество возможных значений непрерывной СВ всегда бесконечно.

Примеры

Дискретные СВ: число попаданий в мишень, количество краж за неделю, число заказов за день, количество бракованных изделий в партии, оценка на экзамене.

Непрерывные СВ: температура воздуха на улице, масса буханки хлеба, яркость освещения, мощность сигнала радиостанции.

Многие дискретные по сути СВ для простоты считают непрерывными: цена (точность до копеек, центов), численность населения.

**Закон распределения** СВ характеризует вероятность появления каждого значения СВ. Сумма всех вероятностей должна быть равна 1, иначе какие-то пропустили.



***Для дискретных СВ*** закон распределения задается:

* таблично;

|  |  |  |  |
| --- | --- | --- | --- |
| *x1* | *x2* | … | *xn* |
| *p1* | *p2* | … | *pn* |

* графически;

Pn

P2

P1

x

x1

x2

…

xn

p

* формулой, например, если все значения равновероятны:



Примеры

Бросание игральной кости (СВ – выпавшее число).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *x* | 1 | 2 | 3 | 4 | 5 | 6 | Сумма |
| *p* | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1 |



Количество законченных высших образований в некоторой стране:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *x* | 0 | 1 | 2 | 3 | 4 | Сумма |
| *p* | 0,2 | 0,5 | 0,2 | 0,099 | 0,001 | 1 |

***Для непрерывных СВ*** нельзя задать вероятность каждого отдельного значения, ведь их бесконечно много. Поэтому для них задают вероятность попадания в некоторый *интервал значений*.

Пример

Вероятность температуры тела здорового человека от 36 до 37 значительно выше, чем вероятность температуры >40.

**Функция распределения (закон распределения, интегральная функция)** *F*(*x*) непрерывной СВ – вероятность того, что значение случайной величины *X* меньше числа *x*:



Задается формулой (часто очень сложной).

Примеры графиков функции распределения:

0

1

x

F(x)

Пример

На следующем графике:

F(2) = p(X < 2) = около 0

F(4) = p(X < 4) = 0,2

F(5) = p(X < 5) = 0,5

F(6) = p(X < 6) = 0,8

F(8) = p(X < 8) = почти 1

Отсюда:

p(X > 4) = 1 – p(X < 4) = 1 – 0,2 = 0,8

p(4 < X < 8) = p(X > 4) – p(X < 8) = 0,8 – 0,2 = 0,6

Но по функции распределения не очень удобно судить о том, какие значения более вероятны, а какие менее вероятны.

**Плотность распределения (дифференциальная функция**) *f*(*x*) – это производная интегральной функции распределения.



Сами значения плотности распределения не являются вероятностями и могут быть больше 1. Но сравнивая между собой два значения *f*(*x*), можно судить, какое из них более вероятное, а какое – менее вероятное. Чем выше график, тем больше вероятность.

Примеры графиков плотности функции:

0

x

f(x)

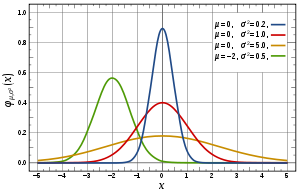
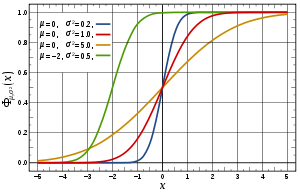
Для предыдущего примера:

Самое вероятное значение = 5. Значения < 2 и >8 почти никогда не появляются. С одинаковой вероятностью могут попасться числа > 5 и < 5.

Существует много разных законов распределения.

**Нормальный (Гауссов) закон распределения** считается одним из наиболее распространенных на практике. Зачастую, если закон распределения неизвестен, его принимают нормальным.

Функция и плотность нормального закона распределения имеют вид:



В предыдущих примерах закон распределения тоже был нормальным.

### Обобщающие числовые характеристики

Делать выводы по отдельным значениям выборки трудно, особенно, когда их много. Поэтому используют различные числовые характеристики, которые позволяют судить **в целом о выборке** (и генеральной совокупности).

**Мощность (объем) выборки** *n* – это число наблюдений в выборке. Чем оно больше, тем ближе выборка к генеральной совокупности. Реальные выборки обычно содержат сотни, тысячи наблюдений. Выборки в 30-50 наблюдений и менее считаются малыми.

**Выборочное среднее** *m* – показывает усредненное значение всех наблюдений. Одна из основных характеристик выборки. Вычисляется как сумма всех значений, деленная на их количество (мощность выборки):



Пример

X: 4; 35; 7; 8; 16

n = 5

*mX* = (4 + 35 + 7 + 8 + 16) / 5 = 14

**Медиана** *Me* – серединное значение упорядоченного ряда вариантов значений. Условный центр выборки.

Чтобы найти медиану, нужно упорядочить значения выборки и взять то, которое находится в середине. Например, если *n* = 15, то нужно взять 8-ое по значению наблюдение *x*8. Если *n* четно, то берется среднее между двумя значениями в центре. Например, если n = 20, то *Me* = (*x*10 + *x*11)/2.

Пример1

Выборка: 6;29; 5; 25; 11; 8; 5

n = 7

Упорядоченная выборка: 5; 5; 6; **8;** 11; 25; 29

*Me* = 8

Пример2

Выборка: 240; 104; 302; 300;160; 440

n = 6

Упорядоченная выборка: 104; 160; **240;** **300;** 302; 440

*Me* = (240 + 300) / 2 = 270

**Мода** *Mo* – наиболее вероятное, часто встречающееся значение выборки.

Пример

Модная вещь – это та, которую чаще всего можно увидеть на ком-нибудь еще.

Чтобы найти моду дискретной СВ, нужно посчитать, сколько раз встречается каждое значение в выборке. Мод может быть несколько, если какие-то значения встречаются одинаково часто. Если все значения встречаются одинаковое число раз (один раз), то мода не определена.

Пример

Выборка: 2; 2; 9; 9; 9; 10; 11; 12; 12; 13

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Значение | 2 | **9** | 10 | 11 | 12 | 13 |
| Частота | 2 | **3** | 1 | 1 | 2 | 1 |

*Mo* = 9

*Me* = 9,5

*m* = 89 / 10 = 8,9

Для непрерывных СВ мода определяется с помощью плотности распределения (самая высокая точка графика).

У ***однородной*** выборки (близкой к нормальному закону распределения) среднее значение, мода и медиана близки между собой. Медиана сильно отличается от среднего, когда в выборке присутствуют аномально большие или маленькие значения. В таком случае медиана предпочтительнее среднего для описания выборки.

Пример

Представьте, что в одной комнате собрали 9 бедняков и одного богача. У бедняков есть по 10 рублей, а у богача – миллион. Сколько у них в среднем денег?

Выборка: 10; 10; 10; 10; 10; 10; 10; 10; 10; 1 000 000

*n* = 10

*m* = 1 000 090 / 10 = 100 009

*Me* = 10

Другая группа показателей характеризуют ***разброс значений*** выборки – насколько далеко находятся ее значения от центра и друг от друга.

Пример

В двух регионах средняя заработная плата одинакова и составляет 1000 д.е. Однако в первом регионе 90% населения получают зарплату от 900 до 1200 д.е., а в другом 90% населения получают зарплату от 50 до 20 000 д.е. Очевидно, что хотя средние зарплаты в регионах равны, разброс у них разный.

**Размах вариации** *V* – разность между максимальным и минимальным значениями выборки. Показывает максимальную амплитуду вариации, в этот диапазон укладываются все значения выборки.

Пример

17; 5; 27; 12; 25; 14; 19

*V* = 27 – 5 = 22

Размах вариации – простая, но неудобная величина: два крайних значения могут быть очень далеки ото всех остальных.

Пример

–4; 1; 105; 106; 106; 107; 110; 112; 125; 126; 528

*V* = 532

Размах вариации получился очень большим, хотя большинство наблюдений лежит в пределах 100-130.

**Дисперсия (вариация)** *D* – условная величина из теории вероятностей, описывающая разброс значений относительно среднего.



Пример

15; 45; 60

*m* = (15 + 45 + 60) / 3 = 120 / 3 = 40

(*xi* – *m*): –25; 5; 20

(*xi* – *m*)2: 625; 25; 400

*D* = (625 + 25 + 400) /(3 – 1) = 1050 / 2 = 525

Есть еще одна, приближенная формула расчета дисперсии:





Дисперсия измеряется в квадратах исходных единиц измерения (руб.2, чел.2 и т.п.) и плохо поддается интерпретации. Поэтому оценивают **стандартное отклонение** *S* – корень из дисперсии. Смысл тот же, но единицы измерения обычные.

Пример (продолжение)

*D* = 525



**Коэффициент вариации** *KV* позволяет измерить вариацию в процентах.



Считается, что:

если *KV* ≤ 40% – вариация выборки небольшая;

40% <*KV* ≤ 60% – вариация средняя (умеренная);

*KV* > 60% – вариация сильная.

Пример (продолжение)

*m* = 40 *S* = 22,9

*KV* = 22,9 / 40 ∙ 100% = 57,25% (умеренная вариация)

**Квартили** похожи на медиану, но разделяют не половины, а четверти выборки. Четверть всех значений выборки меньше нижнего квартиля *Qн*. Четверть значений больше верхнего квартиля *Qв*. И половина значений выборки находится между квартилями.

Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *i* | *1* | *2* | *3* | ***4*** | *5* | *6* | *7* | *8* | *9* | *10* | ***11*** | *12* | *13* | *14* |
| *X* | 10 | 10 | 13 | **14** | 17 | 18 | 18 | 18 | 20 | 21 | **22** | 22 | 25 | 29 |
|  | 1 четверть | | |  | 2 четверть | | | 3 четверть | | |  | 4 четверть | | |

*Qн* = 14 *Qв* = 22

*Me* = 18

Половина всех значений находится между 14 и 22.

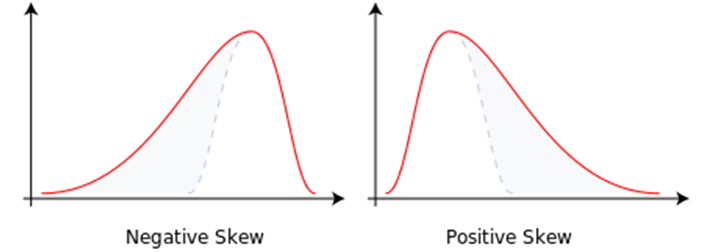
Еще два показателя характеризуют форму распределения в сравнении с нормальным законом.

**Коэффициент асимметрии** KA показывает, с какой стороны от среднего больше значений.

Если он положительный (KA > 0), то больше значений справа, т.е. чаще встречаются значения больше среднего. Медиана больше среднего.

Если отрицательный (KA < 0) – то больше слева, т.е. чаще встречаются значения меньше среднего. Медиана меньше среднего.

Если равен 0, то распределение симметрично относительно среднего (слева и справа от среднего одинаковое количество значений). Медиана равна среднему.



Отрицательная (левая) асимметрия

Положительная (правая) асимметрия

Примеры

В двух регионах средняя заработная плата одинакова и составляет 1000 д.е. Однако в первом регионе коэффициент асимметрии равен –0,5, а во втором +3,4. Это означает, что в первом регионе тех, кто получает зарплату меньше средней и больше средней почти поровну, но тех, кто получает «маленькую» зарплату все же чуть-чуть больше, чем тех, кто получает «большую». А во втором регионе гораздо больше тех, кто получает «большую» заплату (больше средней).

В примере с 9 бедняками и 1 богачом коэффициент асимметрии отрицательный (из 10 наблюдений 9 меньше среднего и только 1 больше, медиана гораздо меньше среднего).

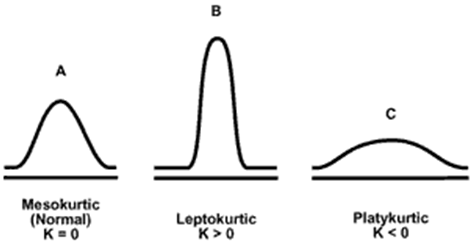
|  |  |  |  |
| --- | --- | --- | --- |
| KA > 0 | правая | *Me > m* | много «богатых», «больших» |
| KA = 0 | симметрично | *Me = m* | «богатых» и «бедных» поровну |
| KA < 0 | левая | *Me < m* | много «бедных», «маленьких» |

**Коэффициент эксцесса** E характеризует остроту пика распределения.

Если E > 0, то пик острый (островершинное), т.е. большинство значений выборки близко к моде, а одно или несколько находятся где-то далеко.

Если E < 0, то значения выборки равномерно распределяются по всему диапазону (плосковершинное), т.е. много и «больших», и «маленьких», и «средних» значений.

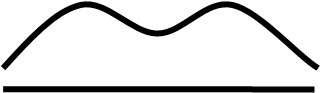
Если E = 0, то распределение близко к нормальному, т.е. чем дальше от среднего, тем меньше таких значений, убывание плавное.



E = 0

E > 0

E < 0

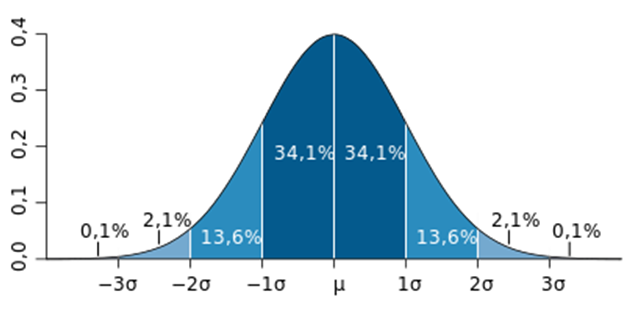


E << 0

Примеры

В четырех регионах средняя заработная плата одинакова и составляет 1000 д.е. Стандартное отклонение тоже одинаковое и равно 200 д.е. Но в первом регионе E = 0,01, во втором E = –1,2, в третьем E = –5,1, а в четвертом E = 3,2.

**Правило трех *S***(только для нормального распределения, E = 0 и KA = 0):



–3*S*

–2*S*

–1*S*

1*S*

2*S*

3*S*

*m*

400

600

800

1200

1400

1600

1000

Это означает, что в первом регионе большинство людей все же получает зарплату, близкую к средней (68% от 800 до 1200 д.е.), а тех, кто получает очень большую или очень маленькую меньше, причем, чем дальше зарплата от среднего, тем меньше людей с такой зарплатой (по 13,5% получают от 600 до 800 или от 1200 до 1400, по 2% от 400 до 600 и по 1400 до 1600, а тех, кто получает меньше 400 или больше 1600 – менее 1%).

В других регионах правило 3*S* не работает.

Во втором регионе почти поровну людей и с большой, и со средней, и с маленькой зарплатой. Правило 3*S* не работает.

В третьем регионе многие получают большую или маленькую зарплату, а людей со средней зарплатой мало.

В четвертом регионе большинство (80-90%) получают среднюю зарплату, и лишь немногие получают очень большую или очень маленькую (но они есть!).

## Задание

Имеется статистическая выборка с наблюдениями некоторого показателя *X*. Необходимо сформировать описательную статистику показателя, рассчитав следующие обобщающие характеристики выборки:

* + мощность выборки *n*;
  + выборочное среднее *m*;
  + медиану *Me*;
  + моду *Mo*;
  + размах вариации *V*;
  + дисперсию *D*;
  + стандартное отклонение *S*;
  + коэффициент вариации *Kv*;
  + верхний и нижний квартили *Q*в и *Q*н.

Охарактеризовать выборку на основе полученных значений. Ответить на вопросы:

* 1. Какова мощность выборки?
  2. Чему равно среднее значение? наиболее ожидаемое значение?
  3. Каков разброс значений (слабый, средний, сильный)?
  4. Какому интервалу принадлежат значения выборки? половина значений?
  5. Какова форма распределения (симметричное или асимметричное, островершинное или плосковершинное)?
  6. По этим результатам можно ли считать выборку однородной?

## Пример

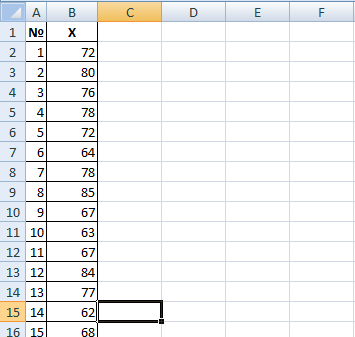
Все расчеты показаны на примере выборки со значениями:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 72 | 85 | 68 | 68 | 83 |
| 80 | 67 | 67 | 71 | 72 |
| 76 | 63 | 78 | 64 | 66 |
| 78 | 67 | 67 | 75 | 75 |
| 72 | 84 | 69 | 64 | 77 |
| 64 | 77 | 66 | 64 | 84 |
| 78 | 62 | 66 | 69 | 77 |

## Подготовка исходных данных

Создайте новый файл Excel и сохраните под именем вида «КАСД группа ФИО.xlsx». Переименуйте «Лист 1» в «Л.р.1».

Скопируйте данные для своего варианта на лист «Л.р.1», начиная со второй строки и столбца B. Добавьте в первый столбец нумерацию наблюдений. Укажите названия столбцов.



## Указания к выполнению работы

Требуемые в задании расчеты можно выполнить несколькими способами. Мы рассмотрим каждый из них и сравним результаты.

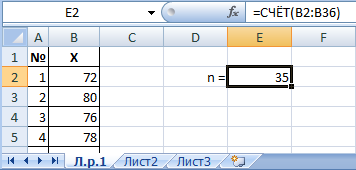
Примечание В работе необходимо выполнить все варианты расчетов.

Все результаты расчетов будем располагать в столбик справа от выборки.

1. Определим количество значений в выборке, т.е. ее **мощность**. Фактически, пронумеровав наблюдения, мы уже узнали их количество. Но есть и формула, по которой можно узнать количество ячеек в диапазоне.

Сначала определим мощность выборки *n*.

|  |  |
| --- | --- |
|  | Функция СЧЁТ(*диапазон*) вычисляет количество непустых ячеек в диапазоне. Диапазон можно ввести вручную или выделить мышкой во время набора формулы. |

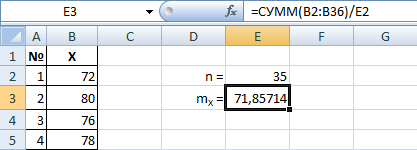


В примере выборка содержит 35 наблюдений.

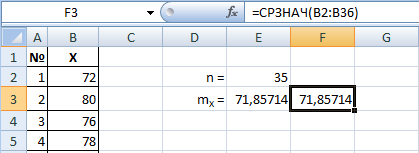
1. Вычислим **выборочное среднее** *m*. Сначала сделаем это в соответствии с формулой:



|  |  |
| --- | --- |
|  | Для расчета суммы большого количества значений используется функция =СУММ(*диапазон*). |



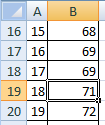
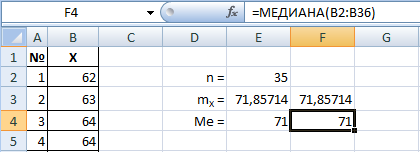
Существует и встроенная формула СРЗНАЧ(*диапазон*). Она должна дать такой же результат.



1. Найдем **медиану**. Это значение должно находиться в середине упорядоченной выборки. В данном примере середина – это 18-ое значение. Необходимо лишь упорядочить выборку.

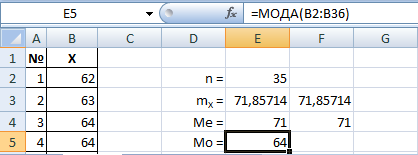
|  |  |
| --- | --- |
|  | Выделите всю выборку (без заголовка) и на ленте, на вкладке «Главная» нажмите «Сортировка и фильтр» – «Сортировка от минимального к максимальному», или кликните по выборке правой кнопкой и найдите этот пункт во всплывающем меню.    Примечание. В некоторых случаях порядок значений в исходной выборке может быть важен. Тогда лучше сделать копию исходной выборки и отсортировать эту копию. В данной работе это не требуется. |

В примере 18-ое наблюдение равно 71. Просто скопируем его в столбец с результатами. Встроенная формула для вычисления – МЕДИАНА(*диапазон*) – должна вернуть то же значение.

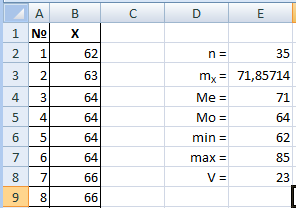
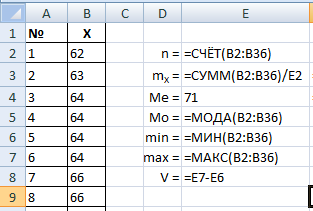
В данном примере медиана мало отличается от среднего значения, что свидетельствует об однородности выборки (нет очень больших или очень маленьких значений, или их поровну).

1. Найдем **моду**. Чтобы найти моду вручную, нужно посчитать, сколько раз встречается каждое значение в выборке. То, которое встречается чаще всех, и есть мода. Но, поскольку это достаточно долго, даже по упорядоченной выборке, сразу воспользуемся формулой МОДА(*диапазон*).



Т.е. наиболее вероятное значение в данной выборке – 64, оно появляется чаще других. Мода немного меньше среднего значения.

1. Найдем **размах вариации**. Это разница между максимальным и минимальным значениями выборки. Поскольку мы уже отсортировали выборку, это первое и последнее значения, но мы воспользуемся формулами МИН(*диапазон*) и МАКС(*диапазон*).



Таким образом, амплитуда колебаний рассматриваемого показателя равна 23.

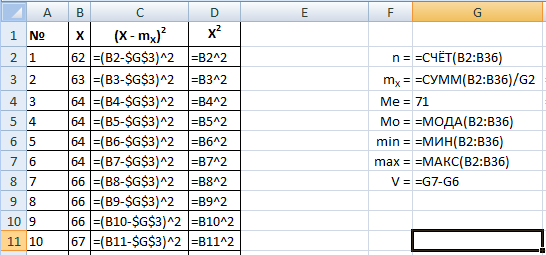
1. Вычислим **дисперсию** по двум формулам:

В этих формулах нам понадобятся промежуточные расчеты:  и . Добавьте два столбца справа от X.

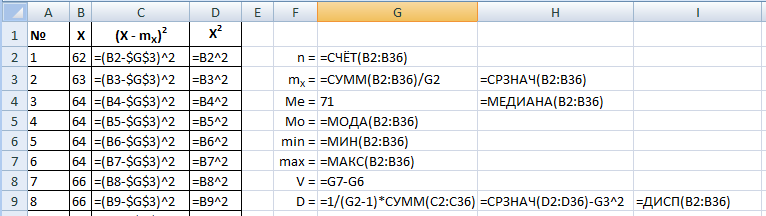
|  |  |
| --- | --- |
|  | Правый клик по названию столбца – «Вставить» (второй пункт, без рисунка). |

Добавьте названия столбцов и заполните их формулами с помощью «растягивания».

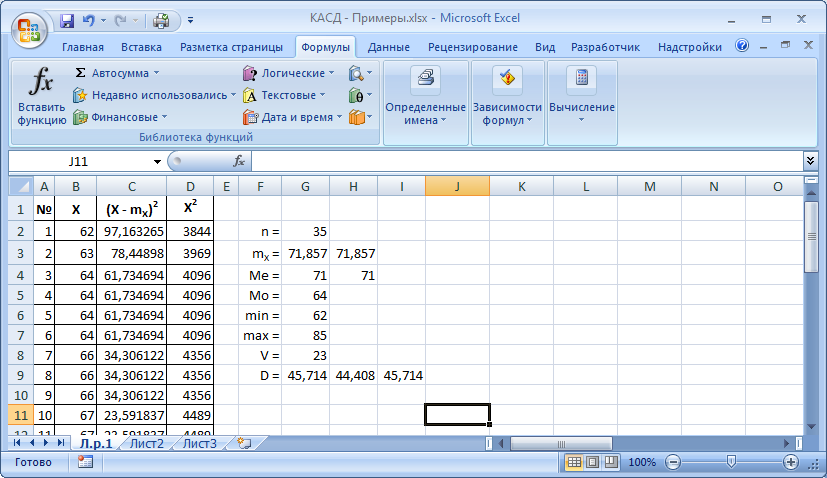


|  |  |
| --- | --- |
|  | Когда вы растягиваете формулы, адреса ячеек в них смещаются. Но mX всегда находится в одной и той же ячейке! Поэтому, чтобы она не смещалась, в адрес ячейки добавляется значок $. Его можно вписать вручную или поставить клавишей F4 на клавиатуре. |
|  | Если в результате вместо чисел в ячейках стоят ##### – это не ошибка, просто число не умещается в ячейку. Увеличьте ширину столбцов. |

Добавляем обе формулы дисперсии и встроенную формулу – ДИСП(*диапазон*).



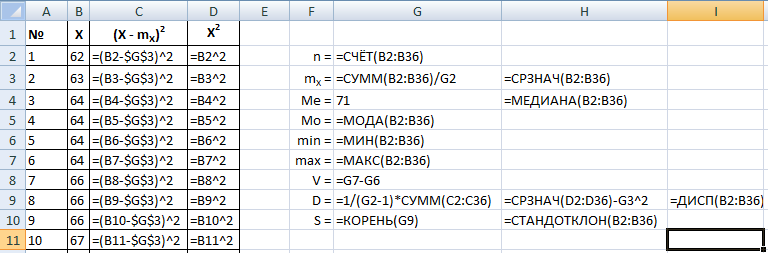
Результат:



Как видим, результат по второй формуле отличается, потому что она приближенная.

Делать выводы о размахе вариации по значению дисперсии трудно, поскольку она возводится в квадрат.

1. Вычислим **стандартное отклонение** двумя способами: извлечем корень из дисперсии (дисперсию берем по первой формуле) и воспользуемся функцией СТАНДОТКЛОН(*диапазон*):

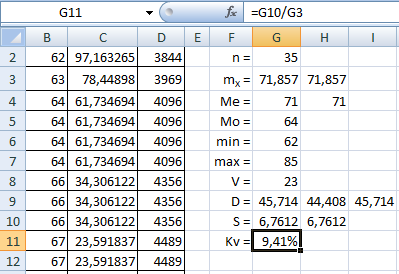




Таким образом, стандартное отклонение равно 6,76, что не очень много по сравнению со средним (71,9).

1. Вычислим **коэффициент вариации** как:



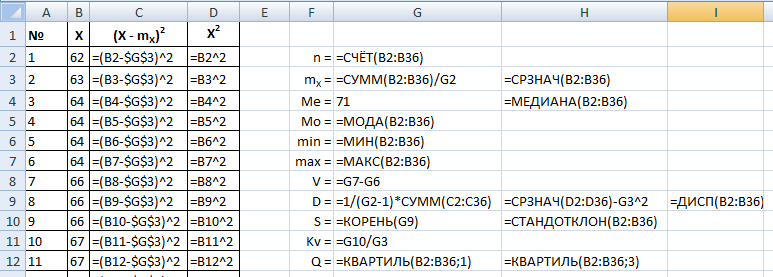


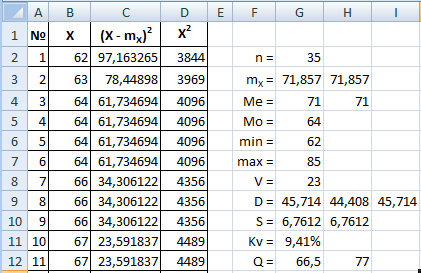
|  |  |
| --- | --- |
|  | Умножать на 100% в Excel не обязательно.  Чтобы показать результат в процентах: правый клик по ячейке – «Формат ячейки» – вкладка «Число» – «Процентный» – «ОК». |

Коэффициент вариации составляет менее 10%, разброс значений исследуемой величины небольшой.

1. Квартили вычислим с помощью встроенной формулы КВАРТИЛЬ(*диапазон*;*часть*).

|  |  |
| --- | --- |
|  | Параметр «часть» показывает, какой квартиль мы хотим получить: 1 - нижний (25%), 2 - медиана (50%); 3 - верхний (75%). |





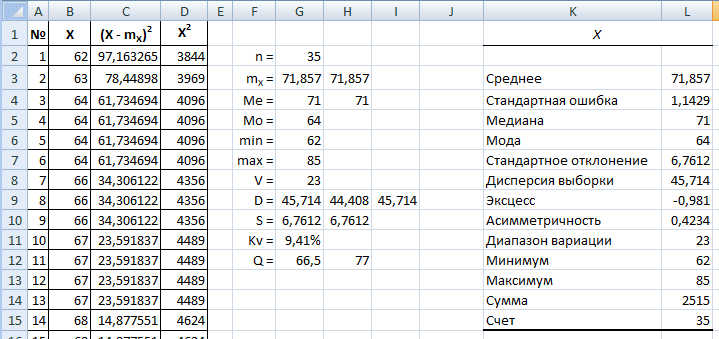
Нижний квартиль равен 66,5, верхний – 77, т.е. половина значений выборки лежит в диапазоне от 66,5 до 77.

1. Те же результаты можно получить с использованием пакета «Анализ данных» (как его включить, см. начало методички).

Кроме того, он выдает значения **асимметрии** и **эксцесса**.

|  |  |
| --- | --- |
|  | На вкладке «Данные» нажмите «Анализ данных». В окне выберите «Описательная статистика»    Укажите диапазон исходной выборки вместе с заголовком “X”, отметьте галочкой «Итоговая статистика» и «Метки в первой строке». Выходной диапазон укажите справа от ваших расчетов. Нажмите ОК. |

Результат:



|  |  |
| --- | --- |
|  | Из-за ошибки Excel вместо «Диапазон вариации» может быть написано ССЫЛКА#. Исправьте текст вручную. |

Коэффициент эксцесса немного меньше нуля, выборка слегка пологая. Коэффициент асимметрии близок к нулю, распределение можно считать симметричным.

1. Сделайте вывод по работе.

**Вывод:**

Мощность выборки составляет 36 значений. Среднее значение выборки равно 71,9, медиана – 71. Наиболее ожидаемое значение (мода) – 64.

Коэффициент вариации равен 9,4%, разброс значений слабый.

Значения выборки находятся в диапазоне от 62 до 85. При этом половина значений попадает в интервал от 66,5 до 77.

Коэффициент асимметрии равен 0,42 и близок к нулю. Выборку можно считать симметричной (с очень слабой положительной (левой) асимметрией).

Коэффициент эксцесса равен –0,98, выборка плосковершинная, но не очень далека от нормальной.

Среднее значение, медиана и мода близки по значениям, выборка симметрична и обладает небольшим отрицательным эксцессом. В целом, выборку можно охарактеризовать как близкую к однородной.

# Лабораторная работа №2 Гистограмма распределения

Гистограмма распределения показывает, насколько часто в выборке встречаются те или иные значения X (аналог плотности распределения). Каждый столбец гистограммы показывает частоту попадания значения выборки в интервал значений – чем выше столбец, тем вероятнее соответствующие значения показателя.

Для построения гистограммы необходимо:

1. Разбить диапазон значений показателя на интервалы.
2. Посчитать число значений, попавших в каждый интервал – *частоты*.
3. Вычислить высоту столбцов гистограммы в соответствии с правилом нормировки: сумма площадей столбцов гистограммы равна 1 (без учета зазоров между столбцами).
4. Построить график.

Рассмотрим каждый шаг подробнее.

### Выделение интервалов

Весь диапазон значений показателя необходимо разбить на части – ***интервалы***, или «***карманы***». Не допускается наличие «дырок» между интервалами и перекрытие интервалов. Чаще всего интервалы имеют одинаковую ширину.

Интервал задается своей нижней и верхней границей: *aj* и *bj*, *j* = 1,2,3... – номер интервала.

Проще всего выделить *интервалы* *с «круглыми» границами*, например, шириной 10:

1. [20;30] *a*1 = 20 *b*1 = 30
2. (30;40] *a*2 = 30 *b*2 = 40
3. (40;50] *a*3 = 40 *b*3 = 50
4. (50;60] *a*4 = 50 *b*4 = 60
5. (60;70] *a*5 = 60 *b*5 = 70

Обратите внимание, что конец одного интервала – это начало следующего интервала, т.е. *bj* = *aj*+1: *b*1 = *a*2, *b*2 = *a*3 и т.д.

Но в этом случае не понятно, какой именно ширины брать интервалы: 10, 50, 100? Если взять слишком маленькие, то в каждый будет попадать мало значений (1-2 или вообще 0), а если слишком большие, то наоборот – слишком много.

Иногда разбиение выполняют в соответствии *с общепринятой традицией*. Например, во всем мире принято распределять население по возрастным группам по 5 лет, как показано в таблице ниже.

**Распределение населения РФ по возрастным группам**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Тыс. человек | | |
| 2001 | 2006 | 2007 |
| **Все  население** | **146304** | **142754** | **142221** |
| в том числе в возрасте, лет**:** |  |  |  |
| 0-4 | 6367 | 7037 | 7223 |
| 5-9 | 7762 | 6418 | 6376 |
| 10-14 | 11789 | 7790 | 7283 |
| 15-19 | 12322 | 11825 | 11088 |
| 20-24 | 11106 | 12405 | 12671 |
| 25-29 | 10451 | 11049 | 11165 |
| 30-34 | 9620 | 10295 | 10442 |
| 35-39 | 11333 | 9417 | 9459 |
| 40-44 | 12651 | 10949 | 10368 |
| 45-49 | 11434 | 12054 | 12067 |
| 50-54 | 9409 | 10645 | 10804 |
| 55-59 | 4995 | 8590 | 8985 |
| 60-64 | 8906 | 4407 | 4336 |
| 65-69 | 5903 | 7609 | 7458 |
| 70 и более | 12256 | 12264 | 12496 |

*Источник: gks.ru*

Третий вариант – выяснить, какое *количество* *интервалов* является оптимальным исходя из мощности выборки (чем больше наблюдений, тем больше интервалов). Для этого используется формула Стерджесса:



Здесь *m* – количество интервалов (не путать со средним значением *mx*).

Примеры

*n* = 10 *m* = 1 + 3,322 lg 10 ≈ 4,3 *m* = 4

*n* = 20 *m* = 1 + 3,322 lg 20 ≈ 5,3 *m* = 5

*n* = 30 *m* = 1 + 3,322 lg 30 ≈ 5,9 *m* = 6

*n* = 40 *m* = 1 + 3,322 lg 40 ≈ 6,3 *m* = 6

*n* = 50 *m* = 1 + 3,322 lg 50 ≈ 6,6 *m* = 7

*n* = 100 *m* = 1 + 3,322 lg 100 ≈ 7,6 *m* = 8

*n* = 1000 *m* = 1 + 3,322 lg 1000 ≈ 10,97 *m* = 11

*n* = 10000 *m* = 1 + 3,322 lg 10000 ≈ 14,8 *m* = 15

Затем весь диапазон значений от минимального до максимального разбивается *на* *равные* *части*. Ширина интервала *w* определяется по формуле:



Первый интервал начинается в *X*min, а последний (с номером *m*) заканчивается в *X*max.

*a*1 = *X*min *b*1 = *a*1 + *w*

*a*2 = b1 *b*2 = *a*2 + *w*

…

*am* = *bm*–1 *bm* = *am* + *w* = *X*max

Пример

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *i* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *X* | 9 | 10 | 10 | 12 | 14 | 15 | 18 | 19 | 21 | 22 | 28 | 37 | 43 | 44 | 47 |

*n* = 15

*m* = 1 + 3,322 lg 15 = 4,9 = 5

*X*min = 9

*X*max = 47

*w* = (47 – 9) / 5 = 38 / 5 = 7,6

*a*1 = *X*min = 9 *b*1 = 9 + 7,6 = 16,6

*a*2 = 16,6 *b*2 = 16,6 + 7,6 = 24,2

*a*3 = 24,2 *b*3 = 24,2 + 7,6 = 31,8

*a*4 = 31,8 *b*4 = 31,8 + 7,6 = 39,4

*a*5 = 39,4 *b*5 = 39,4 + 7,6 = 47 = *X*max

Интервалы в виде таблицы:

|  |  |  |
| --- | --- | --- |
| *j* | *a* | *b* |
| 1 | 9 | 16,6 |
| 2 | 16,6 | 24,2 |
| 3 | 24,2 | 31,8 |
| 4 | 31,8 | 39,4 |
| 5 | 39,4 | 47 |

### Подсчет попаданий в интервалы

**Частота (частость)** *nj* попадания в каждый *j*-тый интервал считается явно: сколько из всех значений находятся внутри интервала, т.е.  и .

Сумма всех частот должна быть равна мощности выборки, иначе какие-то значения «потерялись»:



В примере с населением частота попадания в интервал 20-24 года в 2007г. составила 12671 тыс. чел.

Пример (продолжение)

|  |  |  |  |
| --- | --- | --- | --- |
| *j* | *a* | *b* | *n* |
| 1 | 9 | 16,6 | 6 |
| 2 | 16,6 | 24,2 | 4 |
| 3 | 24,2 | 31,8 | 1 |
| 4 | 31,8 | 39,4 | 1 |
| 5 | 39,4 | 47 | 3 |

Сумма *nj*: 6 + 4 + 1 + 1 + 3 = 15 = *n*

**Высота столбца** гистограммы вычисляется через частоту и ширину интервалов:



При этом выполняется *правило нормировки* – площадь графика плотности распределения равна 1.

Доказательство

Площадь каждого столбца (прямоугольника) – это высота умножить на ширину:



Сумма площадей:



Пример (продолжение)

*h*1 = 6 / (15 ∙ 7,6) = 6 / 114 ≈ 0,05263

*h*2 = 4 / (15 ∙ 7,6) = 4 / 114 ≈ 0,03509

*h*3 = 1 / (15 ∙ 7,6) = 1 / 114 ≈ 0,00877

*h*4 = *h*3 ≈ 0,00877

*h*5 = 3 / (15 ∙ 7,6) = 3 / 114 ≈ 0,02632

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *j* | *a* | *b* | *n* | *h* |
| 1 | 9 | 16,6 | 6 | 0,05263 |
| 2 | 16,6 | 24,2 | 4 | 0,03509 |
| 3 | 24,2 | 31,8 | 1 | 0,00877 |
| 4 | 31,8 | 39,4 | 1 | 0,00877 |
| 5 | 39,4 | 47 | 3 | 0,02632 |

### Анализ графика гистограммы

Для визуального анализа гистограмму наносят на график и сравнивают с известными законами распределения.

Чаще всего гистограмма близка к **нормальному закону распределения**. Большинство значений сосредоточены в центре, к краям высоты столбцов плавно снижается. Пример такой гистограммы в сравнении с теоретическим нормальным законом (красная линия):

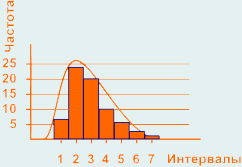
Другой возможный вариант – **равномерное распределение**, когда все значения примерно поровну распределены по столбцам гистограммы, нет заметных максимумов и минимумов:

В идеале, высоты столбцов должны быть почти одинаковыми, но в реальности так получается редко.

Существуют и другие варианты гистограммы распределения:

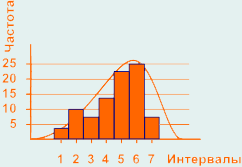
**Положительная асимметрия**

Гистограмма смещена влево, длинный правый «хвост».



**Отрицательная асимметрия**

Гистограмма смещена вправо, длинный левый «хвост».



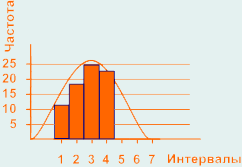
**Бимодальность**



Гистограмма отображает два совмещенных процесса. Такая ситуация может произойти если в исследуемой совокупности смешались два разных по природе типа объектов.

Например, собрана общая статистика спортивных результатов для мужчин и женщин, хотя они сильно различаются.

**Гистограмма усечена**



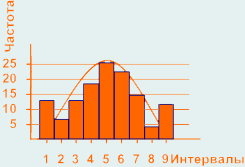
Распределение не является нормальным т.к. нет постепенного снижения частоты результатов измерений от центра к границам. Такой вид гистограммы возникает, если чересчур малые и большие наблюдения невозможны или потеряны.

**Гистограмма не имеет центра**



Достаточно редкая ситуация, когда столбцы в центре диаграммы близки к 0. Может быть случаем бимодальности. Такая гистограмма имеет отрицательный эксцесс.

**«Тяжелые хвосты»**



По краям гистограммы наблюдаются высокие столбцы. Означает сильное «расслоение» показателя, когда много как очень больших, так и очень маленьких значений. Коэффициент эксцесса отрицательный.

Пример (продолжение)

Для нашего примера гистограмма имеет вид:

Она не соответствует ни нормальному, ни равномерному закону распределения: столбцы по краям гистограммы выше, чем посередине. Присутствуют «тяжелые хвосты», гистограмма смещена влево (положительная асимметрия), коэффициент эксцесса отрицательный.

### Кумулятивная кривая

**Кумулята** или **кумулятивная кривая,** в отличие от гистограммы строится по накопленным частотам.

*Накопленная частота* – это сумма частот обычной гистограммы, от первого интервала до текущего:







...



Другими словами, это количество наблюдений, меньше верхней границы интервала (<*bj*). Последняя накопленная частота всегда равна *n*, т.к. это число попаданий во все интервалы, т.е. все наблюдения.

Высота столбца кумуляты рассчитывается как накопленная частота, деленная на общее количество наблюдений:





На графике на горизонтальной оси также отмечают интервалы значений, а на вертикальной оси – накопленные частоты. Кумулята является аналогом закона распределения, и поэтому возрастает от 0 до 1.

Визуально кумуляту труднее анализировать, но с вертикальной оси можно снимать значения вероятностей: высота столбца кумуляты равна вероятности того, что случайная величина меньше верхней границы соответствующего интервала.



Пример (продолжение)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *j* | *a* | *b* | *n* | *h* | *k* | *H* |
| 1 | 9 | 16,6 | 6 | 0,05263 | 6 | 0,400 |
| 2 | 16,6 | 24,2 | 4 | 0,03509 | 10 | 0,667 |
| 3 | 24,2 | 31,8 | 1 | 0,00877 | 11 | 0,733 |
| 4 | 31,8 | 39,4 | 1 | 0,00877 | 12 | 0,800 |
| 5 | 39,4 | 47 | 3 | 0,02632 | 15 | 1,000 |

График кумуляты:

Таким образом, с вероятностью 0,4 (40%) значения исследуемого показателя будут меньше 16,6; с вероятностью 0,667 (66,7%) – меньше 24,2 и т.д.

## Задание

Имеется статистическая выборка с наблюдениями показателя *X –* ежемесячного дохода лиц определенной профессии, тыс. д.е. Необходимо построить гистограмму распределения:

1. Определить мощность выборки, минимальное, максимальное, среднее значения и стандартное отклонение.
2. Определить число интервалов по формуле Стерджесса и их ширину.
3. Определить границы и середины интервалов («карманов»).
4. Вычислить частоты попаданий в интервалы двумя способами: через «Анализ данных» и через формулы.
5. Построить график гистограммы.
6. Построить график кумуляты.
7. Сравнить полученную гистограмму с нормальным законом распределения (визуально).
8. Сравнить полученную гистограмму с равномерным законом распределения (визуально).

Сделать выводы о предположительном законе распределения (нормальный, равномерный, другой). Если это возможно, определить, в каком диапазоне дохода совершение преступления а) наиболее вероятно; б) наименее вероятно. Какова вероятность, что представитель данной профессии имеет доход менее 10 тыс. у.е?

## Пример

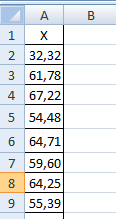
Все расчеты показаны на примере выборки:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 32,32 | 43,69 | 55,62 | 33,28 | 59,92 |
| 61,78 | 39,15 | 39,94 | 51,83 | 55,95 |
| 67,22 | 52,41 | 40,67 | 55,92 | 42,29 |
| 54,48 | 40,87 | 57,46 | 49,89 | 59,24 |
| 64,71 | 45,79 | 30,92 | 34,2 | 56,32 |
| 59,6 | 51,18 | 46,25 | 37,62 | 63,64 |
| 64,25 | 46,88 | 43,76 | 48,74 | 48,74 |
| 55,39 | 56,84 | 40,61 | 51,05 | 56,89 |
| 50,88 | 59,99 | 51,16 | 43,99 | 61,4 |
| 60,91 | 56,38 | 63,22 | 46,57 | 42,99 |
| 49,05 | 58,19 | 44,96 | 53,84 | 37,59 |
| 53,29 | 56,08 | 68,03 | 58,3 | 42,58 |
| 57,67 | 54,15 | 50,71 | 56,01 | 46,73 |
| 31,14 | 59,29 | 57,91 | 44,76 | 28,71 |
| 50,4 | 44,7 | 41,5 | 46,79 | 23,02 |
| 53,17 | 44,9 | 54,87 | 65,09 | 26,01 |
| 55,89 | 62,02 | 59,35 | 36,57 | 32,53 |

## Подготовка исходных данных

Работа выполняется в том же файле, что и первая, но на новом листе. Переименуйте «Лист 2» в «Л.р.2».

Скопируйте данные для своего варианта на лист «Л.р.2», начиная со второй строки и столбца A. Укажите название столбца (X).



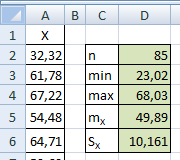
Нумеровать значения в данной работе необязательно.

## Указания к выполнению работы

В описании работы для наглядности все ячейки с расчетами закрашены зеленым. Вам это делать необязательно.

1. Сначала необходимо рассчитать основные числовые характеристики: мощность выборки, минимальное и максимальное значения, среднее и стандартное отклонение. Они потребуются ходе выполнения работы и для выводов.

Числовые характеристики рассчитываются по формулам из лабораторной работы №1 (любым способом).



В примере выборка содержит 85 наблюдений, от 23,02 до 68,03. Средний доход составляет 49,89, стандартное отклонение – 10,16.

1. Начнем построение гистограммы – вычислим количество интервалов и их ширину.

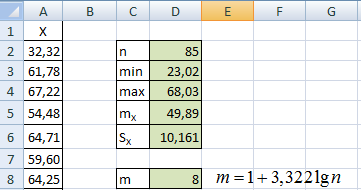
Число интервалов определим по формуле Стерджесса:



Для удобства, формулу можно скопировать на лист Excel.

Полученное число необходимо округлить до большего значения.

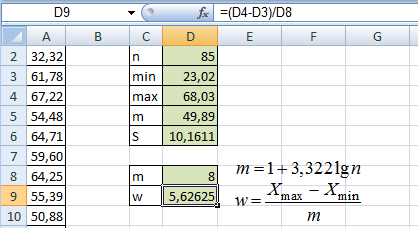
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Функции:   |  |  | | --- | --- | | ОКРУГЛВВЕРХ(*число*; *знаков\_после\_запятой*) | Округлить число в сторону большего до указанного знаков после запятой (0 – до целых). | | LOG10(*число*) | Десятичный логарифм числа (lg). | |



Чтобы определить ширину интервала, воспользуемся формулой:



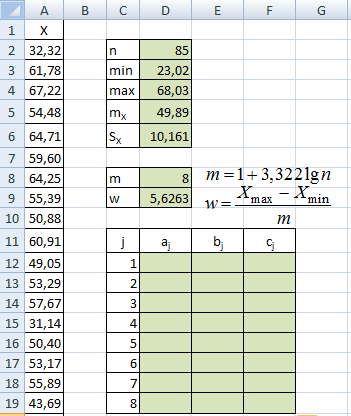
Ширину интервала округлять **не** нужно.



Таким образом, необходимо исходные данные распределить по 8 интервалам шириной 5,626 тыс.д.е.

1. Определим границы интервалов («карманов») и подготовим таблицу для будущих расчетов. Число строк в таблице равно числу интервалов, столбцы мы будем добавлять по мере необходимости (всего их будет 9, пока подготовим 3).

|  |  |
| --- | --- |
|  | Таблицу необходимо просто расчертить, **не надо** вставлять ее через команду на Ленте. |



*aj* – нижняя граница;

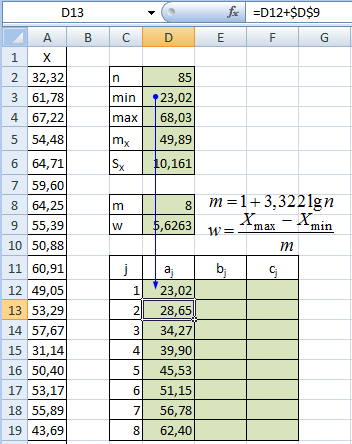
*bj* – верхняя граница;

*cj* – середина интервала (понадобится позже).

Нижняя граница первого интервала – это минимальное значение выборки (в примере 23,02). Нижняя граница каждого следующего интервала – это предыдущий интервал плюс ширина интервала:

.

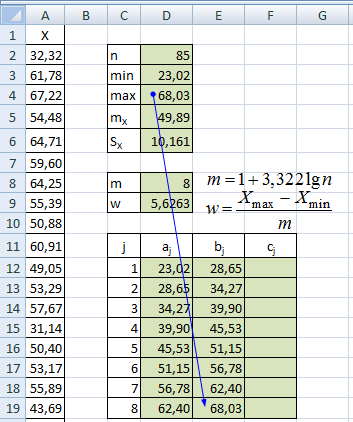
|  |  |
| --- | --- |
|  | Не забудьте зафиксировать ячейку с **w** в формуле, прежде чем растянуть ее! |



Верхняя граница интервала совпадает с нижней границей следующего интервала и рассчитывается по такой же формуле.

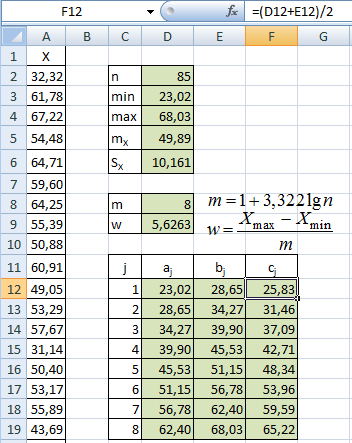
.

Верхняя граница последнего интервала должна быть в точности равна максимальному значению выборки.



Середина интервала рассчитывается как среднее между его границами.





Таким образом, исходную выборку мы разбили на 8 числовых интервалов:

1. [23,02;28,65]
2. (28,65;34,27]
3. (34,27;39,9]
4. (39,9;45,53]
5. (45,53;51,15]
6. (51,15;56,78]
7. (56,78;62,4]
8. (62,4;68,03]

Квадратная скобка означает, что число на границе входит в интервал, круглая – что не входит. Например, 28,65 входит в первый интервал, но не входит во второй. На практике иногда делают наоборот, в зависимости от удобства.

Обратите внимание, в первый интервал входят обе границы.

1. Далее необходимо определить частоты попадания в интервал, т.е. сколько наблюдений из исходной выборки попадает в каждый интервал.

|  |  |
| --- | --- |
|  | В Excel это можно сделать двумя способами: через «Анализ данных» и через формулы. |

Способ 1.

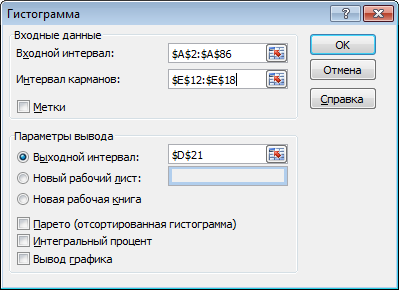
Убедитесь, что на вкладке «Данные» есть кнопка «Анализ данных» (как ее включить см. в л.р.1).

Нажмите «Анализ данных» – «Гистограмма» – «ОК».

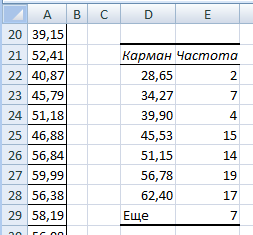
В поле «Входной интервал» укажите все значения X.

В поле «Интервал карманов» – верхние границы интервалов, кроме последней.

В «Параметрах вывода» выберите «Выходной интервал» и укажите любую ячейку под таблицей. В качестве «карманов» указаны верхние границы интервалов, кроме последнего.



В результате получим таблицу вида:

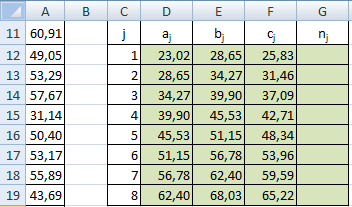


Карман «Еще» соответствует последнему интервалу и считает все значения, не вошедшие в остальные интервалы.

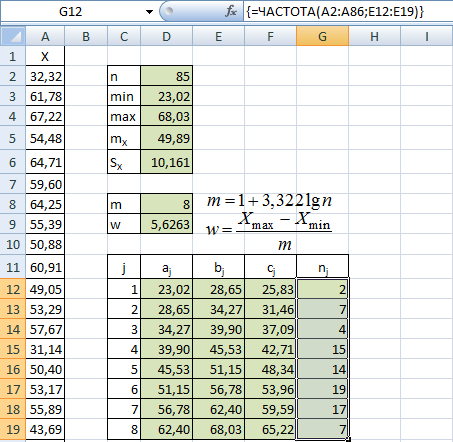
Способ 2.

Способ более трудоемкий, но позволяет автоматически пересчитывать гистограмму при изменении или добавлении новых значений.

Значения по способу 2 мы внесем в следующий столбец нашей таблицы, обозначив его *nj*.

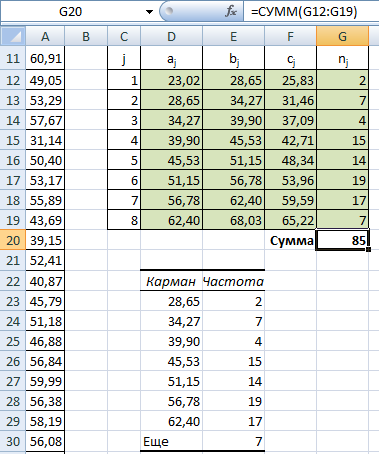


|  |  |
| --- | --- |
|  | Воспользуемся формулой  ЧАСТОТА(*исходные\_данные*;*интервалы*)  *исходные\_данные* – это все значения X  *интервалы* – верхние границы интервалов  Внимание!  Это одна из специальных формул Excel – формула массива. Результат расчета по такой формуле занимает несколько ячеек. Она **не** растягивается, и **не** вводится отдельно в каждую ячейку, а вводится следующим образом.   1. Выделите весь диапазон ячеек, в которых будут значения *nj* (в примере G12:G19). 2. Нажмите на клавиатуре клавишу F2. Включится режим ввода формулы, но выделены будут по-прежнему все ячейки. 3. Введите формулу для расчета. В примере =ЧАСТОТА(A2:A86;E12:E19). **Не** нажимайте Enter. 4. Нажмите сочетание Ctrl+Shift+Enter.   В результате формула заполнит все ячейки, в строке формул она отображается в фигурных скобках {=ЧАСТОТА(A2:A86;E12:E19)}  Чтобы отредактировать такую формулу нужно повторить ту же последовательность действий: выделить все ячейки → F2 → ввод формулы → Ctrl+Shift+Enter.  При попытке изменить отдельную ячейку внутри массива, Excel выдаст ошибку. |



Проверьте себя – сумма *nj* должна быть равна *n*. Добавьте под таблицей строку «Сумма» и вычислите в ней сумму *nj*.

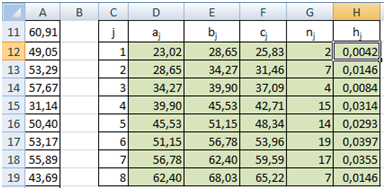
Результат для обоих способов должен совпадать.



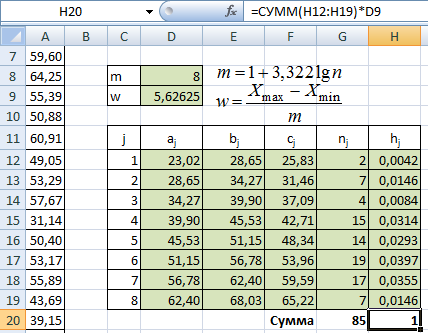
1. Построим график гистограммы. Для этого потребуются некоторые дополнительные расчеты, т.к. частоты – это еще не гистограмма. Для нее должно выполняться правило нормировки – сумма площадей всех столбцов гистограммы должна быть равна 1. Исходя из этого правила, высота столбца гистограммы вычисляется по формуле:



Добавим соответствующие расчеты.



Проверьте себя! Сумма *hj*, умноженная на *w*, должна быть равна 1.



Построим график гистограммы.

|  |  |
| --- | --- |
|  | Выделите значения *hj* вместе с заголовком и на вкладке «Вставка» выберите «Диаграмма» – «Гистограмма» – «Гистограмма с группировкой» (самая первая).  Удалите заголовок диаграммы.  Чтобы под столбцами отображались не номера интервалов, а их границы, нужно кликнуть по гистограмме правой кнопкой – «Выбрать данные...», нажать кнопку «Изменить» справа, в разделе «Подписи горизонтальной оси» и выделить значения *aj* и *bj* (без заголовков). |

В результате получим график вида:

1. Построим кумулятивную кривую по сумме накопленных частот. Для кумуляты также отдельно рассчитываются накопленные частоты (*ki*) и высоты столбцов (*Hj*).

Частоты можно вычислить следующим образом.

Частота первого интервала одинакова для гистограммы и кумуляты:



Остальные частоты рассчитываются как предыдущая частота кумуляты плюс текущая частота гистограммы:



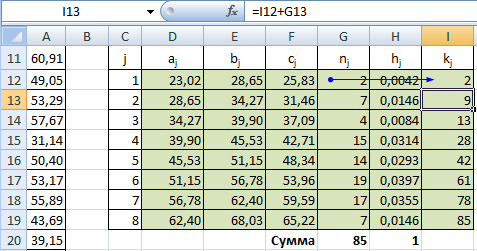
Например, для 2 и 3 строки получится:





и т.д.

Добавим в таблицу столбец *kj*:

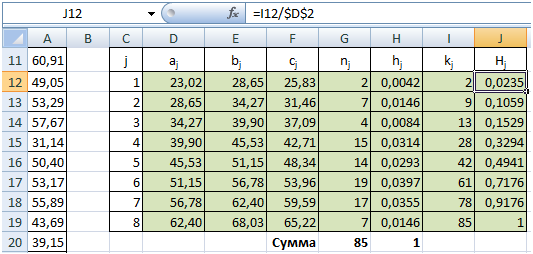


Обратите внимание, в последнем интервале должна получиться мощность выборки *n*.

Теперь можно вычислить высоты столбцов кумуляты:



Для кумуляты правило нормировки не действует, поэтому частоты делятся просто на *n*. В последней строке должна получиться 1.



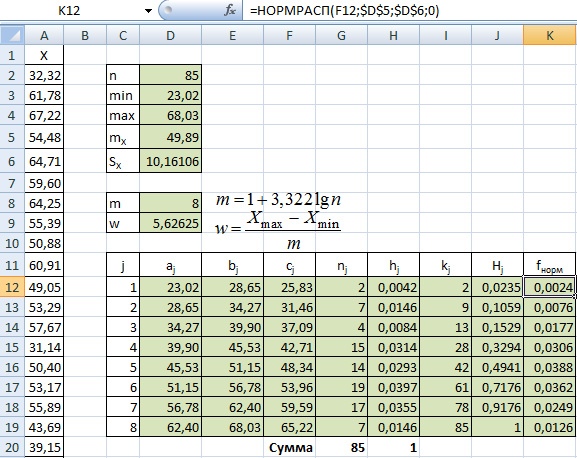
По значениям кумуляты можно сказать, например, что только в 2,35% случаев сумма ежемесячного заработка в данной профессии составляет менее 28,65 тыс.д.е. А почти в половине случаев (49,41%) сумма заработка составляет до 51,15 тыс.д.е.

График кумуляты строится аналогично обычной гистограмме. Можно скопировать уже созданную диаграмму и изменить исходные данные.

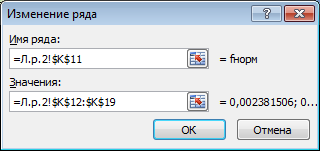
1. Сравним полученную в п.5 гистограмму с нормальным законом распределения, который очень часто встречается на практике.

Добавим в таблицу еще один столбец для расчета плотности нормального распределения.

|  |  |
| --- | --- |
|  | Воспользуемся формулой НОРМРАСП(*x*;*mx*;*Sx*;*тип*)  *x* – для какого значения рассчитываем (середины интервалов *cj*)  *mx*, *Sx* – среднее значение и стандартное отклонение  тип = 0, показывает, что нужно рассчитать именно плотность, а не интегральный закон (кумуляту) |

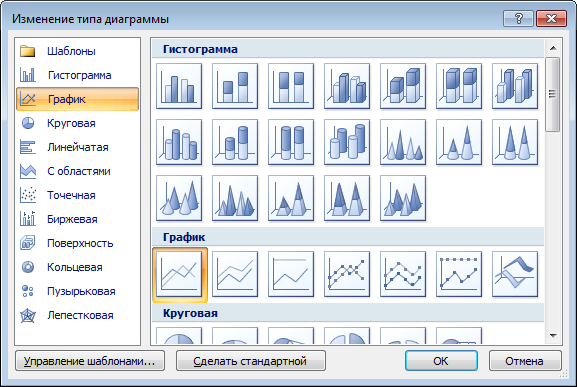


Добавим его на график с гистограммой. Кликните по ней правой кнопкой – «Выбрать данные», нажмите «Добавить» и укажите столбец с fнорм.



В результате диаграмма примет вид:

Но плотность распределения принято изображать не столбцами, а гладкой линией. Для этого выделите новые столбцы (красные), кликните по любому из них правой кнопкой – «Изменить тип диаграммы для ряда». Выберите тип – «График», нажмите «ОК». Столбцы превратятся в линию.



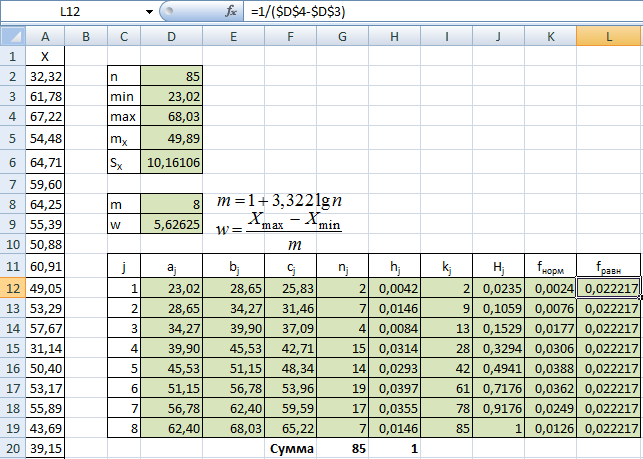
Еще раз кликните по полученной линии правой кнопкой – «Формат ряда данных» – «Тип линии», поставить галочку «сглаженная линия».В результате получим:

**Вывод**: В целом, гистограмма распределения не соответствует нормальному закону: она не симметрична (справа столбцы высокие, слева низкие, мода не в центре распределения). Имеются «провалы» в 3 и 5 интервалах, а 2 и 7 столбцы, наоборот, значительно выше нормальной плотности распределения.

1. Сравним гистограмму с еще одним распространенным законом распределения – равномерным. У равномерного закона отсутствует мода, нет более или менее вероятных интервалов – все допустимые значения равновероятны.

Значения плотности равномерного распределения во всех точках одинаковы и рассчитываются по формуле:





Добавим равномерный закон на график гистограммы аналогично нормальному.

**Вывод**: Гистограмма распределения далека от равномерного закона. Заметно, что высоты столбцов сильно отличаются друг от друга. В правой части гистограммы столбцы выше, т.е. более высокий заработок вероятнее, чем низкий.

**Общий вывод:**

В рассматриваемой профессии наиболее вероятна сумма ежемесячного заработка от 50 до 60 тыс.д.е. (округленно). Маловероятна сумма заработка более 70 и 20 тыс.д.е.

Закон распределения нельзя считать ни нормальным, ни равномерным. Он асимметричный, смещен вправо, с модой в диапазоне от 51,15 до 56,78, что немного больше среднего.

Вероятность получения заработка менее 10 тыс. д.е. в данной профессии близка к 0.

# Лабораторная работа №3 Статистические тесты. Исключение случайных выбросов

## Теоретические сведения

### Проверка статистических гипотез

**Статистическая гипотеза** – это любое предположение о выборке или нескольких выборках, касающееся их статистических свойств (распределение, среднее значение, стандартное отклонение и т.д.)

Примеры гипотез:

1. Средняя урожайность зерна в Самарской области равна 14ц/га.
2. Большинство ИТ-специалистов имеют доход выше среднего.
3. Уровень преступности (среднее число преступлений) в городе N-ске выше в Заречном районе (чем в остальных – Центральном, Заводском и Ленинском).
4. Стандартное отклонение случайной величины *X* меньше 100.
5. Данный судья чаще принимает решение в пользу истца, чем ответчика.
6. Число запросов к сайту за сутки имеет нормальное распределение.

Гипотезы обозначаются буквой H (hypothesis) с цифрой. Основная гипотеза обозначается цифрой 0 и называется **нулевой (H0)**.

Кроме того, обычно описывают другие возможные ситуации, кроме основной. Такие гипотезы называются **альтернативными** и обозначаются H1, H2, H3... Чаще всего выдвигается одна альтернативная гипотеза.

Для предыдущих примеров:

1. H0: mX = 14

H1: mX ≠ 14

Другой вариант:

H0: mX = 14

H1: mX > 14

H2: mX < 14

1. H0: mИТ > mз/п

H1: mИТ ≤ mз/п

1. H0: mЗареч. > mЦентр. и mЗареч. > mЗавод. и mЗареч. > mЛен.

H1: mЗареч. ≤ mЦентр. или mЗареч. ≤ mЗавод. или mЗареч. ≤ mЛен.

1. H0: SX ≥ 100

H1: SX < 100

1. H0: p (1) > p(0)

где 1 означает решение в пользу истца, 0 – решение в пользу ответчика

H1: p (1) ≤ p(0)

1. H0: F(x) ~ N(m, S) (нормальный закон распределения)

H1: F(x) – закон, отличный от нормального

**Решение** заключается в том, чтобы одну из гипотез *принять* (т.е. решить, что она верная), а все остальные *отклонить* (т.е. решить, что они неверные).

При принятии любого решения можно ошибиться. Различают **два рода** **ошибок**:

|  |  |  |
| --- | --- | --- |
| *Гипотеза* | *Решение* | |
| *Принять* H0 | *Отклонить* H0 |
| H0 *верна* | Нет ошибки | Ошибка 1-ого рода  («ложная тревога») |
| H0 *неверна* | Ошибка 2-ого рода  («пропуск цели») | Нет ошибки |

Причем обе ошибки совершаются с некоторой вероятностью, и чем ниже вероятность одной ошибки, тем выше вероятность другой.

Пример 1

H0: подозреваемый невиновен

H1: подозреваемый виновен

Ошибка 1 рода: невиновного осудили

Ошибка 2 рода: преступника оправдали

В каждом конкретном деле есть вероятность совершить одну из этих ошибок. Полностью избежать ошибок 1 рода можно только одним способом – никого никогда не осуждать. Но при этом будет множество ошибок 2 рода.

И наоборот, если осуждать абсолютно всех, то мы никогда не допустим ошибку 2 рода (не отпустим преступника), но будет много ошибок первого рода.

Иными словами, невозможно полностью избежать обеих ошибок. Если стараемся избегать одной, то увеличивается шанс допустить другую.

Пример 2

H0: партия товара бракованная, ее необходимо утилизировать

H1: партия товара качественная, ее можно выпустить в продажу

Для проверки гипотез из всей партии отбирают несколько опытных образцов и проверяют их качество. Как правило, считается допустимым, чтобы из всей партии 1-2% изделий были с браком.

Ошибка 1 рода: все проверенные изделия были с браком и партию товара признали бракованной. Но в реальности весь остальной товар был надлежащего качества.

Ошибка 2 рода: все проверенные изделия были качественными и партию товара выпустили в продажу. Но в реальности оказалось, что половина товара в ней с браком. Фирме пришлось выплатить компенсацию покупателям.

В идеале, необходимо обследовать как можно больше изделий, чтобы убедиться в качестве всей партии. Но на практике это не всегда возможно.

В статистике принято задавать допустимую, приемлемую вероятность ошибки 1 рода и при этом стараться уменьшить вероятность ошибки второго рода.

*α* – вероятность ошибки 1 рода (**уровень значимости**).

(1 – *α*) – уровень *достоверности* результатов

Статистические тесты составляются таким образом, чтобы заданного α (например, 0,01, 0,05, 0,1) минимизировать вероятность ошибки второго рода.

Пример 2 (продолжение)

Пусть α = 0,01, т.е. с вероятностью 1% мы допускаем, что качественный товар будет признан бракованным. В дальнейшем тестирование строится по таким правилам, чтобы гарантированно не выпустить на рынок бракованный товар. Как именно это сделать – уже детали.

Как выбрать *α*? Четкого правила нет. На практике обычно *α* задают от 0,1 до 0,001. Причем, чем больше значений в выборке, тем меньше *α*. Например, для n = 30, лучше взять *α =* 0,05-0,1, а для n = 1000 лучше выбрать *α =*0,01-0,001.

**Статистический тест** – это правило, по которому гипотезы принимаются или отвергаются.

В ходе статистического теста вычисляется некое число – **статистический критерий K** (буква может быть любой). По значению критерия и принимается решение, путем сравнения с критическим значением Kкр. Критическое значение берется из специальных таблиц или вычисляется по специальной формуле.

**Критическая область** – значения критерия, при которых H0 отвергается, а H1 принимается. Все остальные значения называются *допустимыми*.

***Критическая область соответствует альтернативной гипотезе.***

Критическая область может быть левосторонней, правосторонней и двусторонней (примеры см. ниже).

**Левосторонняя критическая область: K < Kкр.**

K

Kкр

критическая область

область допустимых значений

**Правосторонняя критическая область: K > Kкр.**

K

Kкр

критическая область

область допустимых значений

**Двусторонняя критическая область: |K| > Kкр.**

K

Kкр

критическая область

область допустимых значений

–Kкр

0

Например: если K > 1,72, то отвергнуть гипотезу H0 (правосторонняя критическая область).

### Проверка гипотез о величине среднего значения Т-критерий Стьюдента

Применяется, только если случайная величина имеет нормальный закон распределения.

Необходимо сравнить среднее значение *mx* с некоторым числом *a*.

Пример 1 *Конвейер*

Проверяется точность работы автомата на конвейере, фасующего конфеты по пакетам. В каждом пакете должно быть 250г конфет. По результатам 20 контрольных замеров средняя масса конфет в пакете получилась равной 248,5г. Достаточно ли точно автомат фасует конфеты?

В данном случае «вредным» является отклонение от 250г как в большую, так и в меньшую сторону.

Формулировка гипотез:

Н0: mx = 250 – фасовка правильная

Н1: mx ≠ 250 – фасовка неправильная, отклонение от 250г слишком большое

Критическая область двухсторонняя.

Если бы по результатам замеров мы получили среднюю массу 150г или 400г, то было бы очевидно, что автомат работает неправильно. Если бы получили ровно 250г – то очевидно, что фасовка правильная. Но мы получили 248,5г. Отклонение от нормы в 1,5г – это много или мало? Можно ли считать это нарушением нормальной работы? Статистический тест как раз и позволит это проверить.

Пример 2 *Энергосбережение*

Производители ноутбуков утверждают, что разработали принципиально новую технологию энергосбережения, которая позволяет повысить срок работы батареи. При старой системе энергосбережения средний срок работы ноутбука от батареи составлял 2,5ч. Необходимо проверить, увеличился ли срок службы при новой технологии.

В данном примере нас интересует отклонение от среднего только в бо́льшую сторону. Если срок работы не изменился или уменьшился, то технология не приносит пользы.

Гипотезы:

Н0: mx ≤ 2,5 – срок работы не изменился, или уменьшился

Н1: mx > 2,5 – срок работы увеличился

Интересующая нас ситуация (срок работы батареи увеличился) вынесена в альтернативную гипотезу, т.к. она соответствует критической области, существенным отклонениям от среднего.

Если по результатам замеров мы получим среднюю длительность работы от батареи 2ч или 6ч, то ответ очевиден. А если срок работы получился равным 2,55ч? 3ч? Ответ зависит и от того, сколько ноутбуков мы проверили, и насколько разными получились значения.

Пример 3 *Снижение веса*

Человеку с весом 120кг была прописана диета и гимнастика для похудения. Спустя три дня проведено контрольное взвешивание. Т.к. весы могут давать некоторую погрешность, взвешивание провели трижды и взяли средний вес по трем замерам. Он получился равным 117,8кг. Можно ли сказать, что человек похудел?

Иными словами, является ли *критическим* снижение веса со 120кг до 117,8кг? Можно ли с уверенностью утверждать, что вес стал меньше? Если бы по результатам взвешивания вес стал равен 121кг, то проверка гипотезы вообще не была бы нужна – очевидно, что снижения веса не наблюдается.

Гипотезы:

Н0: mx ≥ 120 – вес больше или равен 120кг

Н1: mx < 120 – вес меньше 120кг

**Схема проверки гипотез**

|  |  |  |  |
| --- | --- | --- | --- |
| *Гипотезы* | H0: mx = *a*  H1: mx ≠ *a* | H0: mx ≤ *a*  H1: mx > *a* | H0: mx ≥ *a*  H1: mx < *a* |
| *Критерий* |  | | |
| *Критическое значение* |  |  |  |
| *Критическая область* | Двусторонняя  |T| ≥ *tкр.*  => mx ≠ *a* | Правосторонняя  T > *tкр.*  => mx > *a* | Левосторонняя  T < *tкр.*  => mx < *a* |

 – t-статистика Стьюдента, берется из специальных таблиц или рассчитывается на компьютере.

*d* = *n* – 1 – число степеней свободы.

Фрагмент таблицы t-статистики Стьюдента:

|  |  |  |  |
| --- | --- | --- | --- |
| *d* | *α* | | |
| **0,01** | **0,05** | **0,1** |
| **10** | 2,764 | 1,812 | 1,372 |
| **15** | 2,602 | 1,753 | 1,341 |
| **20** | 2,528 | 1,725 | 1,325 |
| **30** | 2,457 | 1,697 | 1,310 |
| **40** | 2,423 | 1,684 | 1,303 |
| **50** | 2,403 | 1,676 | 1,299 |
| **100** | 2,364 | 1,660 | 1,290 |
| **1000** | 2,330 | 1,646 | 1,282 |

Например, мы выбрали уровень значимости *α* = 0,01 и число наблюдений в выборке n = 51. Тогда *d* = *n* – 1 = 50. По таблице находим *t*0,01;50 = 2,403.

Обратите внимание на формулу *T*: чем больше наблюдений в выборке *n*, тем больше *T*, тем надежнее результат. А чем больше разброс значений *S*, тем *T* меньше и результат менее надежный.

Для проверки статистической гипотезы, кроме среднего значения, необходимо знать объем выборки *n*, разброс значений *S* и выбрать уровень значимости *α*.

Пример

Проверить предположение о том, что *mx* меньше 50 с уровнем значимости 0,1.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***№*** | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** | ***9*** | ***10*** | ***Сумма*** |
| ***X*** | 45 | 32 | 69 | 23 | 75 | 21 | 32 | 41 | 87 | 49 | **474** |
| ***X2*** | 2025 | 1024 | 4761 | 529 | 5625 | 441 | 1024 | 1681 | 7569 | 2401 | **27080** |

1. Гипотеза:

H0: *mx* ≥ 50

H1: *mx* < 50

1. Вычисляем T-критерий

*mx* = 474 / 10 = 47,4

*S2x*= *m*(*x*2) – (*mx*)2= 27080 / 10 – 47,42 = 461,24

*Sx =* 21,47



1. Критическое значение (левосторонняя критическая область)

*α* = 0,1

*d* = *n* – 1 = 10 – 1 = 9

*t*кр. = –*t*0,1;9 = –1,372

1. Вывод:

*T* > *t*кр. – нулевую гипотезу следует принять.

По имеющейся статистике нет оснований считать среднее значение *X* меньше 50 (хотя расчетное значение 47,4 < 50).

### Случайные выбросы

**Случайные выбросы** – аномально большие или маленькие значения, которые сильно отличаются от основной массы выборки.

**Причины** возникновения выбросов:

* ошибки при сборе, записи и хранении исходных данных (неисправность приборов, человеческий фактор, повреждение файла с данными);
* преднамеренное искажение данных;
* присутствуют необычные, нетипичные, исключительные, редкие случаи;
* выборка неоднородна, в нее попали значения из другой генеральной совокупности.

Иногда случайные выбросы можно установить, исходя из природы изучаемого явления (температура тела человека не может быть равна 100 градусов, зарплата не может быть отрицательной и т.д.) или путем простого визуального сравнения значений в выборке.

Примеры

1. Рост студентов, см:

175; 148; 154; 168; **1711**; 182; 205

Причина: опечатка.

1. Вес посетителей курсов по снижению веса:

128, 94, 71, **40**, 210, 88

Возможные причины: ребенок или человек, которому не нужно снижать вес.

1. Доход сотрудников предприятия за последний месяц, тыс. руб:

22,5; 24,3; 18,1; 28,5; 23,7; **1014**; 13,2; 10,5.

Причина: человек выиграл в лотерею.

1. Температура в разных точках жилой комнаты:

19; 21; 22, 20; **65**.

Причина: замерили температуру возле батареи.

1. Число вызовов на пульт экстренной службы в течение часа:

Причина: произошла авария с большим числом пострадавших.

Случайный выброс – не обязательно ошибка, это может быть просто нетипичная ситуация.

Но в любом случае случайные выбросы следует исключать из выборки, иначе они могут сильно исказить статистические расчеты и привести к неверным выводам. Либо использовать специальные методы, которые работают даже при случайных выбросах (*робастные методы*).

Пример

Использование медианы вместо среднего значения, см. пример из 1 л.р. с бедняками и богачом.

Существует несколько статистических тестов для поиска случайных выбросов в разных типах выборок.

Один из самых простых тестов – **тест Граббса** (Grubbs) для случайных величин с нормальным законом распределения.

|  |  |
| --- | --- |
| *Гипотеза* | H0: в выборке нет случайных выбросов  H1: в выборке есть как минимум один выброс |
| *Критерий* |  |
| *Критическое значение* |  |
| *Критическая область* |  |

Если тест показал, что в выборке есть случайные выбросы, из нее удаляется одно значение с наибольшим отклонением от среднего  (либо самое большое число, либо самое маленькое). После этого процедура повторяется, пока не будут удалены все выбросы.

Для расчета необходимо знать среднее значение *mx*, стандартное отклонение *Sx*, объем выборки *n* и задать уровень значимости *α*. Причем *mx*, *Sx*, *n* будут изменяться после каждого удаленного значения.

Пример

Исключить случайные выбросы с достоверностью 95%.

*X*: 5; 10; 12; 12; 13; 15; 17; 18; 19; 21; 22; 75

*n* = 12

Числа 5 и 75 сильно отличаются от остальных. Можно предположить, что это случайные выбросы.

*mX* =19,9

*SX* = 18,0

|*X – mX*|: 14,9; 9,9; 7,9; 7,9; 6,9; 4,9; 2,9; 1,9; 0,9; 1,1; 2,1; 55,1

*G* = 3,057

*α* = 0,05

tα/n; n-2 = 3,277

*Gкр.* = 2,285

*G* > *Gкр.*

Следовательно, в выборке есть случайный выброс. Это значение с наибольшим отклонением от среднего. У значения 75 отклонение равно 55,1. Это и есть случайный выброс. Его необходимо удалить из выборки.

После удаления случайного выброса все расчеты необходимо повторить:

*X*: 5; 10; 12; 12; 13; 15; 17; 18; 19; 21; 22

*n* = 12

*mX* =14,9

*SX* = 5,1

|*X – mX*|: 9,9; 4,9; 2,9; 2,9; 1,9; 0,1; 2,1; 3,1; 4,1; 6,1; 7,1

*G* = 1,940

tα/n; n-2 = 3,310

*Gкр.* = 2,234

*G* < *Gкр.*

Следовательно, больше в выборке нет случайных выбросов. Число 5 не является выбросом.

## Задание

Имеется статистическая выборка с наблюдениями показателя *Q –* объем выпуска продукции фирмой, в тыс. д.е. Необходимо:

1. Найти медиану, среднее значение, стандартное отклонение, эксцесс, асимметрию.
2. Проверить гипотезу о том, что средняя сумма компенсации равна/больше/меньше *q* (см. в варианте задания) с уровнем значимости α.
3. Найти случайные выбросы графическим способом.
4. Удалить из выборки случайные выбросы с помощью теста Граббса.
5. Повторить расчеты из пункта 1-2 по выборке, очищенной от выбросов. Изменились ли результаты?

## Пример

Все расчеты показаны на примере выборки:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *i* | *Qi* | *i* | *Qi* | *i* | *Qi* | *i* | *Qi* |
| 1 | 170 | 9 | 25 | 17 | 140 | 25 | 180 |
| 2 | 160 | 10 | 140 | 18 | 180 | 26 | 250 |
| 3 | 170 | 11 | 150 | 19 | 160 | 27 | 120 |
| 4 | 130 | 12 | 80 | 20 | 100 | 28 | 350 |
| 5 | 90 | 13 | 750 | 21 | 130 | 29 | 120 |
| 6 | 120 | 14 | 200 | 22 | 150 | 30 | 180 |
| 7 | 160 | 15 | 160 | 23 | 165 | 31 | 150 |
| 8 | 420 | 16 | 220 | 24 | 120 | 32 | 170 |

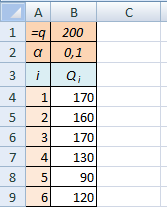
*q* = 200, проверка на равенство: m = q

*α* = 0,1

## Подготовка исходных данных

Работа выполняется в том же файле, что и предыдущие, но на новом листе. Переименуйте «Лист 3» в «Л.р.3».

Скопируйте данные для своего варианта на лист «Л.р.3».

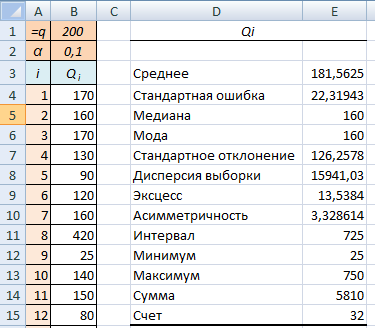


Обратите внимание, сама выборка начинается со строки 4. В строке 1 находится значение *q*, в строке 2 – *α*.

В примере перед *q* стоит =, значит в п.2 задания необходимо проверить среднее значение на равенство: H0: m = q (если >*q*, то проверить H1: m > q; если <*q*, то проверить H1: m < q).

## Указания к выполнению работы

1. Найдем требуемые для расчетов значения с помощью «Анализа данных».



**Вывод:**

Таким образом, среднее число рассмотренных дел составляет около 181,6 шт. При этом медиана равна 160 шт. Разница между ними довольно существенная. Стандартное отклонение равно 126,26 шт., разброс значительный (больше 60%).

Коэффициент асимметрии больше 0, значит, большая часть значений больше среднего, т.е. чаще случается так, что число рассмотренных дел велико.

Коэффициент эксцесса гораздо больше 0, значит, большая часть значений близко к среднему и лишь одно или несколько далеки от среднего (очень много либо очень мало дел).

1. В примере необходимо проверить гипотезу следующего вида:

H0: m = 200

H1: m ≠ 200

Для этого рассчитаем T-статистику:



и критическое значение t-статистики:

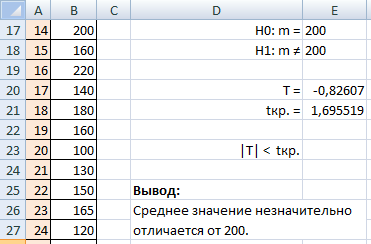
 (при проверке на равенство)

или

 (при проверке на больше/меньше)

|  |  |
| --- | --- |
|  | Для вычисления t-статистики используется функция **СТЬЮДРАСПОБР(α; n-1)**  Обратите внимание, при проверке на равенство в исходной формуле стоит , а в формулу Excel нужно подставлять просто значение α, **без** деления попалам.  Но, если в вашем варианте используется проверка на больше/меньше и в исходной формуле стоит  (без деления пополам), то в формулу Excel нужно вписать СТЬЮДРАСПОБР(α\*2; n-1)  Например, если в вашем варианте α = 0,05 и проверка на равенство, то формула будет:  СТЬЮДРАСПОБР(0,05; ...)  Если α =0,05 и проверка на больше (m > q), то  СТЬЮДРАСПОБР(0,05\*2; ...)  или (0,05\*2 = 0,1)  СТЬЮДРАСПОБР(0,1; ...) |

Затем необходимо сравнить *T* и *tкр*..



**Вывод**: Таким образом, предположение о том, что среднее число рассмотренных за месяц дел равно 200 шт., подтверждается статистическими данными.

1. Покажем исходные значения на графике.

По горизонтальной оси откладываем номера наблюдений *i*, по вертикальной – наблюдения *Qi*. Удалите с диаграммы лишние надписи, добавьте вертикальные линии сетки.

|  |  |
| --- | --- |
|  | Тип диаграммы – точечная.    Добавление линий сетки: выделить диаграмму, на ленте вкладка «Макет», «Сетка» – Вертикальные линии сетки – Основные линии сетки |

В результате получим график вида:

На данном графике по вертикальной оси откладываются значения, а по горизонтальной – номера наблюдений.

Можно заметить, что большинство значений находится в районе 100-200, и лишь несколько чисел значительно больше (около 400, 750, 350). Есть и одно очень маленькое число (меньше 50). Эти числа, предположительно, и будут случайными выбросами.

Обведите на графике прямоугольной рамкой значения, которые кажутся «нормальными» (без выбросов).

Найдите случайные выбросы в исходной выборке и выделите их красным цветом.

**Вывод:**

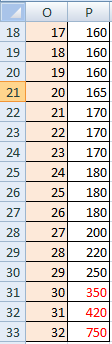
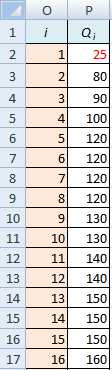
О наличии случайных выбросов косвенно свидетельствует большое значение коэффициента эксцесса (более 10). Положительная асимметрия показывает, что присутствует среди выбросов чаще встречаются выбросы с большими значениями.

По графику можно заметить, что большинство значений находится в районе 100-250 шт. Два значения находятся очень далеко от остальных (420 и 750). Значение 350 тоже заметно больше остальных, а значение 25 – гораздо меньше.

Таким образом, случайными выбросами будем считать: 25, 350, 420, 750.

1. В этом пункте задания нам потребуется удалять значения из выборки, поэтому сделаем ее копию, с которой и будем работать.

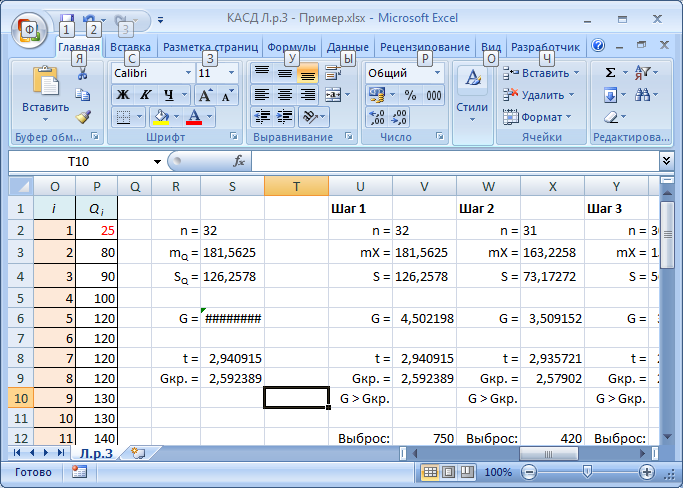
В новой копии отсортируйте значения *Q* по возрастанию. В результате значения, которые были отмечены как выбросы в предыдущем задании, окажутся в начале и в конце выборки.



Для расчетов нам понадобятся:

* объем выборки *n*;
* среднее значение *mQ*;
* стандартное отклонение *SQ*.

Причем они должны пересчитываться, когда мы удаляем значение из выборки. Поэтому их лучше рассчитать через формулы, а не через «Анализ данных». Значения должны совпасть с тем, что получилось в п.1.

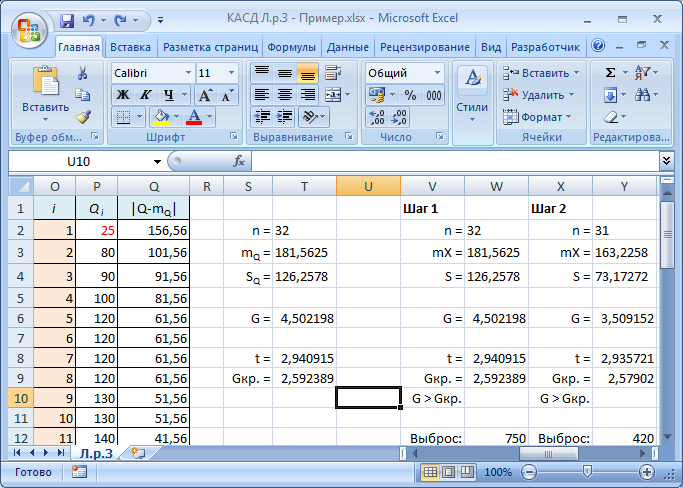


Теперь выполним проверку по тесту Граббса.

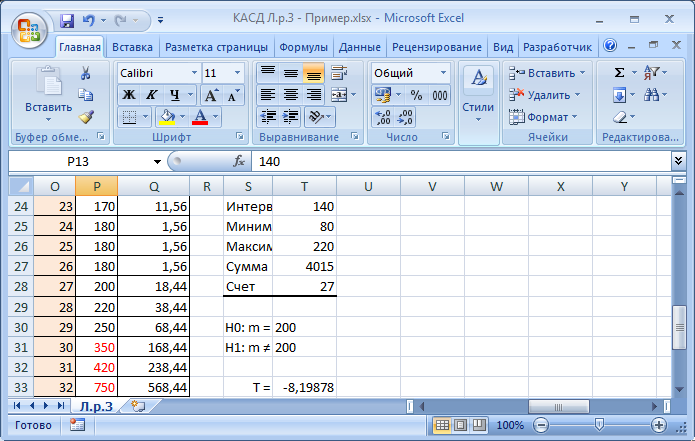


Для этого придется добавить столбец с расчетом отклонений от среднего  для каждого значения.

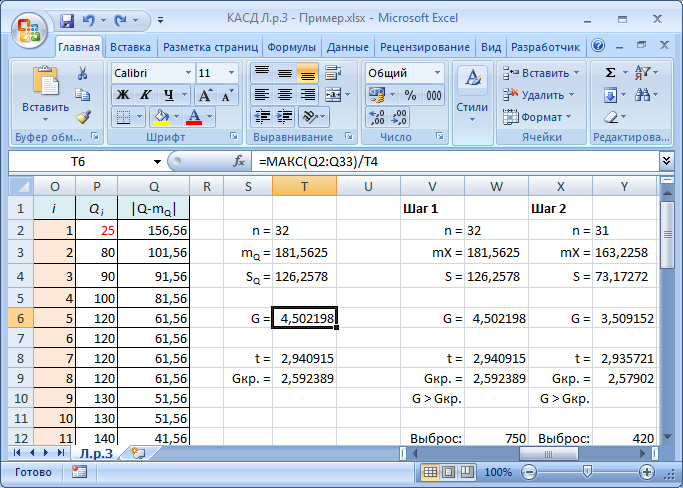
|  |  |
| --- | --- |
|  | Модуль числа вычисляется по формуле ABS(*число*) |



В примере наибольшее отклонение имеет значение 750. Его отклонение равно 568,44.



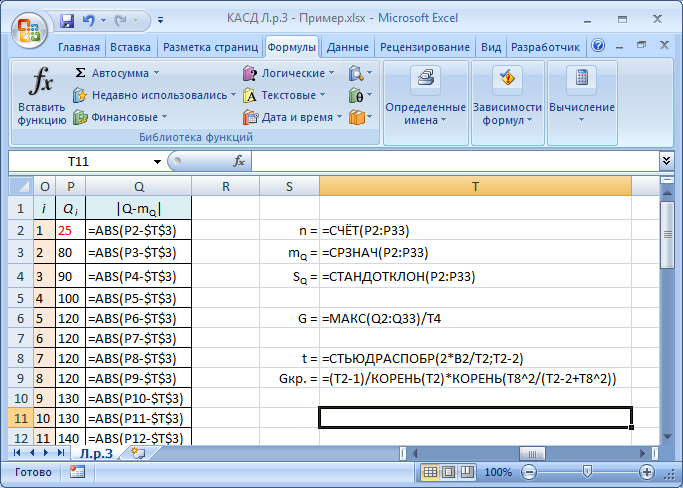
Теперь вычислим значение G по указанной формуле. Необходимо найти максимальное из отклонений от среднего и поделить на стандартное отклонение.

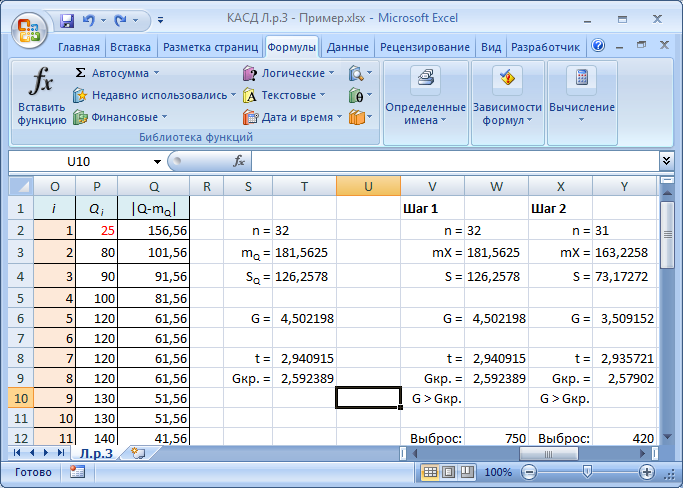


Теперь вычислим критическое значение критерия:



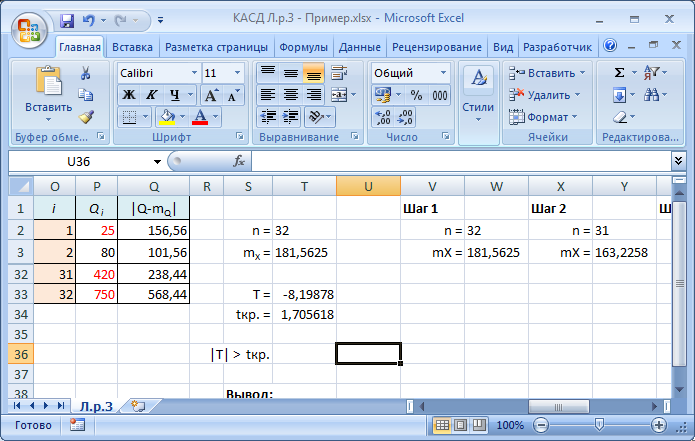
|  |  |
| --- | --- |
|  | Обратите внимание, в формуле стоит . В Excel нужно брать значение вдвое больше, поэтому в СТЬЮДРАСПОБР подставляем .  Чтобы не писать слишком длинную формулу,  посчитаем в отдельной ячейке. |





G > Gкр., значит в выборке есть случайный выброс. Это число с максимальным отклонением от среднего (либо первое, либо последнее).

На скриншоте промежуточные значения скрыты.

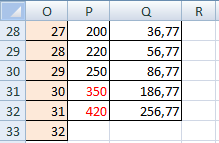


В нашем примере самое маленькое значение *Q*1 = 25 имеет отклонение равное |*Q*1‑ *mQ*| = 156,56, а самое большое значение *Q*32 = 750 – отклонение |*Q*32‑ *mQ*| = 568,44. Следовательно, у 750 отклонение от среднего больше (568,44 >156,56) и именно оно является случайным выбросом.

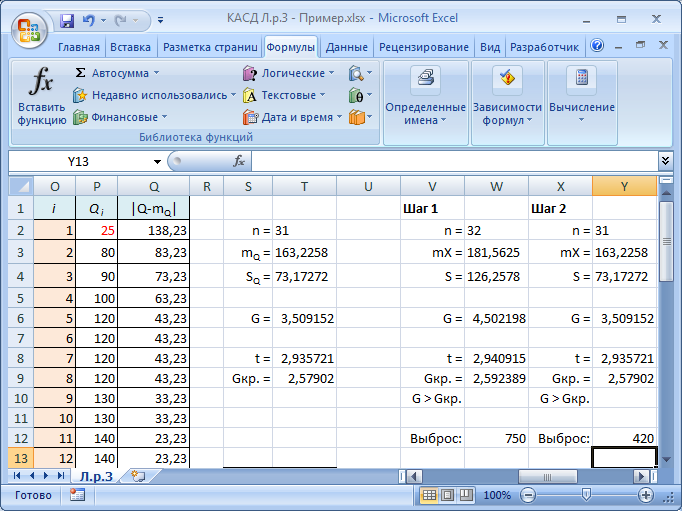
Когда мы удалим это значение из выборки, все показатели пересчитаются автоматически. Поэтому перед удалением сохраним промежуточные расчеты по текущему шагу, скопировав их на пустое пространство справа от вычислений.

|  |  |
| --- | --- |
|  | Если вы просто скопируете и вставите ячейки, то в результате получите нули или ошибки вместо чисел, потому что Excel копирует формулы и смещает их, как при растягивании.  Чтобы вставить числа, а не формулы, используется «Специальная вставка».  Скопируйте ячейки как обычно, а для вставки либо выберите на ленте Вставить – Вставить значения, либо кликните правой кнопкой по ячейке, куда будете вставлять – «Специальная вставка...», отметьте пункт «Значения и форматы чисел» (тогда вставится не только число, но и его оформление» – ОК. |

Теперь удалим случайный выброс из выборки (совсем удалим).

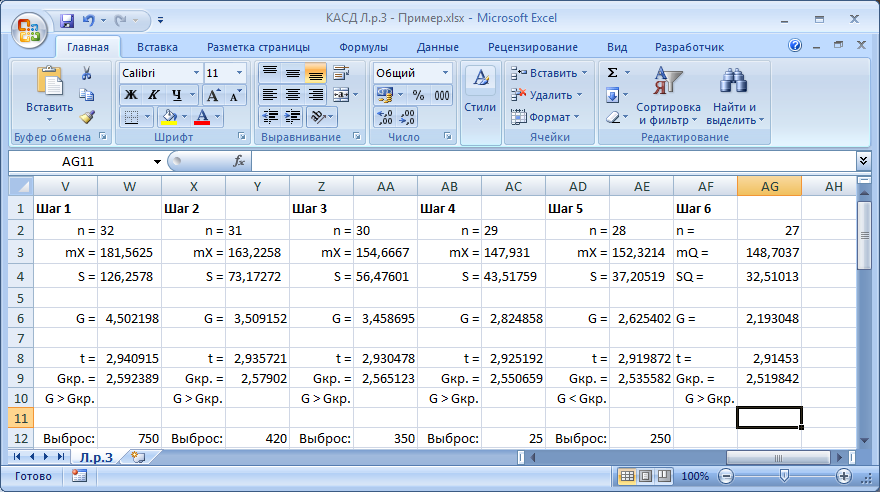


После удаления все формулы пересчитались автоматически, нужно лишь сделать вывод и скопировать полученный результат.



Таким образом, 420 – тоже случайный выброс, его необходимо удалить.

По аналогии проделаем следующие шаги, пока G не станет меньше Gкр.



В примере это произошло на 6 шаге, когда в выборке осталось 27 значений и G = 2,19 стало меньше Gкр. = 2,51.

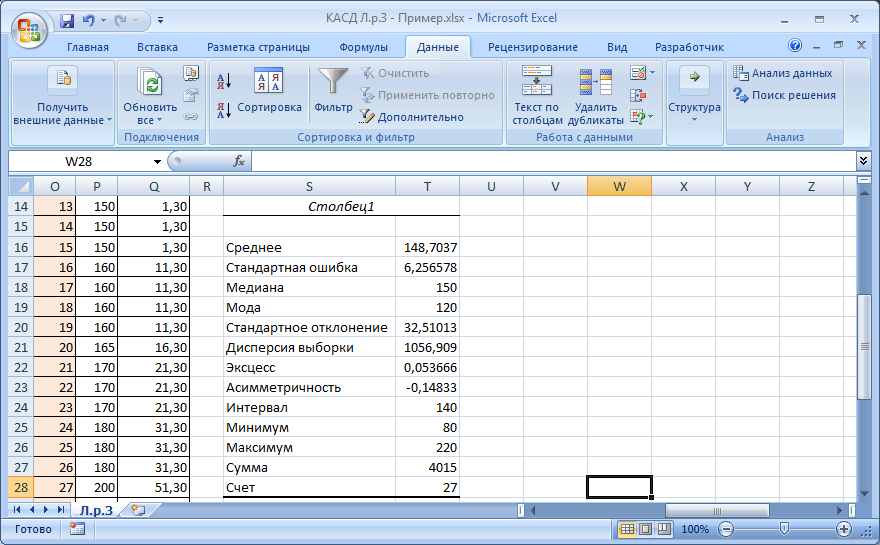
**Вывод**: Из выборки были исключены случайные выбросы 750, 420, 350, 25 и 250. При анализе графика не удалось определить, что 250 является случайным выбросом, другие выбросы были определены верно. Очищенная выборка содержит 27 значений из первоначальных 32.

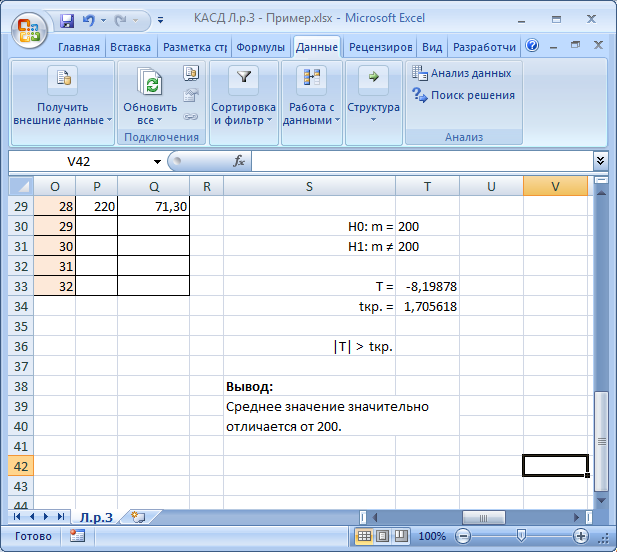
Примечание

Тест Граббса, как и графический анализ, позволяет найти случайные выбросы, но не объясняет, почему они появились – ошибки это, или просто необычная ситуация. Поэтому на практике не всегда нужно удалять случайные выбросы. Все зависит от конкретного случая.

1. После исключения случайных выбросов характеристики выборки изменились, это можно увидеть на предыдущем рисунке (*mQ* уменьшилось с 181,56 до 148,70, а *SQ* с 126,25 до 32,5).

Повторим вычисления из п.1 и 2 для очищенной от выбросов выборки (с удаленными значениями).





**Вывод**: Таким образом, для очищенной выборки выводы изменились. Среднее число рассмотренных за месяц дел нельзя считать равным 200 шт. Расчетное среднее значение составляет около 150 шт.

Изменились и значения эксцесса и асимметрии. Теперь они ненамного отличаются от 0, т.е. выборка симметрична, не является остро- или плосковершинной, а близка к нормальной.

Стандартное отклонение уменьшилось до 32,5 (коэффициент вариации равен 21,8%), т.е. разброс значений стал незначительным.