Федеральное государственное образовательное учреждение
высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»

Департамент анализа данных, принятия решений и финансовых технологий

И.Э. Гурьянова

СБОРНИК ЗАДАНИЙ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА С ПРИМЕНЕНИЕМ EXCEL»

По дисциплине:

«Анализ данных»

Для студентов, обучающихся по направлению подготовки
38.03.01 «Экономика», 38.03.02 «Менеджмент», 39.03.01 «Социология»,
09.03.03 «Прикладная информатика», 38.03.05 «Бизнес-информатика»,
10.03.01 «Информационная безопасность»
(все профили подготовки бакалавров)

Москва, 2020

Федеральное государственное образовательное учреждение
высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»

Департамент анализа данных, принятия решений и финансовых технологий

И.Э. Гурьянова

СБОРНИК ЗАДАНИЙ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА С ПРИМЕНЕНИЕМ EXCEL»

По дисциплине:

«Анализ данных»

Для студентов, обучающихся по направлению подготовки
38.03.01 «Экономика», 38.03.02 «Менеджмент», 39.03.01 «Социология»,
09.03.03 «Прикладная информатика», 38.03.05 «Бизнес-информатика»,
10.03.01 «Информационная безопасность»
(все профили подготовки бакалавров)

Рассмотрено и одобрено на заседании Департамента анализа данных, принятия решений и финансовых технологий (протокол №11 от «12» мая 2020 г.)

Москва, 2020

УДК 519.2

ББК 22.172
Г 95

Рецензент: В.А. Газарян, к.ф-м.н., доцент Департамента анализа данных, принятия решений и финансовых технологий.

Гурьянова И.Э. Сборник заданий «Математическая статистика с применением Excel» М.: Финансовый университет при Правительстве РФ, Департамент анализа данных, принятия решений и финансовых технологий, 2020. 118 с.
Сборник заданий с применением функций Excel по дисциплине «Анализ данных», где каждая глава предваряется подробными теоретическими сведениями, предназначен для бакалавров очной формы обучения по направлению подготовки: 38.03.01 «Экономика», 38.03.02 «Менеджмент», 39.03.01 «Социология», 09.03.03 «Прикладная информатика», 10.03.01 «Информационная безопасность», 38.03.05 «Бизнес-информатика» (все профили подготовки бакалавров).

УДК 519.2
ББК 22.172

Учебное издание
Гурьянова Ирина Эдуардовна

«Сборник задач по математической статистике с применением Excel»

Компьютерный набор **Гурьянова И.Э.**
Компьютерная верстка **Гурьянова И.Э.**
Формат 60x90/16. Гарнитура *Times New Roman*
Усл. п.л. 7,4

Оглавление

Введение

Предлагаемый сборник задач по математической статистике охватывает все разделы этой дисциплины, входящие в учебные программы для студентов вузов, обучающихся по экономическим специальностям. Сборник содержит основные теоретические сведения по математической статистике и предназначен для закрепления теоретических знаний по этому курсу. В основу курса положен учебник В.И. Соловьева «Анализ данных в экономике: теория вероятностей, прикладная статистика и визуализация данных в Microsoft Excel». Курс соответствует программе дисциплины «Анализ данных» для профилей «Бизнес-информатика», «Финансы и кредит», «Мировая экономика» и «Налоги налогообложение». Предлагаемый сборник может быть полезен студентам и преподавателям ВУЗов, а также лицам, изучающим математическую статистику самостоятельно. Предлагаемый материал может быть использован для самостоятельного изучения курса студентами, обучающимися заочно.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Математической статистикой называется наука, которая, основываясь на методах теории вероятностей, занимается систематизацией, обработкой и использованием статистических данных (то есть результатов наблюдений) для получения научных и практических выводов.

Основные задачи математической статистики состоят в разработке методов:

1. Организации и планирования статистических наблюдений, в том числе способов определения числа необходимых испытаний (планирование эксперимента).

2. Сбора статистических данных.

3. «Свертки информации», то есть методов группировки данных и сведения большого числа данных к небольшому числу параметров.

4. Анализа статистических данных.

5. Принятия решений, рекомендаций и выводов на основе анализа статистических данных.

6. Прогнозирования случайных явлений.

В математической статистике признак – то же самое, что функция, но без явной привязки к некоторой области определения— это фиксируемое свойство исследуемого объекта. Вместо термина «признак» используется термин «переменная». Признаки, как и случайные величины, обозначаются буквами X, Y, Z.

Определение.

Совокупностью (статистической совокупностью) называется множество изучаемых объектов, а число ее элементов – объемом.

Объем совокупности может быть как конечным, так и бесконечным. Предположим временно, что объем совокупности конечен. Рассмотрим признак X, заданный на совокупности $\Omega = \{\omega_1, \dots, \omega_n\}$. Пусть $x_1 = X(\omega_1)$, …, $x_n = X(\omega_n)$–его значения.

Эмпирические характеристики признаков

Определение.

Эмпирическим средним или средним значением признака в совокупности Ω называется среднее арифметическое всех его значений в этой совокупности

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Определение.

Эмпирической дисперсией или дисперсией признака X в совокупности Ω называется среднее арифметическое квадратов отклонений его значений от эмпирического среднего

$$D(X) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n};$$

при этом $\sigma = \sqrt{D(X)}$ называется стандартным отклонением X в совокупности Ω.

Эмпирическим начальным моментом k-того порядка признака X называется величина

$$\nu_k(X) = \frac{x_1^k + x_2^k + \cdots + x_n^k}{n}.$$

Эмпирическим центральным моментом k-того порядка называется величина

$$\mu_k(X) = \frac{(x_1 - \bar{x})^k + (x_2 - \bar{x})^k + \cdots + (x_n - \bar{x})^k}{n}.$$

Эмпирическая функция распределения $F(x)$ определяется следующим образом:

$$F(x) = \frac{\{\text{число элементов } \omega \in \Omega, \text{для которых } X(\omega) < x\}}{n}.$$

Если рассматривать признак, как случайную величину с равновероятными возможными значениями, то эмпирические характеристики совпадают с числовыми характеристиками из теории вероятностей.

Например, известная формула $D(X) = E(X^2) - E^2(X)$ приобретает для эмпирической дисперсии вид: $D(X) = \overline{X^2} - \bar{X}^2$.

Вариационный и статистический ряды

1. Вариационный ряд.

Пусть $x_1, \dots, x_n$– значения (возможно, с повторениями) некоторого признака X в совокупности объема n, которые называются элементами.

Элементы, расположенные в возрастающем (неубывающем) порядке, то есть ранжированные $x_{(1)}, \dots, x_{(n)}$, составляют вариационный ряд.

Пример.

6, 9, 10, 12, 13, 14, 14, 15, 16, 16, 16, 17, 17, 18, 18, 19, 20, 21, 22, 24.

Разность между наибольшей и наименьшей вариантой называется размахом признака или размах вариации.

Определение.

Порядковый центр (середина) вариационного ряда называется эмпирической медианой Me.

Если в вариационном ряду нечетное число вариант $n = 2k + 1$, то $Me = x_{(k+1)}$.

Если четное число $n = 2k$ вариант, то

$$Me = \frac{x_{(k)} + x_{(k+1)}}{2}.$$

Примеры.

Пусть признак X– возраст людей.

8; 9; 11; 12; 15; 16; 18; 19; 21

$n = 9$; $Me = x_{(5)} = 15$.

8; 9; 11; 12; 15; 16; 18; 19; 21; 23; 24; 26

$n = 12$;

$$Me = \frac{x_{(6)} + x_{(7)}}{2} = \frac{16 + 18}{2} = 17.$$

2. Статистический ряд.

Если в совокупности Ω значение x_1 наблюдалось n_1 раз, значение x_2 – n_2 раза,…, $x_к$ - $n_к$ раз, то числа $n_1, n_2, \dots, n_k$ называются частотами вариант. $(n_1 + \cdots + n_k = n)$.

Числа $W_i = \frac{n_i}{n}$ называются относительными частотами или частостями. $(W_1 + \cdots + W_k = 1)$. Частоты и частости называются весами.

Таблица частот вида

x_1	x_2	…	x_k
n_1	n_2	…	n_k

называется частотным распределением признака, или статистическим распределением (рядом),

а таблица относительных частот вида

x_1	x_2	…	x_k
W_1	W_2	…	W_k

называется эмпирическим распределением (рядом) признака.

Формулы для введенных ранее эмпирических характеристик приобретают вид:

$$\bar{x} = \frac{x_1 n_1 + \cdots + x_k n_k}{n} = x_1 W_1 + \cdots + x_k W_k \,.$$

$$D(X) = \frac{(x_1 - \bar{x})^2 n_1 + \cdots + (x_k - \bar{x})^2 n_k}{n} = (x_1 - \bar{x})^2 W_1 + \cdots + (x_k - \bar{x})^2 W_k \,.$$

Кроме того, верна формула:

$$D(X) = \left(\sum_{i=1}^{k} n_i x_i^2 \right) \Big/ n - \bar{x}^2.$$

$$\nu_k = \frac{x_1^k n_1 + \cdots + x_k^k n_k}{n}$$

$$\mu_k = \frac{(x_1 - \bar{x})^k n_1 + \cdots + (x_k - \bar{x})^k n_k}{n}$$

Определение.

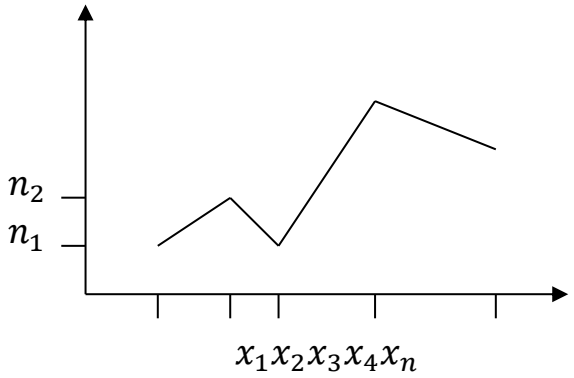
Пусть $W_1, \ldots, W_k$– положительные, а $x_1, \ldots, x_k$–произвольные числа. Отношение

$$\bar{x} = \frac{x_1 W_1 + \cdots + x_k W_k}{W_1 + \cdots + W_k}$$

называется взвешенным средним чисел $x_1, \ldots, x_k$, при этом число W_i называется весом x_i.

Графическим изображением статистического распределения вариационного ряда является полигон частот.

Полигон – это ломаная, соединяющая точки с координатами (x_1, n_1),…, (x_k, n_k).



Определение.

Модой статистического распределения называется значение x_i, которому соответствует наибольшая частота.

Может быть: одна мода – унимодальное распределение,
две моды – бимодальное распределение,
три и более – мультимодальное распределение.

$As = \frac{\mu_3}{\sigma^3}$ – асимметрия вариационного ряда;

$As < 0$ –левосторонняя скошенность;

$As > 0$ – правосторонняя скошенность;

$As = 0$ – симметричный ряд;

Эксцесс $Ex = \frac{\mu_4}{\sigma^4} - 3$.

Эмпирическая ковариация и эмпирический коэффициент корреляции

Пусть $x_i = X(\omega_i), y_i = Y(\omega_i),\ \omega_i \in \Omega$ – значения признаков X и Y на совокупности $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$. Эмпирическая ковариация определяется формулой:

$$cov(X,Y) = \left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)/n\,.$$

Таблицей сопряженности или совместным частотным распределением признаков X и Y называется таблица:

X \ Y	$Y = y_1$	$Y = y_2$	…	$Y = y_k$
$X = x_1$	n_{11}	n_{12}	…	n_{1k}
…	…	…	…	…
$X = x_m$	n_{m1}	n_{m2}	…	n_{mk}

Сумма всех частот n_{ij} равна объему совокупности n.

Формула для вычисления ковариации:

$$cov(X,Y) = \left(\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}(x_i - \bar{x})(y_i - \bar{y})\right)\Big/n = \left(\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}x_i y_i - n\bar{x}\bar{y}\right)\Big/n,$$

или $cov(X,Y) = \overline{xy} - \bar{x}\,\bar{y}$.

Эмпирический коэффициент корреляции определяется формулой:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{D(X)D(Y)}} = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Пример.

В совокупности 16 студентов определены два признака, X – оценка по математике и Y – оценка по иностранному языку. Совместное статистическое распределение оценок задано таблицей

	$X=2$	$X=3$	$X=4$	$X=5$
$Y=3$	1	0	1	0
$Y=4$	2	4	4	2
$Y=5$	0	1	0	1

Найти $\rho(X,Y)$.

Найдем эмпирические распределения признаков:

x_i	2	3	4	5
n_i	3	5	5	3

Y_i	3	4	5
m_i	2	12	2

$$\bar{x} = \frac{2\cdot 3 + 3\cdot 5 + 4\cdot 5 + 5\cdot 3}{16} = \frac{6+15+20+15}{16} =$$

$$= \frac{56}{16} = \frac{7}{2} = 3{,}5.$$

$$D(X) = \frac{\sum n_i x_i^2 - n\bar{x}^2}{n} =$$

$$= \frac{4\cdot 3 + 9\cdot 5 + 16\cdot 5 + 25\cdot 3 - 16\cdot 12{,}25}{16} =$$

$$= \frac{12+45+80+75}{16} - 12{,}25 = 1;$$

$$\sigma(X) = 1.$$

$$\bar{Y} = \frac{3\cdot 2 + 4\cdot 12 + 5\cdot 2}{16} = \frac{6+48+10}{16} = \frac{64}{16} = 4.$$

$$D(Y) = \frac{9\cdot 2 + 16\cdot 12 + 25\cdot 2}{16} - 4^2 = \frac{18+192+50}{16} - 16 = 0{,}25;$$

$$\sigma(Y) = 0{,}5.$$

$$cov(X,Y) =$$

$$= \frac{2\cdot 3\cdot 1 + 0 + 4\cdot 1\cdot 3 + 0 + 2\cdot 4\cdot 2 + 3\cdot 4\cdot 4 + 4\cdot 4\cdot 4 + 5\cdot 4\cdot 2}{16} +$$

$$+ \frac{0 + 3\cdot 5\cdot 1 + 0 + 5\cdot 5\cdot 1}{16} - 3{,}5\cdot 4 =$$

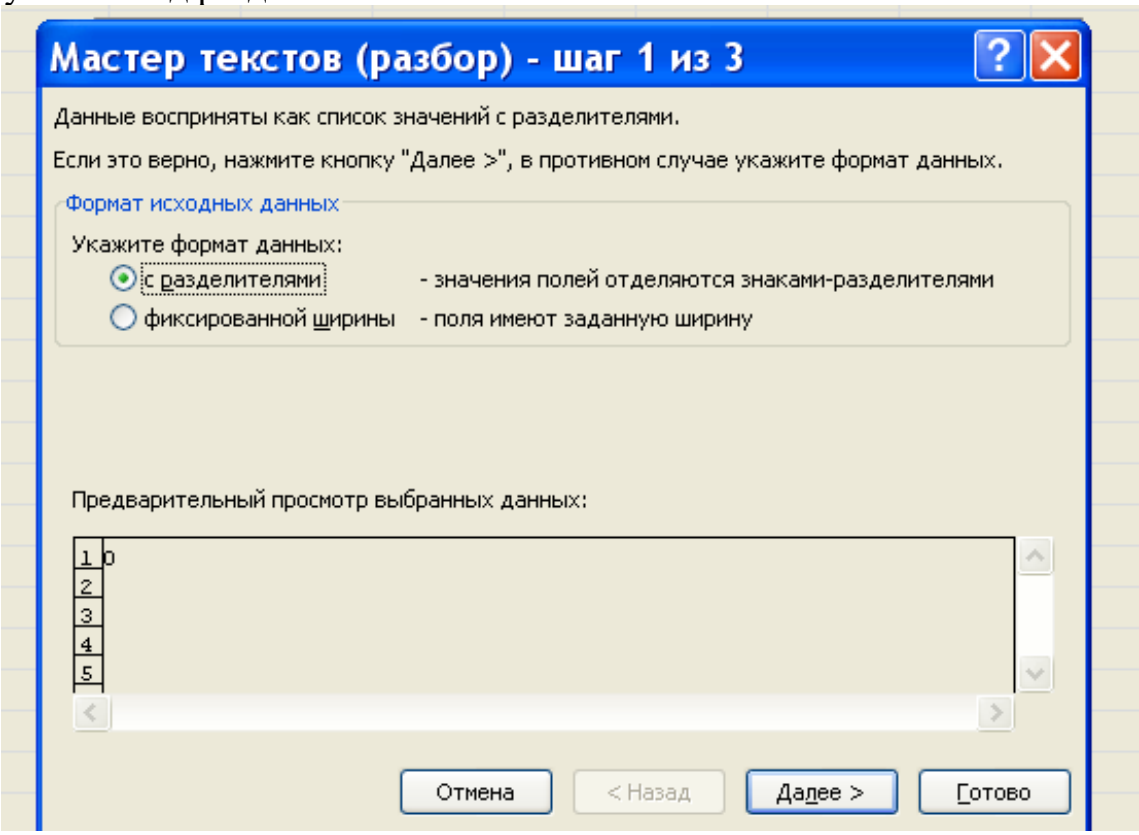
$$= \frac{6 + 12 + 16 + 48 + 64 + 40 + 15 + 25}{16} - 14 = 0{,}125 = \frac{1}{8}.$$

$$\rho(X,Y) = \frac{0{,}125}{1\cdot 0{,}5} = 0{,}25\,.$$

Перенос данных в файл Excel

Прежде, чем приступить к решению задач, научимся переносить данные в файл Excel.

При копировании данных из файла Word в файл Excel все данные оказываются в одной ячейке. Чтобы каждое число расположить в отдельной ячейке, нужно на вкладке «Данные» нажать кнопку «Текст по столбцам» и в появившемся диалоговом окне указать формат данных «с разделителями» и затем указать вид разделителя.



После этого все числа будут в отдельных ячейках одной строки. Чтобы из строки сделать столбец, надо произвести копирование с транспонированием.

После этих действий первая ячейка столбца часто не поддается сортировке А/Я или Я/А. Это можно исправить так:

а) сделать «формат по образцу» любой другой ячейки столбца;

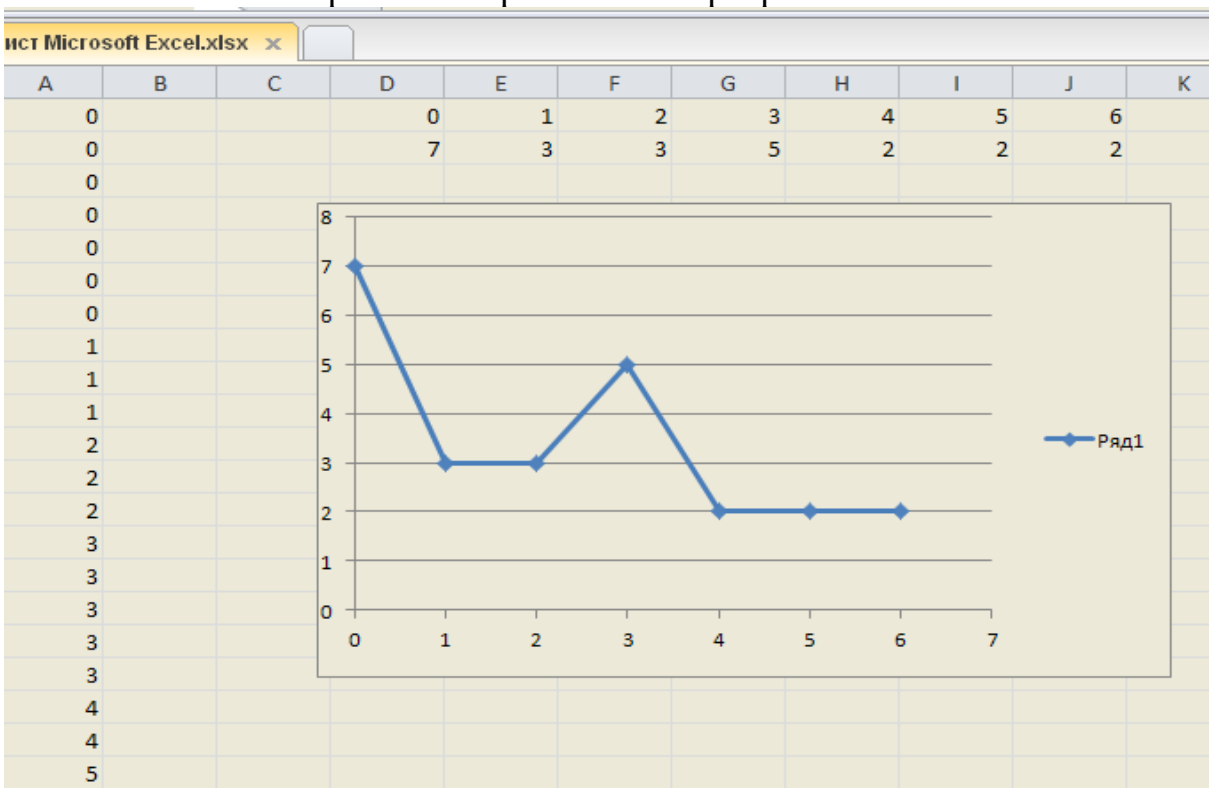
б) «Главная» →Ластик → Щелкнуть по стрелочке списка и выбрать «очистить форматы»;

в) «Данные»→ «Сортировка» → «Мои данные не содержат заголовков» (убрать галочку);

г) удалить пробелы в ячейках, то есть заменить пробел на пустоту, Иногда это мешает ранжированию.

Построение полигона частот покажем на примере нижеследующей задачи 3. Табличку с данными создадим в файле Excel.

После этого, пометив табличку данных, вызываем Мастер диаграмм → «График» → «График с маркерами». Или по-другому «Диаграммы» → «Точечные» → «С прямыми отрезками и маркерами».



Для поиска и удаления **пропущенных значений** из набора данных нужно:

а) меню «Данные» → «Фильтр» снять выделение со строки «выделить все», прокрутить вниз до конца и выделить строку «Пустые». После этого покажет пустые строки;

б) пометить область данных и «Расположить по убыванию» сверху окажутся пустые ячейки, их сразу удалить.

Задачи для самостоятельного решения.

1. Записать в виде вариационного и статистического рядов выборку 5, 3, 7, 10, 5, 5, 2, 10, 7, 2, 7, 7, 4, 2, 4.

Решение. Объем выборки n = 15. Упорядочив элементы выборки по величине, получим вариационный ряд

2, 2, 2, 3, 4, 4, 5, 5, 5, 7, 7, 7, 7, 10, 10 .

Статистический ряд записывается в виде таблицы

x_i	2	3	4	5	7	10
n_i	3	1	2	3	4	2

2. Записать эмпирическую функцию распределения для выборки, представленной статистическим рядом.

x_i	15	16	17	18	19
n_i	1	4	5	4	2

Ответ: $F^*(x) = \begin{cases} 0, & x \le 15, \\ 0{,}0625, & 15 < x \le 16, \\ 0{,}3125, & 16 < x \le 17, \\ 0{,}625, & 17 < x \le 18, \\ 0{,}875, & 18 < x \le 19, \\ 1, & x > 19. \end{cases}$

3. С помощью журнала посещаемости собраны данные о числе пропусков по математике у 25 студентов 2 курса. В итоге получены значения: **2, 5, 0, 1, 6, 3, 0, 1, 5, 4, 0, 3, 3, 2, 1, 4, 0, 0, 2, 3, 6 ,0, 3, 0, 1.**

Построить вариационный ряд и статистический ряд;

выписать таблицу частот и относительных частот – статистический ряд и эмпирическое частотное распределение;

найти медиану и моду;

построить полигон частот и относительных частот;

вычислить эмпирическое среднее, эмпирическую дисперсию и стандартное отклонение;

построить эмпирическую функцию распределения.

Ответ:

$\overline{x} = 2{,}2, D = 3{,}76; \sigma = 1{,}939, MoX = 0, MeX = 2$

4. Получены данные о числе телевизоров, продаваемых ежедневно в магазине электроники в течение 26 дней:

16; 12; 15; 15; 23; 9; 15; 13; 14; 14; 21; 15; 14; 17; 27; 15; 16; 12; 16; 19; 14; 16; 17; 13; 14; 14.

Построить вариационный ряд и статистический ряд;

выписать таблицу частот и относительных частот – статистический ряд и эмпирическое частотное распределение;

найти медиану и моду;

построить полигон частот и относительных частот;

вычислить эмпирическое среднее, эмпирическую дисперсию и стандартное отклонение;

построить эмпирическую функцию распределения.

Ответ:

$\overline{x} = 15{,}615, D = 12{,}71; \sigma = 3{,}565; MeX = 15 : MoX = 14$

5. На практическом занятии по математике отобрали случайным образом 10 студентов и произвели хронометраж затрат времени на решение 1 задачи. В результате наблюдений: **2,7; 3,1; 1,4; 1,6; 2,3; 3,7; 2,8; 3,5; 2,6; 1,9 минут.**

Построить вариационный ряд и статистический ряд;

выписать таблицу частот и относительных частот – эмпирическое частотное распределение;

найти медиану и моду;

построить полигон частот и относительных частот;

вычислить эмпирическое среднее, эмпирическую дисперсию и стандартное отклонение;

построить эмпирическую функцию распределения.

Ответ:

$\overline{x} = 2{,}56, D = 0{,}53244$

6. При обследовании численности 50 семей сотрудников крупного предприятия установлено следующее количество членов семьи 5, 3, 2, 1, 4, 6, 3, 7, 9, 1, 3, 2, 5, 6, 8, 2, 5, 2, 3, 6, 8, 3, 4, 4, 5, 6, 5, 4,7, 5, 6, 4, 8, 7, 4, 5, 7, 8, 6, 5, 7, 5, 6, 6, 7, 3, 4, 6, 5, 4. Построить вариационный ряд и статистический ряд;

выписать таблицу частот и относительных частот – эмпирическое частотное распределение;

найти медиану и моду;

построить полигон частот и относительных частот;

вычислить эмпирическое среднее, эмпирическую дисперсию и стандартное отклонение;

построить эмпирическую функцию распределения.

Ответ: Мо = 5;

$\overline{x} = 4{,}94, D = 3{,}6964$

7. Имеется распределение 80 предприятий по числу работающих на них (чел.):

x_i	150	250	350	450	550	650	750

n_i	1	3	7	30	19	15	5 .

Найти числовые характеристики распределения предприятий по числу работающих. Ответ:

$\bar{x} = 510, D = 15400$

8. Построить эмпирическую функцию распределения по выборке:

x_i	2	6	10
n_i	12	18	30

Ответ: $F^*(x) = \begin{cases} 0, & x \le 2, \\ 0{,}2, & 2 < x \le 6, \\ 0{,}5, & 6 < x \le 10, \\ 1, & x > 10. \end{cases}$

9. Совместное частотное распределение признаков X и Y задано таблицей.

	X = 1	X = 3	X = 5	X = 6
Y = 1	4	2	2	4
Y = 2	1	0	0	1
Y = 3	0	1	1	0

Вычислить эмпирический коэффициент корреляции.

Ответ:

$\bar{x} = 3{,}6875, D(X) = 4{,}3398\ \bar{y} = 1{,}375, D(Y) = 0{,}484$

Cov = 0,0547, ρ = 0,0379.

10. Совместное частотное распределение признаков X и Y задано таблицей

	X = 0	X = 2	X = 4
Y = 1	1	4	0
Y = 3	0	3	2
Y = 4	1	5	2
Y = 6	3	3	0

Найти эмпирический коэффициент корреляции.

Ответ:

$\bar{x} = \frac{23}{12}, D(X) = 1{,}493\ \bar{y} = \frac{11}{3}, D(Y) = 2{,}972$

Cov = - 0,4444, (или -0,446) ρ = - 0,2121.

11. Совместное частотное распределение признаков X и Y задано таблицей

	X = 2	X = 4	X = 6	X = 8
Y = 1	2	4	3	0
Y = 3	1	2	4	3
Y = 5	0	3	2	1

Найти эмпирический коэффициент корреляции.

$\overline{x} = 5{,}12, D(X) = 3{,}226\ \overline{y} = 2{,}76, D(Y) = 2{,}342$

Cov = 0,7488, ρ = 0,2724.

12. Совместное частотное распределение признаков X и Y задано таблицей

	X = 2	X = 3	X = 4	X = 5
Y = 3	1	0	1	0
Y = 4	2	4	4	2
Y = 5	0	1	0	1

Найти эмпирический коэффициент корреляции.

Ответ: $\overline{x} = 3{,}5$, $\overline{y} = 4$, D(X)=1, D(Y)=0,25, cov(X,Y) = 1/8, ρ = 0,25.

13. Измерен рост (с точностью до сантиметра) 30 наудачу отобранных студентов. Результаты измерений:

178, 160, 154, 183, 155, 153, 167, 186, 163, 155, 157, 175, 170, 166, 159, 173, 182, 167, 171, 169, 179, 165, 156, 179, 158, 171, 175, 173, 164, 172.

Построить интервальное статистическое распределение и гистограмму.

14. Значения признаков X и У заданы на множестве Ω = {1,2,…,34} таблицей частот:

	X = 2	X = 4	X = 5	X = 7
Y = 1	7	1	5	8
Y = 2	3	0	1	0
Y = 3	6	0	3	0

Найти эмпирический коэффициент корреляции.

Ответ: $\bar{X}$= 4,0294; D(X)=4,264; σ(X)=2,06…; $\bar{Y}$=1,647; D(Y) = 0,758; σ(Y)=0,8701; cov(X,Y)=--0,695; ρ(X,Y)=--0,386.

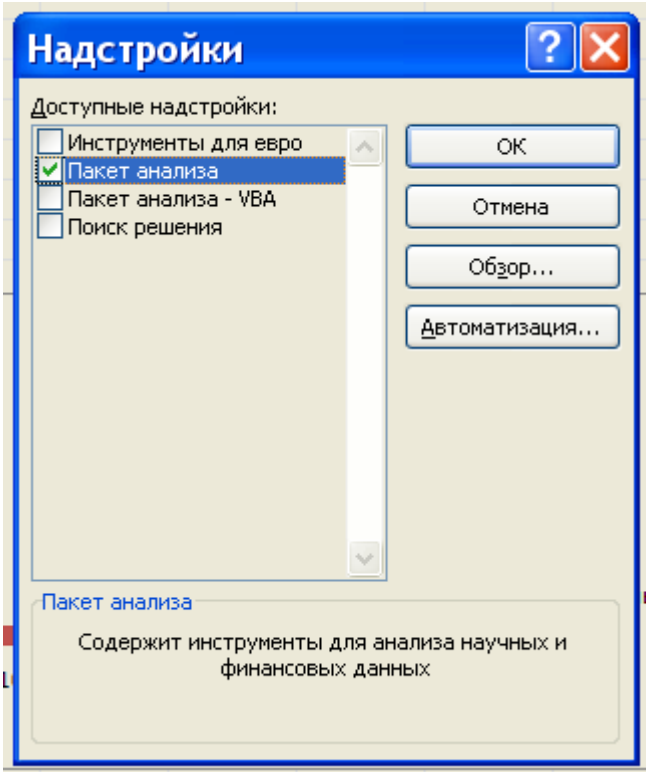
15. Значения признаков X и У заданы на множестве Ω = {1,2,…,2000} таблицей частот:

	Y = 1	Y = 3	Y = 5
X = 4	500	100	100
X = 6	400	300	600

Из Ω **с возвращением** извлекается 700 элементов. Пусть $\bar{X}$ и $\bar{Y}$ - средние значения признаков в выборочной совокупности. Найти $Cov(\bar{X},\bar{Y})$. Ответ: $\bar{X}$= 5,3; $\bar{Y}$=2,8; cov(X,Y)= 0,66; cov($\bar{X}$, $\bar{Y}$)= 0,000943.

Пакет анализа Microsoft Excel

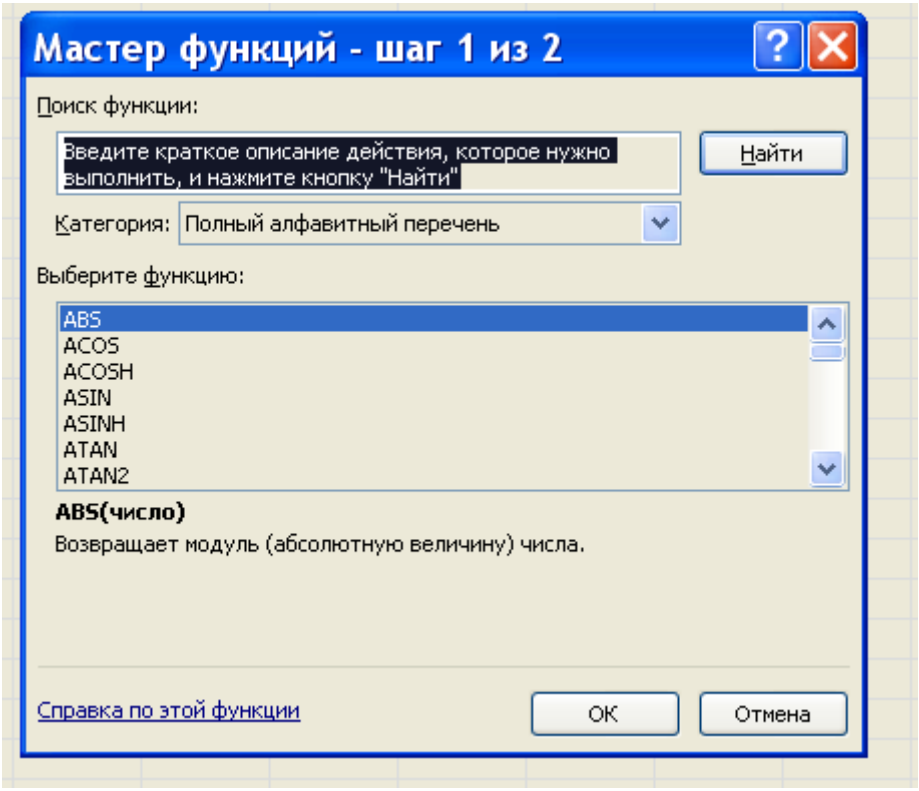
В первую очередь установим важную надстройку Microsoft Excel **Пакет анализа.** Меню «Файл» → «Параметры» → «Надстройки» → «Надстройки Excel» , нажать кнопку «Перейти» и поставить галочку в клетке Пакет анализа.



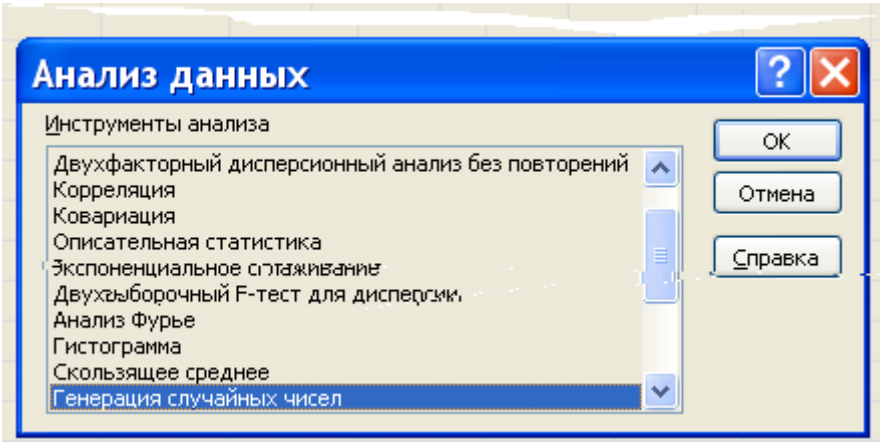
После этого на вкладке «Данные» справа появится раздел «Анализ данных».

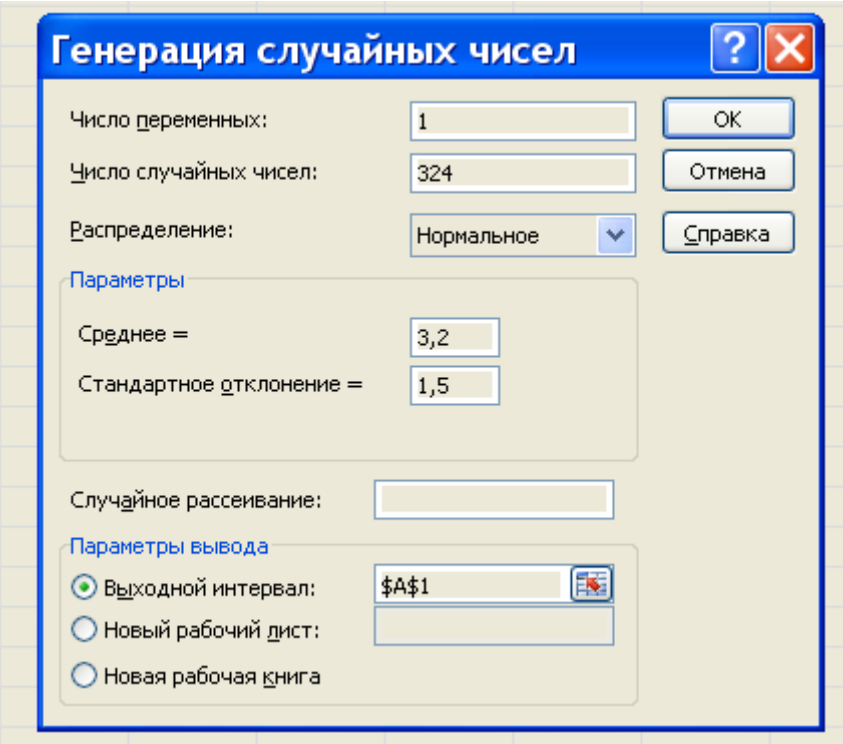
Кроме этого, мы будем много работать с **Мастером функций**. Для удобства кнопку «Вставить функцию»

вынесем на панель быстрого доступа. Не все необходимые нам функции есть в списке «Статистические» поэтому будем использовать «Полный алфавитный перечень» мастера функций.



С помощью Пакета анализа научимся генерировать случайные выборки. На вкладке «Анализ данных» выберем «Генерация случайных чисел».





В поле Число переменных введем 1 . Число переменных это число столбцов.

В поле Число случайных чисел введем любое число, например, 324.

В поле распределение введем Нормальное.

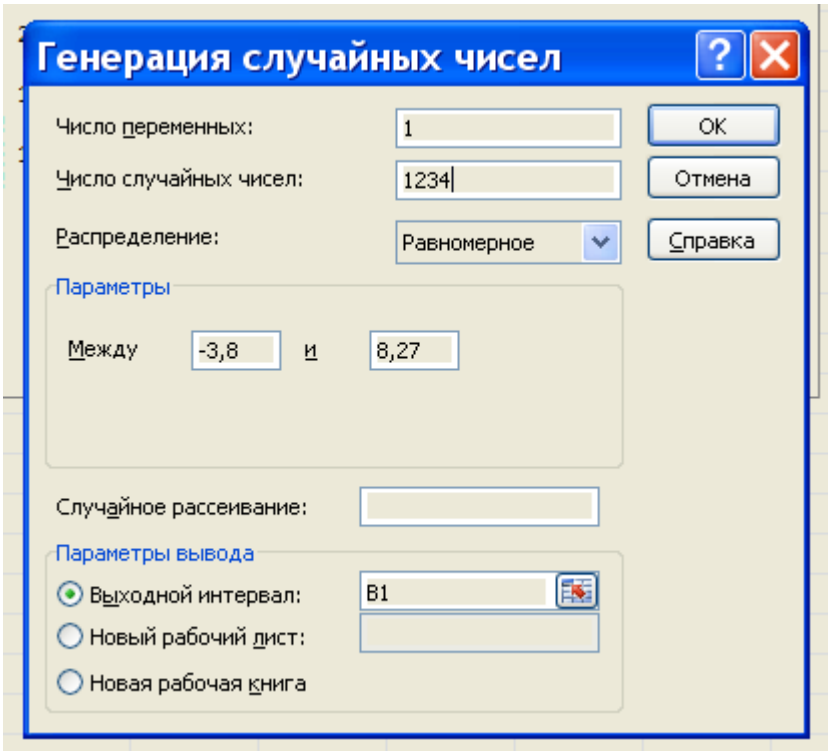
В поле среднее введем 3,2; (тут можно вводить и отрицательные числа)

В поле Стандартное отклонение введем 1,5.

В поле Случайное рассеивание ничего не вводим, это поле предназначено для закрепления выборки, чтобы выборку с этими же параметрами можно было вызвать еще раз.

В поле выходной интервал введем ячейку А1. ОК.

Или выберем Равномерное распределение между -3,8 и 8,27.



Кроме того можно генерировать любое число данных с помощью **функции СЛЧИС**, просто протягивая крестик в углу ячейки вниз на любое число ячеек. Там генерируются только числа от 0 до1.

Также можно генерировать любое число данных с помощью **функции** СЛУЧМЕЖДУ, между -4 и 6. просто протягивая крестик вниз на любое число ячеек. Там будут появляться только целые числа.

Задачи для самостоятельного решения.

1.Сгенерируйте выборку нормального распределения объема 1438 с параметрами -2,7 и 0,6.

2.Сгенерируйте выборку равномерного распределения объема 456 между -11,16 и 27,9.

3. Сгенерируйте выборку распределения Пуассона объема 753 с параметром 2.

4. Сгенерируйте выборку биномиального распределения объема 198 с параметрами 100 и 0,78.

5. С помощью функции СЛЧИС сгенерируйте выборку объема 87.

6. С помощью функции СЛУЧМЕЖДУ сгенерируйте выборку целых чисел объема 145 между числами 112 и 278.

7. Сгенерируйте выборку объема 74 чисел, находящихся между 0 и 1.

8. Сгенерируйте выборку объема 171 из целых чисел между -11 и 14.

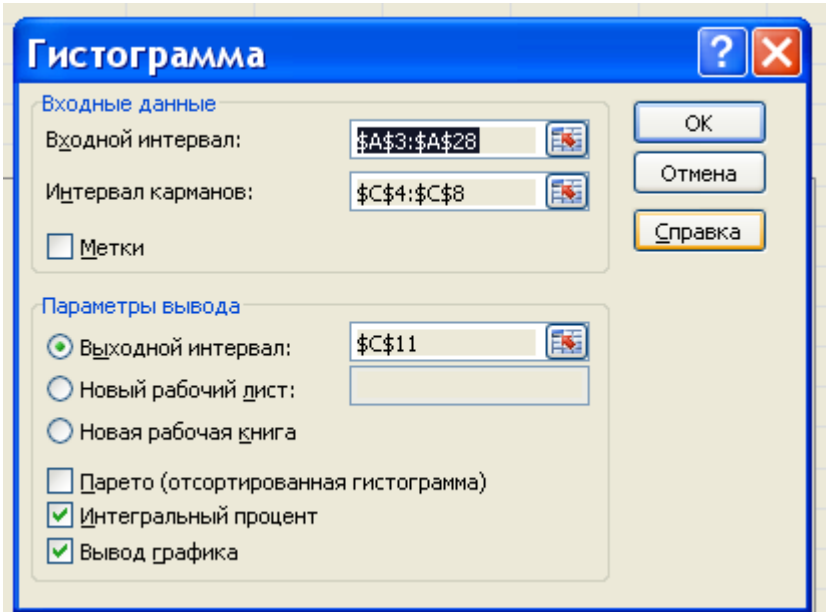
График эмпирической функции распределения

Научимся переносить данные из файлов Word в Excel и строить график эмпирической функции распределения F(x) с помощью *пакета* анализа. Сделаем это на примере задачи 4 стр. 14.

Откроем файл Excel и скопируем в него данные из задачи. Все числа окажутся в одной ячейке. Поместим каждое число в отдельную ячейку, выбрав на вкладке «Данные» → «Текст по столбцам». Поставим точку в поле «с разделителями»

Нажмем кнопку «Далее» и выберем в качестве разделителя точку с запятой. Затем превратим строку данных в столбец, используя копирование с транспонированием. Отранжируем выборку с помощью кнопки А↓Я на вкладке «Данные».

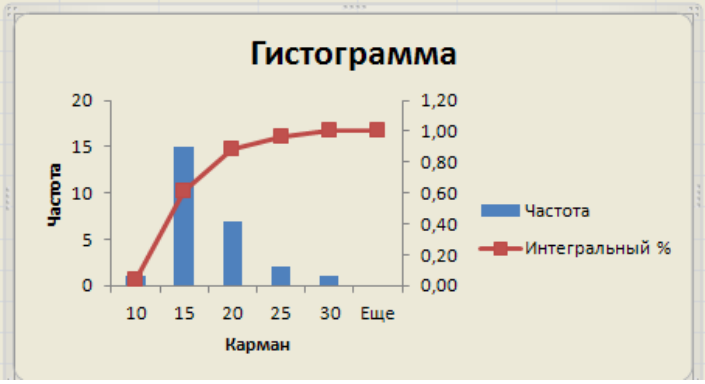
Видно, что данные - это числа от 9 до 27. Все данные помещаются на расширенном отрезке [5;30] Разделим этот промежуток на несколько равных отрезков и *правые* границы малых интервалов разместим в отдельном столбце. Это будет так называемый *интервал карманов*. Эти значения –*правые* границы будущих интервалов гистограммы. Затем перейдем на вкладку Данные → Анализ данных → Гистограмма



Входной интервал – это столбец А, интервал карманов – это интервал С4-С8. Обязательно поставить галочки в полях Вывод графика и Интегральный процент. ОК. После вывода графика следует формат ячеек в столбце *Интегральный %* поменять с процентного на числовой.

10
15
20
25
30

Карман	Частота	Интегральный %
10	1	0,04
15	15	0,62
20	7	0,88
25	2	0,96
30	1	1,00
Еще	0	1,00



Можно сделать карманов больше и гистограмму с большим числом столбцов. Получится немного другая гистограмма и график.

Межгрупповая дисперсия

Пусть X – признак в совокупности Ω объема n, разбитой на k групп объема n_i:

$$\Omega_i = \{\omega_{i1}, \omega_{i2}, \dots, \omega_{in_i}\}, \quad i = 1, \dots, k$$

$x_{ij} = X(\omega_{ij})$– значение признака на j-том элементе i-той группы

$$\bar{x}_i = \left(\sum_{j=1}^{n_i} x_{ij}\right) \Big/ n_i$$ – эмпирическое среднее в i-той группе или i-тое групповое среднее;

$$D_i(X) = \sigma_i^2 = \left(\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2\right) \Big/ n_i$$ –i-тая групповая дисперсия;

$$\bar{x} = \left(\sum_{i=1}^{k} \bar{x}_i n_i\right) \Big/ n = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$ –эмпирическое среднее во всей совокупности Ω или общее среднее.

Величина

$$\bar{\sigma}^2 = \left(\sum_{i=1}^{k} \sigma_i^2 n_i\right) \Big/ n$$ –называется средняя групповая дисперсия (это – средняя из дисперсий).

Величина

$$\delta^2 = \left(\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2\right) \Big/ n$$ –называется межгрупповая дисперсия (это – дисперсия средних).

Общая дисперсия или эмпирическая дисперсия признака в Ω

$$D(X) = \sigma^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}\right)^2$$

может быть представлена виде суммы $\sigma^2 = \bar{\sigma}^2 + \delta^2$, где первое слагаемое $\bar{\sigma}^2$– характеризует среднюю изменчивость признака в каждой группе $\Omega_1, \Omega_2, \ldots, \Omega_k$, а второе слагаемое δ^2 характеризует разброс групповых средних.

$$\eta = \frac{\delta^2}{\sigma^2}$$ –коэффициент детерминации.

Задачи для самостоятельного решения.

1. Данные о средних дисперсиях заработной платы двух групп рабочих приведены в таблице

Группа рабочих	Число рабочих	Средняя заработная плата одного рабочего в группе, р.	Дисперсия заработной платы
Работающие на одном станке	40	2400	180000
Работающие на двух станках	60	3200	200000

Найти общую дисперсию распределения числа рабочих по заработной плате.

Ответ: $\bar{X}$=2880; $\bar{\sigma}^2$=192000; δ^2=153600; D(X)= 345600.

2. Статистические данные о результатах экзамена в трех группах приведены в таблице:

№ группы	Число студентов	Средний балл	Среднее квадрат. отклонение
1	20	60	4
2	26	58	16
3	25	62	20

При проведении экзамена студенты случайным образом размещались в нескольких аудиториях. В одной из них находилось 22 студентов. Найти математическое ожидание и дисперсию среднего балла по результатам, полученным в данной аудитории.

Ответ: $\bar{X}$=59,9718; $\bar{\sigma}^2$=239,09859; δ^2=2,8724; D(X)= 241,971; D($\bar{X}$) = 7,699.

3. Статистические данные о результатах экзамена в трех группах приведены в таблице:

№ группы	Число студентов	Средний балл	Среднее квадрат. отклонение
1	27	64	12
2	25	60	8
3	24	62	14

При проведении экзамена студенты случайным образом размещались в нескольких аудиториях. В одной из них находилось 26 студентов. Найти математическое ожидание и дисперсию среднего балла по результатам, полученным в данной аудитории.

Ответ: $\bar{X}$=62,053; $\bar{\sigma}^2$=134,106; δ^2=2,688; D(X)= 136,794; D($\bar{X}$) = 3,509.

Интервальные характеристики признака

Часто, особенно, если объем совокупности велик, строится интервальный вариационный ряд. Для этого диапазон значений признака разбивается на несколько интервалов, и указывают количество вариант, попавших в каждый интервал. Получаем интервальное статистическое распределение вида:

интервалы	(a_1, a_2)	(a_2, a_3)	…	(a_k, a_{k+1})
частоты	n_1	n_2	…	n_k

$$\sum_{i=1}^{k} n_i = n$$

Согласно формуле Стэрджеса рекомендуемое число интервалов

$$L \approx 1 + 3{,}322 \lg n = 1 + 3{,}322 \frac{\ln n}{\ln 10},$$

а оптимальная величина интервала вычисляется по формуле

$$h \approx k_i = \frac{x_{max} - x_{min}}{1 + 3{,}322 \lg n} = \frac{x_{max} - x_{min}}{L}.$$

Формула Стэрджеса — оценочная, поэтому её можно и нужно округлять. Значение 3,322 можно заменить на 3,3 или даже на 3. Ширину интервала удобно выбирать или целую или с минимальным количеством значащих цифр.

Обозначив x_i^*– середину i-того интервала, получим формулы эмпирических интервальных характеристик:
интервальное среднее–

$$\bar{x}^* = \left(\sum_{i=1}^{k} n_i x_i^* \right) \Big/ n \ ;$$

интервальная дисперсия –

$$D^*(X) = \frac{1}{n} \sum_{i=1}^{k} (x_i^* - \bar{x}^*)^2 \, n_i \ ;$$

интервальное стандартное отклонение –

$$\sigma^*(X) = \sqrt{D^*(X)}$$

В типичном случае с одинаковыми интервалами длины h применяется поправка Шеппарда

$$\sigma^2 \approx \sigma^{*^2} - \frac{h^2}{12}.$$

Поправка Шеппарда чаще всего применяется в тех случаях, когда эмпирическая функция закона распределения хорошо приближается функцией распределения нормального закона.

Графическим изображением интервального статистического распределения является гистограмма. Для ее построения на оси OX откладывают интервалы шириной h, на каждом интервале строят прямоугольник высотой n_i/h. Эта величина называется плотностью частоты. Или строится прямоугольник высотой $n_i/(n \cdot h)$– это плотность частости.

Площадь гистограммы частот равна объему выборки, а площадь гистограммы частостей равна 1.

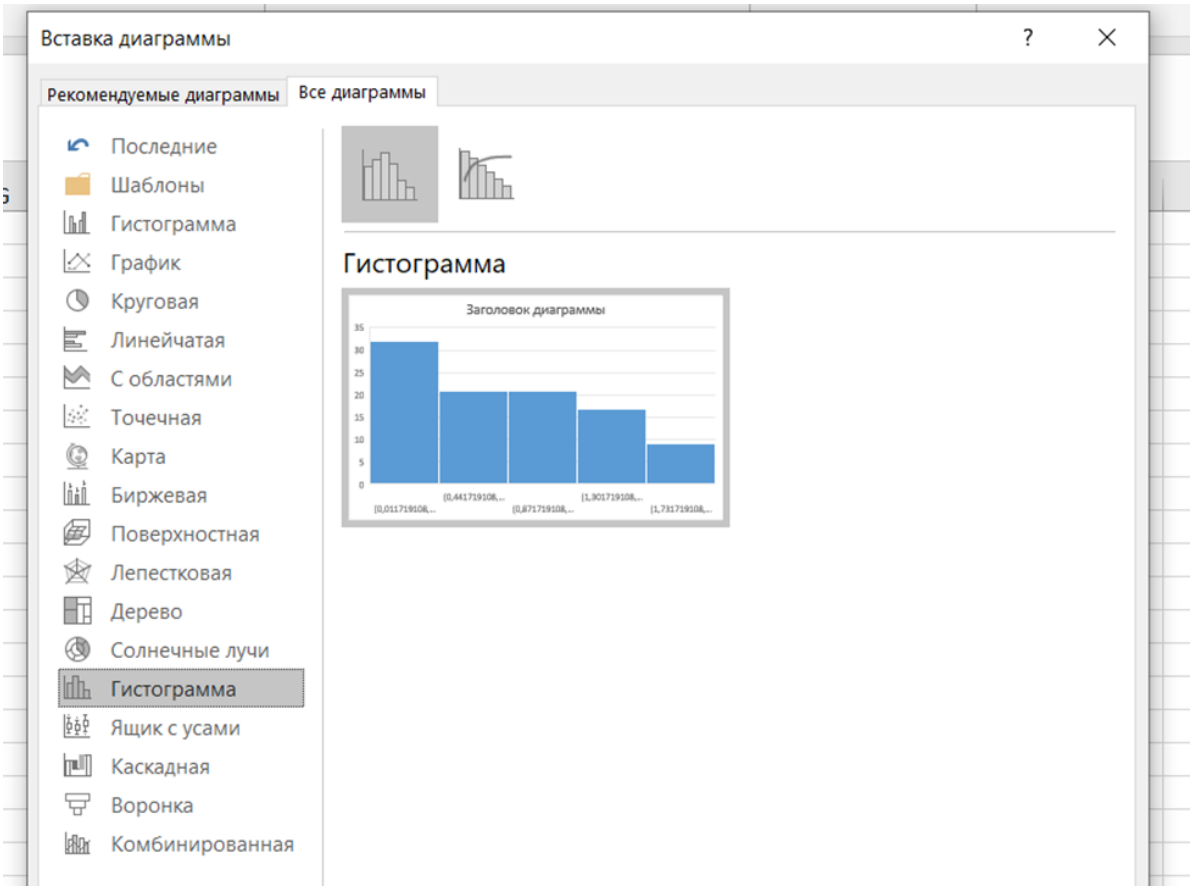
Гистограмма плотности частостей является статистическим аналогом функции плотности вероятности.

Замечание.

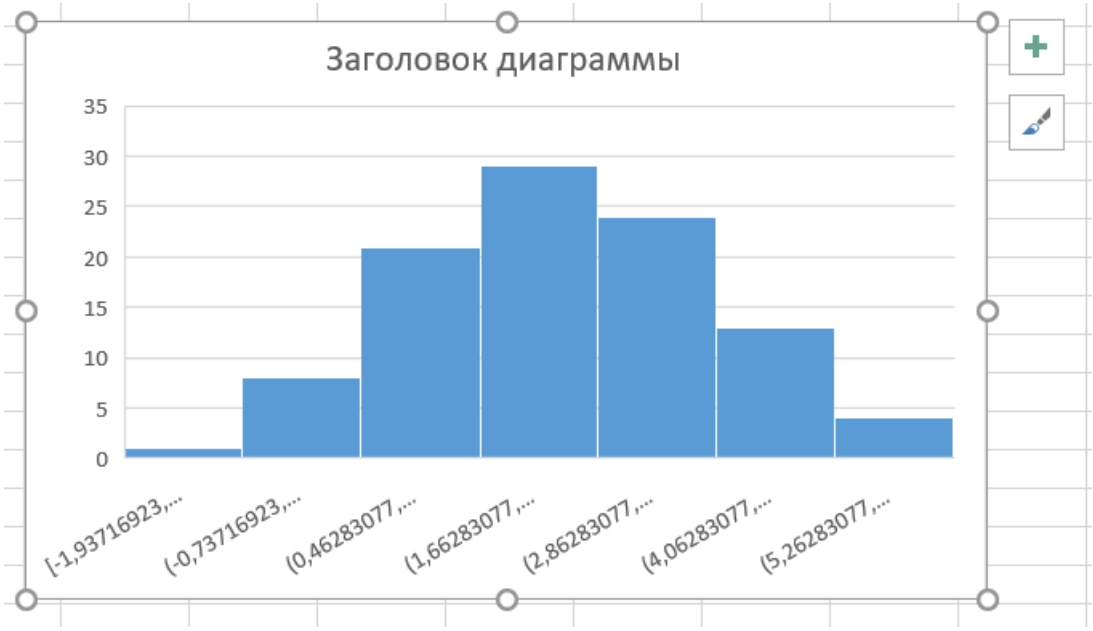
За начало первого интервала рекомендуется брать величину

$$x_{\text{нач.}} = x_{min} - \frac{h}{2}.$$

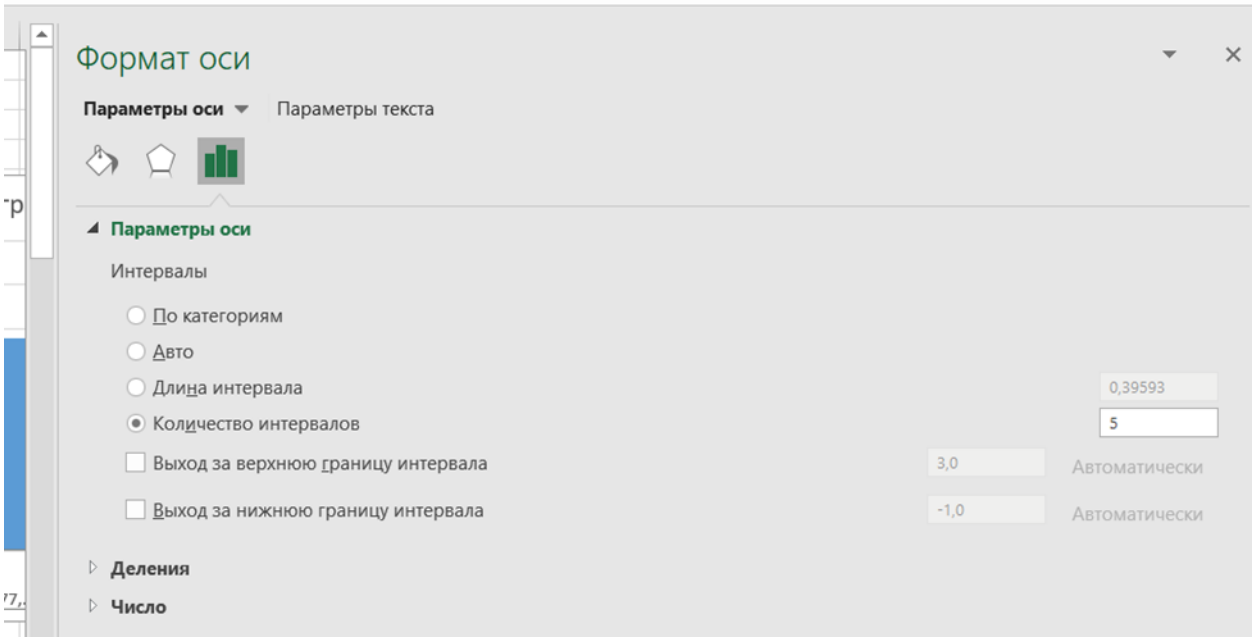
Построение гистограммы с помощью Пакета анализа - лишь один из способов визуализации данных. Более наглядным способом визуализации данных выборки является построение статистической гистограммы с помощью мастера диаграмм (статистическая гистограмма есть только в Excel 2016 и более поздних версиях). Для удобства добавим значок Мастера диаграмм на панель быстрого доступа.



Для построения гистограммы нужно выделить область данных и нажать кнопку Гистограмма в списке Мастера диаграмм.



В уже построенной гистограмме можно делать любое количество интервалов и любую длину интервала. Для этого щелкнуть правой кнопкой мыши по ее нижней строке и их открывшегося списка выбрать «Формат оси». Там открывается вкладка и есть графы «Длина интервала» и «Количество интервалов».

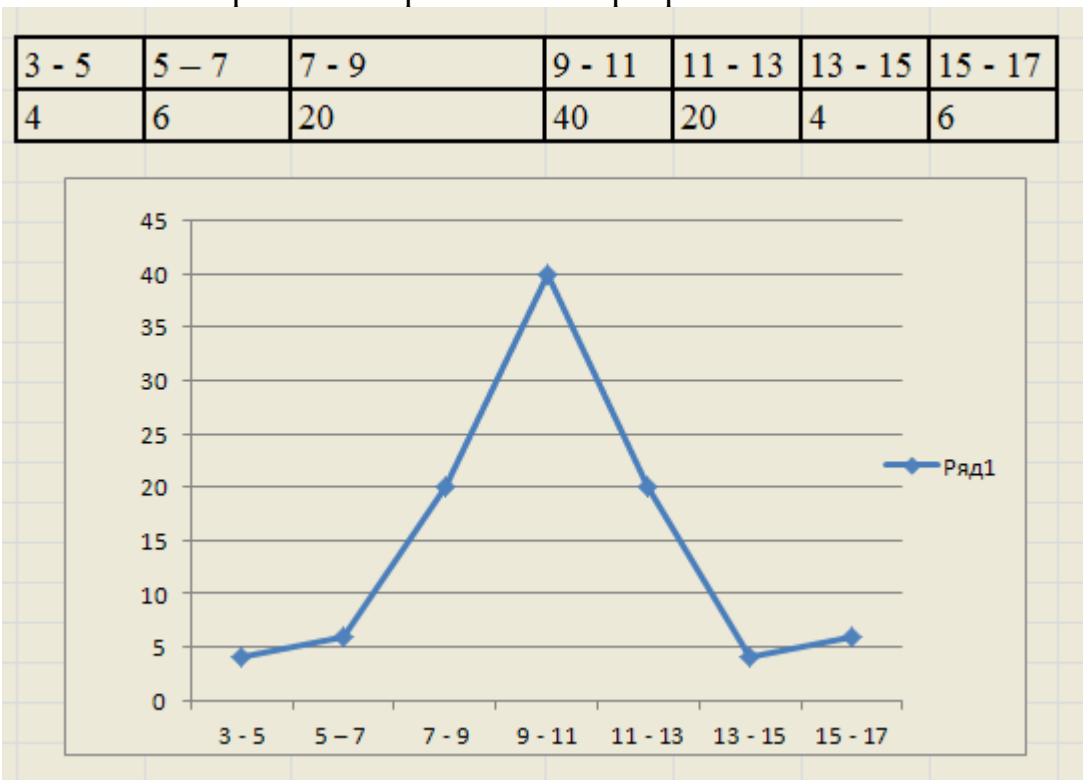


Чтобы сделать границы интервалов целыми, надо поставить галочки в полях «Выход за верхнюю границу интервала» и «Выход за нижнюю границу интервала». Затем щелкнуть по надписи «Число» и в поле «Категории», щелкнув по стрелке, из открывшегося списка выбрать «Числовой»; в поле «Число десятичных знаков» поставить ноль. После этого границы интервалов на гистограмме станут целыми.

В правом верхнем углу гистограммы нажмите на большой +. Из открывшегося списка выберите «Метки данных» и появятся все числовые значения над столбцами. Можно там же добавлять названия осей. Гистограмма теперь смотрится гораздо «опрятнее».

Построение полигона частот покажем на примере нижеследующей задачи 4. Табличку с данными копируем в файл Excel. Вызываем Мастер диаграмм → «График» → «График с маркерами». Или по-другому «Диаграммы» → «Точечные» → «С прямыми отрезками и маркерами».

3 - 5	5 – 7	7 - 9	9 - 11	11 - 13	13 - 15	15 - 17
4	6	20	40	20	4	6



Замечание. В верхней строке таблички данных формат ячеек поставим «текстовый».

Задачи для самостоятельного решения.

1. Время решения контрольной работы студентами (в минутах):
38 60 41 51 33 42 45 21 53 60
68 52 47 46 49 49 14 57 54 59
77 47 28 48 58 32 42 58 61 30
61 35 47 72 41 45 44 55 30 40
67 65 39 48 43 60 54 42 59 50

Найти размах выборки, число и длину равных интервалов, если первый интервал 14 - 23. Составить группированную выборку, построить гистограмму и полигон, вычислить эмпирические характеристики.

Ответ:

Интервалы	14-23	23-32	32-41	41-50	50-59	59-68	68-77
w_i	0,00	0,007	0,013	0,038	0,022	0,020	0,007

x=48,34; D=160,6244; σ=12,673768; Me X = 48; Mo X = 42; 47; 60.

2. Задано интервальное распределение признака

1 - 5	5 – 9	9 - 13	13 - 17	17 - 21
10	20	50	12	8

вычислить эмпирические характеристики., построить гистограмму частот и относительных частот; построить кумуляту.

Ответ: x*=10,52; D*=16,4096; σ^2=15,0763.

3. Задано интервальное распределение признака

2 - 7	7 – 12	12 - 17	17 - 22	22 - 27
5	10	25	6	4

вычислить эмпирические характеристики. построить гистограмму частот и относительных частот; построить кумуляту.

Ответ: x*=13,9; D*=25,64; σ^2=23,5567.

4. Задано интервальное распределение признака

3 - 5	5 – 7	7 - 9	9 - 11	11 - 13	13 - 15	15 - 17
4	6	20	40	20	4	6

вычислить эмпирические характеристики построить гистограмму частот и относительных частот;

построить кумуляту.

Ответ: x*=10,04; D*=6,7984; σ^2=6,6317.

5. На основе интервального вариационного ряда темпов роста производства предприятий легкой промышленности

Темп роста объема проиводства	102 – 104	104 - 106	106 - 108	108 - 110	110 - 112
Количество предприятий	7	14	13	9	7

вычислить эмпирические характеристики .

Ответ: 106 ,8; 6,28; 2,506.

6. Значение признака X в генеральной совокупности задано следующей таблицей:

Значение	3 - 23	23 - 43	43 - 63
частоты	20	60	20

Из этой совокупности извлекается повторная выборка объема 25. Пусть $\overline{x}_0$ - генеральное, а $\overline{X}$ - выборочное среднее. Найдите среднеквадратичную ошибку в приближенном равенстве $\overline{x}_0 \approx \overline{X}$. При вычислении генеральной дисперсии используйте поправку Шеппарда.

Ответ: 0,869.

7.Сгенерируйте выборку нормального распределения объема 1678 с параметрами 4,5 и 0,8 и постройте для нее статистическую гистограмму так , чтобы границы интервалов были целыми числами.

8.Сгенерируйте выборку равномерного распределения объема 880 между -12,14 и 22,4 . и постройте для нее статистическую гистограмму так , чтобы границы интервалов были целыми числами.

Выборочный метод

В математической статистике понятие генеральной совокупности трактуется как совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий.

Понятие генеральной совокупности в определенном смысле аналогично понятию случайной величины.

Тот набор объектов (наблюдений), который случайным образом был зафиксирован в процессе наблюдений, называется выборочной совокупностью или просто случайной выборкой. В дальнейшем будем говорить просто «выборка».

Сущность выборочного метода состоит в том, чтобы по выборке выносить суждение о свойствах генеральной совокупности.

Выборка называется репрезентативной (представительной) если она достаточно хорошо отражает свойства генеральной совокупности.

Основной недостаток выборочного метода – *неустранимая* ошибка репрезентативности, связанная с тем, что мы выносим суждение о целом по его части.

Повторные и бесповторные выборки из конечной генеральной совокупности

Есть два способа получения выборки:

1) повторная выборка, когда каждый элемент, случайно отобранный и исследованный, возвращается в генеральную совокупность и может быть отобран повторно;
2) бесповторная выборка, когда отобранный элемент не возвращается в генеральную совокупность.

Пусть из генеральной совокупности Ω объема N извлекается выборка $\widetilde{\Omega}$ объема n. Пусть X– некоторый признак, имеющий количественную природу (например, рост человека, его вес, размер обуви и т.д.).
Обозначим $x_{01}, x_{02}, \dots, x_{0N}$ – значения признака X в генеральной совокупности. Это – числа, совокупность которых составляет множество возможных значений признака. $X_1, X_2, \dots, X_n$– значения признака в выборке, они рассматриваются как случайные (от выборки к выборке) величины, совокупность возможных значений которых принадлежит множеству возможных значений признака.

Определение.

Генеральными характеристиками признака X называются его эмпирические характеристики в генеральной совокупности.

$\bar{x}_0$–генеральное среднее равное $\dfrac{1}{N}(x_{01} + \cdots + x_{0N})$.

$D(X) = \dfrac{1}{N}\sum_{i=1}^{N}(x_{0i} - \bar{x}_0)^2$ – генеральная дисперсия.

$\sigma(X)$ – это числа.

Выборочными характеристиками признака X называются его эмпирические характеристики в выборочной совокупности.

$\bar{X} = \dfrac{1}{n}(X_1 + \cdots + X_n)$ –выборочное среднее;

$\widehat{D}(X) = \hat{\sigma}^2(X) = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ –выборочная дисперсия – случайные величины.

Важнейшей задачей выборочного метода является оценка характеристик генеральной совокупности по результатам выборки.

Теоретическую основу применимости выборочного метода составляет закон больших чисел, согласно которому при неограниченном увеличении объема выборки практически достоверно, что выборочные характеристики как угодно близко приближаются (сходятся по вероятности) к параметрам генеральной совокупности.

ТЕОРЕМА.

Пусть $X_1, \dots, X_n$– значения признака X в выборке, $\bar{x}_0$– генеральное среднее, $D(X)$– генеральная дисперсия. Тогда для выборочного среднего $\bar{X}$ имеем:

1) в случае как повторной так и бесповторной выборки

$$E(\bar{X}) = \bar{x}_0.$$

2) в случае повторной выборки

$$D(\bar{X}) = \frac{D(X)}{n}.$$

3) в случае бесповторной выборки

$$D(\bar{X}) = \frac{D(X)}{n} \cdot \frac{N-n}{N-1},$$

где N – объем генеральной совокупности.

Определение.

Средней ошибкой выборки (или ошибкой средней) назовем среднеквадратичную ошибку в равенстве $\bar{x}_0 \approx \bar{X}$, равную $\sigma(\bar{X})$.

$$\sigma(\bar{X}) = \sqrt{D(\bar{X})}.$$

Для повторной и бесповторной выборки выполняется неравенство

$$\sigma(\bar{X}) \leq \frac{\sigma(X)}{\sqrt{n}},$$

поэтому при $n \to \infty$,$\sigma(\bar{X}) \to 0$.

Из теоремы можно получить

Следствие.

Пусть $X_1, \dots, X_n, Y_1, \dots, Y_n$ - значения признаков X и Y в выборке объема n, $cov(X,Y)$– ковариация признаков X и Y в генеральной совокупности объема N. Тогда для ковариации выборочных средних справедливы соотношения:

- в случае повторной выборки

$$cov(\bar{X}, \bar{Y}) = \frac{cov(X,Y)}{n};$$

- в случае бесповторной выборки

$$cov(\bar{X}, \bar{Y}) = \frac{cov(X,Y)}{n} \cdot \frac{N-n}{N-1}.$$

Выборочная доля признака

Одновременно с распределением признака X в генеральной совокупности

x_1	x_2	…	x_k
N_1	N_2	…	N_k

$, N_1 + \dots + N_k = N$

рассмотрим распределение признака X в выборке

x_1	x_2	…	x_k
n_1	n_2	…	n_k

$, n_1 + \dots + n_k = n.$

Определение.

Отношение $p_i = N_i/N$ называется генеральной долей значения x_i признака X. Отношение $\hat{p}_i = n_i/n$ называется выборочной долей значения x_i признака X.

ТЕОРЕМА.

Пусть p – генеральная, а $\hat{p}$ – выборочная доля значения x_i признака X; $q = 1 - p$. Тогда:

1) для повторной или бесповторной выборки

$$E(\hat{p}) = p;$$

2) для повторной выборки

$$D(\hat{p}) = \frac{pq}{n};$$

3) для бесповторной выборки

$$D(\hat{p}) = \frac{pq}{n} \cdot \frac{N-n}{N-1} \approx \frac{pq}{n}\left(1 - \frac{n}{N}\right).$$

Следовательно,

$$\sigma(\hat{p}) \approx \sqrt{\frac{pq}{n}\left(1 - \frac{n}{N}\right)}\ .$$

Статистические оценки параметров распределения

Пусть генеральное распределение признака X зависит от некоторого параметра $\theta \in \Theta$.

Статистической оценкой параметра θ называется функция $\widehat{\Theta} = \widehat{=\Theta}(X_1, X_2, \dots, X_n)$ результатов наблюдений, предназначенная для приближенного вычисления неизвестного параметра θ по выборочным значениям признака.

Статистическая оценка называется точечной, если она выражается одним числом.

Для того чтобы статистические оценки были «хорошими», то есть давали хорошие приближения они должны удовлетворять определенным требованиям.

Определение.

Величина $\hat{\theta}$ называется несмещенной оценкой параметра θ, если $E(\hat{\theta}) = \theta$, то есть ее математическое ожидание равно оцениваемому параметру. В противном случае оценка называется смещенной.

Выполнение требования несмещенности гарантирует отсутствие систематических ошибок при оценивании.

Определение.

Оценка $\hat{\theta}$ называется эффективной в некотором классе оценок, если в этом классе при фиксированном объеме выборки она имеет наименьшую среднюю квадратическую ошибку.

Определение.

Оценка $\hat{\theta}$ называется состоятельной, если она удовлетворяет закону больших чисел, то есть сходится по вероятности к оцениваемому параметру

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1 \quad \text{для любого } \varepsilon > 0\ .$$

ТЕОРЕМА (достаточное условие состоятельности)

Для того, чтобы оценка $\hat{\theta}$ была состоятельной достаточно, чтобы выполнялись условия:

1) $\lim\limits_{n\to\infty} D(\hat{\theta}) = 0$,

2) $\lim\limits_{n\to\infty} E(\hat{\theta}) = \theta$.

Для несмещенных оценок достаточно, только первого условия, поскольку второе выполняется автоматически.

Пример – теорема Бернулли:

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1 \quad \text{для любого } \varepsilon > 0 .$$

Из нее следует, что относительная частота $\widehat{p} = \frac{m}{n}$ является состоятельной оценкой вероятности, а также несмещенной и эффективной в классе линейных оценок ($p = c_1x_1 + \ldots.+ c_nx_n$).

Кроме того,

$$D\left(\frac{m}{n}\right) = D(\hat{p}) = \frac{pq}{n} .$$

Замечание.

Отметим, что на практике довольно часто используются смещенные и неэффективные статистические оценки, тогда как несостоятельные оценки обычно не применяются.

ТЕОРЕМА.

Пусть $X_1, \ldots, X_n$– выборка из генеральной совокупности (повторная или бесповторная). Тогда выборочное среднее

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

является несмещенной и состоятельной оценкой генеральной средней

$$\bar{x}_0 = \frac{1}{N}\sum_{i=1}^{N} x_{0i} .$$

При нормальном распределении генеральной совокупности эта оценка является и эффективной.

ТЕОРЕМА.

Пусть $X_1, \dots, X_n$– выборка (повторная или бесповторная) из генеральной совокупности. Тогда выборочная дисперсия

$$\widehat{D}(X) = \hat{\sigma}^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 ,$$

является смещенной оценкой генеральной дисперсии

$$D(X) = \frac{1}{N}\sum_{i=1}^{N}(x_{0i} - \bar{x}_0)^2 ,$$

так как можно доказать, что

$$E\left(\widehat{D}(X)\right) = \frac{n-1}{n}D(X).$$

Поэтому выборочную дисперсию $\widehat{D}(X)$ исправляют, умножая на $\frac{n}{n-1}$ и получая формулу $S^2 = \frac{n}{n-1}\widehat{D}(X)$ исправленной выборочной дисперсии.

Следствие.

Исправленная выборочная дисперсия

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

является несмещенной состоятельной оценкой генеральной дисперсии.

! (эффективной не является). Несмещенную оценку S^2 удобно вычислять по формуле $S^2 = \frac{\sum_{i=1}^{k} n_i {X_i}^2 - n\bar{X}^2}{n-1}$.

Величина S^2 имеет $n-1$ степень свободы.

Если генеральное среднее $\mu = E(X)$ известно, используется другая несмещенная состоятельная оценка

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 .$$

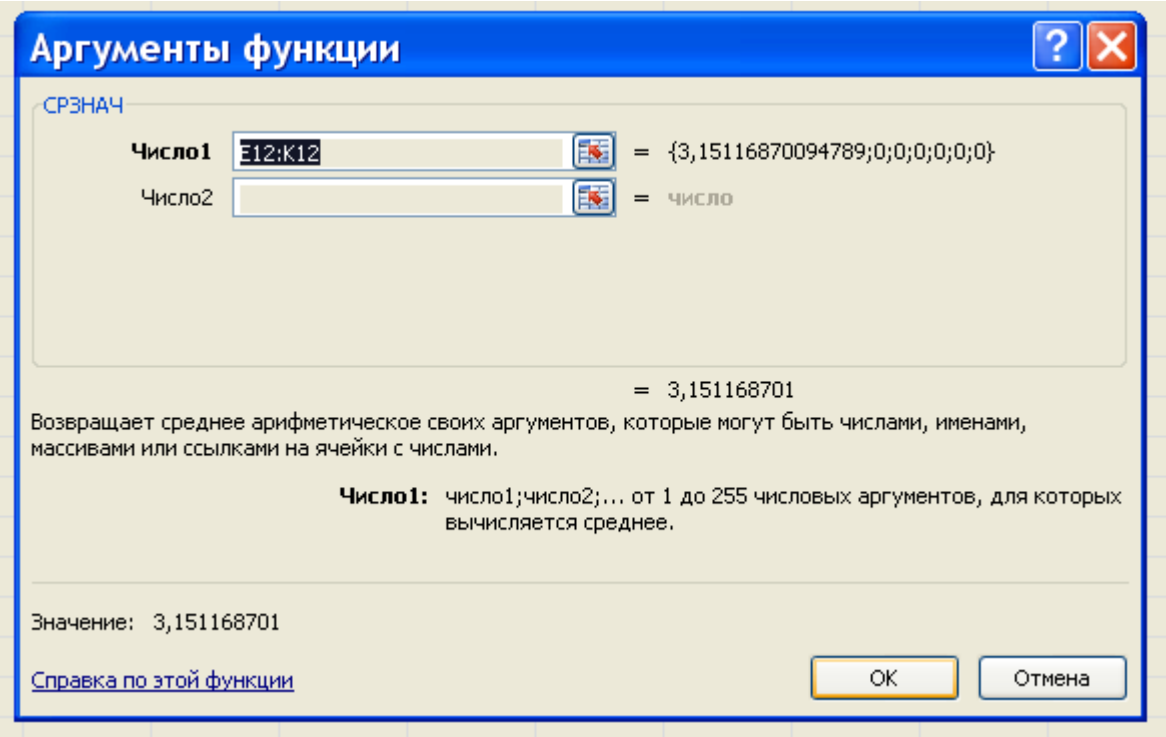
Замечание.

Если S_0^2 характеризует вариацию значений признака относительно генеральной средней μ, то S^2– относительно выборочной средней $\bar{X}$.

Вычисление оценок параметров с помощью Excel

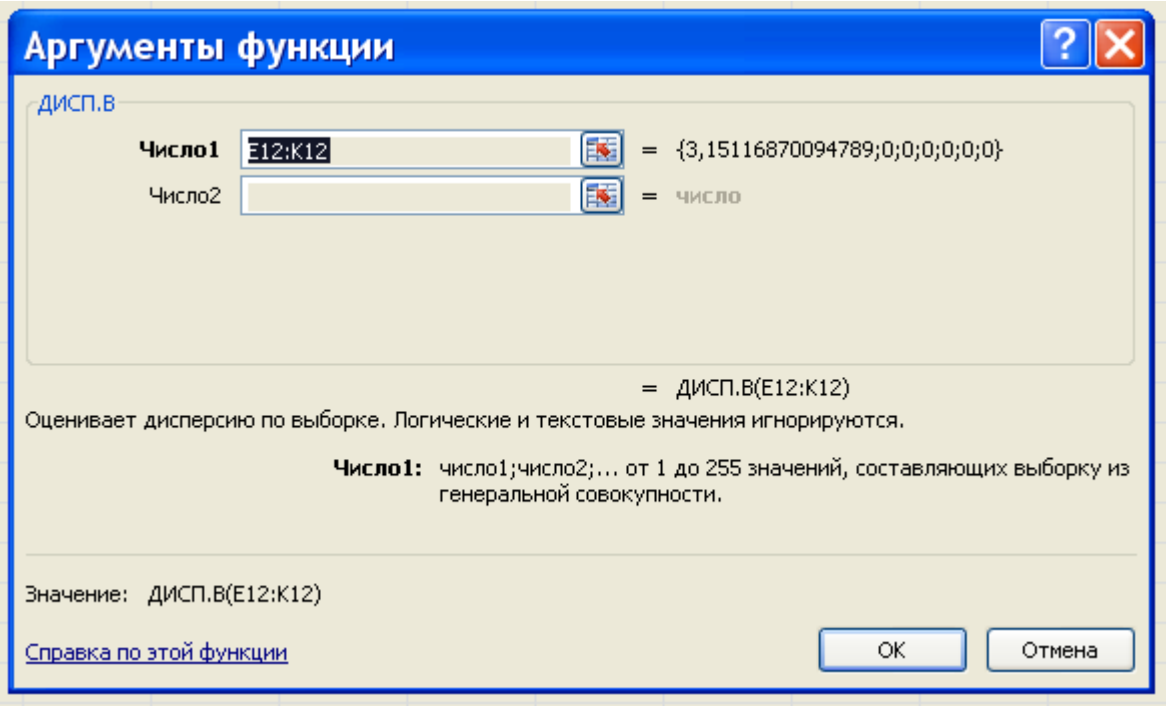
Оценки параметров можно вычислять с помощью функций Excel.

Функция СРЗНАЧ (массив) вычисляет выборочное среднее $\bar{X}$.



Функция ДИСП.В (массив) вычисляет исправленную выборочную дисперсию S^2.

Функция ДИСП это функция ДИСП.В для старых версий Excel.



Функция ДИСП.Г (массив) вычисляет эмпирическую дисперсию, рассматривая выборку как генеральную совокупность.

Функция ДИСПР это функция ДИСП.Г для старых версий Excel.

Функция ДИСПРА – генеральная дисперсия с учетом логических и текстовых значений.

Функция СТАНДОТКЛОН вычисляет исправленное выборочное среднее квадратическое отклонение S = $\sqrt{S^2}$

Функция МЕДИАНА (диапазон) вычисляет медиану.

Функция МОДА.ОДН(диапазон) - вычисляет единственное значение моды.

Функция МОДА.НСК (диапазон) дает все значения моды данного вертикального диапазона. Только при этом для вставки функции нужно выделить не одну ячейку, а несколько, и нажимать не Enter, а Ctrl + Shift + Enter . (сначала Ctrl + Shift потом Enter).

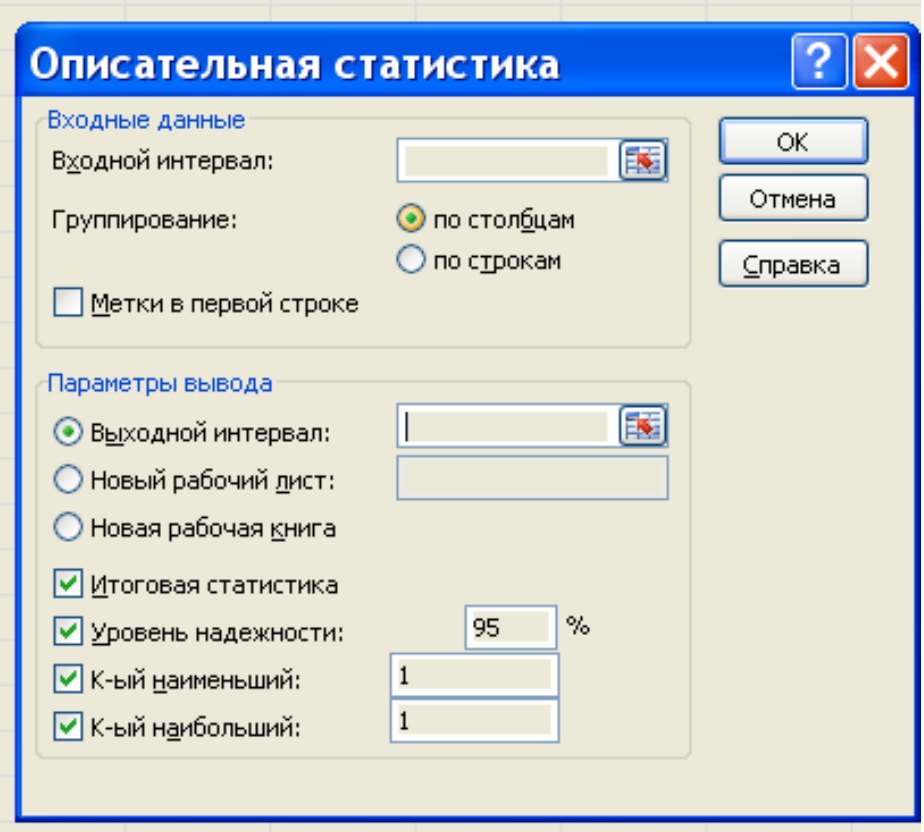
Если диапазон горизонтальный, то используется функция ТРАНСП(МОДА.НСК).

Функция ЭКСЦЕСС (массив) вычисляет эксцесс множества данных.

Функция СКОС(массив) вычисляет асимметрию множества данных.

Функция СКОС.Г (массив) вычисляет асимметрию по генеральной совокупности.

Для нахождения всех этих характеристик сразу нужно использовать **«Описательную статистику»** из «Пакета анализа».



В «Описательной статистике» флажок в поле «Метки» надо ставить, если первая строка входного интервала содержит заголовок.

Флажок «Уровень надежности» (этот доверительная вероятность в процентах) устанавливается, когда нужно найти доверительный интервал для математического ожидания. В таблице результатов появится число, равное половине длины доверительного интервала.

Поля «к-ый наименьший» и «к-ый наибольший» это элементы упорядоченной выборки к-ый с начала и к-ый с конца. В эти поля надо ввести желаемое значение. По умолчанию там стоит 1. Тогда это просто минимум и максимум.

Столбец1	
Среднее	0,945690479
Стандартная ошибка	0,219780715
Медиана	1,045981778
Мода	#Н/Д
Стандартное отклонение	2,197807153
Дисперсия выборки	4,830356281
Эксцесс	0,99923665
Асимметричность	0,03447329
Интервал	12,90056389
Минимум	-5,046029739
Максимум	7,854534149
Сумма	94,56904789
Счет	100
Наибольший(1)	7,854534149
Наименьший(1)	-5,046029739
Уровень надежности(95,0%)	0,436092621

«Интервал» - это размах выборки, «Счет» – это объем выборки.

Задачи для самостоятельного решения.

1. Записать в виде вариационного ряда и определить выборочные среднее, дисперсию, моду и медиану для выборки 5, 6, 8, 2, 3, 1, 4,1 .

Ответ:

$$\bar{x} = 3{,}75, S^2 = 6{,}21, MoX = 1, MeX = 3{,}5$$

2. Записать в виде вариационного ряда и определить выборочные среднее, дисперсию, моду и медиану для выборки 3,1; 3,0; 1,5; 1,8; 2,5; 3,1; 2,4; 2,8; 1,3.

Ответ:

$$\bar{x} = 2{,}39, S^2 = 0{,}49, MoX = 3{,}1, MeX = 2{,}5$$

3. В итоге пяти измерений длины стержня одним прибором (без систематических ошибок) получены следующие результаты (в мм): 92; 94; 103; 105; 106. Найти выборочную среднюю длину стержня выборочную и исправленную дисперсии ошибок прибора.

Ответ: $\overline{X}$=100; $D_в$ = 34; S^2=42,5.

4. Найти дисперсию и исправленную дисперсию по данному распределению выборки:

x_i	1	2	3	4
n_i	20	15	10	5

Ответ: $\overline{X}$=2; S^2=1,0204.

5. В 28 независимых испытаниях случайная величина X значение 4 приняла 17 раз, а значение 6 – 11 раз. Найти несмещенную оценку дисперсии.

Ответ: $\overline{X}$=4,786; S^2=0,987.

6. Даны результаты 8 независимых измерений одной и той же величины прибором, не имеющим систематических ошибок: 365, 374, 317, 423, 385, 404, 376, 383 м. Найти несмещенную оценку дисперсии ошибок измерений, если истинная длина неизвестна.

Ответ: $\overline{X}$=378,375; S^2=954,8393.

7. Даны результаты 8 независимых измерений одной и той же величины прибором, не имеющим систематических ошибок: 387, 378, 402, 365, 388, 399, 372, 361 м. Найти несмещенную оценку дисперсии ошибок измерений, если истинная длина известна и равна 375 м.

Ответ: 242

8. Измерен рост (с точностью до сантиметра) 30 наудачу отобранных студентов. Результаты измерений:

178, 160, 154, 183, 155, 153, 167, 186, 163, 155, 157, 175, 170, 166, 159, 173, 182, 167, 171, 169, 179, 165, 156, 179, 158, 171, 175, 173, 164, 172.

Найти выборочную среднюю величину роста студента и исправленную выборочную дисперсию.

Ответ: $\overline{X}$=167,8333; S^2=89,59195.

9. На обслуживание каждого из 4 посетителей районного отделения Пенсионного фонда затрачено, соответственно, 51 мин., 49 мин., 52 мин. и 48 мин. По этой выборке требуется вычислить несмещенные оценки математического ожидания, дисперсии, среднего квадратического отклонения времени обслуживания посетителей, а также выборочные центральные моменты третьего и четвертого порядков.

Ответ: 50; 3,333; 1,8258; 0; 8,5.

10. Для оценки стрелковой подготовки личного состава батальона было отобрано 50 человек, каждый из которых произвел 10 выстрелов по мишени. Результаты стрельбы представлены в таблице :

Число попаданий	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Число стрелков	*1*	*0*	*2*	*1*	*4*	*6*	*8*	*11*	*11*	*2*	*4*

Построить полигон относительных частот , эмпирическую функцию распределения , вычислить выборочные среднее и исправленную выборочную дисперсию .

Ответ: $\overline{X}$=6,48; S^2=4,58122.

11. Студентам был предложен тест из 24 вопросов. По числу правильных ответов студенты распределились следующим образом:

Количество верных ответов	10-12	12-14	14-16	16-18	18-20	20-22	22-24
Количество студентов	2	4	8	12	16	10	3

Вычислить выборочные коэффициенты асимметрии и эксцесса данного распределения.

Ответ: As = - 0,39; Ex = - 0,30.

12. На одном из участков шоссе было проведено измерение средней скорости движения автомобилей. Результаты были сведены в следующую таблицу:

Скорость	61-65	65-69	69-73	73-77	77-81
Количество автомобилей	1	4	5	8	14

Вычислить выборочные коэффициенты асимметрии и эксцесса распределения.

Ответ: As = 0,89; Ex = 7,55.

13.Сгенерируйте выборку нормального распределения объема 1421 с параметрами 2,1 и 0,6 и вычислите для нее оценки параметров с помощью Описательной статистики.

14.Сгенерируйте выборку нормального распределения объема 728 с параметрами 0,2 и 2,8 и вычислите для нее оценки параметров с помощью Описательной статистики.

15.Сгенерируйте выборку равномерного распределения объема 564 между -9,6 и 11,48 . и вычислите для нее оценки параметров с помощью Описательной статистики.

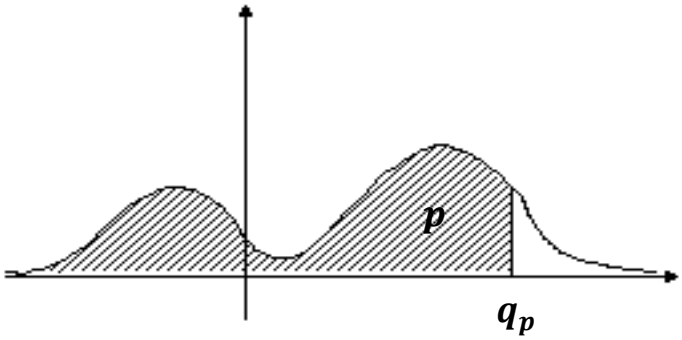
Вычисление квартилей. Диаграмма «Ящик с усами»

Квантилем порядка (уровня) p называется такое значение q_p абсолютно непрерывной случайной величины, что

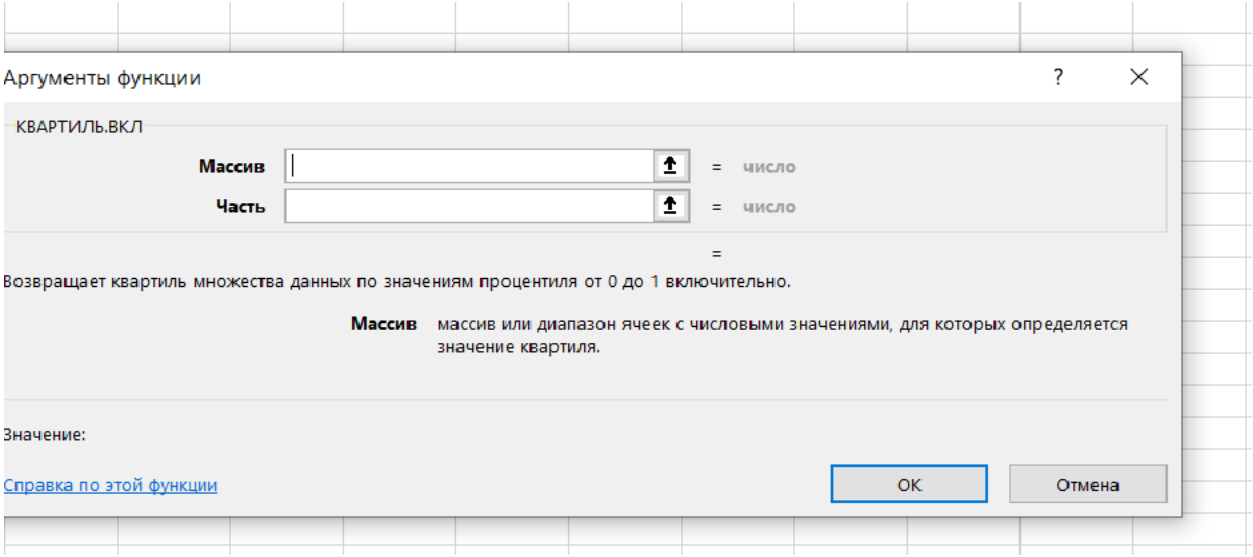
$$\int_{-\infty}^{q_p} f(x)dx = p\,.$$

С понятием квантиля тесно связано понятие α·100%-процентной точки – это квантиль порядка $1-\alpha$.

Пусть $\alpha = 0{,}05$; 5%-ная точка – это квантиль порядка 0,95.



Важными числовыми характеристиками являются также ***квартили*** распределений. Это квантили порядка 0,25 (первый квартиль), порядка 0,5 (второй квартиль или медиана), порядка 0,75 (третий квартиль). Их вычисляют с помощью функции КВАРТИЛЬ. ВКЛ (КВАРТИЛЬ в более ранних версиях Excel) или КВАРТИЛЬ.ИСКЛ.



В поле «Часть» нужно ввести целое число от 0 до 4.

Если ввести 0, то функция даст наименьшее значение массива;

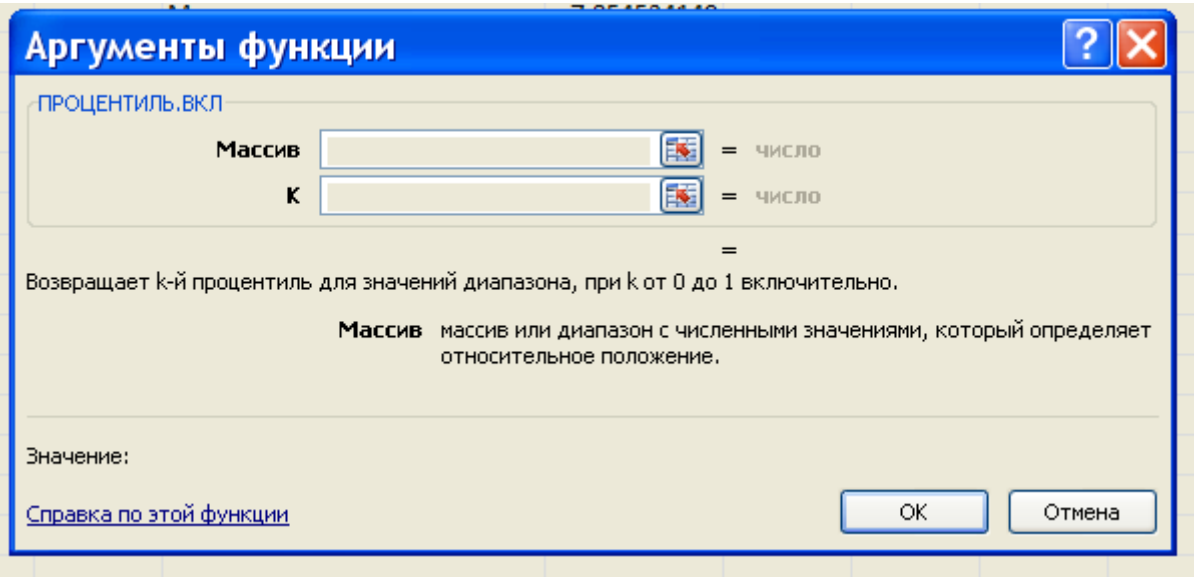
если число равно 1, даст левый квартиль $q_{0,25}$;

если число равно 2 , даст медиану $q_{0,5}$;

если число равно 3, то даст правый квартиль $q_{0,75}$;

если число равно 4, то даст наибольшее значение массива.

Функция КВАРТИЛЬ.ВКЛ является частным случаем функции ПРОЦЕНТИЛЬ.ВКЛ



С ее помощью также можно вычислить квартили , если в поле К вводить значения 0,25; 0,5 и 0,75. Она также вычисляет квантили любого порядка по выборке.

Межквартильным размахом IQR называется разность между третьим и первым квартилями. **$IQR = q_{0,75} - q_{0,25}$**. Это характеристика разброса значений в выборке

Выбросы - это элементы выборки, находящиеся **вне отрезка**

$[q_{0,25} - 1,5 \cdot IQR;\ q_{0,75} + 1,5 \cdot IQR]$.

Для удаления выбросов следует

а) использовать «Данные» → «Фильтр» → «Числовые фильтры»;

б) функция СЧЕТЕСЛИ**(диапазон, “<=”& значение).**

Кроме того, в Excel есть функции КВАРТИЛЬ.ИСКЛ и ПРОЦЕНТИЛЬ. ИСКЛ.

Замечание. В сообществе статистиков нет единого мнения, по какому алгоритму считать квартили. В Excel используется два алгоритма вычисления, две функции КВАРТИЛЬ.ВКЛ и КВАРТИЛЬ.ИСКЛ. Первая использует алгоритм инклюзивной медианы, вторая – эксклюзивной. Алгоритм расчета квартилей с помощью функции КВАРТИЛЬ.ИСКЛ дает значения чуть более далекие от медианы по сравнению с КВАРТИЛЬ.ВКЛ.

В Мастере диаграмм Excel 2016 и в более поздних версиях появился приобретающий в настоящее время широкую популярность новый вид визуализации ***статистических*** данных - **Диаграмма «ящик с усами».**

Ящик с усами – это диаграмма размаха.

В ящике с усами:

крестик в центре ящика – это среднее значение;

горизонтальная линия по центру ящика – это медиана;

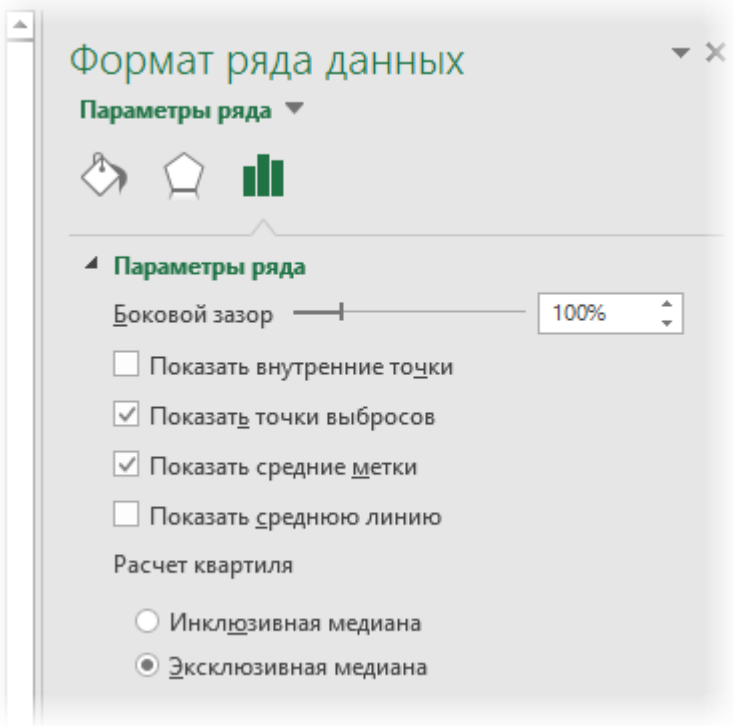
числа справа от ящика это медиана и первый и третий квартили, вычисленные с помощью функции КВАРТИЛЬ.ИСКЛ;

границы усов – это концы отрезка $[\max\{x_{min}; Q_{0,25} - 1{,}5IQR\}, \min\{x_{max}; Q_{0,75} + 1{,}5IQR\}]$, где x_{min} и x_{max} - это не минимум и максимум выборки, это минимальное и максимальное значения выборки, принадлежащие отрезку $[Q_{0,25} - 1{,}5IQR; Q_{0,75} + 1{,}5IQR]$,

если за границами «усов» есть значения выборки – это выбросы.

Чтобы справа от ящика стояли не КВАРТИЛИ.ИСКЛ, а КВАРТИЛИ.ВКЛ надо щелкнуть правой кнопкой мыши по самому ящику, из списка выбрать «Формат ряда данных»→ «Параметры ряда»

Справа появится панель настроек



Там по умолчанию стоит «Эксклюзивная медиана», надо ее заменить на «Инклюзивная медиана». Тогда в ящике с усами справа будут стоять значения квартилей, вычисленные с помощью функции КВАРТИЛЬ.ВКЛ

Замечание. **Инклюзивная медиана**: медиана включается в вычисления, если N (число значений в данных) — нечетное число.

Эксклюзивная медиана: медиана исключается из вычислений, если N (число значений в данных) — нечетное число (support.office.com).

Еще одной важной числовой характеристикой является коэффициент вариации.

Коэффициент вариации – это отношение стандартного отклонения S к среднему значению, выраженному в процентах:

$$V = \frac{S}{X} 100\%$$

Коэффициент вариации и среднеквадратическое отклонение могут использоваться как меры риска, например, при финансовых операциях.

Коэффициент вариации может быть использован при сравнении стандартных отклонений, которые вычислены по данным, имеющим различные средние.

Пример. Предположим, что цены на ценные бумаги широко колеблются. Инвестор, который покупает акции по низкой цене, а продает по высокой, имеет хороший доход. Однако если цены на акции падают ниже стоимости, по которой инвестор купил, то он теряет доход.

Чтобы оценить меру риска, инвестор может использовать коэффициент вариации и среднеквадратическое отклонение.

Какую информацию о степени риска может дать коэффициент вариации по сравнению со среднеквадратическим отклонением?

Допустим, за пять недель цены:

на акции 1 представлялись в виде \$57, 68, 64, 71, 62;

на акции 2 представлялись в виде \$12, 17, 8, 15, 13.

Средняя цена на акции 1 $\bar{X} = \$64.40$ и $S = \$4.84$.

Средняя цена на акции 2 $\bar{X} = \$13.00$ и $S = \$3.03$.

Со среднеквадратическим отклонением как мерой риска акции 1 более рискованные. Однако среднее арифметическое акций 1 почти в 5 раз больше среднего арифметического акций 2. Коэффициент вариации, используемый в данном случае, дает следующие результаты: $V_1 = \frac{4,84}{64,40} 100\% = 7,52\%$

$$V_2 = \frac{3,03}{13,00} 100\% = 23,31\%$$

Для акций 2 коэффициент вариации почти в три раза больше, чем коэффициент вариации для акций 1.

Используя коэффициент вариации в данном случае, можно сделать заключение, что покупать акции 2 более рискованно.

Методы нахождения точечных оценок

Метод моментов

Метод моментов для нахождения точечных оценок неизвестных параметров заданного распределения состоит в приравнивании теоретических моментов распределения соответствующим эмпирическим моментам, найденным по выборке. При этом: 1) теоретические моменты выражаются через неизвестные параметры; 2) находятся выборочные значения моментов; 3) приравниваются одни другим, составив уравнение или систему уравнений; 4) находятся решения и выбираются те, которые подходят по смыслу задачи. Обычно выражаются моменты низших порядков в количестве, которое соответствует количеству неизвестных параметров распределения.

(Отметим, что математическое ожидание является начальным моментом первого порядка, а дисперсия – центральным моментом второго порядка).

Метод моментов является наиболее простым методом оценки параметров. Он был предложен в 1894 году Пирсоном. Оценки метода моментов обычно состоятельны, однако их эффективность часто значительно меньше единицы.

Задачи для самостоятельного решения.

1. Случайная величина X распределена по закону Пуассона $P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$. Результаты 436 независимых наблюдений X отражены в таблице:

Значение X	0	1	2	3
Частота	198	134	80	24

Найдите методом моментов точечную оценку $\hat{\lambda}$.
Ответ: 0,83944.

2. Случайная величина X (время бесперебойной работы устройства) распределена по показательному закону с плотностью $f(x)=\lambda e^{-\lambda x}$ x ≥ 0. По эмпирическому распределению времени работы:

Время работы	0 -10	10 - 20	20 - 30	30 - 40
Число устройств	127	69	27	19

Найдите методом моментов точечную оценку $\hat{\lambda}$.
Ответ: $\bar{X}$=12,438; $\hat{\lambda}$ = 0,0804.

3. Случайная величина X (число семян сорняков в пробе зерна) распределена по закону Пуассона: $P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$,
где x_i — число семян в одной пробе. Ниже приведено распределение семян сорняков в n= 1000 пробах зерна (в первой строке указано количество семян сорняков в одной пробе ; во второй строке указано число таких проб):

x_i	0	1	2	3	4	5	6
n_i	405	366	175	40	8	4	2

Найдите методом моментов точечную оценку $\hat{\lambda}$.
Ответ: 0,9.

Некоторые законы распределения, используемые в математической статистике

Рассмотрим распределение некоторых случайных величин, представляющих собой функции нормальных случайных величин, используемые в математической статистике.

Распределение χ^2 (хи-квадрат или Пирсона)

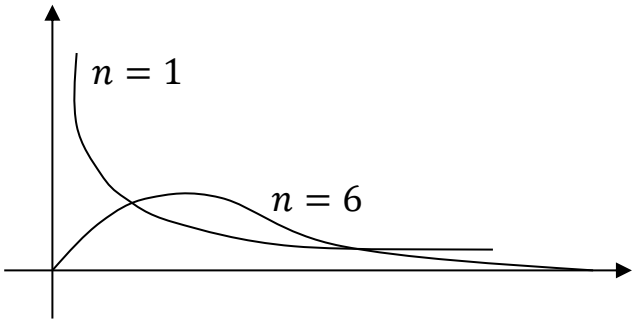
Определение.

Распределением χ^2 с $k = n$ степенями свободы называется распределение суммы квадратов n независимых случайных величин, распределенных по стандартному нормальному закону

$$Z = Y_1^2 + Y_2^2 + \cdots + Y_n^2.$$

$$Y_i \sim N(0,1) \qquad Z \sim \chi^2(n).$$

График плотности распределения χ^2 имеет вид:



Процентные точки распределения χ^2 затабулированы.

Отметим, что $\chi_\alpha^2(k) \approx \left(Z_\alpha + \sqrt{2k-1}\right)^2/2$ при $k > 30$.

Распределение Стьюдента (*t*– распределение)

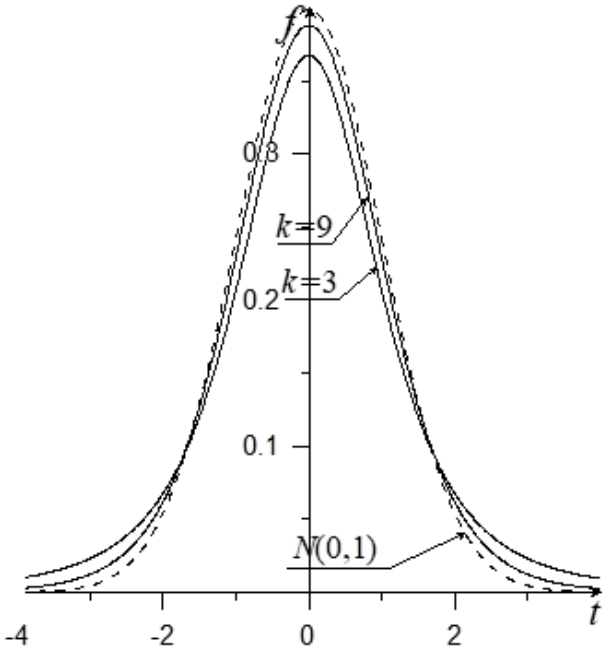
Определение.

Распределением Стьюдента или t – распределением с k степенями свободы называется распределение случайной величины

$$Z = \frac{X}{\sqrt{\frac{Y}{k}}},$$

где X и Y независимы, $X \sim N(0,1)$, $Y \sim \chi^2(k)$, обозначается $Z \sim t(k)$.

График плотности распределения Стьюдента имеет вид:



Кривая распределения Стьюдента по сравнению с нормальной более пологая.

Числовые характеристики распределения Стьюдента:

$$E(Z) = 0; \qquad D(Z) = \frac{k}{k-2} (\text{при } k > 2).$$

При $k \to \infty$ t – распределение приближается к стандартному нормальному. При $k > 30$ они уже практически неразличимы.

Процентные точки t – распределения затабулированы.

Отметим, что $t_\alpha(k) \approx Z_\alpha$ при $k > 30$; $t_{1-\alpha}(k) = -t_\alpha(k)$.

Распределение Фишера – Снедекора

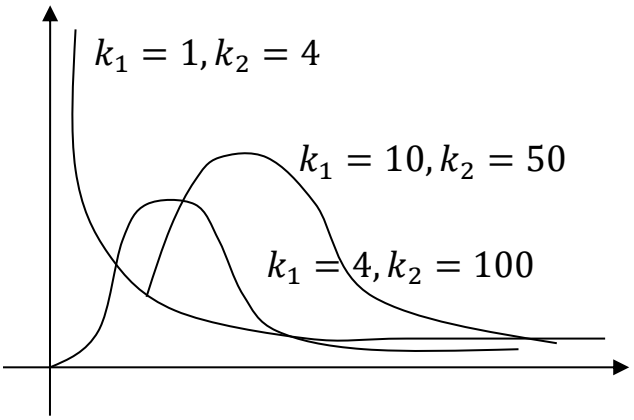
Определение.

Распределением Фишера или F – распределением с k_1,k_2 степенями свободы называется распределение случайной величины

$$Z = \frac{X/k_1}{Y/k_2},$$

где X и Y– независимые случайные величины, $X \sim \chi^2(k_1)$, $Y \sim \chi^2(k_2)$.

График плотности распределения Фишера имеет вид:



С ростом числа степеней свободы распределение Фишера приближается к нормальному.

Верхние процентные точки распределения Фишера затабулированы. Отметим, что

$$F_{1-\alpha}(k_1, k_2) = [F_\alpha(k_2, k_1)]^{-1}.$$

Замечание. Квантили и процентные точки этих распределений будем находить теперь не с помощью таблиц, а с помощью Мастера функций Excel.

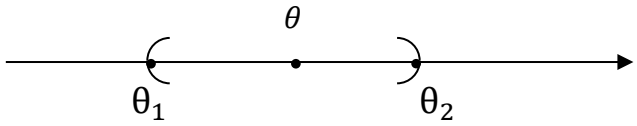
Интервальные оценки параметров

Точечная оценка параметра не позволяет ответить на вопрос, какую ошибку мы совершаем, принимая вместо точного значения неизвестного параметра θ его оценку $\hat{\theta}$. При выборке малого объема точечная оценка может давать грубую ошибку. Поэтому предпочтительнее пользоваться интервальными оценками.

Оценка неизвестного параметра называется интервальной, если она определяется двумя числами – концами интервала.

Определение.

Интервальной оценкой параметра θ называется числовой интервал (θ_1, θ_2), который с заданной вероятностью γ накрывает неизвестное значение параметра θ.



Границы интервала (θ_1, θ_2) находятся по выборочным данным и потому являются случайными величинами, в отличии от параметра θ– величины неслучайной, поэтому предпочтительнее говорить, что интервал (θ_1, θ_2) «накрывает», а не «содержит» значение θ.

Такой интервал (θ_1, θ_2) называется *доверительным*, а вероятность – *доверительной вероятностью, уровнем доверия или надежностью оценки.* Величина $\alpha = 1 - \gamma$ называется *уровнем значимости*.

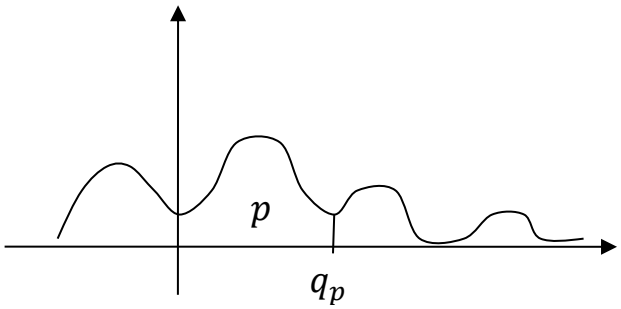
Величина доверительного интервала существенно зависит от объема выборки n (уменьшается с ростом n) и от значения доверительной вероятности γ (увеличивается с приближением γ к единице).

Очень часто (но не всегда) доверительный интервал выбирается симметричным относительно несмещенной точечной оценки параметра $\hat{\theta}$ параметра θ, то есть $(\hat{\theta} - \varepsilon, \hat{\theta} + \varepsilon)$. Величина $\varepsilon > 0$ характеризует точность оценки.

Величина γ выбирается заранее, ее выбор зависит от конкретно решаемой задачи.

Если q_p – квантиль распределения порядка p, то он также называется верхней $100 \cdot (1 - p)$-процентной точкой распределения.

Если $p = 0{,}95$, то q_p– верхняя 5%-ная точка распределения.



Доверительные интервалы для параметров нормального распределения

Пусть $X_1, \dots, X_n$– выборка из генеральной совокупности, имеющей нормальное распределение $N(\mu, \sigma^2)$. Назовем μ– генеральным средним, а σ^2– генеральной дисперсией.

Выборочное среднее $\bar{X}$ также будет распределено по нормальному закону, причем $E(\bar{X}) = \bar{x}_0 = \mu$; $D(\bar{X}) = \sigma^2/n$, то есть $\bar{X} \sim N(\mu, \sigma^2/n)$. Тогда случайная величина

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0;1) -$$

имеет стандартное нормальное распределение.

Доверительный интервал для математического ожидания нормального распределения при известной дисперсии

Используем формулу $P(|X - E(X)| < \varepsilon) = 2\Phi(\varepsilon/\sigma)$

В нашем случае:

$$P(|\bar{X} - \mu| < \varepsilon) = \gamma = 2\Phi\left(\frac{\varepsilon}{\sigma/\sqrt{n}}\right).$$

Обозначив $Z = \frac{\varepsilon\sqrt{n}}{\sigma}$, тогда $\varepsilon = \frac{Z \cdot \sigma}{\sqrt{n}}$, с вероятностью γ

$$|\bar{X} - \mu| < \frac{Z \cdot \sigma}{\sqrt{n}};$$

$$\bar{X} - \frac{Z \cdot \sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{Z \cdot \sigma}{\sqrt{n}}.$$

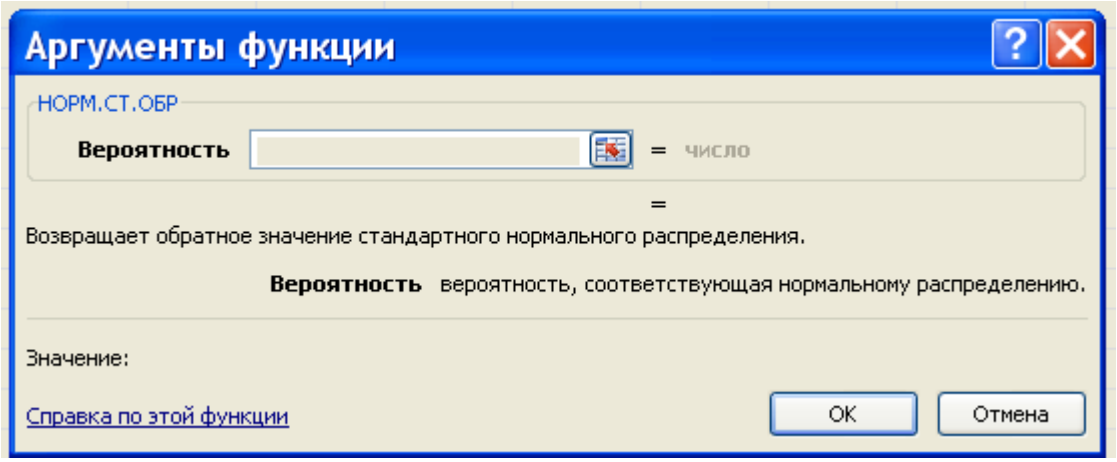
Здесь $\mathrm{Z} = \mathrm{Z}_{(1+\gamma)/2}$ – квантиль уровня $\frac{1+\gamma}{2}$ стандартного нормального распределения.

Доверительный интервал для мат. ожидания при известной генеральной дисперсии имеет вид

$$\bar{X} - \frac{\mathrm{Z}_{(1+\gamma)/2} \cdot \sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{\mathrm{Z}_{(1+\gamma)/2} \cdot \sigma}{\sqrt{n}},$$

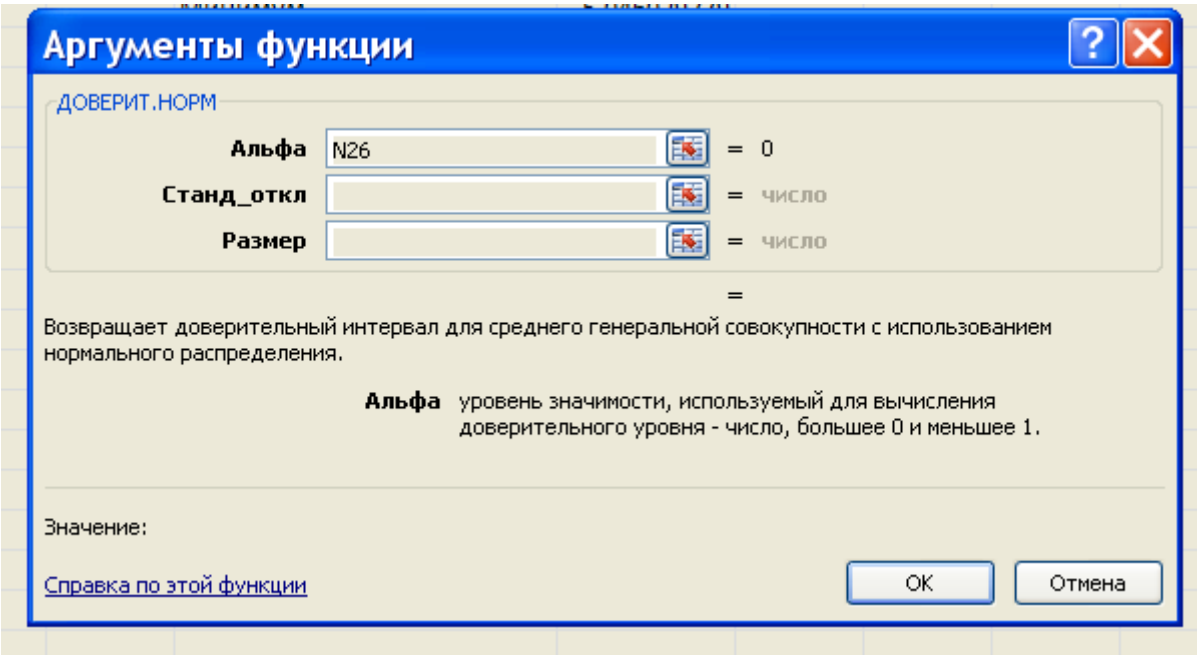
где γ – доверительная вероятность, $\mathrm{Z}_{(1+\gamma)/2}$ – квантиль уровня $\frac{1+\gamma}{2}$ стандартного нормального рспределения.

$\mathrm{Z}_{(1+\gamma)/2} =$ НОРМ.СТ.ОБР($\frac{1+\gamma}{2}$).



Видно, что с ростом n число ε убывает, значит, точность оценки увеличивается.

Другой способ вычисления доверительного интервала – сразу вычислить ε с помощью функции ε = ДОВЕРИТ.НОРМ (α,σ,n) = $\frac{Z_{(1+\gamma)/2}\cdot\sigma}{\sqrt{n}} = \frac{Z_{\alpha/2}\cdot\sigma}{\sqrt{n}}$.



Задачи для самостоятельного решения.

1. Пусть из нормальной генеральной совокупности извлечена выборка объема 25. Выборочное среднее равно 122, генеральная дисперсия известна и равна 400. Найти доверительный интервал для неизвестной генеральной средней с надежностью $\gamma = 0{,}95$.

Ответ: (114,16; 129,84).

2. Предположим, что менеджер по поставке краски хотел оценить точное количество краски, содержащейся в канистре емкостью в один галлон. Из технических условий производителя известно, что стандартное отклонение количества краски равняется 0.02 галлона. По выборке из 50-ти канистр емкостью в один галлон определено среднее количество краски на каждую канистру, равное 0.995 галлона. Построить 99 %-й доверительный интервал для

оценки генерального среднего количества краски, находящегося в канистре емкостью в один галлон.

Ответ: (0.987, 1.0022).

3. Содержание никотина в сигаретах (в мг) подчинено нормальному закону. В результате случайной выборки из 5 сигарет выборочное среднее равно 18,307. Генеральное среднее квадратическое отклонение известно и равно $\sigma = 0{,}0029$. С доверительной вероятностью 0,95 найти доверительный интервал для истинного содержания никотина.

Ответ: (18,30446; 18,3095).

4. Найти доверительный интервал для оценки с надежностью 0,95 неизвестного мат. ожидания нормально распределенного признака X генеральной совокупности, если генеральное среднее квадратическое отклонение $\sigma = 5$, выборочная средняя $\overline{X} = 14$ и объем выборки $n = 25$.

Ответ: (12,04; 15,96).

5. Найти доверительный интервал для оценки с надежностью 0,99 неизвестного мат. ожидания нормально распределенного признака X генеральной совокупности, если известны генеральное среднее квадратическое отклонение σ, выборочная средняя $\overline{X}$ и объем выборки n:

а) $\sigma = 4$, $\overline{X} = 10{,}2$, $n = 16$;

б) $\sigma = 5$, $\overline{X} = 16{,}8$, $n = 25$.

Ответ: а) (7,63; 12,77);

б) (14,23; 19,37).

6. Выборка из большой партии электроламп содержит 150 ламп. Средняя продолжительность горения ламп оказалась равной 1400 ч. Найти приближенный 0,994 – доверительный интервал для средней продолжительности горения лампы всей партии, если известно, что среднеквадратичное отклонение продолжительности горения лампы в партии равно 18 ч.

Ответ: (1395,958; 1404,0417).

7. С целью определения среднего трудового стажа на предприятии методом случайной повторной выборки проведено обследование трудового стажа рабочих. Из всего коллектива рабочих завода случайным образом выбрано 400 рабочих, данные о трудовом стаже которых и составили выборку. Средний по выборке стаж оказался равным 9,4 года. Считая, что трудовой стаж рабочих имеет нормальный закон распределения, определить с вероятностью 0,97 границы, в которых окажется средний трудовой стаж для всего коллектива, если известно, что
$\sigma = 1{,}7$ года.

Ответ: (9,18; 9,62).

Доверительный интервал для математического ожидания нормального распределения при неизвестной дисперсии

Неизвестную дисперсию σ^2 заменим на исправленную выборочную дисперсию.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

а неизвестное стандартное отклонение σ на $S = \sqrt{S^2}$.

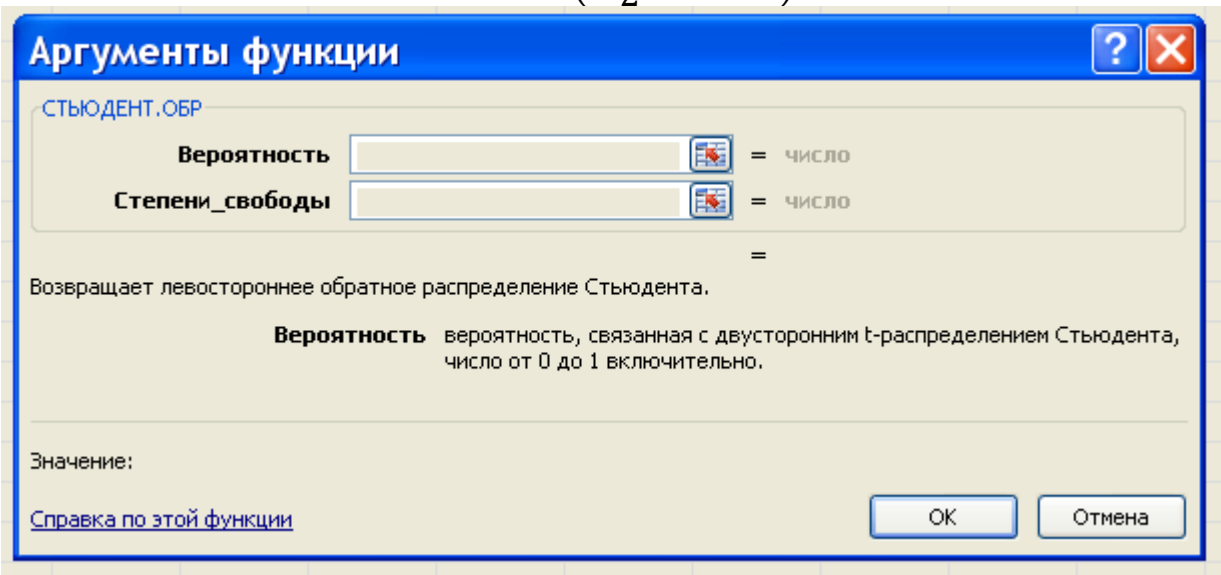
Тогда случайная величина $Y = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ имеет распределение Стьюдента с $k = n-1$ степенями свободы, то есть $Y \sim t(n-1)$. Таким образом, $|Y| < t_{(1+\gamma)/2}(n-1)$.

Доверительный интервал для μ примет вид:

$$\bar{X} - t_{(1+\gamma)/2}(n-1)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{(1+\gamma)/2}(n-1)\frac{S}{\sqrt{n}},$$

где $t_{(1+\gamma)/2}(n-1)$ квантиль распределения Стьюдента с n-1 степенями свободы порядка (1+γ)/2,

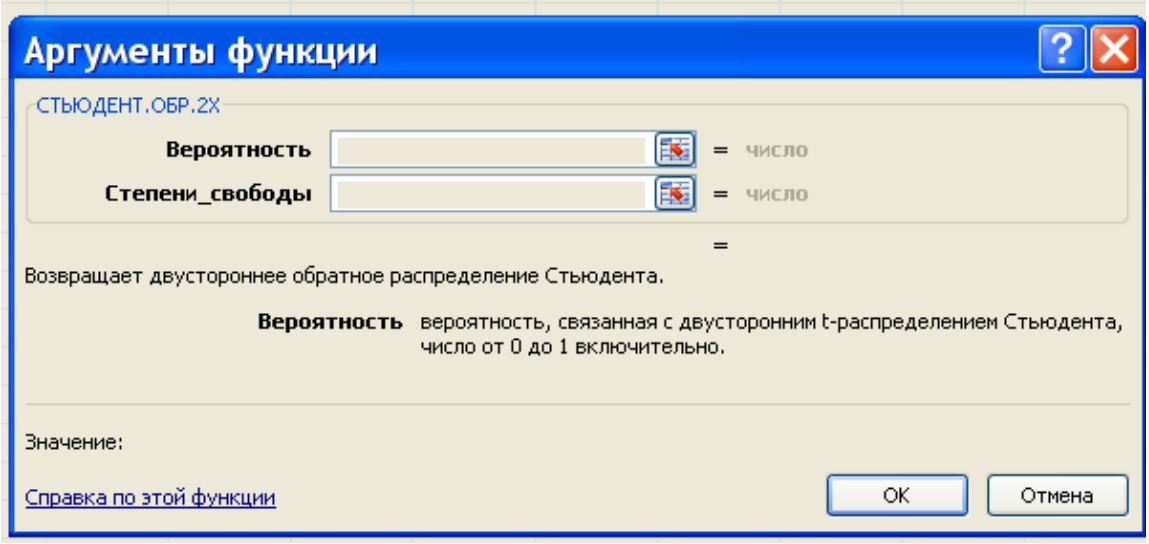
$$t_{(1+\gamma)/2}(n-1) = \text{СТЬЮДЕНТ.ОБР}\left(\frac{1+\gamma}{2}, n-1\right).$$

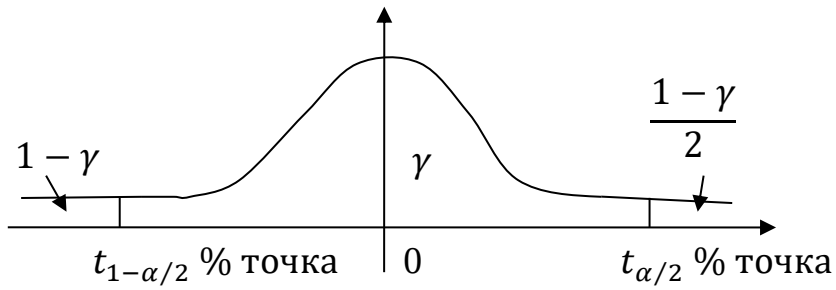


Или

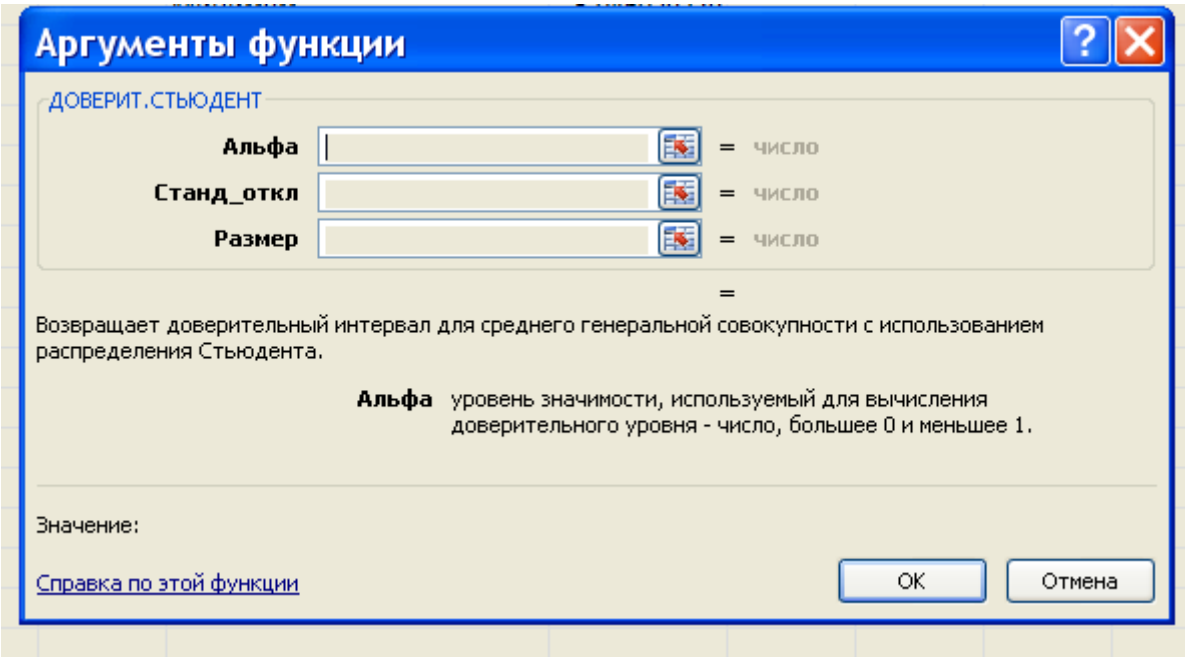
$$\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},$$

$t_{\alpha/2}(n-1)$ = СТЬЮДЕНТ.ОБР.2Х(α, n-1) - **α/2 ·100% -ная** точка распределения Стьюдента с n-1 степенями свободы уровня α.





Есть другой способ с использованием функции ДОВЕРИТ.СТЬЮДЕНТ ε = ДОВЕРИТ.СТЬЮДЕНТ(α, S,n) = $t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}$.



Третий способ построения Доверительного интервала для мат. ожидания при неизвестной генеральной дисперсии: в «Описательной

статистике» *Пакета анализа* поставить флажок «Уровень надежности» (этот доверительная вероятность в процентах). В таблице результатов появится число, равное ε половине длины доверительного интервала.

Задачи для самостоятельного решения.

1. С целью определения средней продолжительности рабочего дня на предприятии методом случайной повторной выборки проведено обследование продолжительности рабочего дня сотрудников. Из всего коллектива завода случайным образом выбрано 30 сотрудников. Данные табельного учета о продолжительности рабочего дня этих сотрудников и составили выборку. Средняя по выборке продолжительность рабочего дня оказалась равной 6,85 часа, а S = 0,7 часа. Считая, что продолжительность рабочего дня имеет нормальный закон распределения, с надежностью $\gamma = 0{,}95$ определить, в каких пределах находится действительная средняя продолжительность рабочего дня для всего коллектива данного предприятия.

Ответ: (6,59; 7,11).

2. Аналитик фондового рынка оценивает среднюю доходность определенных акций. Случайная выборка 15 дней показала, что средняя (годовая) доходность равна 10,37% со средним квадратическим отклонением S = 3,5%. Предполагая, что доходность акций подчиняется нормальному закону распределения, построить 95%-ный и 99%-ный доверительные интервалы для средней доходности интересующего аналитика вида акций.

Ответ: (8,4316; 12,308).; (7,6797; 13,0603).

3. Владелец магазина заметил, что в хлебном отделе у него ежедневно остается некоторое количество непроданных батонов хлеба, и он решил оценить реальную потребность в этом сорте хлеба. В течение месяца он ежедневно записывал число проданных батонов и по данным этой выборки из 30 дней установил, что в среднем за день продается 120 батонов, с исправленным средним квадратическим отклонением 10,1709 батонов. С надежностью 90% построить доверительный интервал для дневного количества батонов.

Ответ: (117; 123).

4. Служба контроля Энергосбыта провела в течение одного из летних месяцев выборочную проверку расхода электроэнергии жителями 10 квартир многоквартирного дома. В результате выяснилось, что расход составил 125; 78; 102; 140; 90; 45; 50; 125; 115; 112. кВт/час. С надежностью 0,95 определите доверительный интервал для оценки среднего расхода электроэнергии на одну квартиру во всем доме.

Ответ: (75,2269; 121,1731).

5. Определить 90 %-й доверительный интервал для генерального среднего по выборке объема n = 18 из нормально распределенной генеральной совокупности для $\bar{X}$= 13,56; S = 7,8.

Ответ: (10,36; 16,76)

6. Предположим, что величина X нормально распределена. Используйте следующую выборку :

313 320 319 340 325 310 321 329 317 311 307 318,

чтобы построить 90 %-й доверительный интервал для математического ожидания и определить ошибку оценивания.

Ответ: (313.376; 324.956); 5.79.

Построить 99%-й доверительный интервал. Определить ошибку оценивания.

Ответ: (309.9676; 328.3644); 9.1984.

Сравнить ширину 90 %-го интервала и 99 %-го интервала и сделать вывод, какая из оценок хуже.

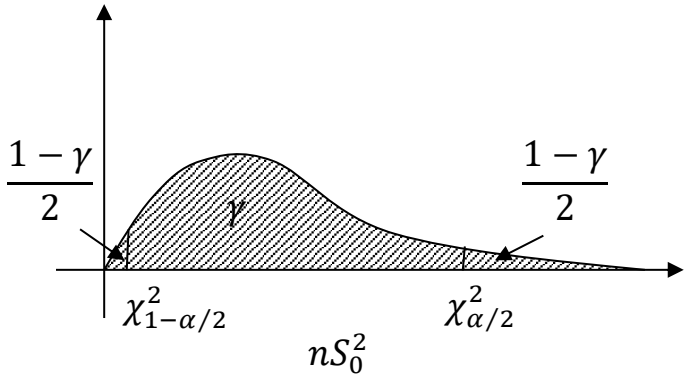
Доверительный интервал для дисперсии при известном генеральном среднем (математическом ожидании)

При известном μ используется несмещенная точечная оценка дисперсии

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Случайная величина $\frac{nS_0^2}{\sigma^2} \sim \chi^2(n)$ – имеет распределение χ^2 с $k = n$ степенями свободы.

При доверительной вероятности γ



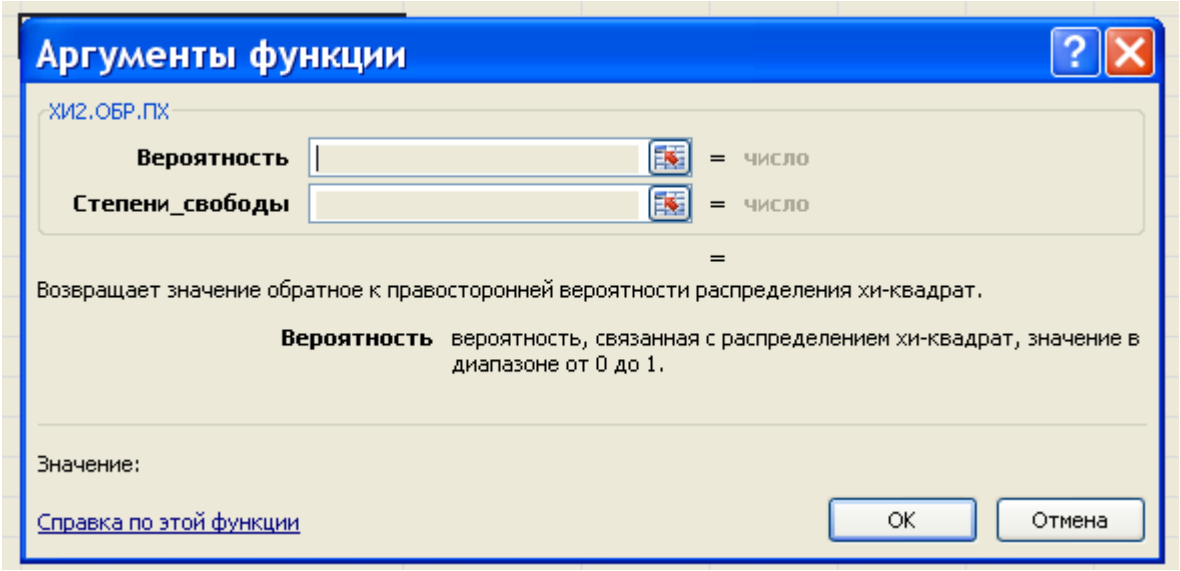
$$\chi^2_{1-\alpha/2}(n) < \frac{nS_0^2}{\sigma^2} < \chi^2_{\alpha/2}(n)$$

$$\frac{1}{\chi^2_{1-\alpha/2}(n)} > \frac{\sigma^2}{nS_0^2} > \frac{1}{\chi^2_{\alpha/2}(n)}$$

Доверительный интервал имеет вид:

$$\frac{nS_0^2}{\chi^2_{\alpha/2}(n)} < \sigma^2 < \frac{nS_0^2}{\chi^2_{1-\alpha/2}(n)},$$

где $\chi^2_{\alpha/2}(n)$ и $\chi^2_{1-\alpha/2}(n)$–$\alpha/2 \cdot 100\%$ точка распределения χ^2 с n степенями свободы вычисляются с помощью функций Excel
$\chi^2_{\alpha/2}(n) =$ ХИ2.ОБР.ПХ(α/2; n); $\chi^2_{1-\alpha/2}(n) =$ ХИ2.ОБР.ПХ(1 − α/2; n)
(ПХ означает правый хвост).

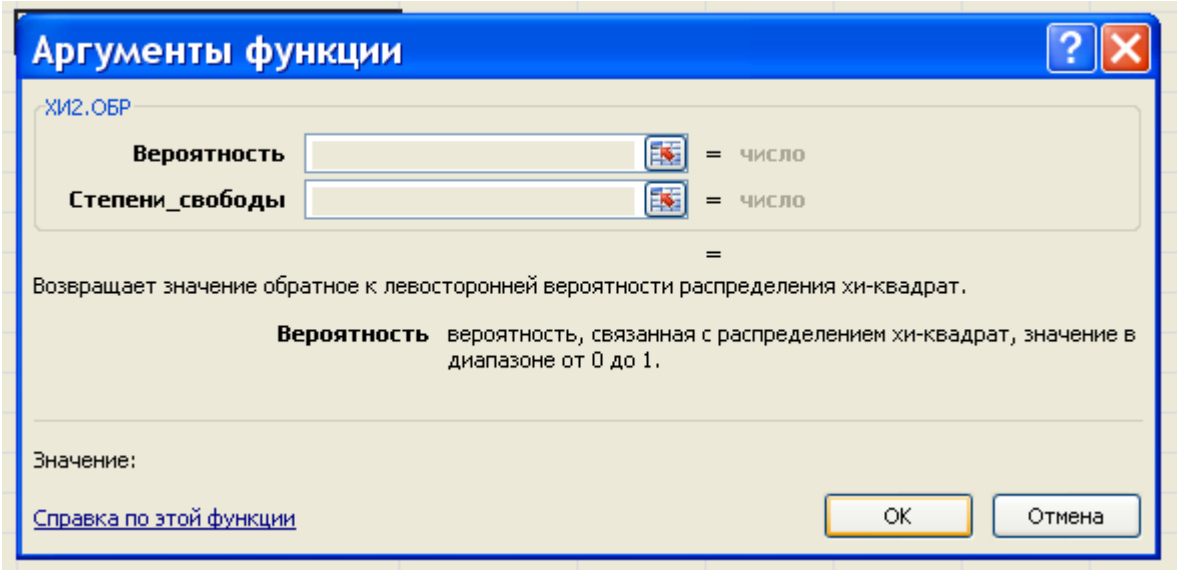


Или

$$\frac{nS_0^2}{\chi^2_{(1+\gamma)/2}(n)} < \sigma^2 < \frac{nS_0^2}{\chi(^2_{1-\gamma)/2}(n)},$$

Где $\chi^2_{(1+\gamma)/2}(n)$ = ХИ2.ОБР($\frac{1+\gamma}{2}$; n); это квантиль распределения хи-квадрат уровня $\frac{1+\gamma}{2}$; с n степенями свободы;

$\chi^2_{(1-\gamma)/2}(n)$ = ХИ2.ОБР($\frac{1-\gamma}{2}$; n); это квантиль распределения хи-квадрат уровня $\frac{1-\gamma}{2}$; с n степенями свободы;



Задачи для самостоятельного решения.

1. На фабрике работает автоматическая линия по фасовке растворимого кофе в 100-граммовые банки. Согласно техническим характеристикам автомата средний вес банки (мат. ожидание) 100г. Если дисперсия веса банок превышает заданное значение, то линия должная быть остановлена на переналадку. С линии в случайном порядке отобрано 30 банок с кофе и оценка дисперсии $S_0^2 = 18{,}54$. Построить 95%-ный доверительный интервал для генеральной дисперсии σ^2.

Ответ: (11,839; 33,1249).

2. Автомат фасует чай в 400-граммовые пачки. Средний вес пачки согласно техническим характеристикам автомата равен 400г. В случайном порядке отобрано 28 пачек чая и вычислена оценка дисперсии $S_0^2 = 16{,}34$ Построить 98%-ный интервал для генеральной дисперсии σ^2.

Ответ: (9,476734033; 33,72869809).

3. Автомат фасует чай в 200-граммовые пачки. Средний вес пачки согласно техническим характеристикам автомата равен 200г. В случайном порядке отобрано 60 пачек чая и вычислена оценка дисперсии $S_0^2 = 21{,}34$ Построить 95%-ный интервал для генеральной дисперсии σ^2.

Ответ: (15,3746; 31,62887).

4. Автомат фасует муку в пакеты по 1000г. Средний вес пакета согласно техническим характеристикам автомата равен 1000г. В случайном порядке отобрано 50 пакетов муки и вычислена оценка дисперсии $S_0^2 = 68{,}7$.
Построить 98%-ный интервал для генеральной дисперсии σ^2.

Ответ: (45,1060339; 115,630548).

Решить предыдущую задачу при $\gamma = 0{,}95$.

Ответ: (48,109; 106,19).

5. Автомат фасует чай в 150-граммовые пачки. Средний вес пачки согласно техническим характеристикам автомата равен 150г. В случайном порядке отобрано 40 пачек чая и вычислена оценка дисперсии $S_0^2 = 9{,}34$ Построить 98%-ный интервал для генеральной дисперсии σ^2.

Ответ: (5,8658; 16,8561)

Доверительный интервал для дисперсии при неизвестном генеральном среднем (математическом ожидании)

При неизвестном μ используется точечная несмещенная оценка дисперсии:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X}_i - \bar{X})^2 .$$

Случайная величина $\frac{(n-1)S^2}{\sigma^2}$ имеет распределение χ^2 с $k = n-1$ степенями свободы.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)$$

Получаем доверительный интервал:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}$$

где $\chi^2_{\alpha/2}(n-1)$ и $\chi^2_{1-\alpha/2}(n-1)$–$\alpha/2 \cdot 100\%$ точка распределения χ^2 с n-1 степенями свободы вычисляются с помощью функций Excel

$\chi^2_{\alpha/2}(n-1) =$ ХИ2.ОБР.ПХ(α/2; n-1); $\chi^2_{1-\alpha/2}(n-1) =$ ХИ2.ОБР.ПХ(1 − α/2; n-1).

Или

$$\frac{(n-1)S^2}{\chi^2_{(1+\gamma)/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(1-\gamma)/2}(n-1)},$$

где

$\chi^2_{(1+\gamma)/2}(n-1)$ = ХИ2.ОБР($\frac{1+\gamma}{2}; n-1$); это квантиль распределения хи-квадрат уровня $\frac{1+\gamma}{2}$; с $n-1$ степенями свободы;

$\chi^2_{(1-\gamma)/2}(n-1)$ = ХИ2.ОБР($\frac{1-\gamma}{2}$; $n-1$); это квантиль распределения хи-квадрат уровня $\frac{1-\gamma}{2}$ с $n-1$ степенями свободы.

Задачи для самостоятельного решения.

1. В нескольких мелких магазинах проведена проверка качества 101 изделия, после чего осуществлена обработка полученных данных. В результате вычислено несмещенное значение выборочного среднего квадратического отклонения S = 4. Считая распределение качественных изделий нормальным, найти с надежностью 0,95 доверительный интервал для оценки генерального среднего квадратичного отклонения.

Ответ: (3,514167963; 4,642948614).

2. Как изменится ответ предыдущей задачи, если число проверяемых изделий уменьшится до 61?

Ответ: (3,394838956; 4,869742392).

3. По данным выборки объема n = 25 найдено несмещенное значение выборочного среднего квадратического отклонения S = 3. Найти с надежностью 0,98 доверительный интервал для оценки генерального среднего квадратичного отклонения.

Ответ: (2,241788199; 4,460512163).

4. На основании выборочных наблюдений производительности труда 20 работниц было установлено, что среднее квадратическое отклонение суточной выработки составляет 15 м ткани в час. Предполагая, что производительность труда работницы имеет нормальное распределение, найти границы, в которых с надежностью 0,9 заключена генеральная дисперсия суточной выработки работниц.

Ответ: (141,8214919; 422,5555481).

Решить ту же задачу при объеме выборки 101.

Ответ: (180,9523691; 288,7226282).

5. По данным выборки объема n = 16 найдено «исправленное» среднее квадратическое отклонение S = 1. Найти с надежностью 0,95 доверительный интервал для оценки генерального среднего квадратичного отклонения.

Ответ: (0,738704858; 1,547691223).

6. По данным 16 независимых равноточных измерений некоторой физической величины найдены среднее арифметическое результатов измерений $\overline{X} = 42{,}8$ и «исправленное» среднее квадратическое отклонение

S = 8. Оценить истинное значение измеряемой величины и ее среднее квадратическое отклонение с помощью доверительного интервала с надежностью γ = 0,95. Предполагается, что результаты измерений распределены нормально.

Ответ: (38,538; 47,062), (5,908; 12,384).

7. По данным девяти независимых равноточных измерений некоторой физической величины найдены среднее арифметическое результатов измерений $\overline{X} = 30{,}1$ и «исправленное» среднее квадратическое отклонение S = 6. Оценить истинное значение измеряемой величины с помощью доверительного интервала с надежностью γ = 0,99. Предполагается, что результаты измерений распределены нормально.

Ответ: (23,39; 36,81).

8. По данным 16 независимых равноточных измерений некоторой физической величины найдены среднее арифметическое результатов измерений $\overline{X} = 42{,}8$ и «исправленное» среднее квадратическое отклонение S = 8. Оценить истинное значение измеряемой величины с помощью доверительного интервала с надежностью γ = 0,999. Предполагается, что результаты измерений распределены нормально.

Ответ: (34,654; 50,946).

9. Решить предыдущую задачу при γ = 0,9.

Ответ: (39,294; 46,306).

10. Случайная величина X распределена по нормальному закону. Статистическое распределение выборки представлено в таблице

x_i	3	5	7	8	10	12	14
n_i	3	7	4	6	7	5	9

Найти с надежностью 0,95 доверительный интервал для оценки математического ожидания и для оценки среднего квадратического отклонения.

Ответ: $\bar{x} = 9{,}171$; $S^2 = 12{,}940$; S= 3,5971; (8,036; 10,306), (2,953; 4,603).

11. Случайная величина X распределена по нормальному закону. Статистическое распределение выборки представлено в таблице

x_i	1	3	5	7	9
n_i	2	5	4	6	3

Найти с надежностью 0,95 доверительный интервал для оценки математического ожидания и для оценки среднего квадратического отклонения.

Ответ: $\overline{X} = 5{,}3$; S^2=6,432; S = 2,536; (4,113; 6,487) ; (1,929; 3,704).

12. Случайная величина X распределена по нормальному закону. Статистическое распределение выборки представлено в таблице

x_i	-2	1	2	3	4	5	
n_i	2	1	2	2	2	1	

Найти с надежностью 0,95 доверительный интервал для оценки математического ожидания для оценки среднего квадратического отклонения.

Ответ: (0,3; 3,7), - для мат. ожидания

Доверительный интервал для генеральной доли признака

Пусть X признак в генеральной совокупности Ω, распределен по закону Бернулли:

X	0	1
P	q	p

, где q = $1-p$.

То есть, доля тех элементов $\omega \in \Omega$, для которых $X(\omega) = 1$ равна p. p– вероятность или генеральная доля признака.

Пусть $X_1, \dots, X_n$– выборка объема n.

X	0	1
Относит частота	$\hat{q}$	$\hat{p}$

$\hat{p}$ – доля тех значений выборки, которые приняли значение 1. Поэтому

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \hat{p}\,.$$

Дисперсия распределения Бернулли равна $\hat{\sigma}^2 = \hat{p} \cdot \hat{q}$.

Величина $\sum_{i=1}^{n} X_i$ подчиняется биномиальному закону распределения. Если $npq \gg 1$, то применима приближённая формула Лапласа.

Применим формулу для математического ожидания с известной дисперсией нормального распределения:

$$\hat{p} - \mathrm{Z}_{(1+\gamma)/2}\frac{\sqrt{\hat{p}\cdot\hat{q}}}{\sqrt{n}} < p < \hat{p} + \mathrm{Z}_{(1+\gamma)/2}\frac{\sqrt{\hat{p}\cdot\hat{q}}}{\sqrt{n}}$$

где γ – доверительная вероятность, $\mathrm{Z}_{(1+\gamma)/2}$ – квантиль уровня $\frac{1+\gamma}{2}$

$\mathrm{Z}_{(1+\gamma)/2} =$ НОРМ.СТ.ОБР($\frac{1+\gamma}{2}$).

$$\hat{p} - \mathrm{Z}_{(1+\gamma)/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + \mathrm{Z}_{(1+\gamma)/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

[В других обозначениях:

$$\hat{p} - \mathrm{Z}_{(1+\gamma)/2} \cdot \sigma(\hat{p}) < p < \hat{p} + \mathrm{Z}_{(1+\gamma)/2} \cdot \sigma(\hat{p}).$$

Причем отметим, что

$$\sigma(\hat{p}) = \sqrt{\frac{pq}{n}} \quad \text{–для повторной выборки;}$$

$$\sigma(\hat{p}) = \sqrt{\frac{pq}{n}\left(1 - \frac{n}{N}\right)} \quad \text{–для бесповторной выборки.]}$$

В других обозначениях, где $\hat{p} = \frac{m}{n} = W$ – статистическая вероятность, доверительный интервал приобретает вид:

$$W - \mathrm{Z}_{(1+\gamma)/2}\sqrt{\frac{W(1-W)}{n}} < p < W + \mathrm{Z}_{(1+\gamma)/2}\sqrt{\frac{W(1-W)}{n}}.$$

Это интервальная оценка вероятности по относительной частоте.

То же самое:

$$\frac{m}{n} - \mathrm{Z}_{(1+\gamma)/2}\frac{\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)}}{\sqrt{n}} < p < \frac{m}{n} + \mathrm{Z}_{(1+\gamma)/2}\frac{\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)}}{\sqrt{n}}.$$

Задачи для самостоятельного решения.

1. На крупном предприятии изучается качество производимых изделий. Из 2000 отобранных для изучения изделий 150 оказались бракованными. С надежностью 0,95 найти границы, в которых заключена доля брака всей продукции предприятия.

Ответ: (0,063456578; 0,086543422).

2. Насколько надо увеличить объем выборки в предыдущей задаче, чтобы разница между генеральной и выборочной долей брака не превышала 0,005?

Ответ: на 8660 деталей.

3. Среди 900 пользователей Интернет-ресурса оказалось 270 женщин. С надежностью 90% построить доверительный интервал для генеральной доли женщин среди всех пользователей ресурса.

Ответ: (0,274874446; 0,325125554).

4. Компания, занимающаяся пассажирскими перевозками, предполагает открыть новый автобусный маршрут. Среди 50 пассажиров, выбранных в случайном порядке, 18 заявили, что будут регулярно пользоваться новым маршрутом. С надежностью 90% построить доверительный интервал для генеральной доли пассажиров, которые будут использовать новый маршрут.

Ответ: (0,248343633; 0,471656367).

5. По результатам городского социологического опроса, для которого была составлена выборка из 789 избирателей по избирательным спискам, выяснилось, что 48% из них собираются голосовать против нынешнего мэра города. С надежностью 99% построить доверительный интервал доли горожан, которые не будут поддерживать на выборах действующего мэра.

Ответ: (0,43418572; 0,52581428).

6. Мэрия города хотела бы знать мнение горожан о работе нового автовокзала. Из 500 опрошенных горожан, одобрительно отозвались о работе нового автовокзала 350 человек. С надежностью 90% оцените долю горожан, которым понравилась работа нового автовокзала.

Ответ: (0,666290532; 0,733709468).

7. Изготовлен экспериментальный игровой автомат, который должен обеспечить появление выигрыша в одном случае из 100 бросаний монеты в автомат. Для проверки пригодности автомата произведено 400 испытаний, причем выигрыш появился 5 раз. Найти доверительный интервал, покрывающий неизвестную вероятность появления выигрыша с надежностью 0,999.

Ответ: (-0,005779276; 0,030779276) → (0; 0,030779276).

8. Выборочно обследовали качество кирпича. Из $n = 1600$ проб в $m = 42$ случаях кирпич оказался бракованным. В каких пределах заключается доля брака для всей продукции, если результат гарантируется с надежностью 0,95 ?

Ответ: (0,018416128; 0,034083872).

9. Для определения процента людей, нашедших себе супруга через брачное агентство, была организована случайная выборка, объем которой составлял 500 человек из обратившихся в брачное агентство. Среди них 75 нашли себе супруга. Найти 90%-ный доверительный интервал, накрывающий неизвестный процент людей, нашедших себе супруга через брачное агентство.

Ответ: (0,123733794; 0,176266206).

10. При проведении анализа эффективности рекламы товара, размещенной в Интернете, была организована случайная выборка объемом 500 человек из покупателей. Выяснилось, что 200 из них узнали о товаре из рекламы в Интернете. С надежностью 0,95 определить границы вероятности того, что случайно отобранный покупатель воспользовался рекламой в Интернете.

Ответ: (0,357059341; 0,442940659).

11. Событие A в серии из n=100 испытаний Бернулли произошло m=7 раз. Найти интервальную оценку вероятности p появления события A в одном испытании с надежностью 0.9.

Ответ: (0,711862547; 0,848137453).

12. Событие A в серии из n=5 испытаний Бернулли произошло m=3 раза. Найти интервальную оценку вероятности p появления события A в одном испытании с надежностью 0.95.

Ответ: (0,170593406; 1,029406594) → (0,170593406;1) .

Рассмотрим задачу об определении объема выборки, соответствующего допустимой величине погрешности генеральной дли признака. Найдем минимально достаточный объем выборки n_{min}, который обеспечивает **погрешность выборки** не более ε с доверительной вероятностью γ, $\alpha = 1 - \gamma$.

Нужно, чтобы

$$\frac{Z_{\alpha/2}\sqrt{\frac{m}{n}\left(1-\frac{m}{n}\right)}}{\sqrt{n}} < \varepsilon$$

Здесь $Z_{\alpha/2}$ это α/2 ▪100% -ная точка стандартного нормального распределения.

$$Z_{\alpha/2}^2 \frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n} < \varepsilon^2$$

$$Z_{\alpha/2}^2 \cdot \frac{m}{n} \cdot \left(1-\frac{m}{n}\right) < \varepsilon^2 \cdot n\,;$$

$$n > \frac{Z_{\alpha/2}^2}{\varepsilon^2} \cdot \frac{m}{n} \cdot \left(1-\frac{m}{n}\right)\,;$$

обозначим $\frac{m}{n} = x$;

$$n > \frac{Z_{\alpha/2}^2}{\varepsilon^2} \cdot x \cdot (1-x) = n(x)$$

$$n(x) = Ax - Ax^2; \qquad n'(x) - A - 2Ax = 0;$$

$$1 - 2x = 0; x = 0{,}5; \ \frac{m}{n} = 0{,}5;$$

$$n_{min} = \frac{Z^2_{\alpha/2}}{\varepsilon^2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{Z^2_{\alpha/2}}{4\varepsilon^2}.$$

Задачи для самостоятельного решения.

1. Производится выборочное обследование возраста читателей массовых библиотек. Сколько карточек необходимо взять для обследования, чтобы с вероятностью 0,95 можно было бы утверждать, что средний возраст в выборочной совокупности отклонится от генерального среднего не более, чем на 2 года? Генеральное среднее квадратичное отклонение принять равным 6 годам.

Ответ: n≥35.

2. Определить необходимый объем выборки для интервального оценивания математического ожидания, если известно, что генеральное среднеквадратическое отклонение равно 1,9, ошибка оценивания равна 0,7, доверительная вероятность равна 0,95.

Ответ: 28.

3. Найти минимальный объем выборки, на основе которой можно было бы оценить математическое ожидание генеральной совокупности с ошибкой, которая не превышает 0.2 и надежностью 0.98, если допускается что генеральная совокупность имеет нормальное распределение с $\sigma = 4$.

Ответ: 2171.

4. Рекламное агентство, обслуживающее местную радиостанцию, хотело бы оценить среднее время прослушивания передач станции. Какой объем выборки необходим, если агентство желает быть уверено в результатах на 90% с предельной ошибкой не более 5 мин.? Из прошлого опыта известно, что среднее квадратическое отклонение времени прослушивания радиопередач равно 45 мин.

Ответ: 219.

Статистическая проверка гипотез

Предположим, что на основании данных выборки $X_1, \dots, X_n$, извлеченной из генеральной совокупности с функцией распределения $F(x, \theta_i)$, делаются

некоторые предположения либо о виде функции $F(x, \theta_i)$, либо о параметрах θ_i. Предположения такого рода называются статистическими гипотезами.

Статистическая гипотеза называется параметрической, если в ней сформулированы предположения относительно значений параметров функции распределения известного вида.

Статистическая гипотеза называется непараметрической, если в ней сформулированы предположения относительно вида функции распределения.

Статистические гипотезы подразделяются на нулевые и альтернативные.

Нулевой гипотезой называют основную (проверяемую) гипотезу. Ее обозначают H_0.

Часто нулевые гипотезы утверждают, что различие между сравниваемыми величинами отсутствует, а наблюдаемые отклонения объясняются лишь случайными ошибками.

Пример.

$$H_0: \sigma_1^2 = \sigma_2^2.$$

В качестве H_0 можно высказать предположение о том, что генеральная средняя $\mu = E(X)$ равна заданному числу или лежит в определенных границах; о том, что различие между вероятностью p появление события в одном испытании и его относительной частотой $W = \frac{m}{n}$ несущественно; о том, что исследуемый признак имеет то или иное распределение.

Альтернативной называется гипотеза, конкурирующая с нулевой гипотезой в том смысле, что если нулевая гипотеза отвергается, то принимается альтернативная. Если альтернативная гипотеза сложная, то, возможно, окажется приемлемой одна из множества гипотез, ее составляющих. Ее часто обозначают символом H_a или H_1.

Пример.

$$H_1: \sigma_1^2 \neq \sigma_2^2.$$

Параметрическая гипотеза называется простой, если она имеет вид $\theta = \theta_0$. Гипотеза вида $\theta \in \Theta$, где Θ– какое-либо множество, содержащее, по меньшей мере, два элемента, называется сложной.

Проверка статистических гипотез осуществляется на основе данных выборки.

Для проверки статистических гипотез используют некоторую случайную величину (выборочную статистику), являющуюся функцией данных выборки, точное или приближенное распределение которой известно. Эту выборочную

статистику обозначают различными буквами в зависимости от закона ее распределения.

Статистическим критерием (тестом) называется правило, по которому гипотеза H_0 отвергается или принимается.

Возможны четыре случая:

Гипотеза H_0	**Принимается**	**Отвергается**
верна	Правильное решение	Ошибка 1-го рода
неверна	Ошибка 2-го рода	Правильное решение

Определение.

Вероятность α допустить ошибку 1-го рода, то есть отвергнуть гипотезу H_0, когда она верна, называется уровнем значимости критерия.

Определение.

Вероятность допустить ошибку 2-го рода, то есть принять гипотезу H_0, когда она неверна обозначается β.

Число $1-\beta$– то есть вероятность не допустить ошибку 2-го рода, называется мощностью критерия.

Например, в ряде прикладных исследований ошибка 1-го рода α означает, что предназначавшийся наблюдателю сигнал тревоги не будет им принят, а ошибка 2-го рода β – вероятность того, что наблюдатель примет ложный сигнал.

Очевидно, желательно сделать сколь угодно малыми α и β. Однако, при фиксированном объеме выборки с уменьшением α растет β и наоборот. Снижения вероятности обеих ошибок α и β можно добиться путем увеличения объема выборки.

При фиксированном объеме выборки можно сделать как угодно малой лишь одну из величин α и β.

Обычно фиксируется α, задается уровень доверия $\gamma = 1-\alpha$. (Если очень важна β, то нужно переименовать гипотезы: основную сделать альтернативной и наоборот).

Очевидно, из нескольких критериев с одним и тем же уровнем значимости α следует выбрать тот, которому соответствует меньшая β, то есть б*о*льшая мощность.

Следует помнить, что подтверждение гипотезы вовсе не означает, что она единственно верна; гипотеза просто не противоречит имеющимся у нас статистическим данным и вполне возможно, что при другой выборке она окажется несостоятельной. Иными словами, статистически проверенную

гипотезу следует расценивать не как абсолютно верный факт, а лишь как достаточно правдоподобное утверждение, не противоречащее данному опыту.

Наблюдаемым значением статистики критерия называется ее значение, вычисленное по данным выборки.

Множество всех возможных значений статистики критерия делится на два подмножества:

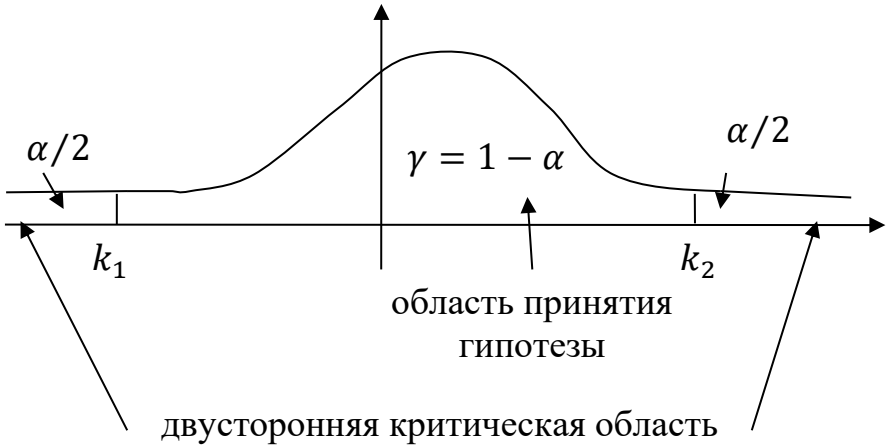
- область принятия гипотезы – совокупность значений, при которых гипотеза принимается;
- критическая область – совокупность значений, при которых гипотеза отвергается.

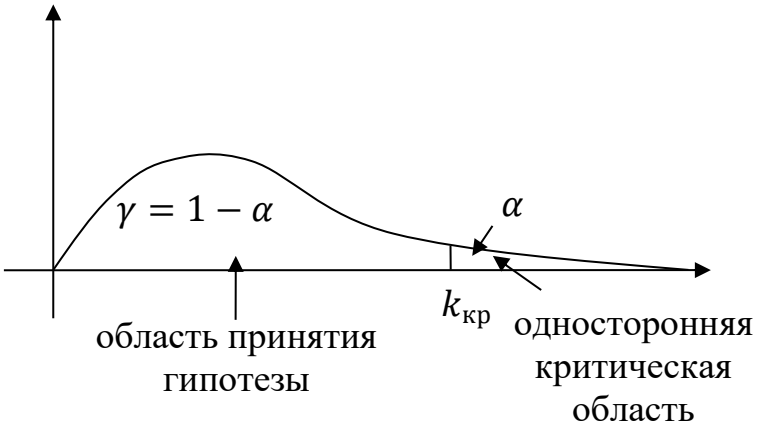
Критическими точками называются точки, отделяющие критическую область от области принятия гипотезы.

Основной принцип проверки статистических гипотез: если наблюдаемое значение статистики критерия принадлежит области принятия гипотезы, гипотезу принимают, если критической области – отвергают.

Двусторонней называют критическую область, определяемую неравенствами: $k < k_1$ и $k > k_2 (k_2 > k_1)$.

Односторонней называют критическую область, определяемую неравенствами $k > k_{\text{кр}}$ или $k < k_{\text{кр}}$.





Проверка гипотез с помощью p - значения (p – value)

Внедрение в образовательный процесс компьютерных технологий (в частности, программы Excel) сделало возможным применение на практике более надежного способа проверки гипотез с использованием p-значения. Ранее это было затруднительно из-за сложности вычислений.

Определение. Достигаемый уровень значимости (p –значение, англ. p-value) — это наименьшая величина уровня значимости, при которой нулевая гипотеза отвергается для данного значения статистики критерия T:

$$p(T) = \min\{\alpha: T \in \Omega_\alpha\},$$

где Ω_α— критическая область критерия.

Пусть T — статистика, используемая при тестировании некоторой нулевой гипотезы H_0. Предполагается, что если нулевая гипотеза справедлива, то распределение этой статистики известно. Обозначим функцию распределения F(t) = P(T ≤ t), P-значение наиболее часто (при правосторонней альтернативной гипотезе) определяется как:

p(t) = P(T > t) = 1 – F(t).

При левосторонней альтернативной гипотезе,

p_0(t) = P(T ≤ t) = F(t).

В случае двусторонней альтернативной гипотезы p-значение равно:

p(t) = 2 min (p_0(t), p(t)).

Решение о принятии или отклонении нулевой гипотезы принимается в результате сравнения p-значения с выбранным уровнем значимости. Если оно превышает указанный уровень

значимости, то для отклонения нулевой гипотезы (принятия альтернативной) нет достаточных оснований.

Преимуществом данного подхода является то, что видно, при каком уровне значимости нулевая гипотеза будет отвергнута, а при каких принята, то есть виден уровень надежности статистических выводов, точнее, вероятность ошибки при отвержении нулевой гипотезы. При любом уровне значимости больше p -значения нулевая гипотеза отвергается, а при меньших значениях — нет.

Если p-значение меньше заданного уровня значимости, то нулевая гипотеза отвергается в пользу альтернативной. В противном случае она не отвергается.

Проверка гипотез об определенном значении параметров нормального распределения

Пусть $X_1, \dots, X_n$– выборка из распределения $N(\mu, \sigma^2)$.

Структура критерия при проверке гипотезы $\mu = \mu_0$ зависит от того известна или нет генеральная дисперсия σ^2, а также от вида альтернативной гипотезы.

Аналогично, способ проверки гипотезы $\sigma^2 = \sigma_0^2$ зависит от того, известно или нет генеральное среднее μ, а также от вида альтернативной гипотезы.

Во всех случаях прослеживается сходство с построением доверительных интервалов для соответствующих параметров нормального распределения.

Проверка гипотезы об определенном значении генеральной средней при известной дисперсии

Пусть генеральная дисперсия σ^2 известна.

Для проверки гипотезы $H_0: \mu = \mu_0$ при любой из трех альтернативных гипотез 1) $H_1: \mu > \mu_0$; 2) $H_2: \mu < \mu_0$; 3) $H_3: \mu \neq \mu_0$ используется статистика

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Она имеет стандартное нормальное распределение $N(0; 1)$.

Критическая область выбирается в соответствии с таблицей.

H_1	критическая область
$\mu > \mu_0$	$Z > Z_{1-\alpha}$
$\mu < \mu_0$	$Z < -Z_{1-\alpha}$
$\mu \neq \mu_0$	$\lvert Z\rvert > Z_{1-\alpha/2}$

$Z_{1-\propto}$ = НОРМ.СТ.ОБР(1-α) - квантиль уровня 1- α стандартного нормального распределения.

При этом наблюдаемый уровень значимости (р- значение) равен

$P(Z > Z_{набл.}) = 1 -$ НОРМ.СТ.РАСП.($|Z_{набл.}|$; ИСТИНА) при односторонней альтернативной гипотезе и

$P(Z > Z_{набл.}) = 2($ 1 – НОРМ.СТ.РАСП.(| $Z_{набл.}$ |; ИСТИНА)) при двусторонней альтернативной гипотезе.

Пример.

Автоматическая линия фасует пакеты с мукой весом в 1 кг = 1000г. Производитель утверждает, что точность наладки линии высока и средний вес пакетов равен 1 кг. Покупатель же партии фасованной муки сомневается в точности веса и выдвигает предположение, что вес пакета меньше 1 кг, то есть происходит обвес.

$H_0: \mu = 1000$ г; $H_1: \mu < 1000$ г.

Произведена случайная выборка 100 пакетов с мукой. Полученная выборочная средняя $\bar{X} = 995$ г. Будем считать, что $\sigma_{генерал.}$ известно, $\sigma = 10$ г. Проверим при уровне значимости$\alpha = 0{,}05$ гипотезу о равенстве веса пакета с мукой одному килограмму. Альтернативная гипотеза утверждает, что вес пакета с мукой менее 1 килограмма.

Решение.

$$Z = \frac{995 - 1000}{10/\sqrt{100}} = -5\,.$$

$Z_{набл.} = -5$. При односторонней альтернативе $Z_{крит}$ = НОРМ.СТ.ОБР(1-0,05)=

= 1,644853627.

Гипотеза H_0 отклоняется, так как $Z_{набл.}$ принадлежит критической области.

Мы отвергаем утверждение производителя об отсутствии обвеса. При этом, возможно, мы совершаем ошибку 1-го рода с вероятностью α = 0,05.

2 способ проверки гипотезы с помощью р-значения:

Р-значение = 1 – НОРМ.СТ.РАСП.(5; ИСТИНА) = 2,86652E-07 – меньше уровня значимости 0,05, гипотеза отклоняется.

Тот же результат получается, если альтернативная гипотеза имеет вид $H_1: \mu \neq 1000$ г.

$Z_{крит}$ = НОРМ.СТ.ОБР(1-0,05/2)= 1,959963985. Гипотеза отклоняется.

Р-значение = 2(1 – НОРМ.СТ.РАСП.(5; ИСТИНА)) = 5,73303E-07 – меньше уровня значимости 0,05, гипотеза отклоняется.

Задачи для самостоятельного решения.

1. Дисперсия генеральной совокупности $\sigma^2 = 100$. Выборка 25 объектов это совокупности дала среднюю арифметическую, равную 17. Можем ли мы отклонить $H_0: \mu = 21$; при конкурирующей гипотезе $H_1: \mu \neq 21$? Уровень значимости $\alpha = 0{,}05$.

Ответ: $z_{набл.} = -2$; $z_{крит.} = 1{,}959963985$. Гипотеза H_0 отклоняется.

2. Инженер по контролю качества проверяет среднее время горения нового вида электроламп. Для проверки в случайном порядке было отобрано 100 ламп, среднее время горения которых оказалось 1075 часов. Среднее квадратическое отклонение времени горения в генеральной совокупности известно и равно 100 часов. При уровне значимости $\alpha = 0{,}05$ проверьте гипотезу о том, среднее время горения ламп больше 1000 часов.

Ответ: $z_{набл.} = 7{,}5$; $z_{крит.} = 1{,}645$. Время горения ламп больше 1000 часов, имеющиеся различия не случайны, значимы.

3. Ежедневная заработная плата в определенной отрасли есть случайная величина, распределенная по нормальному закону со средней 13,2 € и $\sigma = 2{,}5$ €.Если компания в этой отрасли нанимает 40 рабочих и платит им в среднем 12,2 €, может ли эта компания быть обвиненной в том, что она платит слишком низкую зарплату? Уровень значимости принять $\alpha = 0{,}05$.

Ответ: $z_{набл.} = -2{,}53$; $z_{крит.} = 1{,}645$. Да, может.

Проверка гипотезы об определенном значении генеральной средней при неизвестной дисперсии

Если генеральная дисперсия σ^2 неизвестна, то в качестве статистики используется величина

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

которая отличается от статистики Z заменой σ на S, где

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{X}_i - \bar{X})^2.$$

Эта величина t имеет распределение Стьюдента с $k = n - 1$ степенями свободы.

Критическая область выбирается в соответствии с таблицей.

H_1	критическая область
$\mu > \mu_0$	$t > t_{1-\alpha}(n-1)$
$\mu < \mu_0$	$t < -t_{1-\alpha}(n-1)$
$\mu \neq \mu_0$	$\lvert t\rvert > t_{1-\alpha/2}(n-1),$

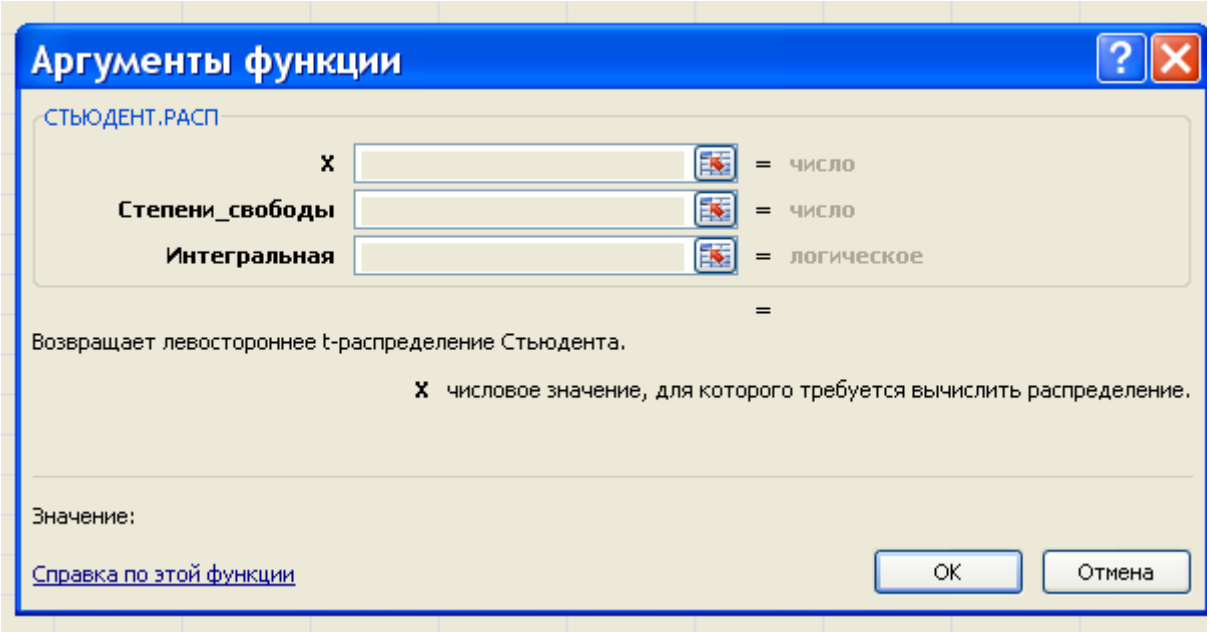
где $t_{1-\alpha}(n-1)$–квантиль порядка 1-α распределения Стьюдента с $k = n - 1$ степенями свободы.

При этом наблюдаемый уровень значимости (p- значение) равен

P(t > $t_{набл.}$) = 1 – СТЬЮДЕНТ.РАСП.(| $t_{набл.}$ |; n-1; ИСТИНА) = = СТЬЮДЕНТ.РАСП.ПХ(|$t_{набл.}$ | ; n-1) при односторонней альтернативной гипотезе.

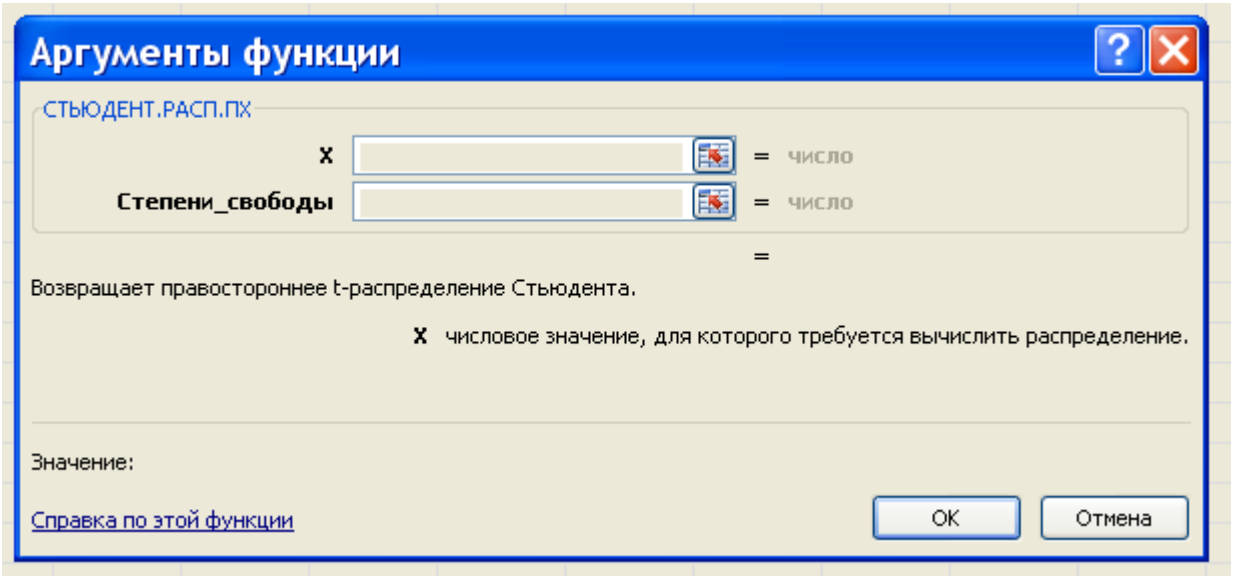
Это значит, что при любом уровне значимости большем, чем этот , есть основание отвергнуть гипотезу H_0

То есть вероятность ошибиться, отвергнув гипотезу H_0 , не превосходит 1 – СТЬЮДЕНТ.РАСП.(| $t_{набл.}$ |; n-1; ИСТИНА).



При двусторонней альтернативе наблюдаемый уровень значимости (p-значение) равен

$P(t > t_{набл.}) = 2(1 - $ СТЬЮДЕНТ.РАСП.$(|t_{набл.}|; n-1;$ ИСТИНА$)) =$
$= 2$ СТЬЮДЕНТ.РАСП.ПХ$(|t_{набл.}|; n-1)$.



Пример.

Менеджер кредитного отдела некоторого учреждения решил выяснить, является ли среднемесячный баланс владельцев кредитных карточек равным 75$. Аудитор случайным образом отобрал 100 счетов и нашел, что среднемесячный баланс владельцев составляет 83,4$ с выборочным стандартным отклонением, равным 23,65$. Определить при 5% уровне значимости, может ли этот аудитор утверждать, что средний баланс равен 75$.

Решение.

$H_0: \mu_0 = 75; H_1: \mu_0 \neq 75.$

$$t_{\text{набл.}} = \frac{83{,}4 - 75}{23{,}65/\sqrt{100}} = \frac{8{,}4}{2{,}365} = 3{,}55\,.$$

$$t_{\text{крит.}}\left(1 - \frac{\alpha}{2}; 100 - 1\right) = \text{СТЬЮДЕНТ.ОБР}\left(1 - \frac{0{,}05}{2}; 99\right) = 1{,}984216952.$$

$$t_{\text{набл.}} = 3{,}55 > t_{\text{крит.}}$$

Нулевая гипотеза не принимается.

Р-значение = 2(1-СТЬЮДЕНТ.РАСП(3,55; 99; 1) =0,000591234 – меньше уровня значимости, нулевая гипотеза отклоняется.

Задачи для самостоятельного решения.

1. Компания, производящая средства для потери веса, утверждает, что прием таблеток в сочетании со специальной диетой позволяет сбросить в среднем в неделю 400 г веса. Случайным образом отобраны 25 человек, использующих эту терапию, и обнаружено, что в среднем они сбросили 430 г при исправленном среднем квадратическом отклонении 110 г. Проверьте гипотезу о том, что средняя потеря веса составляет 400 г. Уровень значимости принять α = 0,05.

Ответ: $t_{набл.}$= 1,36; $t_{крит.}$ = 2,063898562 . Гипотеза принимается.

2. Проверка, проведенная в отделе фасованных продуктов, показала, что средний вес 121 штуки случайно отобранных 60-граммовых пакетиков с маком составил 59 г с исправленным средним квадратическим отклонением 5,0208 г. Проверьте при уровне значимости α = 0,05 гипотезу о том, является полученная разница в весе случайной, или в действительности вес пакетиков с маком меньше 60 г.

Ответ: $t_{набл.}$= -2,19; $t_{крит.}$ = 1,66 . Гипотеза отклоняется, то есть полученная разница в весе не является случайной.

3. Производители стирального порошка утверждают, что средний вес коробки с порошком составляет 325 г. Случайная выборка 25 коробок с порошком обнаружила, что средний вес коробки составляет 323,8 г, а исправленное среднее квадратическое отклонение равно 11,941 г. На 1%-ном уровне значимости определите, отличается ли средний вес коробок от 325 г.

Ответ: $t_{набл.}$= -0,502; $t_{крит.}$ = 2,8 . Гипотеза принимается, то есть полученное различие в весе коробок случайно, незначимо.

Проверка гипотезы об определенном значении генеральной дисперсии

Способ проверки гипотезы $H_0: \sigma^2 = \sigma_0^2$ зависит от того, известно или нет генеральное среднее μ.

Если μ известно, то используется статистика

$$\chi^2 = \frac{nS_0^2}{\sigma_0^2},$$

где

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 .$$

Она имеет распределение χ^2 с $k = n$ степенями свободы.

Критическая область выбирается по таблице

H_1	критическая область
$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{1-\alpha}(n)$
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{\alpha}(n)$
$\sigma^2 \neq \sigma_0^2$	$\{\chi^2 < \chi^2_{\alpha/2}(n)\} \cup \{\chi^2 > \chi^2_{1-\alpha/2}(n)\}$

где $\chi^2_{1-\alpha}(n)$ и $\chi^2_{\alpha}(n)$ – квантили распределения χ^2 порядка α и 1-α с $k = n$ степенями свободы. $\chi^2_{\alpha}(n) =$ ХИ2.ОБР$(\alpha; n)$.

При этом наблюдаемый уровень значимости (p- значение) равен

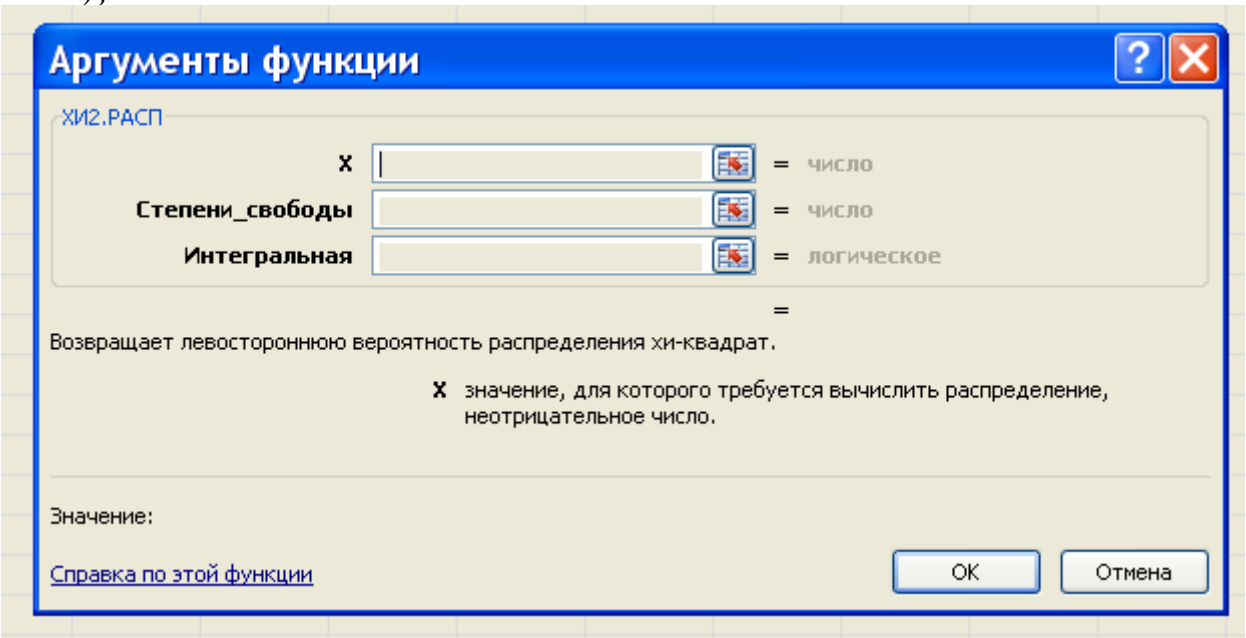
при H_1: $\sigma^2 > \sigma_0^2$, P(χ^2>$\chi^2_{\text{набл.}}$) = 1- ХИ2.РАСП ($\chi^2_{\text{набл.}}$; n; ИСТИНА) = ХИ2.РАСП.ПХ ($\chi^2_{\text{набл.}}$; n; ИСТИНА);

при H_1: $\sigma^2 < \sigma_0^2$, P(χ^2>$\chi^2_{\text{набл.}}$) = ХИ2.РАСП ($\chi^2_{\text{набл.}}$; n; ИСТИНА) ;

при H_1: $\sigma^2 \neq \sigma_0^2$, P(χ^2>$\chi^2_{\text{набл.}}$) =

= 2min{1- ХИ2.РАСП ($\chi^2_{\text{набл.}}$; n; ИСТИНА); ХИ2.РАСП ($\chi^2_{\text{набл.}}$; n; ИСТИНА)}=

= 2 min {ХИ2.РАСП .ПХ($\chi^2_{\text{набл.}}$; n; ИСТИНА) ; ХИ2.РАСП ($\chi^2_{\text{набл.}}$; n; ИСТИНА)}.



Если генеральная средняя μ неизвестна, используется статистика

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2},$$

где

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

Критическая область определяется той же таблицей, но с другим числом степеней свободы $k = n - 1$, так как случайная величина χ^2 имеет распределение χ^2 с $k = n - 1$ степенями свободы. Р-значение определяется аналогично, лишь число степеней свободы меняется с n на n-1.

Используя данные из предыдущего примера проверить при уровне значимости $\alpha = 0{,}05$ гипотезу $H_0: \sigma = 25\$$ $(\sigma^2 = 625)$при конкурирующей гипотезе $H_1: \sigma \neq 25\$$.

$$\chi^2 = \frac{99 \cdot 23{,}65^2}{625} = 88{,}597 = \chi^2_{\text{набл.}}$$

$$\chi^2_{1-\alpha/2}(99) = \chi^2_{0{,}975}(99) = 128{,}4219886.$$

$$\chi^2_{\alpha/2}(99) = \chi^2_{0{,}025}(99) = 73{,}36108019$$

$$73{,}36108019 < \chi^2_{\text{набл.}} = 88{,}597 < 128{,}4219886.$$

Гипотеза принимается.

Р-значение равно
2min{ 1-ХИ2.РАСП(88,597; 99; 1); ХИ2. РАСП(88,597; 99; 1) }=
=2min {0,763957049; 0,236042951} = 0,472085902 - больше уровня значимости, гипотеза принимается.

Задачи для самостоятельного решения.

1. На фабрике работает автоматическая линия по фасовке растворимого кофе в 100-граммовые банки. Согласно техническим характеристикам автомата средний вес банки (мат. ожидание) 100г. Если дисперсия веса банок превышает значение 16, то линия должная быть остановлена на переналадку. С линии в случайном порядке отобрано 30 банок с кофе и оценка дисперсии $S_0^2 = 18{,}54$. При 5%-ном уровне значимости проверить гипотезу о равенстве генеральной дисперсии σ^2 числу 16.

Ответ: $\chi^2_{набл.} = 34{,}7625$; $\chi^2_{крит.} = 43{,}773$ Гипотеза принимается.

2. Автомат фасует чай в 400-граммовые пачки. Средний вес пачки согласно техническим характеристикам автомата равен 400г. Если дисперсия веса пачек превышает значение 14, то автомат должен быть остановлен на переналадку. В случайном порядке отобрано 28 пачек чая и вычислена оценка дисперсии $S_0^2 = 16{,}34$ При 1%-ном уровне значимости проверить гипотезу о равенстве генеральной дисперсии σ^2 числу 14.

Ответ: $\chi^2_{набл.} = 32{,}68;\ \chi^2_{крит.} = 48{,}278$ Гипотеза принимается.

3. На станке вытачиваются детали для сборки механизмов. Партия деталей принимается контролером, если среднее квадратическое отклонение диаметра деталей не превосходит 3 мм. По произведенной выборке из 36 деталей получено среднее квадратическое отклонение 3,14 мм. При 1%-ном уровне значимости определите, можно ли принять партию.

Ответ: $\chi^2_{набл.} = 38{,}343;\ \chi^2_{крит.} = 57{,}2915$ принимается.

4. Инвестор намерен вложить деньги в некоторый вид акций, если дисперсия цены этих акций (риск) не превосходит значения 20 у.е.^{2} Стоимость данного вида акций подчинена нормальному закону. Случайная выборка за 15 дней цены акции дала величину выборочной дисперсии, равную 26 у.е.^{2}. При уровне значимости 5% выяснить, согласится ли инвестор вложиться в данный вид акций.

Ответ: $\chi^2_{набл.} = 18{,}2;\ \chi^2_{крит.} = 23{,}685$ согласится..

4. На основании выборочных наблюдений производительности труда 20 работниц было установлено, что среднее квадратическое отклонение суточной выработки составляет 15 м ткани в час. Предполагая, что производительность труда работницы имеет нормальное распределение, при уровне значимости 0,05 проверить гипотезу о том, что среднее квадратическое отклонение суточной выработки равно 20 м ткани в час.

5. Содержание никотина в сигаретах (в мг) подчинено нормальному закону. В результате случайной выборки из 15 сигарет вычислено выборочное среднее квадратическое отклонение 0,0029. При 5%-ном уровне значимости проверить гипотезу о том, что среднее квадратическое отклонение в генеральной совокупности равно 0,003.

Ответ: $\chi^2_{крит.}\textit{нижнее} = 5{,}629 < \chi^2_{набл.} = 13{,}082 < 26{,}119 = \chi^2_{крит.}\textit{верхнее}$

6. Автомат фасует чай в 200-граммовые пачки. Средний вес пачки согласно техническим характеристикам автомата равен 200г. Если дисперсия веса пачек превышает значение 15, то автомат должен быть остановлен на переналадку. В случайном порядке отобрано 60 пачек чая и вычислена оценка дисперсии $S_0^2 = 21{,}34$. При 1%-ном уровне значимости и при 5%-ном уровне значимости проверить гипотезу о равенстве генеральной дисперсии σ^2 числу 15.

Ответ: При $1\% -$ ном уровне значимости $\chi^2_{набл.} = 85{,}36;\ \chi^2_{крит.} = 88{,}379$ Гипотеза принимается.

При 5%-ном уровне значимости $\chi^2_{крит.} = 79{,}082$ Гипотеза противоречит имеющимся данным.

7. Автомат фасует муку в пакеты по 1000г. Средний вес пакета согласно техническим характеристикам автомата равен 1000г. Если дисперсия веса пакетов превышает значение 45, то автомат должная быть остановлен на переналадку. В случайном порядке отобрано 50 пакетов муки и вычислена оценка дисперсии $S_0^2 = 68{,}7$. При 5%-ном уровне значимости проверить гипотезу о равенстве генеральной дисперсии σ^2 числу 45.

Ответ: $\chi^2_{набл.} = 76{,}333;\ \chi^2_{крит.} = 67{,}42$ Гипотеза противоречит имеющимся данным.

8. Инвестор намерен вложить деньги в некоторый вид акций, если дисперсия цены этих акций (риск) не превосходит значения 30 у.е.^{2}. Случайная выборка за 25 дней цены акции дала величину выборочной дисперсии, равную 40 у.е.^{2}. . При уровне значимости 5% проверить гипотезу.

Ответ: $\chi^2_{набл.} = 32;\ \chi^2_{крит.} = 36{,}415$ согласится

9. В цеху изготовляются детали для сборки автомобилей. Партия деталей принимается контролером, если среднее квадратическое отклонение диаметра деталей не превосходит 2 мм. По произведенной выборке из 30 деталей получено среднее квадратическое отклонение 3,24 мм. При 1%-ном уровне значимости определите, можно ли принять партию.

Ответ: $\chi^2_{набл.} = 76{,}1076;\ \chi^2_{крит.} = 49{,}588$ партия не принимается.

10. Точность работы станка-автомата проверяется по дисперсии контролируемого размера изделий, которая не должна превышать 0,1. Взята проба из 25 случайно отобранных изделий; результаты сведены в таблицу:

Контролируемый размер	3,0	3,5	3,8	4,4	4,5
Частота	2	6	9	7	1

При уровне значимости 0,05 проверить, обеспечивает ли станок требуемую точность.

Ответ: $\chi^2_{набл.} = 48;\ \chi^2_{крит.} = 36{,}4$ не обеспечивает.

11. Менеджер нового отделения банка желает выяснить, что время ожидания клиентами обслуживания не является слишком длительным. Опросив 30 клиентов, он выяснил, что среднее значение времени ожидания равнялось 8 минутам, а исправленная выборочная дисперсия времени ожидания равна 16 мин. ^{2}, в то время как в других отделениях банка ее значение равно 9. Предполагая, что время ожидания распределено нормально, при 5%-ном уровне значимости проверить гипотезу о равенстве генеральной дисперсии σ^2 числу 9.

Ответ: $\chi^2_{набл.} = 51{,}65;\ \chi^2_{крит.} = 42{,}56$ Гипотеза противоречит имеющимся данным.

12. Из нормальной генеральной совокупности извлечена выборка объема $n = 31$:

x_i	10,1	10,3	10,6	11,2	11,5	11,8	12,0
n_i	1	3	7	10	6	3	1

Проверить при уровне при уровне значимости 0,05 нулевую гипотезу

H_0 : D = □0,18, приняв в качестве конкурирующей гипотезы H_1 : D > 0,18.
Ответ: Гипотеза противоречит имеющимся данным.

13. Из нормальной генеральной совокупности извлечена выборка объема n = □20:

x_i	56	58	60	62	64
n_i	1	4	10	3	2

Проверить при уровне значимости 0,05 нулевую гипотезу H_0 : D = 2, приняв в качестве конкурирующей гипотезы H_1 : D ≠ □2.
Ответ: Гипотеза противоречит имеющимся данным.

Сравнение параметров двух нормальных распределений

Пусть $\vec{X} = (X_1, \dots, X_m)$ – выборка из нормально распределенной генеральной совокупности $N(\mu_X, \sigma_X^2)$, а $\vec{Y} = (Y_1, Y_2, \dots, Y_n)$ – выборка из нормального распределения $N(\mu_Y, \sigma_Y^2)$. Считаем выборки $\bar{X}$ и $\overline{Y}$ независимыми, что означает независимость в совокупности $m + n$ случайных величин $X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n$.

Проверка гипотезы о равенстве дисперсий двух нормально распределенных генеральных совокупностей по критерию Фишера

Считаем, что все параметры $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ неизвестны. В качестве нулевой гипотезы примем $H_0: \sigma_X^2 = \sigma_Y^2$.

Пусть по выборкам $\vec{X}$ и $\vec{Y}$ вычислены выборочные средние $\bar{X}$ и $\bar{Y}$ и несмещенные оценки дисперсий:

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \quad \text{и} \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 .$$

В качестве статистики используется величина

$$F = S_1^2/S_2^2; \text{ (предполагается, что } S_1^2 > S_2^2).$$

Эта величина имеет распределение Фишера (Фишера – Снедекора) с k_1,k_2 степенями свободы, где k_1 – число степеней свободы числителя, k_2 – число степеней свободы знаменателя.

В соответствии с видом альтернативной гипотезы H_1 критическая область определяется неравенствами:

$\boldsymbol{H_1}$	**критическая область**
$\sigma_X^2 > \sigma_Y^2$	$\frac{S_X^2}{S_Y^2} > F_{1-\alpha}(m-1, n-1)$
$\sigma_X^2 < \sigma_Y^2$	$\frac{S_Y^2}{S_X^2} > F_{1-\alpha}(n-1, m-1)$
$\sigma_X^2 \neq \sigma_Y^2$	$\frac{S_1^2}{S_2^2} > F_{1-\alpha/2}(k_1, k_2)$,

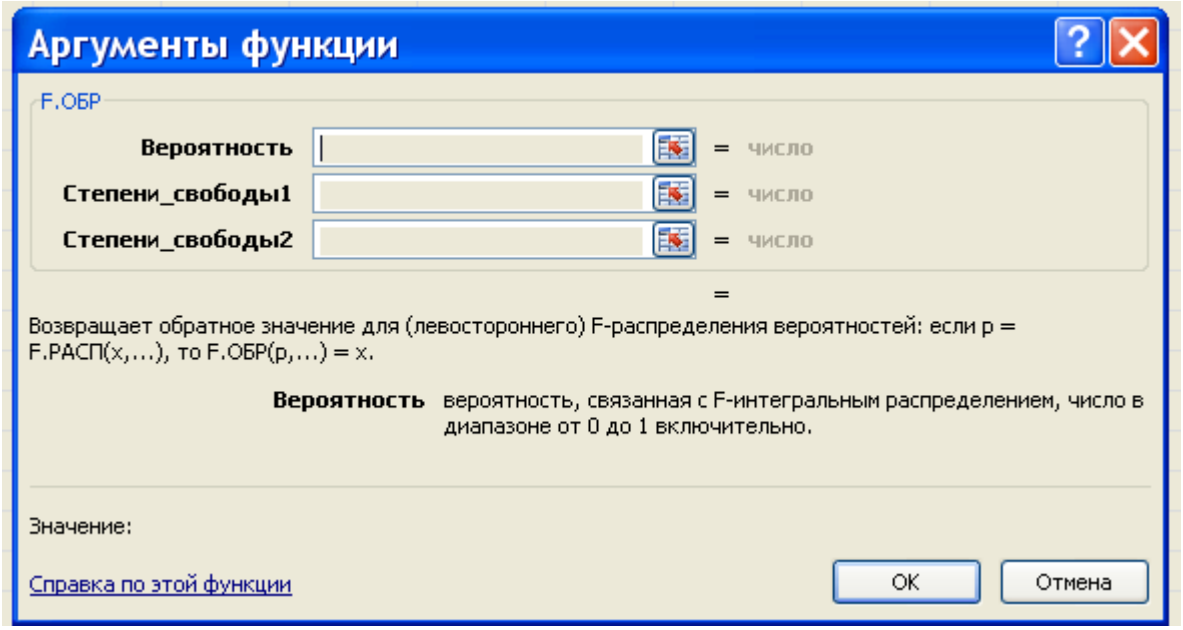
где символы S_1^2,S_2^2, k_1, k_2 в зависимости от соотношения между S_X^2 и S_Y^2 определяются таблицей:

символ	$\boldsymbol{S_X^2 \geq S_Y^2}$	$\boldsymbol{S_X^2 < S_Y^2}$
S_1^2	S_X^2	S_Y^2
S_2^2	S_Y^2	S_X^2
k_1	$m-1$	$n-1$

k_2	$n-1$	$m-1$

Здесь $F_{1-\alpha}(m-1, n-1)$ это квантиль распределения Фишера порядка 1-α с числом степеней свободы k_1=m-1 , k_2 = n-1.

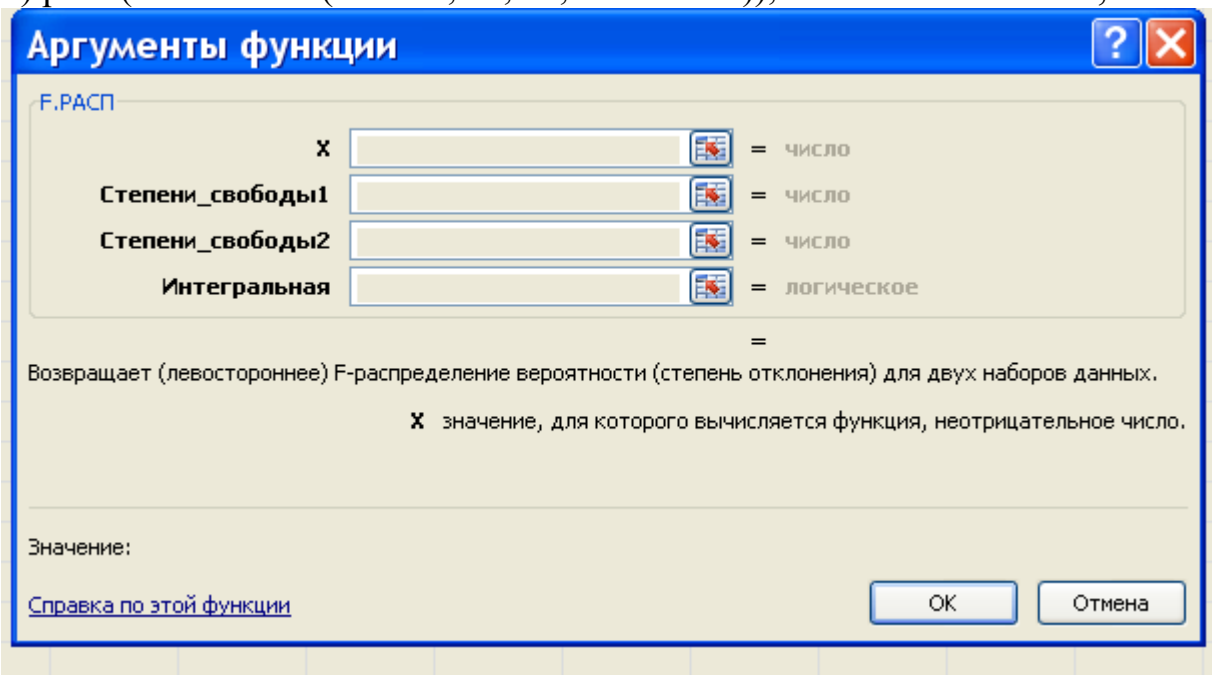
$F_{1-\alpha}(k_1,k_2)$ = F.ОБР(1-α; k_1,k_2)



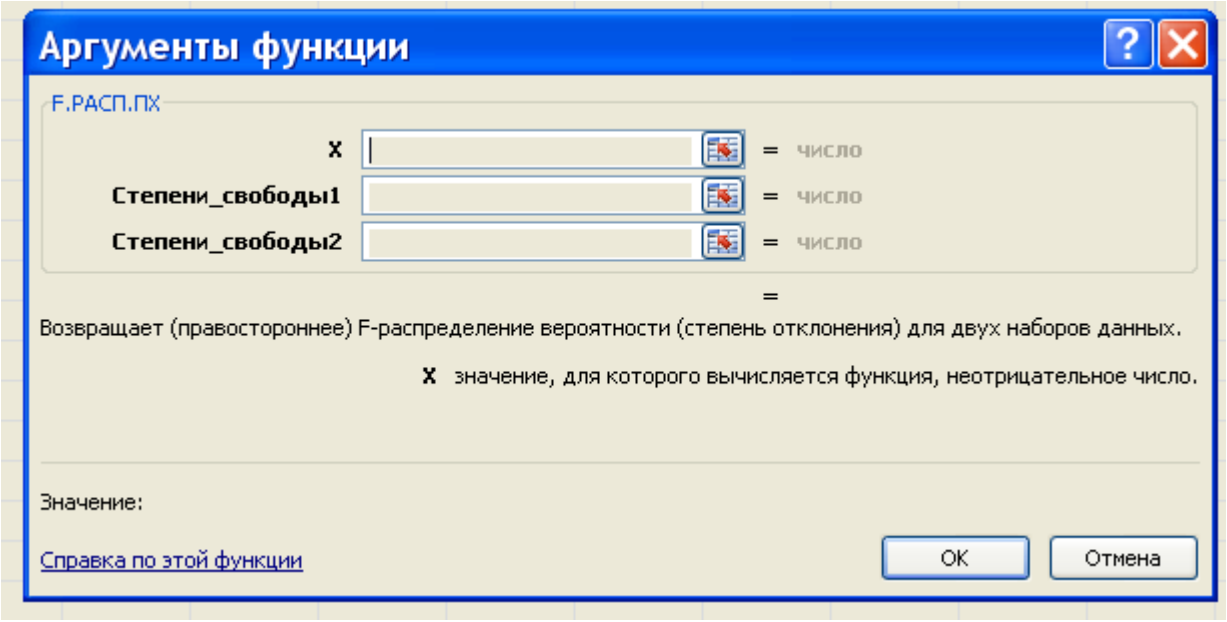
Р-значение для проверки гипотезы о равенстве дисперсий при двусторонней альтернативе вычисляется несколькими способами:

а) p = F.РАСП.(F набл., k1, k2, ИСТИНА), умноженное на 2, если Fнабл. < 1;

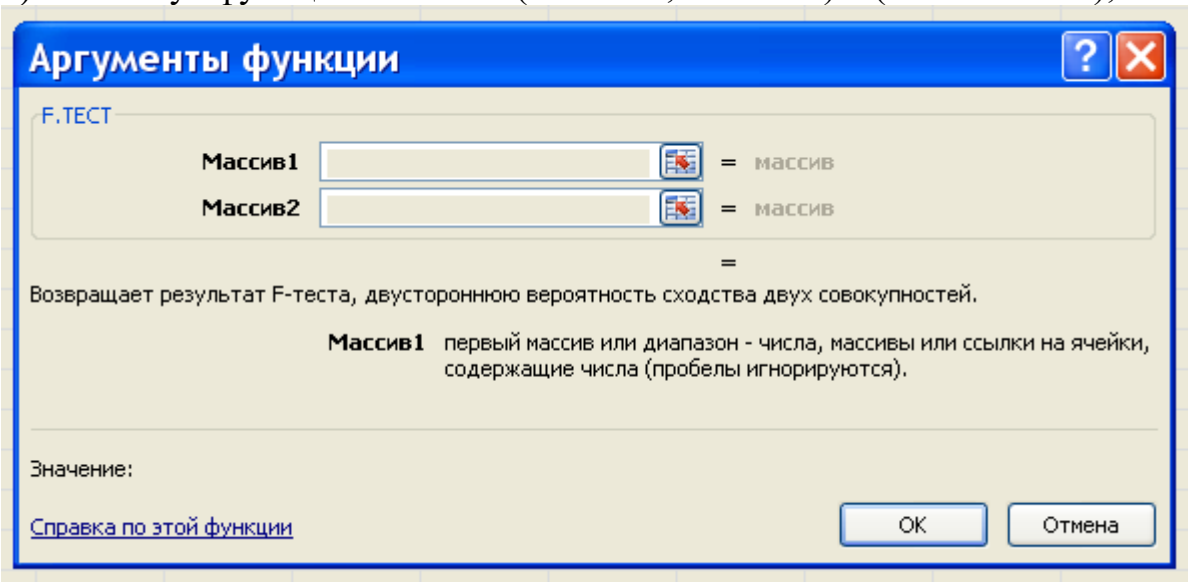
б) p = 2(1 - F.РАСП.(F набл., k1, k2, ИСТИНА.)), если F набл. > 1;



в) p = 2F.РАСП.ПХ(F набл., k1, k2), если F набл. > 1;



г) используя функцию Ф.ТЕСТ (массив 1, массив 2). (Или F.ТЕСТ.);



д) p–значение, это P(F<=f) одностороннее из двухвыборочного F-теста для дисперсий Пакета анализа, умноженное на 2.

Двухвыборочный F-тест для дисперсии

	Переменная 1	*Переменная 2*
Среднее	0,945690479	0,872651029
Дисперсия	4,830356281	15,16408992
Наблюдения	100	100
df	99	99
F	0,318539148	
P(F<=f) одностороннее	**1,5735E-08**	
F критическое одностороннее	0,717328593	

Пример.

На двух токарных станках обрабатываются втулки. Отобраны две пробы: из втулок, сделанных на первом станке $m = 18$ штук, на втором станке $n = 15$ штук. По данным этих выборок рассчитаны выборочные дисперсии диаметров втулок $S_1^2 = 6{,}3$ (для первого станка), $S_2^2 = 8{,}5$ для второго станка. Полагая, что диаметры втулок подчиняются нормальному закону, при уровне значимости $\alpha = 0{,}05$ выяснить, можно ли считать, что станки обладают различной точностью.

Решение.

$H_0: \sigma_1^2 = \sigma_2^2;$

$H_1: \sigma_1^2 \neq \sigma_2^2.$

$$F_{\text{набл.}} = \frac{8{,}5}{6{,}3} = 1{,}349$$

$$F_{\text{крит.}} = F_{1-\alpha/2}(15 - 1{,}18 - 1) = F_{1-\alpha/2}(14; 17) = F_{0{,}975}(14; 17) = 2{,}752640707$$

$1{,}349 < 2{,}752640707.$ Гипотеза не отвергается.

Р−значение = 2(1 - F.РАСП.(1,349, 14, 17, ИСТИНА.)) = 0,551546722 .

Р-значение больше уровня значимости, гипотеза принимается.

Задачи для самостоятельного решения.

1. По двум независимым выборкам, объемы которых m = 12 и n = 6, извлеченным из нормальных генеральных совокупностей X и У, найдены несмещенные оценки генеральных дисперсий $S_x^2 = 18$ и $S_y^2 = 27$. При уровне значимости α = 0,02 проверьте гипотезу H_0 : D(X) = D(Y) о равенстве генеральных дисперсий при альтернативной гипотезе H_1 : D(X) ≠ D(Y).

Ответ: Fнабл. = 1,5; Fкрит. = $F_{0,01}(5;11)$=5,32.

2. Руководство универсама решило упорядочить очередь к кассам и выяснить, является ли дисперсия времени ожидания в очереди к кассам одинаковой. Для этого были организованы 2 независимые выборки по 12 наблюдений времени ожидания в очереди к двум кассам. Получены результаты S_1 = 2,5 мин. и S_2= 3,1 мин. Проверьте гипотезу на уровне значимости α= 0,01.

Ответ: $F_{набл}$ =1,54; $F_{крит}$=4,46 , гипотеза не отвергается.

3. По двум независимым выборкам, объемы которых m = 10 и n =16, извлеченным из нормальных генеральных совокупностей X и У, найдены

несмещенные оценки генеральных дисперсий $S_x^2 = 87$ и $S_y^2 = 58$. При уровне значимости α = 0,05 проверьте гипотезу H_0 : D(X) = D(Y) о равенстве генеральных дисперсий при альтернативной гипотезе H_1 : D(X) > D(Y). Ответ: принимается.

4. Биржевой маклер исследует возможности двух инвестиций А и В. Ожидаемая ежегодная прибыль от этих инвестиций одинакова и равна 17,8 %. Инвестиция А предполагается на срок 10 лет, а инвестиция В – на 8 лет. Исправленные выборочные средние квадратические отклонения от этих двух инвестиций равны соответственно 7,14% и 3,21%. Предполагается, что распределения ежегодных прибылей подчиняются нормальному закону. При 1%-ном уровне значимости выяснить, одинаковы ли риски инвестиций А и В, определяемые дисперсиями ежегодных прибылей.

Ответ: $F_{набл}$ =4,81; $F_{крит}$=6,72 , гипотеза не отвергается, риски одинаковы.

5. Данные, представляющие собой случайную выборку числа ежедневных продаж нового сорта стирального порошка в магазинах города до и после показа его рекламы по местному телевидению, отражены в таблице

До рекламы	329	234	423	328	400	399	326	452	541	680	456	220
После рекламы	212	630	276	112	872	788	345	544	110	129	776	-----

На уровне значимости α = 0,05, выяснить, увеличилась ли дисперсия числа ежедневных продаж порошка после размещения рекламы.

Ответ: $F_{набл}$ =5,2985; $F_{крит}$=2,86 , гипотеза отвергается, дисперсия после рекламы значительно больше.

6. Важной мерой, ассоциируемой с риском акции, является стандартное отклонение или дисперсия цены акции. Финансовый аналитик проверяет одностороннюю гипотезу о том, что акция А имеет больший риск (большую дисперсию цены), чем акция В. Случайная выборка за 13 дней цены акции А дала величину выборочной дисперсии, равную $S_A^2 = 6{,}52\$^2$ и случайная выборка за 18 дней акции В дала выборочную дисперсию, равную $S_B^2 = 3{,}47\$^2$. Проверьте эту гипотезу при уровне значимости α=0,01.

Ответ: принимается.

7. Финансовый аналитик проверяет гипотезу о том, что акция В имеет больший риск (большую дисперсию цены), чем акция А. Случайная выборка за 14 дней цены акции А дала величину выборочной дисперсии, равную 0,38 и случайная выборка за 11 дней акции В дала выборочную дисперсию, равную 0,76 . Проверьте эту гипотезу при уровне значимости α=0,1.

Ответ: нет оснований отвергнуть гипотезу о равенстве генеральных дисперсий.

Проверка гипотезы о равенстве генеральных средних двух нормально распределенных совокупностей с известными дисперсиями

Пусть параметры σ_X^2 и σ_Y^2 известны, а генеральные средние неизвестны. В качестве основной гипотезы примем $H_0: \mu_X = \mu_Y$.

Пусть по выборкам $\vec{X}$ и $\vec{Y}$ вычислены выборочные средние $\bar{X}$ и $\bar{Y}$. В качестве статистики используется величина

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}.$$

Если гипотеза H_0 верна, то эта величина имеет стандартное нормальное распределение.

В соответствии с видом альтернативной гипотезы H_1 критическая область определяется по таблице:

$\boldsymbol{H_1}$	**критическая область**
$\mu_X > \mu_Y$	$Z > Z_{1-\alpha}$
$\mu_X < \mu_Y$	$Z < -Z_{1-\alpha}$
$\mu_X \neq \mu_Y$	$\lvert Z\rvert > Z_{1-\alpha/2}$

где $Z_{1-\alpha}$ – квантиль стандартного нормального распределения порядка 1-α.

$Z_{1-\alpha}$ = НОРМ.СТ. ОБР(1-α).

При этом наблюдаемый уровень значимости (р- значение) равен

P(Z > $Z_{\text{набл.}}$) = 1 – НОРМ.СТ.РАСП.(|$Z_{\text{набл.}}$ |; ИСТИНА) при односторонней альтернативной гипотезе и

P(Z > $Z_{\text{набл.}}$) = 2(1 – НОРМ.СТ.РАСП.(| $Z_{\text{набл.}}$ |; ИСТИНА)) при двусторонней альтернативной гипотезе.

Пример.

Для проверки эффективности новой технологии отобраны две группы рабочих: в первой группе численностью $m = 50$ человек, где применялась новая технология, выборочная средняя выработка составила $\bar{X} = 85$ (изделий), во второй группе численностью $n = 70$ человек выборочная средняя составила $\bar{Y} = 78$ (изделий). Предварительно установлено, что дисперсии выработки в группах

равны соответственно $\sigma_X^2 = 100$ и$\sigma_Y^2 = 74$. На уровне значимости $\alpha = 0{,}05$ выяснить влияние новой технологии на среднюю производительность.

Решение.

Проверяемая гипотеза $H_0: \mu_X = \mu_Y$, то есть средние выработки рабочих одинаковы по старой и новой технологии. В качестве альтернативной гипотезы возьмем $H_1: \mu_X > \mu_Y$– то есть средняя выработка больше при новой технологии.

$$Z_{\text{набл.}} = \frac{85-78}{\sqrt{\frac{100}{50}+\frac{74}{70}}} = \frac{7}{\sqrt{2+1{,}0571}} = \frac{7}{1{,}748} = 4{,}0035\,.$$

$Z_{\text{крит.}} = \text{НОРМ.СТ.ОБР}\,(1-0{,}05) = \ 1{,}644853627$

$Z_{\text{набл.}} = 4{,}0035 > 1{,}644853627 = Z_{\text{крит.}}$

Гипотеза H_0 отвергается, то есть при 5% уровне значимости можно сделать вывод о том, что новая технология позволяет увеличить среднюю выработку рабочих.

Р-значение равно 1- НОРМ.СТ.РАСП$(4{,}0035;\ \text{ИСТИНА}) =$ 0,000031206100581937500 – меньше заданного уровня значимости. Гипотеза отвергается.

Задачи для самостоятельного решения.

1. Имеются данные о результатах проверки прочности деталей

Партия деталей	Объем партии	Средняя прочность (кг/см^2)	Дисперсия генеральная σ^2
По старой технологии	100	40	250
По новой технологии	100	44	150

На уровне значимости $\alpha = 0{,}05$ выяснить, является ли повышение средней прочности деталей 40 кг/см^2 до 44 кг/см^2 следствием внедрения новой технологии или это результат случайно колеблемости показателей и новую технологию нельзя считать эффективной.

Ответ: Z набл. = - 2; Zкрит. = Z_α= 1,645. Гипотеза отвергается, новая технология эффективна.

2. По двум независимым выборкам, объемы которых m = 29 и n = 17, извлеченным из нормальных генеральных совокупностей X и У, найдены выборочные средние: $\bar{x} = 410$
и $\bar{y} = 401$. Генеральные дисперсии известны: D(X) = 71 и D(Y) = 54. Требуется при уровне значимости α = 0,004 проверить нулевую гипотезу H_0 : E(X) = E(Y) при альтернативной гипотезе H_1 : E(X) > E(Y).
Ответ: Z набл. = 3,794; Zкрит. = Z_α= 2,65. Гипотеза отвергается.

Проверка гипотезы о равенстве генеральных средних двух нормально распределенных совокупностей при неизвестных равных генеральных дисперсиях

Предположим, что генеральные дисперсии σ_X^2 и σ_Y^2 неизвестны, но одинаковы, то есть $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. В качестве основной гипотезы примем $H_0: \mu_X = \mu_Y$.

[Предварительно следует проверить гипотезу о равенстве дисперсий по критерию Фишера].

Пусть по выборкам $\vec{X}$ и $\vec{Y}$ найдены выборочные средние $\bar{X}$ и $\bar{Y}$ и исправленные выборочные дисперсии

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2\,;\;\; S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\,.$$

Составим так называемую сводную дисперсию

$$S_{\text{св.}}^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}.$$

В качестве статистики используется величина

$$t = \frac{\bar{X} - \bar{Y}}{S_{\text{св.}} \cdot \sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Если верна гипотеза H_0, то эта величина имеет распределение Стьюдента с $m + n - 2$ степенями свободы.

В соответствии с видом альтернативной гипотезы критическая область определяется по таблице.

H_1	критическая область
$\mu_X > \mu_Y$	$t > t_{1-\alpha}(m+n-2)$
$\mu_X < \mu_Y$	$t < -t_{1-\alpha}(m+n-2)$
$\mu_X \neq \mu_Y$	$\lvert t\rvert > t_{1-\alpha/2}(m+n-2)$

где $t_{1-\alpha}(m+n-2)$–квантиль порядка 1-α распределения Стьюдента с $k = m + n - 2$ степенями свободы.

При этом наблюдаемый уровень значимости (p- значение) равен

P(t > t $_{\text{набл.}}$) = 1 – СТЬЮДЕНТ.РАСП.(| t $_{\text{набл.}}$ |; m+n-2; ИСТИНА) = = СТЬЮДЕНТ.РАСП.ПХ(|t $_{\text{набл.}}$ | ; m+n-2) при односторонней альтернативной гипотезе.

При двусторонней альтернативе наблюдаемый уровень значимости (p-значение) равен

P(t > t $_{\text{набл.}}$) = 2(1 – СТЬЮДЕНТ.РАСП.(| t $_{\text{набл.}}$ | ; m+n-2; ИСТИНА)) = = 2 СТЬЮДЕНТ.РАСП.ПХ(|t $_{\text{набл.}}$ | ; m+n-2).

Пример.

Менеджер предприятия, работающего в две смены, решил выяснить, существует ли разница в производительности труда рабочих дневной и вечерней смены. Случайно организованная выборка 10 рабочих дневной смены показала, что средний выпуск продукции составляет 74,3 ед./ч., а исправленная выборочная дисперсия оказалась равной $S^2 = 16$ ед.^{2}/ч. Выборка же 10 рабочих вечерней смены выявила, что средний выпуск продукции равнялся 69,7 ед./ч., а $S^2 = 18$ ед.^{2}/ч. При двухпроцентном уровне значимости определить, существует ли разница в производительности труда рабочих дневной и вечерней смены.

Решение.

Проверим предварительно нулевую гипотезу $H_0: \sigma_1^2 = \sigma_2^2$ о равенстве генеральных дисперсий, пользуясь критерием Фишера. Альтернативная гипотеза $H_1: \sigma_1^2 \neq \sigma_2^2$.

$$F_{\text{набл.}} = \frac{18}{16} = \frac{9}{8} = 1\frac{1}{8} = 1{,}125\,.$$

$$F_{\text{крит.}} = F.\text{ОБР}\left(1 - \frac{\alpha}{2}; 9; 9\right) = F.\text{ОБР}(0{,}99; 9; 9) = 5{,}351128861\,.$$

Так как $F_{\text{набл.}} = 1{,}125 < 5{,}351128861 = F_{\text{крит.}}$, то гипотеза о равенстве генеральных дисперсий не отклоняется.

Составим сводную дисперсию

$$S^2_{\text{св.}} = \frac{9 \cdot 16 + 9 \cdot 18}{9 + 9} = \frac{306}{18} = 17\,.$$

$$S_{\text{св.}} = \sqrt{S^2_{\text{св.}}} = \sqrt{17} = 4{,}123.$$

Проверим теперь гипотезу о равенстве генеральных средних $H_0: \mu_X = \mu_Y$ при альтернативной гипотезе $H_1: \mu_X \neq \mu_Y$.

$$t_{\text{набл.}} = \frac{74{,}3 - 69{,}7}{4{,}123\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{4{,}6}{4{,}123 \cdot 0{,}4472} = 2{,}495\,.$$

$t_{\text{крит.}} = t_{1-\alpha/2}(9+9) = t_{0{,}99}(18) =$
$\text{СТЬЮДЕНТ.ОБР}(0{,}99; 18) = 2{,}55237963.$

Гипотеза не отвергается, так как 2,495 <2,55237963.
Р-значение равно 2 СТЬЮДЕНТ.РАСП.ПХ($|t_{\text{набл.}}|$; m+n-2) =
=2 СТЬЮДЕНТ.РАСП.ПХ(2,495; 18) = 0,02254101. Это больше, чем уровень значимости 2%, поэтому гипотеза о равенстве средних не отвергается.

Задачи для самостоятельного решения.

1. Из двух партий изделий, изготовленных на двух одинаково настроенных станках , извлечены малые выборки , объемы которых n = 10 и m = 12. Получены следующие результаты измерения контролируемого параметра изделий :

x_i	*3.4*	*3.5*	*3.7*	*3.9*
n_i	*2*	*3*	*4*	*1*

y_i	*3.2*	*3.4*	*3.6*
m_i	*2*	*2*	*8*

Требуется при уровне значимости α = 0,005 проверить гипотезу H_0 : E(X) = E(Y) о равенстве средних размеров изделий при конкурирующей гипотезе H_1 : E(X) ≠ E(Y). Предполагается , что случайные величины X и Y распределены нормально .

Ответ: $F_{\text{набл}}$ =1,05; $F_{\text{крит}}$=5,54; $t_{\text{набл.}}$ = 1,45; $t_{\text{крит}}$ = 2,85. Гипотеза о равенстве средних принимается. Таким образом, средние размеры изделий существенно не различаются.

2. Менеджер компании, производящей овсяные хлопья, решил выяснить, привело ли появление новой формы упаковки к увеличению сбыта хлопьев. Для этого он организовал случайную выборку из 30 однотипных магазинов, 18 из которых продавали овсяные хлопья в новой упаковке, а 12 – в старой упаковке. Им были получены следующие результаты

Новая упаковка	Старая упаковка
$n_1 = 18$ магазинов	$n_2 = 12$ магазинов
$\bar{X} = 130$ коробок	$\bar{Y} = 117$ коробок
$S_1 = 12$ коробок	$S_2 = 16$ коробок

При 5%-ном уровне значимости проверьте утверждение менеджера, что внедрение новой формы упаковки привело к увеличению сбыта овсяных хлопьев.

Ответ: $F_{набл.} = 1{,}78$; $F_{крит.} = 2{,}41$; $t_{набл.} = 2{,}63$; $t_{крит} = 1{,}71$. Гипотеза о равенстве средних отклоняется. Действительно, внедрение новой формы упаковки привело к увеличению сбыта хлопьев.

3. Произведены две выборки урожая пшеницы: при своевременной уборке и уборке с некоторым опозданием. В первом случае, при наблюдении восьми участков, выборочная средняя урожайность составила $\bar{X} = 16.2$ ц/га, $S_X = 3.2$ ц/га. Во втором случае, при наблюдении девяти участков, $Y = \square 13.9$ ц/га, $S_Y = 2.1$ ц/га. На уровне значимости $\alpha = \square 0.05$ выяснить влияние своевременности уборки урожая на увеличение среднего значения урожайности.

4. Отдел маркетинга автотранспортного предприятия провел обследование стоимости топлива на бензоколонках по трассе между городами А и В. Результаты показали, что средняя цена одного литра топлива на 52 заправках фирмы Торойл равна 30,5 руб. со стандартными отклонениями 0,085 руб., а на 58 заправках других фирм средняя цена одного литра – 30 руб. со стандартным отклонением 0,075 руб. Проверьте при уровне значимости $\alpha=0{,}05$ гипотезу о том, что средняя цена одного литра топлива на заправках Торойл выше цены этого же топлива на заправках других фирм.

Ответ: нулевая гипотеза отклоняется, действительно, средняя цена выше.

5. Техническая норма предусматривает в среднем 40 секунд на выполнение определенной технологической операции на конвейере. От работниц, работающих на этой операции, поступила жалоба, что они в действительности затрачивают на эту операцию больше времени. Для проверки данной жалобы произведены хронометрические измерения времени выполнения этой операции у 16 работниц, занятых на этой операции, получено среднее время операции 42 секунды и исправленное среднее квадратическое отклонение $S = 3{,}5$ секунды. При уровне значимости $\alpha = 0{,}01$ проверить гипотезу о том, что время выполнения операции соответствует норме.

Ответ: $t_{набл.} = 2{,}2131$; $t_{крит.} = 2{,}6$. Гипотеза принимается, жалобы работниц необоснованны.

6. На основании сделанного прогноза средняя дебиторская задолженность однотипных предприятий региона должна составить 120 ден.ед. Выборочная проверка 10 предприятий показала среднюю задолженность величиной 135 ден.ед. и среднее квадратическое отклонение S = 20 ден.ед. При уровне значимости 0,05 выяснить, можно ли принять данный прогноз.

Ответ: $t_{набл.} = 2{,}25$; $t_{крит.} = 1{,}83$. Прогноз отклоняется, реальная задолженность оказалась больше.

7. На основании сделанного прогноза средняя задолженность за электричество владельцев однокомнатных квартир в данном регионе должна составить 18,2 руб. Выборочная проверка 30 квартир показала среднюю задолженность величиной 18,9 руб. и среднее квадратическое отклонение S = 2,18 руб. При уровне значимости 0,05 выяснить, можно ли принять данный прогноз, или задолженность превысит ожидаемое значение.

Ответ: $t_{набл.} = 0{,}5$; $t_{крит.} = 1{,}699$. Прогноз принимается.

8. Из двух партий изделий, изготовленных на двух одинаково настроенных станках , извлечены малые выборки , объемы которых n = 10 и m = 12. Получены следующие результаты измерения контролируемого параметра изделий

x_i	3.4	3.5	3.7	3.9
n_i	2	3	4	1

y_i	3.2	3.4	3.6
m_i	2	2	8

Требуется при уровне значимости $\alpha = 0.005$ проверить гипотезу H_0: E(X)=E(Y) о равенстве средних размеров изделий при конкурирующей гипотезе H_1: E(X)≠E(Y). Предполагается, что случайные величины X и Y распределены нормально.

Ответ: $t_{набл.} = 1{,}62$; $t_{крит.} = 1{,}75$. Так как $1.62 < 1.75$, то H_0 принимается. Это означает, что для данных выборок сроки уборки не влияют на урожайность.

9. По двум независимым выборкам, объемы которых m = 24 и n = 15, извлеченным из нормальных генеральных совокупностей X и У, найдены выборочные средние: $\bar{x} = 395$ и $\bar{y} = 388$ и исправленные дисперсии $S_x^2 = 64{,}3$ и $S_y^2 = 58{,}8$. Требуется при уровне значимости $\alpha = 0{,}002$ проверить нулевую гипотезу H_0 : E(X) = E(Y) при альтернативной гипотезе H_1 : E(X) ≠ E(Y).

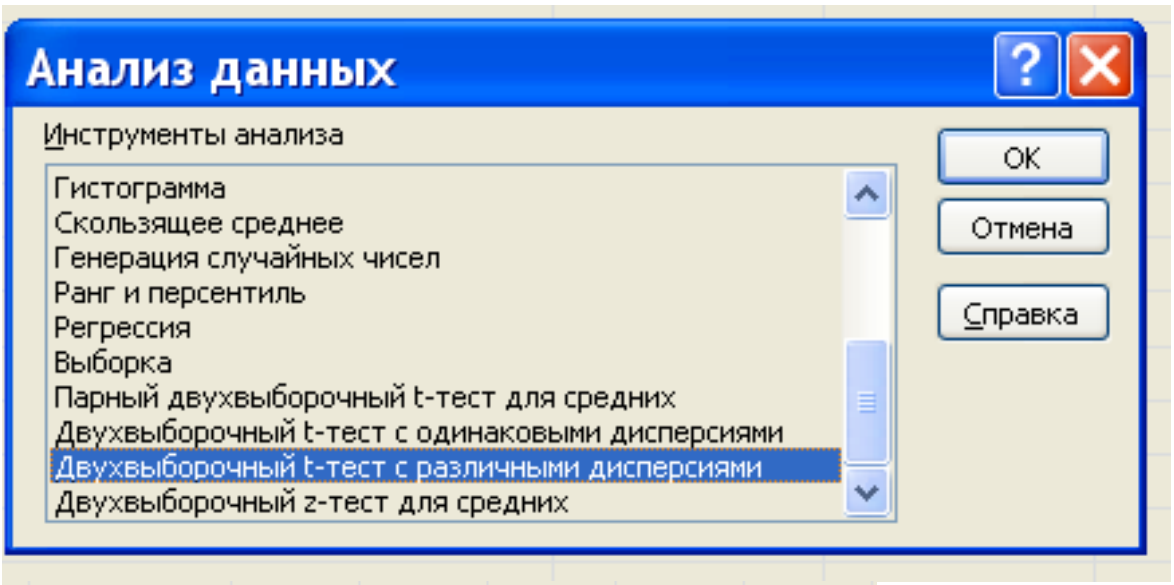
Ответ: принимается; $S_{CB}^2 = 62{,}2189$; $t_{набл.} = 2{,}596$; $t_{крит} = t_{0{,}001}(37) = 3{,}3304$.

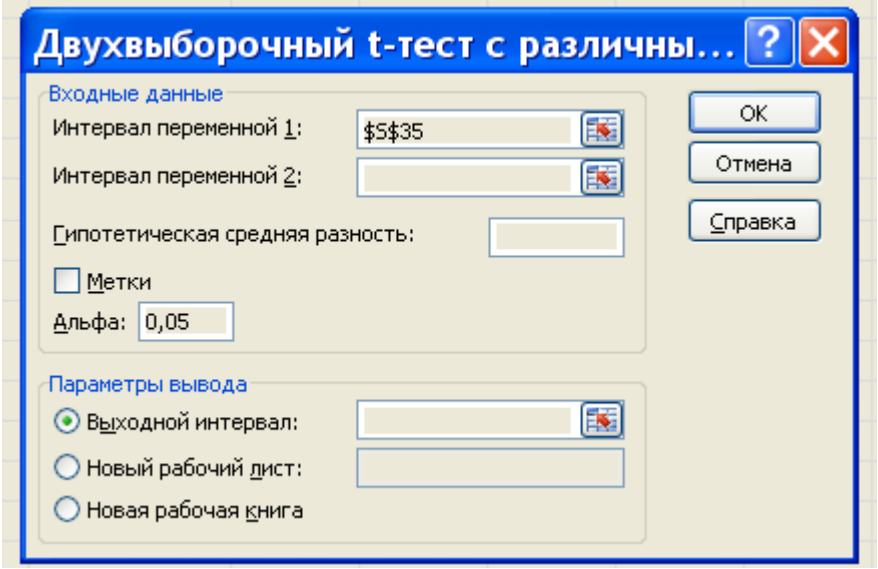
10. По двум независимым выборкам, объемы которых m = 14 и n = 12, извлеченным из нормальных генеральных совокупностей X и У **с равными дисперсиями**, найдены выборочные средние: $\bar{x} = 416{,}7$ и $\bar{y} = 431{,}6$ и исправленные дисперсии $S_x^2 = 22{,}4$ и $S_y^2 = 26{,}5$. Требуется при уровне значимости α = 0,01 проверить нулевую гипотезу H_0 : E(X) = E(Y) при альтернативной гипотезе H_1 : E(X) ≠ E(Y).

Ответ: отвергается.

Проверка гипотезы о равенстве генеральных средних двух нормально распределенных совокупностей при неизвестных генеральных дисперсиях

Можно использовать Двухвыборочный t-тест с различными дисперсиями из Пакета анализа.





В поле «Гипотетическая средняя разность» надо ввести 0.

Там содержится и p–значение для двусторонней альтернативы, это P(T ≤ t) двустороннее, и для односторонней альтернативы это это P(T ≤ t) одностороннее.

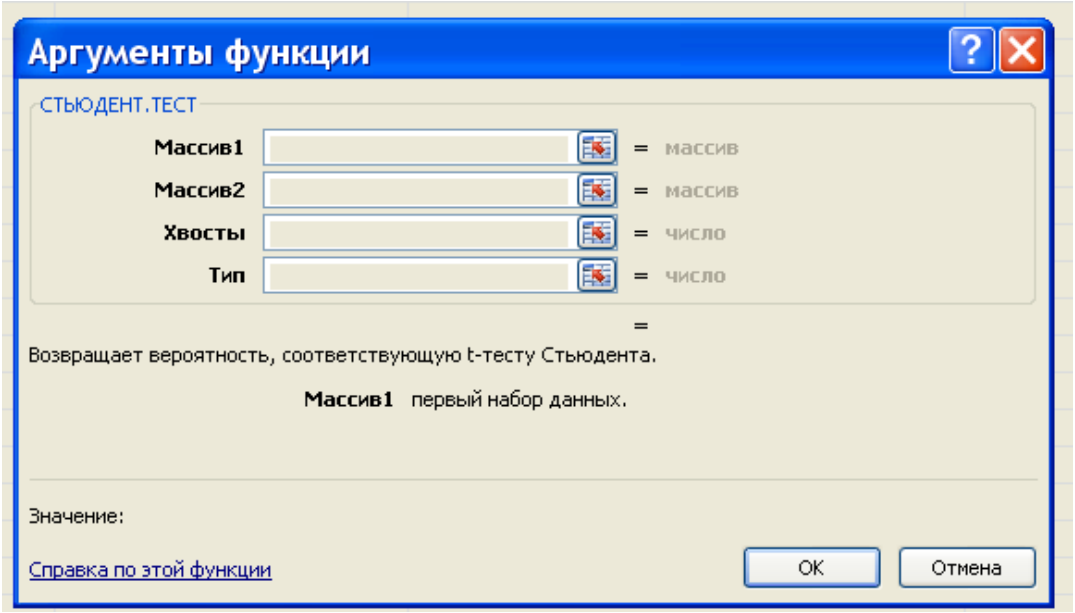
Двухвыборочный t-тест с различными дисперсиями		
	Переменная 1	*Переменная 2*
Среднее	0,945690479	0,872651029
Дисперсия	4,830356281	15,16408992
Наблюдения	100	100
Гипотетическая разность средних	0	
df	**156**	
t-статистика	**0,163343856**	
P(T<=t) одностороннее	**0,435229562**	
t критическое одностороннее	1,654679996	
P(T<=t) двухстороннее	**0,870459124**	
t критическое двухстороннее	1,975287508	

В этом тесте t-статистика , это

$$t\text{набл.} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

Число степеней свободы df (degrees of freedom) это частное от числа $(\frac{S_X^2}{m} + \frac{S_Y^2}{n})^2$, деленного на число $\frac{(\frac{S_X^2}{m})^2}{m-1} + \frac{(\frac{S_Y^2}{n})^2}{n-1}$, и округленное до ближайшего целого.

P-значение, полученное с помощью Двухвыборочного t-теста с различными дисперсиями из Пакета *анализа является округленным, неточным, так как число степеней свободы df находится приближенно, округленно.* ***Поэтому для получения точного p-значения следует использовать функцию ТТЕСТ или СТЬЮДЕНТ.ТЕСТ(массив 1; массив 2).***



В поле «Хвосты» надо ввести 2 для нахождения двустороннего p-значения или 1 для одностороннего p-значения. В поле «Тип» надо ввести 3. Это номер t-теста в списке Пакета анализа.

Пример.

Заданы выборки:

1 выборка

8,929632184	7,836203092	5,642894773	6,532145942
6,086538104	4,891105998	8,2560022	7,501303496
6,535418077	5,863914264	6,843757926	4,83427031
7,478797595	5,381205863	5,460819619	5,250085149
7,604153705	7,501160251	4,535836904	4,968941464

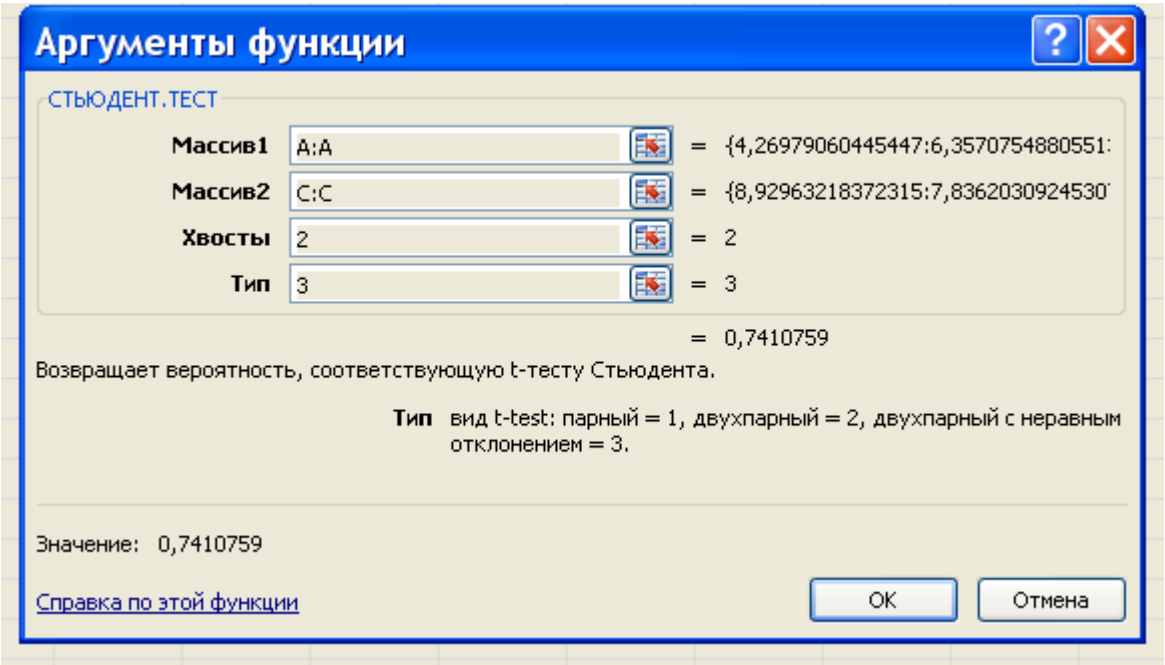
2—я выборка

4,269790604	6,357075488	4,750953319	6,089457582
7,484564246	8,063082003	7,19261099	7,292343905
8,593941327	6,849404524	4,625306626	5,12538865
5,315023752	7,835763807	3,678151568	6,034821052
5,330406946	4,229087079	5,295258614	7,06164921
6,837315975	6,537029929	6,093821792	6,826151018
6,527609837	6,068335022	6,065472842	6,655323357
6,235396011	8,878533133		

На уровне значимости 0,06 проверить гипотезу о равенстве генеральных средних при альтернативной гипотезе об их неравенстве.

Решение.

Скопируем обе выборки в файл Excel и используем функцию СТЬЮДЕНТ.ТЕСТ



Р-значение равно 0,7410759 - больше заданного уровня значимости, гипотеза не отвергается.

Задачи для самостоятельного решения.

1. Заданы выборки:

1 выборка

6,305917469 ; 5,701345418 ; 7,756481482; 5,546320227;
6,552945325; 6,469117198; 6,657840077; 6,081315024;
7,533038643; 7,93548823 ; 5,743165803; 7,476096999;
6,781625034; 6,335539711; 6,12833499 ; 6,285177351;
7,122028892; 6,406921403; 6,445460331; 5,807105099;
6,591846232; 7,55250183; 5,818522896; 7,277283276;
4,637383955; 7,36352884 ; 7,080070264;

2-я выборка

6,02850202; 5,869314072; 4,705659257; 6,796649422;
7,826974132; 3,331289071; 5,308616045; 5,105409463;
7,258576324; 8,669431799; 5,733408278; 5,830250365;
9,149563147; 6,234308847; 6,659087438; 7,78986402;
6,173867468; 6,438953341; 6,054730938; 8,937794581;
5,350432792; 7,020027687; 5,199615833; 6,651412213;
6,232974619; 6,185093815; 7,805817467; 7,818921012;
6,152200575; 8,475889507; 6,558463673; 6,016725882;
7,97152922 ; 9,903399649; 8,021787898; 7,649576281;
7,620108653; 7,284657903.

На уровне значимости 0,04 проверить гипотезу о равенстве генеральных средних при альтернативной гипотезе, утверждающей, что генеральное среднее второй выборки больше, чем у первой..

Ответ: p-значение равно 0,21134034, больше заданного уровня значимости, гипотеза не отвергается.

2. Заданы выборки:

1 выборка

6,086538104; 4,891105998; 8,2560022; 7,501303496;
6,535418077; 5,863914264; 6,843757926; 4,83427031;
7,478797595; 5,381205863; 5,460819619; 5,250085149;
7,604153705; 7,501160251; 4,535836904; 4,968941464;
8,929632184; 7,836203092; 5,642894773; 6,532145942;

2-я выборка

7,826974132; 3,331289071; 5,308616045; 5,105409463;
7,258576324; 8,669431799; 5,733408278; 5,830250365;
9,149563147; 6,234308847; 6,659087438; 7,78986402;
6,173867468; 6,438953341; 6,054730938; 8,937794581;
5,350432792; 7,020027687; 5,199615833; 6,651412213;
6,232974619; 6,185093815; 7,805817467; 7,818921012;
6,152200575; 8,475889507; 6,558463673; 6,016725882;
7,97152922 ; 9,903399649; 8,021787898; 7,649576281;
7,620108653; 6,02850202; 5,869314072; 4,705659257;
6,796649422; 7,284657903.

На уровне значимости 0,03 проверить гипотезу о равенстве генеральных средних при альтернативной гипотезе об их неравенстве.

Ответ: p-значение равно 0,115209986, больше заданного уровня значимости, гипотеза не отвергается.

Проверка гипотез о законе распределения генеральной совокупности

Одной из важнейших задач математической статистики является установление теоретического закона случайной величины, характеризующей изучаемый признак по опытному (эмпирическому) распределению, представляющему вариационный ряд.

Предположение о виде закона распределения может быть выдвинуто, исходя из теоретических предпосылок, опыта аналогичных предшествующих исследований или, наконец, на основании графического изображения эмпирического распределения.

Параметры распределения, как правило, неизвестны, поэтому их заменяют оценками, вычисленными по выборке.

Требуется сделать заключение: согласуются ли результаты наблюдений, то есть выборочные данные с высказанным предположением о виде закона распределения.

Критерием согласия называется статистический критерий о предполагаемом законе неизвестного распределения. В отличие от ранее рассмотренных критериев, альтернативная гипотеза явным образом не выдвигается, и задача ставится так: согласуются ли данные выборки с предполагаемым законом распределения?

Существуют различные критерии согласия Пирсона, Колмогорова, Фишера, Смирнова и других. Критерий согласия Пирсона – наиболее часто употребляемый критерий для проверки простой гипотезы о законе распределения.

Критерий χ^2 (хи-квадрат) Пирсона

Пусть (Ω, L, P)– вероятностное пространство, $A_1, A_2, \dots, A_l \in L$ – попарно несовместные события, такие, что $A_1 + A_2 + \cdots + A_l = \Omega$.

В качестве основной гипотезы H_0 примем гипотезу, состоящую в том, что вероятности этих событий заданы таблицей:

Событие	A_1	…	A_l
Вероятность	p_1	…	p_l

Пусть n_i – эмпирическая частота события A_i, то есть число испытаний, в которых A_i наступило. Исходными данными для критерия χ^2 Пирсона является таблица эмпирических частот

Событие	A_1	…	A_l
Частота	n_1	…	n_l

Если основная гипотеза верна, согласно статистическому определению вероятности $\hat{p}_i \approx p_i$, где $\hat{p}_i = n_i/n$– относительная частота события A_i. В качестве меры одновременной близости l пар чисел $(\hat{p}_i, p_i)$ можно принять любую сумму вида

$$c_1(\hat{p}_1 - p_1)^2 + c_2(\hat{p}_2 - p_2)^2 + \cdots + c_l(\hat{p}_l - p_l)^2,$$

в которой $c_i > 0$ – какие-либо положительные числа. К.Пирсон обнаружил, что если придать б*о*льшие веса маловероятным событиям, положив $c_i = n/p_i$, то при неограниченном увеличении n распределение статистики

$$\chi^2 = \sum_{i=1}^{l} \frac{n}{p_i}(\hat{p}_i - p_i)^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}$$

$$\left[\frac{n}{p_i}(\hat{p}_i - p_i)^2 = \frac{n}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 = \frac{n}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 = \frac{n}{p_i}\left(\frac{n_i - np_i}{n}\right)^2 =\right.$$

$$\left. = \frac{n}{p_i}\frac{(n_i - np_i)^2}{n^2} = \frac{(n_i - np_i)^2}{np_i}\right]$$

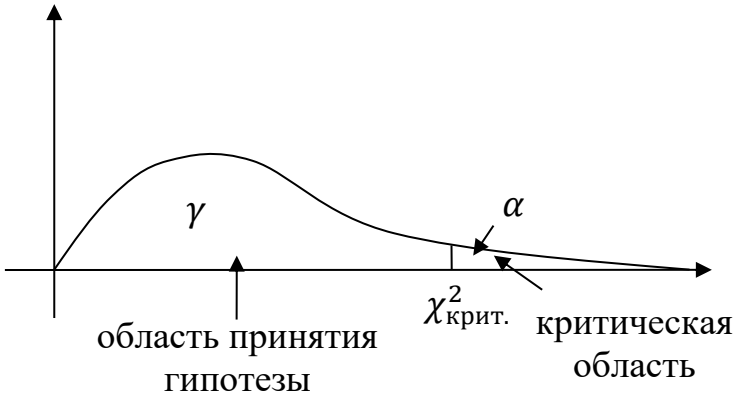
перестает зависеть от конкретных значений вероятностей p_i и стремится к распределению χ^2 с $k = l - 1$ степенями свободы.

Заметим, что при верной гипотезе H_0 случайные величины n_i (частоты) распределены по биноминальному закону с параметрами n и p_i, вследствие чего $np_i = E(n_i)$ называется ожидаемой (теоретической) частотой события A_i.

Таким образом, при проверке гипотезы о виде распределения с помощью сравнения выборочного распределения долей признака с теоретическим распределением в качестве статистики используется величина

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i},$$

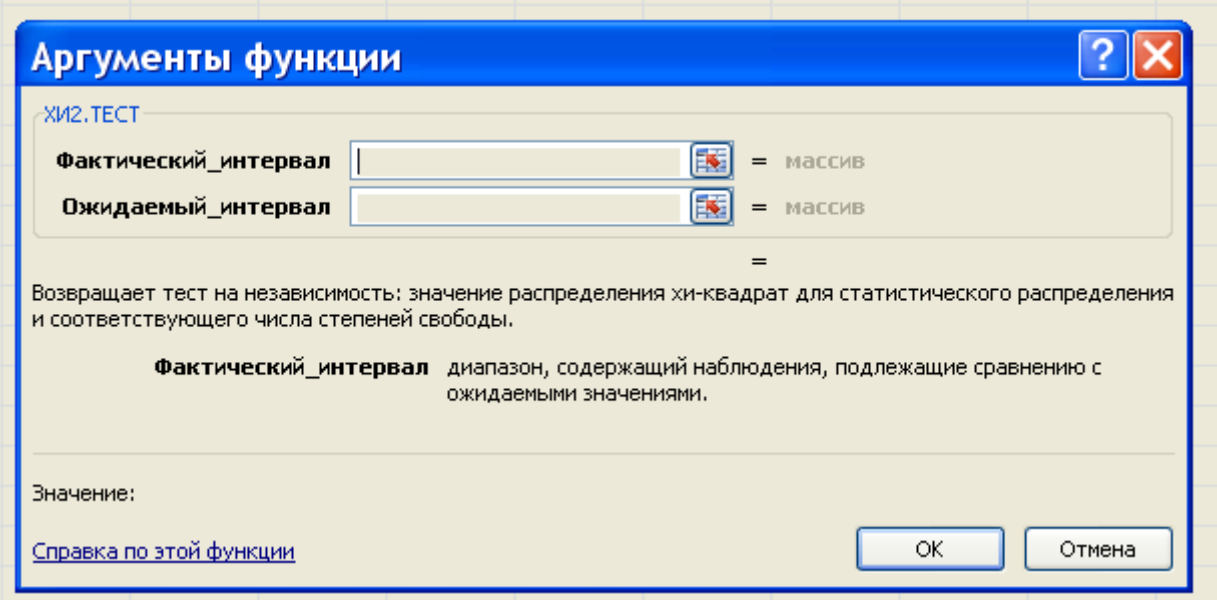
имеющая распределение χ^2 с $k = l - 1$ степенями свободы. Используется односторонняя критическая область $\chi^2 > \chi^2_{1-\alpha}(l-1)$. Здесь $\chi^2_{1-\alpha}(l-1)$ – квантиль распределения χ^2 уровня 1-α с $k = l - 1$ степенями свободы. $\chi^2_{1-\alpha}(l-1)$ = ХИ2.ОБР(1-α; *l*-1).



На практике данный критерий Пирсона применяется, если объем выборки $n > 50$ и все ожидаемые частоты $np_i > 5$. Несоблюдение этих условий приводит к значительному отклонению фактического уровня значимости $P_{H_0}\left(\chi^2 > \chi^2_{1-\alpha}(l-1)\right)$ от требуемого уровня α.

Р-значение для проверки гипотезы по критерию Пирсона (односторонняя критическая область): p = 1- ХИ2.РАСП ($\chi^2_{\text{набл.}}$; L- 1; ИСТИНА) = ХИ2.РАСП.ПХ ($\chi^2_{\text{набл.}}$; L – 1).

Или использовать функцию ХИ2.ТЕСТ(фактический диапазон частот; ожидаемый диапазон частот). ХИ2.ТЕСТ годится только для проверки полного распределения, где число степеней свободы L- 1!



Пример.

По результатам переписи населения установлен следующий возрастной состав

	до 50 лет	от 50 лет и старше
женщины	35%	20%
мужчины	35%	10%

Спустя несколько лет после переписи было отобрано случайным образом 1000 человек, возрастной состав которых оказался следующим

	до 50 лет	от 50 лет и старше
женщины	343	212
мужчины	343	102

Проверим гипотезу о том, что возрастной состав не изменился при уровне значимости $\alpha = 0{,}05$.

События A_1, A_2, A_3, A_4 состоят в том, что отобранный человек оказался:

A_1– женщиной до 50 лет;

A_2– женщиной от 50 лет и старше;

A_3– мужчиной до 50 лет;

A_4– мужчиной от 50 лет и старше.

Таким образом, $n_1 = 343$, $n_2 = 212$, $n_3 = 343$, $n_4 = 102$– эмпирические частоты.

Теоретические частоты:

$np_1 = 1000 \cdot 0{,}35 = 350$; $np_2 = 1000 \cdot 0{,}2 = 200$;

$np_3 = 1000 \cdot 0{,}35 = 350$; $np_4 = 1000 \cdot 0{,}1 = 100$.

Составим статистику χ^2

$$\chi^2 = \frac{(343-350)^2}{350} + \frac{(212-200)^2}{200} + \frac{(343-350)^2}{350} +$$

$$+ \frac{(102-100)^2}{100} = \frac{49}{350} + \frac{144}{200} + \frac{49}{350} + \frac{4}{100} =$$

$$= \frac{49}{175} + 0{,}72 + 0{,}04 = 1{,}04$$

$\chi^2_{\text{набл.}} = 1{,}04$; $\chi^2_{\text{крит.}} = \chi^2_{0{,}95}(4-1) = \chi^2_{0{,}95}(3) =$ $\text{ХИ2.ОБР}(0{,}95; 3) = 7{,}81472790$.

Так как, $\chi^2_{\text{набл.}} = 1{,}04 < 7{,}81472790 = \chi^2_{\text{крит.}}$, то гипотеза о том, что возрастной состав не изменился, принимается.

Замечание.

Число степеней свободы равно $k = l - 1$, так как присутствует одно уравнение связи $n_1 + n_2 + \cdots + n_l = n$.

Задачи для самостоятельного решения.

1. Сотрудники компании Горгаз, по опыту знают, что в некотором районе 80% домохозяйств полностью оплачивают счета за газ, 10% имеют месячную задолженность, 6% имеют задолженность в 2 месяца и 4% - более, чем 2 месяца. В конце прошедшей зимы компания проверила счета 400 домохозяйств и обнаружила, что в 287 из них счета полностью оплачены, 49 имеют месячную задолженность, 30 – двухмесячную задолженность и 34 – белее, чем двух месячную задолженность. При 5%-ном уровне значимости выясните, подтверждают ли данные по задолженностям тенденцию, сложившуюся в прошлые годы.

Ответ: Так как, $\chi^2_{\text{набл.}} = 27{,}1781 > 7{,}815 = \chi^2_{\text{крит.}}$, то проверка гипотезы указывает на увеличение доли домохозяйств с задолженностями за прошедшую зиму.

2. При 4040 бросаниях монеты французский естествоиспытатель Бюффон получил 2048 выпадений герба и 1992 выпадения цифры. На уровне значимости $\alpha = 0{,}05$ проверим гипотезу о том, что монета была правильной.

Ответ: Так как, $\chi^2_{\text{набл.}} = 0{,}778 < 3{,}8 = \chi^2_{\text{крит.}}$, то гипотеза принимается.

3. Фирма владеет тремя магазинами. Руководство фирмы решило выяснить, посещают покупатели все три магазина одинаково охотно, либо имеется некоторое различие. Для проверки была собрана информация о количестве покупателей, сделавших покупки в течение недели. Оказалось, что в первом магазине это число составляет 160 человек, во втором - 225, в третьем 215.На уровне значимости $\alpha = 0{,}01$ выяснить, можно ли объяснить различия только случайностью.

Ответ: Так как, $\chi^2_{\text{набл.}} = 12{,}25 > 9{,}2 = \chi^2_{\text{крит.}}$, то гипотеза о равномерном распределении числа покупателей отвергается, объяснить различия в посещаемости магазинов нельзя объяснить только случайностью.

4. Используя критерий Пирсона при уровне значимости 0,05, установить, случайно или значимо расхождение между эмпирическими и теоретическими частотами, которые вычислены, исходя из предположения о нормальном распределении признака X генеральной совокупности:

n_i	14	18	32	70	20	36	10
n_i'	10	24	34	80	18	22	12.

Ответ: Так как, $\chi^2_{\text{набл.}} = 13{,}93 > 9{,}5 = \chi^2_{\text{крит.}}$, то гипотеза отвергается.

5. Используя критерий Пирсона, при уровне значимости $\alpha = 0{,}05$ установить, случайно или значимо расхождение между эмпирическими m_i и теоретическими n_i частотами, которые вычислены из гипотезы о нормальном распределении генеральной совокупности:

m_i	5	10	20	8	7
n_i^*	6	14	18	7	5

Ответ: Так как, $\chi^2_{\text{набл.}} = 2{,}47 < 6{,}0 = \chi^2_{\text{крит.}}$, то расхождения между эмпирическими и теоретическими частотами случайны,

Критерий χ^2 с оценкой параметров распределения

Случай полностью определенного гипотетического распределения встречается достаточно редко. Чаще встречаются предположения о теоретическом распределении, параметры которого не известны. Тогда приходится накладывать еще два условия: равенство выборочного среднего и математического ожидания и равенство выборочной и теоретической дисперсии.

Поэтому число степеней свободы статистики

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i} \text{ равно } l - 3.$$

В общем случае число степеней свободы $k = l - 3 = l - r - 1$, где r– число параметров теоретического распределения, оценки которых вычислены по экспериментальным данным. Используется односторонняя критическая область $\chi^2 > \chi^2_{1-\alpha}(l - r - 1)$.

Р-значение для проверки гипотезы по критерию Пирсона (односторонняя критическая область): p = 1- ХИ2.РАСП ($\chi^2_{\text{набл.}}$; L - r -1; ИСТИНА) = ХИ2.РАСП.ПХ ($\chi^2_{\text{набл.}}$; L – r -1); где r – число параметров предполагаемого распределения.

Пример.

Проверка гипотезы о распределении случайной величины по закону Пуассона

Проведено наблюдение за числом вызовов на телефонной станции. С этой целью в течение 100 случайно выбранных 5-секундных интервалов времени регистрировалось число вызовов. Получен следующий статистический ряд:

Число вызовов x_i	n_i
0	8
1	28
2	31
3	18
4	9
5	6
$\sum$	n = 100

Проверить гипотезу о том, что распределение числа вызовов согласуется с законом Пуассона. Уровень значимости принять $\alpha = 0{,}05$.

Решение.

Для распределения Пуассона $P(X = m) = \frac{\lambda^m}{m!}e^{-\lambda}$, один параметр $\lambda = E(X) = D(X), p_i = \frac{\lambda^i}{i!}e^{-\lambda}$.

Найдем точечную оценку параметраλ.

$$\bar{X} = \frac{\sum n_i x_i}{n} = \frac{8 \cdot 0 + 28 \cdot 1 + 31 \cdot 2 + 18 \cdot 3 + 9 \cdot 4 + 6 \cdot 5}{100} =$$

$$= \frac{28 + 62 + 54 + 36 + 30}{100} = \frac{210}{100} = 2{,}1\,.$$

$\lambda = 2{,}1.$

Таким образом, $p_i = \frac{2{,}1^{\mathrm{i}}}{\mathrm{i}!}e^{-2{,}1}$.

Вычислим $\chi^2_{\text{набл.}}$, заполнив таблицу

Число вызовов x_i	n_i	p_i	np_i
0	8	0,122	12,2

1	28	0,257	25,7
2	31	0,270	27
3	18	0,189	18,9
4	9	0,099	9,9
5	6	0,063	6,3
∑	n = 100	1	100

$$\chi^2_{\text{набл.}} = \frac{(8-12{,}2)^2}{12{,}2} + \frac{(28-25{,}7)^2}{25{,}7} + \frac{(31-27)^2}{27} +$$

$$+\frac{(18-18{,}9)^2}{18{,}9} + \frac{(9-9{,}9)^2}{9{,}9} + \frac{(6-6{,}3)^2}{6{,}3} =$$

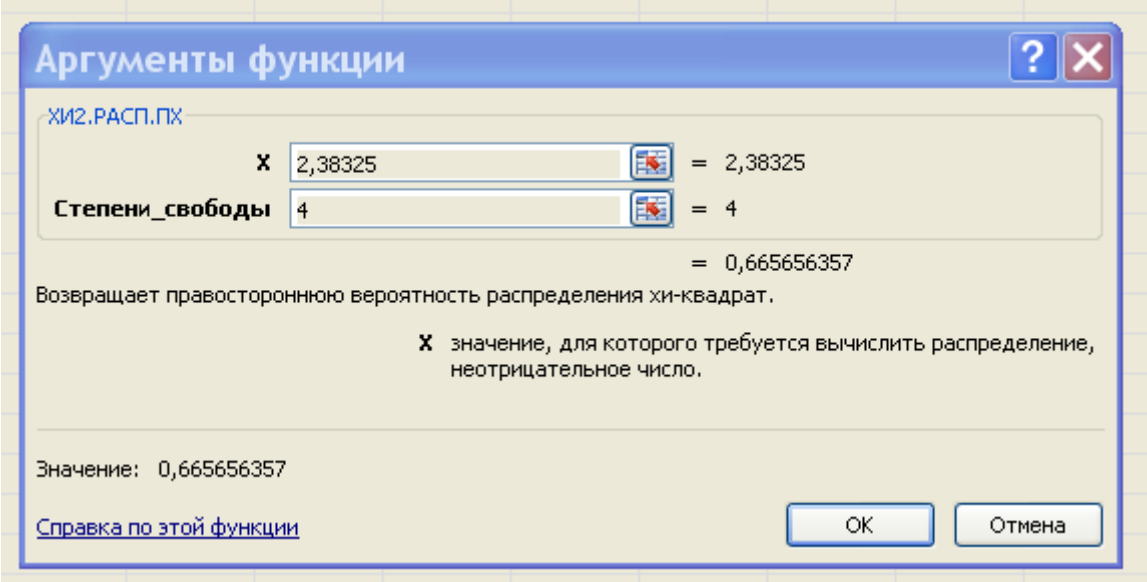
$$= 1{,}4459 + 0{,}2058 + 0{,}59259 + 0{,}042857 + 0{,}081818 +$$

$$+0{,}0142857 = 2{,}38325 .$$

$$\chi^2_{\text{крит.}} = \chi^2_{0{,}95}(6-1-1) = \chi^2_{0{,}95}(4) = \text{ХИ2.ОБР}(0{,}95; 4) = 9{,}487729037 .$$

Так как $\chi^2_{\text{набл.}} = 2{,}38325 < 9{,}487729037 = \chi^2_{\text{крит.}}$, то нет снований для отклонения гипотезы.

Р-значение (односторонняя критическая область равно ХИ2.РАСП.ПХ (2,38325; 4) = 0,665656357.



Так как р-значение больше заданного уровня значимости, гипотеза не отклоняется.

Пример.

Проверка гипотезы о нормальном распределении генеральной совокупности

Имеются сгруппированные данные о дневной выручке в магазине электротоваров (в тыс.руб.).

Суммы продаж	Число продаж
190 – 200	10
200 – 210	26
210 – 220	56
220 – 230	64
230 – 240	30
240 – 250	14

Требуется проверить нулевую гипотезу о том, что сумма продаж X есть случайная величина, распределенная по нормальному закону при уровне значимости $\alpha = 0{,}05$.

Объем выборки $n = 200$; интервальное среднее равно

$$\bar{X} = 221;\ S^2 = 152;\ S = 12{,}33.$$

№	интерва	левfz гр	правая гр	ni	F(Zi)	F(Zi+1)	pi	npi	(ni-npi)^2	(ni-npi)^2/npi
	1	190	200	10	0,005965256	0,044268783	0,038303527	7,660705401	5,4722992	0,714333594
	2	200	210	26	0,044268783	0,186160822	0,141892039	28,37840784	5,6568239	0,199335491
	3	210	220	56	0,186160822	0,46768002	0,281519198	56,30383952	0,0923185	0,001639648
	4	220	230	64	0,46768002	0,767282599	0,299602579	59,92051578	16,642191	0,277737788
	5	230	240	30	0,767282599	0,938336376	0,171053778	34,2107555	17,730462	0,518271568
	6	240	250	14	0,938336376	0,990663287	0,052326911	10,46538216	12,493523	1,193795231
	Сумма			200			0,984698031	196,9396062		**2,90511332**
										7,814727903

$$\chi^2_{\text{набл.}} = 2{,}90511332$$

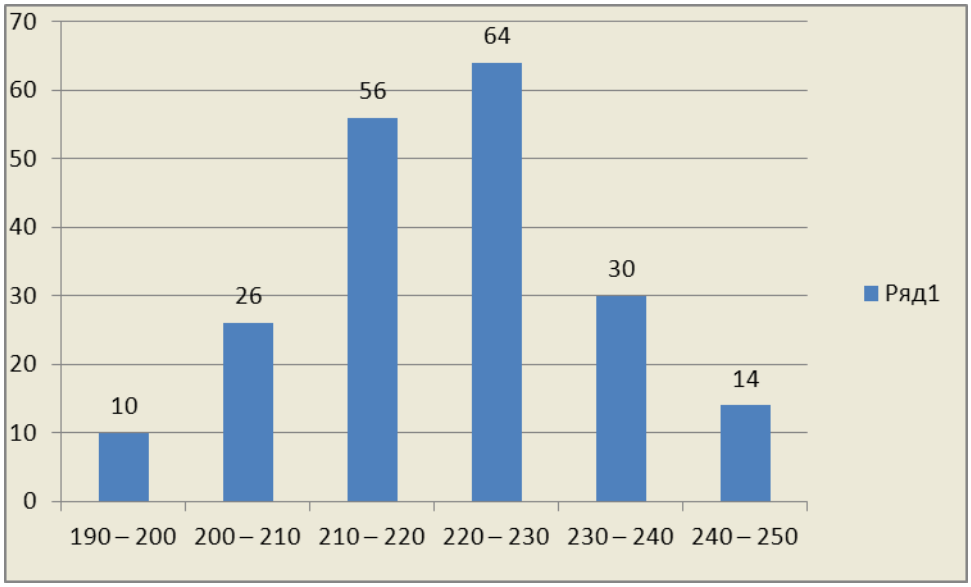
Число степеней свободы для χ^2 равно $k = l - r - 1 = 6 - 2 - 1 = 3$.

$$\chi^2_{\text{крит.}} = \chi^2_{0{,}95}(3) = \text{ХИ2.ОБР}(0{,}95;\ 3) = 7{,}814727903\ .$$

Так как $\chi^2_{\text{набл.}} = 2{,}90511332 < 7{,}814727903 = \chi^2_{\text{крит.}}$, то нет снований для отклонения нулевой гипотезы о нормальном законе распределения.

Р-значение (односторонняя критическая область) равно
ХИ2.РАСП.ПХ (2,90511332; 3) = 0,406487386.

Так как р-значение больше заданного уровня значимости, гипотеза не отклоняется.



Задачи для самостоятельного решения.

1. На уровне значимости $\alpha = 0{,}025$ проверить гипотезу о нормальном распределении генеральной совокупности, если известны эмпирические и теоретические частоты:

Эмпирические частоты	5	10	20	25	14	3
Теоретические частоты	6	14	28	18	8	3

Ответ: $\chi^2_{набл.} = 10{,}8175$; $\chi^2_{крит.} = 9{,}4$ Данные наблюдений не согласуются с гипотезой о нормальном распределении генеральной совокупности.

2. На уровне значимости $\alpha = 0{,}05$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объемом $n = 200$:

x_i	5	7	9	11	13	15	17	19	21
m_i	15	26	25	30	26	21	24	20	13

Ответ: $\chi^2_{набл.} = 22{,}2$; $\chi^2_{крит.} = 12{,}6$ Данные наблюдений не согласуются с гипотезой о нормальном распределении генеральной совокупности.

ЛИТЕРАТУРА

1. В.И. Соловьев. Анализ данных в экономике: теория вероятностей, прикладная статистика и визуализация данных в Microsoft Excel : учебник. М.: КНОРУС, 2019.
2. В.Н. Калинина, В.И. Соловьев. Анализ данных. Компьютерный практикум : учебное пособие. М.: КНОРУС, 2017.
3. А.В. Браилов, В.И. Глебов, С.Я. Криволапов, П.Е. Рябов. Теория вероятностей и математическая статистика. М.-Ижевск : НИЦ «Регулярная и хаотическая динамика»; Институт компьютерных исследований, 2016.
4. А.С. Солодовников, В.А. Бабайцев, А. В. Браилов. Математика в экономике. Часть 3. Теория вероятностей и математическая статистика. М.: Финансы и статистика, 2008.
5. А.В. Браилов. Лекции по математической статистике. М.: Финакадемия, 2007.
6. В.И. Глебов, С.Я. Криволапов. Практикум по математической статистике. Проверка гипотез с использованием Excel, MatCalc, R и Python: учебное пособие. М.: Прометей, 2109.
7. Л.В. Рудикова. Microsoft Office Excel 2016. СПб.: БХВ-Петербург, 2017.
8. Р.Н. Вадзинский. Статистические вычисления в среде Excel. Библиотека пользователя. СПб.: Питер, 2008.
9. Браилов А. В., Рябов П. Е. Теория вероятностей и математическая статистика. Методические рекомендации по самостоятельной работе. Часть 5.
10. Браилов А. В., Люлько Я.А., Рябов П. Е. Теория вероятностей и математическая статистика. Методические рекомендации по самостоятельной работе. Часть 6.
11. И.Е. Денежкина, М.Г. Орлова, Ю.Н. Швецов. Основы математической статистики. Учебно-методическое пособие для самостоятельной работы бакалавров. М.: Финуниверситет, кафедра «Теория вероятностей и математическая статистика», 2010.
12. Браилов А.В., Гончаренко В.М., Зададаев С.А., Коннов В.В. Вопросы и задачи по теории вероятностей. Для студентов бакалавриата экономики. М.: Финансовый университет при Правительстве РФ, кафедра «Теория вероятностей и математическая статистика», 2010.
13. А.В. Браилов, А.С. Солодовников. Сборник задач по курсу «Математика в экономике». Часть 3. М.: Финансы и статистика, ИНФРА-М, 2010.
14. Математика для экономистов: от Арифметики до Эконометрики. Н.Ш. Кремер [и др.]; под ред. проф. Н.Ш. Кремера. М.: Издательство Юрайт; ИД Юрайт, 2011.
15. Гмурман В. Е. Теория вероятностей и математическая статистика - М., Высш.шк., 2003.
16. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике. - М., Высш.шк., 1979.

17. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов. — 2-е изд., перераб. и доп.— М.: ЮНИТИ-ДАНА, 2004.
18. Письменный Д.Т. Конспект лекций по теории вероятностей и математической статистике. - М.: Айрис-пресс, 2004.
19. В.С. Мхитарян, Л.И. Трошин, Е.В. Астафьева, Ю.Н. Миронкина. Теория вероятностей и математическая статистика: учеб пособие. Под ред. В.С. Мхитаряна. М.: Маркет ДС, 2010.
20. Л.И. Ниворожкина, З.А. Морозова, И.Э. Гурьянова. Математическая статистика с элементами теории вероятностей в задачах с решениями: Учебное пособие. Москва: Издательско-торговая корпорация «Дашков и К°», 2015.
21. Ниворожкина Л.И., Морозова З.А. Теория вероятностей и математическая статистика. М.: Эксмо, 2008.
22. В.А. Колемаев, О.В. Староверов, В.Б. Турундаевский. Теория вероятностей и математическая статистика. Под ред. В.А. Колемаева. М.: Высш. шк.. 1991.
23. Л.И. Константинова. Теория вероятностей и математическая статистика: учебное пособие. Томск: Изд-во Национального исследовательского Томского политехнического университета, 2010.
24. Маценко П. К., Селиванов В. В. Руководство к решению задач по теории вероятностей. Учебное пособие. - Ульяновск: УлГТУ, 2000.
25. Попов В.А., Бренерман М.Х. Руководство к решению задач по теории вероятностей и математической статистике. - Казань: Издательство КГУ, 2008.
26. Г.И. Просветов. Теория вероятностей и математическая статистика: задачи и решения. – М. Издательство «Альфа-Пресс», 2009.
27. И.П. Кирилловская, Г.В. Егорова. Теория вероятностей и математическая статистика. – Тольятти: Изд-во ТГУС, 2006.
28. В.А. Карасев, С.Н. Богданов, Г.Д. Левшина. Теория вероятностей и математическая статистика. Раздел 2. Математическая статистика: учебно-методическое пособие. М.: МИСиС, 2005.