МИНОБРНАУКИ РОССИИ

Орский гуманитарно-технологический институт (филиал)
федерального государственного бюджетного образовательного учреждения
высшего образования «Оренбургский государственный университет»
(Орский гуманитарно-технологический институт (филиал) ОГУ)

Кафедра программного обеспечения

Методические указания по выполнению и защите лабораторных работ по дисциплине «Б1.Д.В.14 Обработка экспериментальных данных на электронно-вычислительных машинах»

Уровень высшего образования

БАКАЛАВРИАТ

Направление подготовки

09.03.01 Информатика и вычислительная техника
(код и наименование направления подготовки)

Программное обеспечение средств вычислительной техники и автоматизированных систем
(наименование направленности (профиля) образовательной программы)

Тип образовательной программы

Программа бакалавриата

Квалификация

Бакалавр

Форма обучения

Очная

Год начала реализации программы (набора)

2019

г. Орск 2018

Методические указания предназначены для обучающихся очной формы обучения направления подготовки 09.03.01 Информатика и вычислительная техника профилю Программное обеспечение средств вычислительной техники и автоматизированных систем по дисциплине **«Б1.Д.В.14 Обработка экспериментальных данных на электронно-вычислительных машинах»**

Составитель _____ О.В. Подсобляева

Методические указания рассмотрены и одобрены на заседании кафедры программного обеспечения, протокол № 1 от «01» сентября 2018 г.

Заведующий кафедрой _____ Е.Е. Сурина

Согласовано:
Председатель методической комиссии по направлению подготовки 09.03.01 Информатика и вычислительная техника
_____ Е.Е.Сурина
«12» сентября 2018 г.

Пояснительная записка

В результате изучения дисциплины «Б1.Д.В.14 Обработка экспериментальных данных» у обучающихся должны быть сформированы знания, умения и навыки:

- сформировать навыки и умения связанные с проведением исследований;
- применять необходимые для построения моделей знания принципов действия и описания составных частей программы (информационных, методологических, алгоритмических и средств вычислительной техники);
- реализовывать программу средствами вычислительной техники; определять характеристики объектов профессиональной деятельности по разработанным моделям.

Одной из наиболее эффективных форм закрепления теоретических знаний и выработки навыков самостоятельной работы являются практические занятия.

Целью проведения лабораторных занятий является:

- закрепление знаний студентов по основам проектной деятельности,
- формирование у студентов навыков использования современных технических средств и технологий для решения проектных и исследовательских задач.

Тематический план

Таблица 1 – Тематический план выполнения лабораторных работ по дисциплине «Б1.Д.В.14 Обработка экспериментальных данных» для обучающихся направления подготовки 09.03.01 Информатика и вычислительная техника профиль подготовки Программное обеспечение средств вычислительной техники и автоматизированных систем

Лабораторные работы

в 6 семестре

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов
1	2	Метод скользящего среднего	2
2	3	Обработка экспериментальных данных	2
3	4	Поиск параметров распределений случайных величин	2
4	5	Расчет статистических оценок прогнозов	2
5	6	Построение и оценка парной регрессии	2
6	6	Построение и оценка множественной регрессии	2
7	7	Подготовка эталонов распознавания печатных знаков	2
8	8	Визуализация данных с помощью диаграмм	2
		Итого:	16

в 7 семестре

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов
1	5	Экспериментальные исследования.	4
2	5	Случайные величины и законы распределения.	2
3	6	Метод наименьших квадратов.	2
4	6	Постановка обратных задач и формализация.	2
5	7	Ошибки эксперимента и их оценивание.	2
6	7	Элементарная теория корреляции.	2
7	8	Интервальные оценки.	2
		Итого:	16

Методические указания по выполнению и оформлению лабораторных работ

Практические работы по дисциплине «Обработка экспериментальных данных» предполагают решение задач по темам, представленным в тематическом плане.

В практической работе должны быть выполнены все предусмотренные задания. В работе должна просматриваться логическая последовательность и взаимная увязка основных частей работы.

Рекомендуемая структура лабораторных работ:

1) цель практической работы;

2) задание в соответствии с выбранным вариантом;

3) теоретическая часть, включающая краткое изложение теоретических положений по теме практической работы, формулы для решения задания;

4) практическая часть, включающая решение задания по теме практической работы. Дополнительно для наглядности расчетный материал может быть представлен в виде таблиц, графиков;

5) выводы по лабортаорной работе;

6) список использованной литературы.

Практические работы могут быть оформлены:

- машинописным текстом на листах формата А4.

Титульный лист оформляется на основе СТО 02069024. 101 – 2014 «РАБОТЫ СТУДЕНЧЕСКИЕ. Общие требования и правила оформления».

Работа защищается устно и принимается к зачету, если нет замечаний по ее выполнению и оформлению. При отсутствии зачтенных лабораторных работ студент не допускается к зачету по дисциплине «Б1.Д.В.14 Обработка экспериментальных данных».

Лабораторная работа №1
Метод скользящего среднего

Цель: научится строить прогнозы с применением метода скользящего среднего

Общие сведения:

Метод скользящего среднего применять достаточно просто, однако он не всегда позволяет сделать точный прогноз. При использовании этого метода прогноз любого периода представляет собой не что иное, как получение среднего показателя нескольких результатов наблюдений временного ряда. Например, если вы выбрали скользящее среднее за три месяца, прогнозом на май будет среднее значение показателей за февраль, март и апрель. Выбрав в качестве метода прогнозирования скользящее среднее за четыре месяца, вы сможете оценить майский показатель как среднее значение показателей за январь, февраль, март и апрель.

Вычисления с помощью этого метода довольно просты и достаточно точно отражают изменения основных показателей предыдущего периода. Иногда при составлении прогноза они эффективнее, чем методы, основанные на долговременных наблюдениях.

Например, вы составляете прогноз объема продаж давно и хорошо освоенной вашим предприятием продукции, причем средний показатель объема за последних несколько лет составляет 1000 единиц. Если ваша компания планирует значительное сокращение штата торговых агентов, логично предположить, что среднемесячный объём реализации будет сокращаться, по крайней мере, на протяжении нескольких месяцев.

Если, для прогнозирования объема продаж в будущем месяце вы воспользуетесь средним значением данного показателя за последние 24 месяца, то, вероятно, получите результат, несколько завышенный по сравнению с фактическим. Но если прогноз будет составлен на основании данных всего лишь за три последних месяца, то он намного точнее отразит последствия сокращения штата торговых агентов. В данном случае прогноз будет отставать по времени от фактических результатов на один-два месяца, как это показано на рисунке 1.

Разумеется, это происходит потому, что при применении скользящего среднего за три последних месяца каждый из трех показателей (за этот временной период) отвечает за одну треть

значения прогноза. При 24-месячном скользящем среднем показатели этих же последних месяцев отвечают только за 1/24 часть значения прогноза. Таким образом, чем меньше число результатов наблюдений, на основании которых вычислено скользящее среднее, тем точнее оно отражает изменения в уровне базовой линии.

Порядок выполнения лабораторной работы

Пример: обслуживание клиентов

Предположим, вы — менеджер отдела обслуживания клиентов фирмы, специализирующейся на разработке программного обеспечения. На днях вы получили от внештатной сотрудницы сообщение по электронной почте, в котором она известила вас, что в последнее время ей постоянно звонят клиенты с жалобами на новые программы вашей фирмы. Вы просите ее зарегистрировать все жалобы, поступающие в течение двух недель и сообщить вам результаты.

Полученный по истечении этого времени отчёт включает ежедневное количество звонков с жалобами на конкретный программный продукт. Эти данные вы вводите в рабочий лист Excel, расположив их в ячейках A1:A10, как показано на рисунке 2. Чтобы понять, существует ли какая-либо определенная тенденция поступления жалоб, вы создаете на основе средних данных о полученных звонках скользящее среднее.

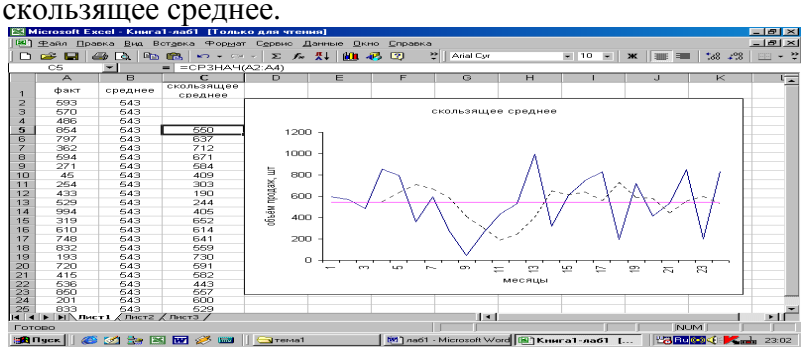


Рисунок 1 - Скользящее среднее объема продаж за три месяца позволяет отслеживать фактический объем продаж с большей точностью, нежели аналогичные наблюдения, но за длительный период времени

Вы решаете воспользоваться трехдневным скользящим средним. Почему за трехдневный период? Ответ на этот вопрос таков: скользящее среднее за меньший период может не отразить тенденцию, а за более продолжительный период слишком сгладит ее. Одним из способов создания скользящего среднего в Excel является прямое введение формулы.

Таким образом, чтобы получить трехдневное скользящее среднее количества телефонных звонков, вы вводите в ячейку B4:

= СРЗНАЧ(A1:A3) (Результат: 10,33).

Затем с помощью средства *Автозаполнение* копируете и вставляете эту формулу в ячейки B5:B10. В данном случае (и это видно из рисунка 2) показатель скользящего среднего действительно имеет тенденцию к увеличению, поэтому поставьте в известность о тревожной ситуации руководство отдела тестирования продукции вашей компании.

При использовании средства *Автозаполнение* для горизонтального или вертикального перетаскивания выбранной ячейки или диапазона ячеек, надо установить указатель в маркер заполнения, который представляет собой маленький крестик в нижнем правом углу выбранного диапазона.

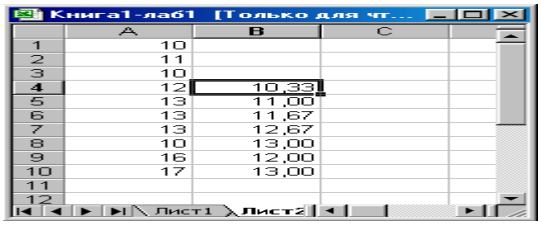


Рисунок 2 - Прогнозы с использованием скользящего среднего приводят к потере некоторых данных в начальном периоде базовой линии

Лабораторная работа №2
Обработка экспериментальных данных

Нормальное распределение

Нормальным распределением (или законом Гаусса) называется распределение непрерывной случайной величины X, плотность которой определяется по формуле

$$f(x) = \frac{1}{o\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad (1.1)$$

где m и σ– параметры распределения. Можно доказать, что параметр m равен математическому ожиданию, а параметр σ– стандартному отклонению случайной величины X.

Функция (интегральная) нормального распределения

$$F(x) = \frac{1}{o\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt. \quad (1.2)$$

Для краткой записи нормального распределения с параметрами m и σ используют обозначение N (m, σ). В частном случае параметры m= 0, σ = 1. Нормальное распределение N (0, 1) называется стандартным нормальным распределением. В этом случае плотность распределения

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (1.3)$$

Функция стандартного нормального распределения иногда называется функцией Лапласа, она имеет специальное обозначение

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

Для вычисления значений плотности и функции нормального распределения в Excel используется встроенная статистическая функция НОРМ.РАСП (рис. 1.1). Чтобы открыть эту функцию необходимо нажать на значок *f(x)* и диалоговом окне выбрать «Статистические».

Синтаксис функции:

НОРМ.РАСП (X; Среднее; Стандартное_откл; Интегральная)

X - значение аргумента, на основе которого вычисляется нормальное распределение.
Среднее - среднее значение распределения.
Стандартное_откл - стандартное отклонение распределения.
При Интегральной = 0 рассчитывается функция плотности, а при 1 рассчитывается функция распределения.

Пример:

= НОРМ.РАСП (70; 63; 5; 1) возвращает0,92.

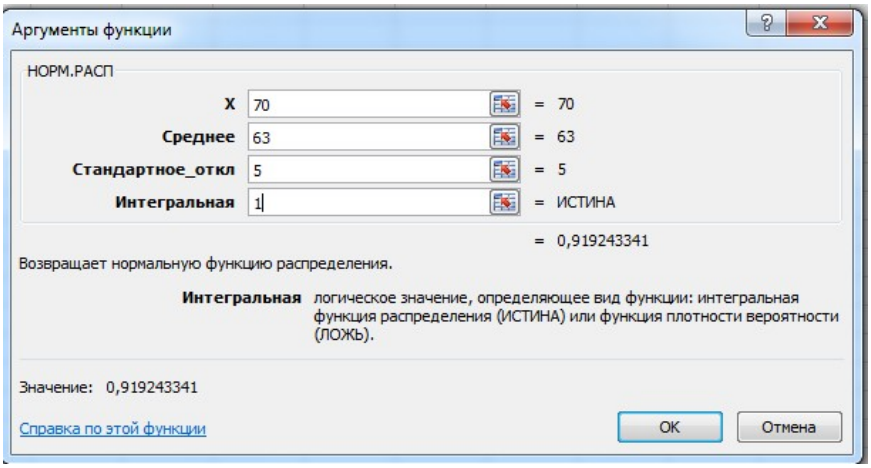


Рис. 1.1. Встроенная функция НОРМ.РАСП .

ЗАДАНИЕ

1. Введите в таблицу значения аргумента x в диапазоне от –3 до 5 с шагом 0,2.
2. Вычислите значение плотности стандартного нормального распределения, а также плотности нормального распределения с параметрами m= 2, σ = 1; m= 0, σ = 0,5; m= 1, σ = 2.
3. Используя мастер диаграмм, постройте соответствующие кривые распределения.
4. Отредактируйте графики в соответствии с образцом оформления (рис.1.2).
5. Для заданных параметров нормального распределения постройте семейство графиков функции распределения.

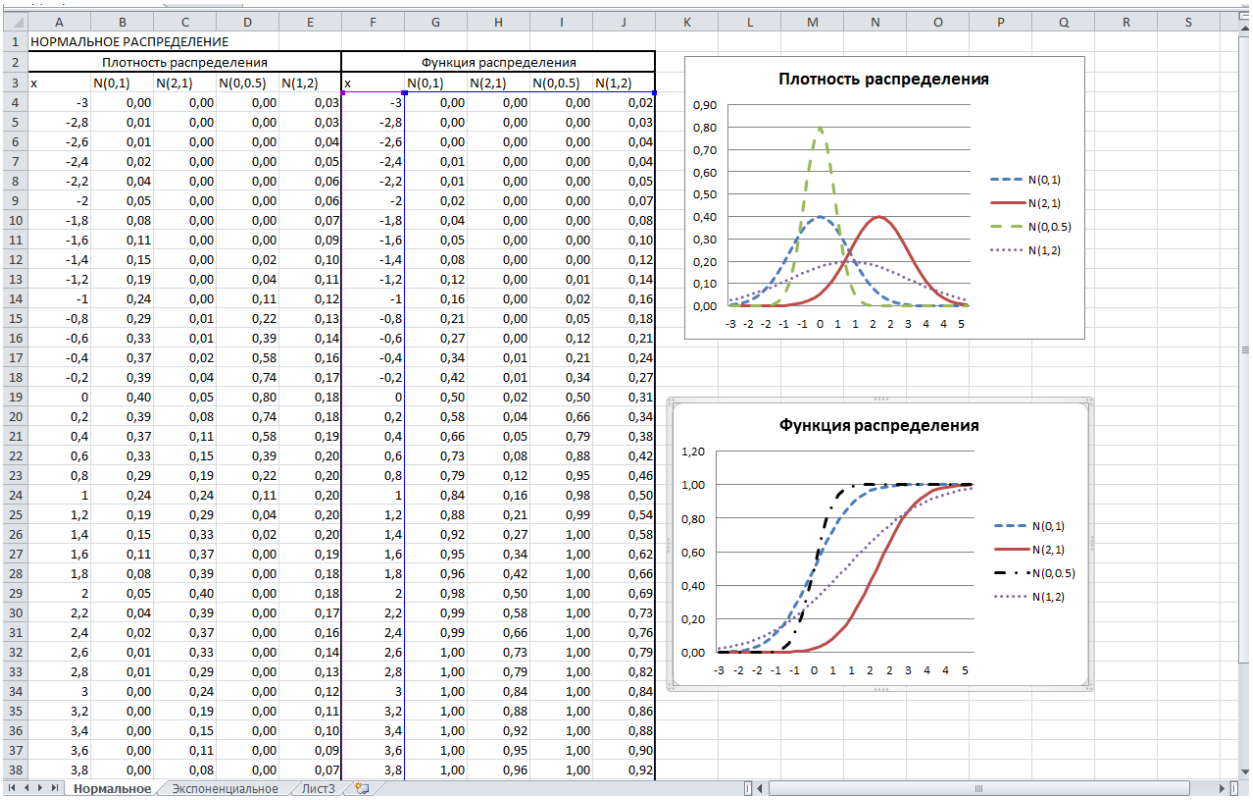


Рис. 1.2. Образец оформления рабочего листа «Нормальное распределение»

Экспоненциальное распределение

Экспоненциальным(или показательным) называется распределение непрерывной случайной величины X, плотность которой

$$f(x) = \lambda e^{-\lambda x} \tag{1.5}$$

при x> 0 (при x ≤0 f(x) = 0).

Функция экспоненциального распределения

$$F(x) = 1 - \lambda e^{-\lambda x} \tag{1.6}$$

Математическое ожидание случайной величины X, имеющей экспоненциальное распределение, равно

$$m_X = 1 / \lambda, \tag{1.7}$$

а дисперсия

$$D_X = 1 / \lambda^2. \tag{1.8}$$

Для вычисления значений плотности и функции экспоненциального распределения в Excel используется встроенная статистическая функция ЭКСП.РАСП (рис. 1.3).

Синтаксис:

ЭКСП.РАСП (X; Лямбда; Интегральная)

X - значение аргумента функции.

Лямбда - значение параметра.

Интегральная: логическое значение, которое определяет форму функции. При 0 рассчитывается плотность, а при 1 рассчитывается функция распределения.

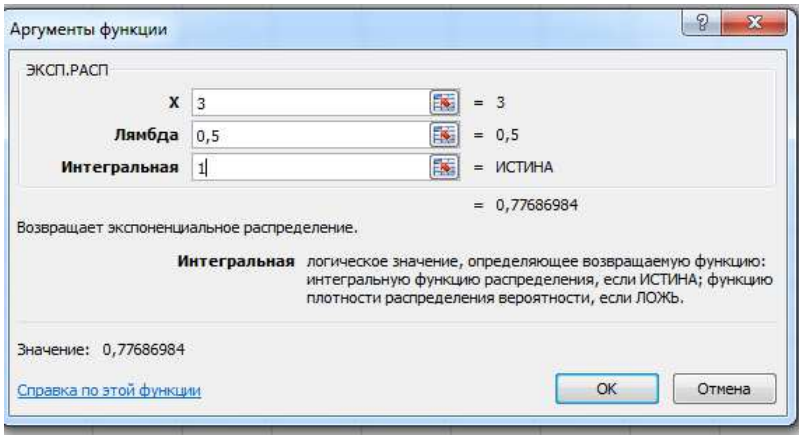
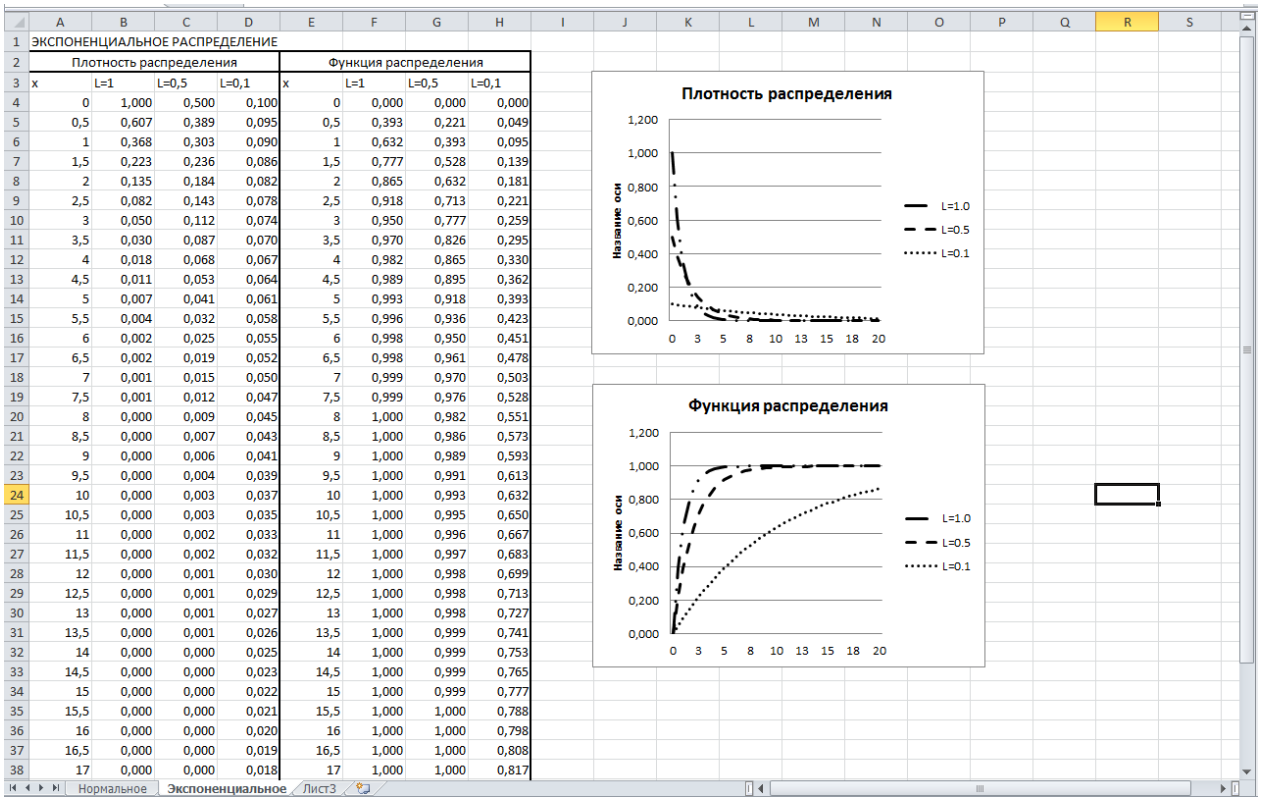


Рис. 1.3. Встроенная функция ЭКСП.РАСП

ЗАДАНИЕ

1. Введите в таблицу значения аргумента x в диапазоне от 0 до20 с шагом 0,5.
2. Вычислите значение плотности экспоненциального распределения при λ= 1; λ= 0,5; λ= 0,1.
3. Используя мастер диаграмм, постройте соответствующие кривые распределения.
4. Отредактируйте графики в соответствии с образцом оформления (рис.1.4).

5. Для заданных значений параметра λ постройте семейство графиков функции экспоненциального распределения.



ЭКСПОНЕНЦИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

	Плотность распределения				Функция распределения			
x		L=1	L=0,5	L=0,1	x	L=1	L=0,5	L=0,1
	0	1,000	0,500	0,100	0	0,000	0,000	0,000
	0,5	0,607	0,389	0,095	0,5	0,393	0,221	0,049
	1	0,368	0,303	0,090	1	0,632	0,393	0,095
	1,5	0,223	0,236	0,086	1,5	0,777	0,528	0,139
	2	0,135	0,184	0,082	2	0,865	0,632	0,181
	2,5	0,082	0,143	0,078	2,5	0,918	0,713	0,221
	3	0,050	0,112	0,074	3	0,950	0,777	0,259
	3,5	0,030	0,087	0,070	3,5	0,970	0,826	0,295
	4	0,018	0,068	0,067	4	0,982	0,865	0,330
	4,5	0,011	0,053	0,064	4,5	0,989	0,895	0,362
	5	0,007	0,041	0,061	5	0,993	0,918	0,393
	5,5	0,004	0,032	0,058	5,5	0,996	0,936	0,423
	6	0,002	0,025	0,055	6	0,998	0,950	0,451
	6,5	0,002	0,019	0,052	6,5	0,998	0,961	0,478
	7	0,001	0,015	0,050	7	0,999	0,970	0,503
	7,5	0,001	0,012	0,047	7,5	0,999	0,976	0,528
	8	0,000	0,009	0,045	8	1,000	0,982	0,551
	8,5	0,000	0,007	0,043	8,5	1,000	0,986	0,573
	9	0,000	0,006	0,041	9	1,000	0,989	0,593
	9,5	0,000	0,004	0,039	9,5	1,000	0,991	0,613
	10	0,000	0,003	0,037	10	1,000	0,993	0,632
	10,5	0,000	0,003	0,035	10,5	1,000	0,995	0,650
	11	0,000	0,002	0,033	11	1,000	0,996	0,667
	11,5	0,000	0,002	0,032	11,5	1,000	0,997	0,683
	12	0,000	0,001	0,030	12	1,000	0,998	0,699
	12,5	0,000	0,001	0,029	12,5	1,000	0,998	0,713
	13	0,000	0,001	0,027	13	1,000	0,998	0,727
	13,5	0,000	0,001	0,026	13,5	1,000	0,999	0,741
	14	0,000	0,000	0,025	14	1,000	0,999	0,753
	14,5	0,000	0,000	0,023	14,5	1,000	0,999	0,765
	15	0,000	0,000	0,022	15	1,000	0,999	0,777
	15,5	0,000	0,000	0,021	15,5	1,000	1,000	0,788
	16	0,000	0,000	0,020	16	1,000	1,000	0,798
	16,5	0,000	0,000	0,019	16,5	1,000	1,000	0,808
	17	0,000	0,000	0,018	17	1,000	1,000	0,817

Рис. 1.4. Образец оформления рабочего листа «Экспоненциальное распределение»

Биномиальное распределение

Пусть проводится эксперимент, в результате которого нас интересует, произошло событие A или не произошло. Случай, в котором событие A произошло, назовем успехом, вероятность этого события P(A) = p. Если же событие A не произошло, то его вероятность

$$P(\bar{A}) = 1 - p = q.$$

Предположим теперь, что серия независимых испытаний такого типа проводится n раз. Нас интересует вероятность события, состоящего в том, что успех произошел ровно m раз, или вероятность того, что дискретная случайная величина X, равная числу успехов, примет значение m. Решение этой задачи имеет вид:

$$P(X = m) = C_n^m p^m q^{n-m}, \tag{1.9}$$

где

$$C_n^m = \frac{n!}{m!(n-m)!} \tag{1.10}$$

– число сочетаний из n элементов по m. Формула (1.9) и задает биномиальный закон распределения дискретной случайной величины X (в ее правой части– разложение бинома$(p+q)^n$).

Математическое ожидание случайной величины X, имеющей биномиальное распределение, равно

$$m_X = np, \tag{1.11}$$

а дисперсия

$$D_X = npq. \tag{1.12}$$

Для вычисления значений биномиального распределения в Excel используется встроенная статистическая функция БИНОМ.РАСП (рис. 1.5).

Синтаксис:

БИНОМ.РАСП (Число_успехов; Число_испытаний; Вероятность успеха; Интегральная)

Число_успехов - количество успешных испытаний.

Число_испытаний - количество независимых испытаний.

Вероятность успеха - вероятность успеха каждого испытания.

При Интегральной = 0 рассчитывается вероятность отдельного события, а при Интегральной = 1 рассчитывается интегральная вероятность.

Пример:

= БИНОМ.РАСП (A1; 12; 0,5; 0) показывает(если вA1 введены значения от 0 до12), что для 12 бросков монеты вероятность выпадения орла равна числу, указанному вA1.

= БИНОМ.РАСП (A1; 12; 0,5; 1) рассчитывает интегральную вероятность для тех же условий. Например, еслиA1 = 4, интегральная вероятность выпаде-ния орла равна0, 1, 2, 3 или4 разам(не исключающее ИЛИ)

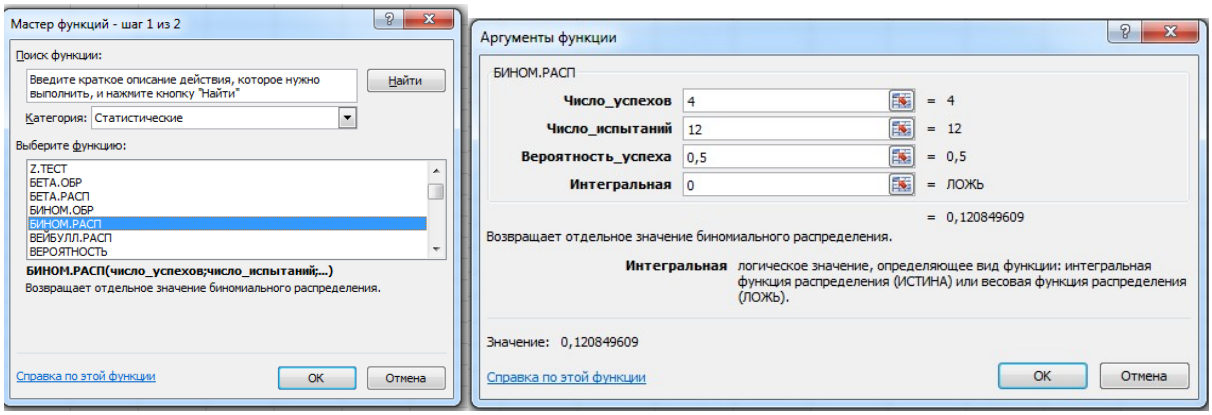


Рис. 1.5. Встроенная функция БИНОМ.РАСП.

ЗАДАНИЕ

1. Введите в таблицу значения аргумента x в диапазоне от 0 до 25 с шагом 1.
2. Вычислите вероятности того, что успех в серии из 25 испытаний про-изойдет ровно x раз(x от 0 до25) при вероятности успеха p= 0,7; p= 0,5; p= 0,2.
3. Используя мастер диаграмм, постройте соответствующие графики распределения.
4. Отредактируйте графики в соответствии с образцом оформления (рис.1.6).

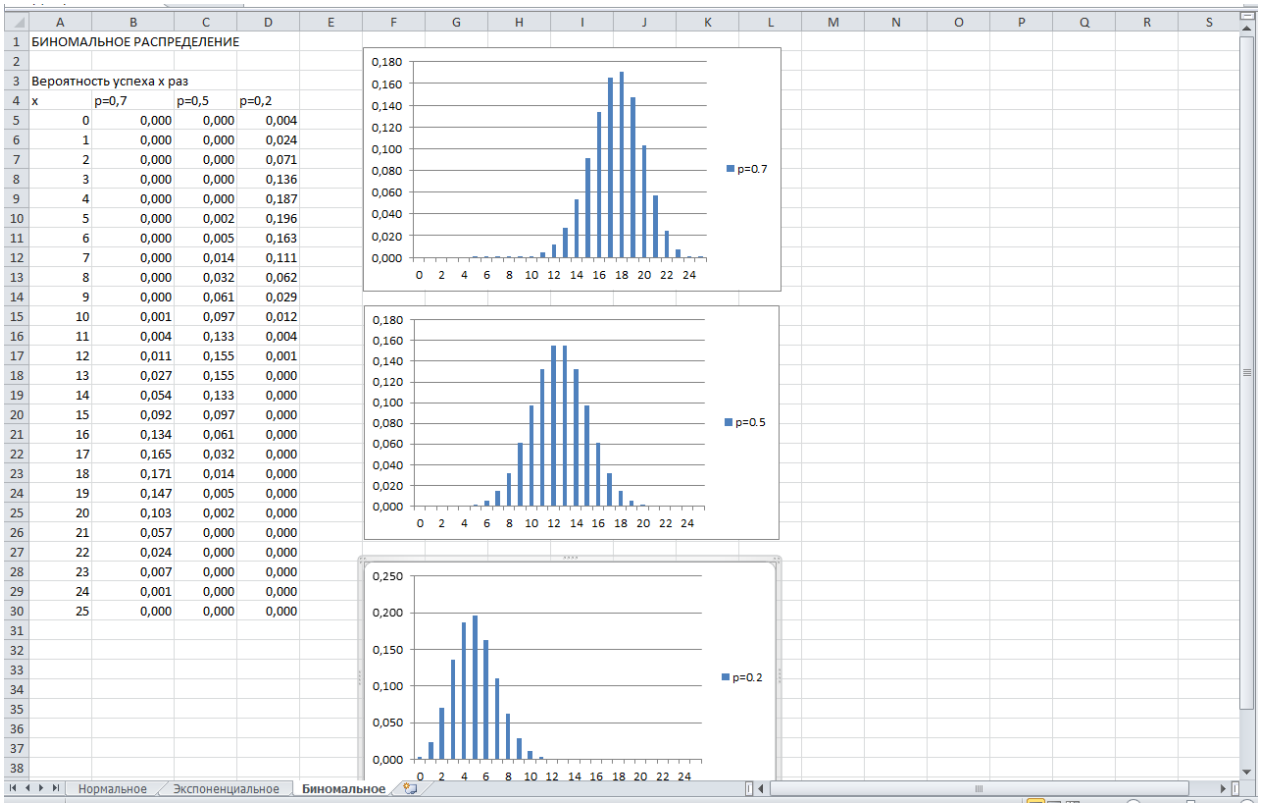


Рис. 1.6. Образец оформления рабочего листа «Биномиальное распределение»

Распределение Пуассона

Пусть в условиях биномиального распределения число испытаний n велико, а вероятность успеха p мала. Если при этом np= λ= const, то можно показать, что (при $n \to \infty$, $p \to 0$)

$$C_n^m p^m q^{n-m} \to \frac{\lambda^m e^{-\lambda}}{m!}.$$

Дискретная случайная величина X имеет распределение Пуассона с параметром λ, если

$$P(X = m) = \frac{\lambda^m e^{-\lambda}}{m!}, \tag{1.13}$$

где параметр λ= np> 0. Учитывая, что вероятность p мала, распределение Пуассона часто интерпретируют как *закон редких явлений*. Математическое ожидание и дисперсия случайной величины X, имеющей распределение Пуассона, одинаковы и равны параметру λ:

$$m_X = D_X = \lambda. \tag{1.14}$$

Для вычисления значений распределения Пуассона в Excel используется встроенная статистическая функция ПУАССОН.РАСП (рис. 1.7).

Синтаксис:

ПУАССОН.РАСП (x; Среднее; Интегральная)

x - значение, на основе которого вычисляется распределение Пуассона.
Среднее - среднее значение распределения Пуассона.
При Интегральной= 0 рассчитывается вероятность отдельного события, а при Интегральной = 1 рассчитывается интегральная вероятность.

Пример:

= ПУАССОН.РАСП (60;50;1) возвращает0,93.

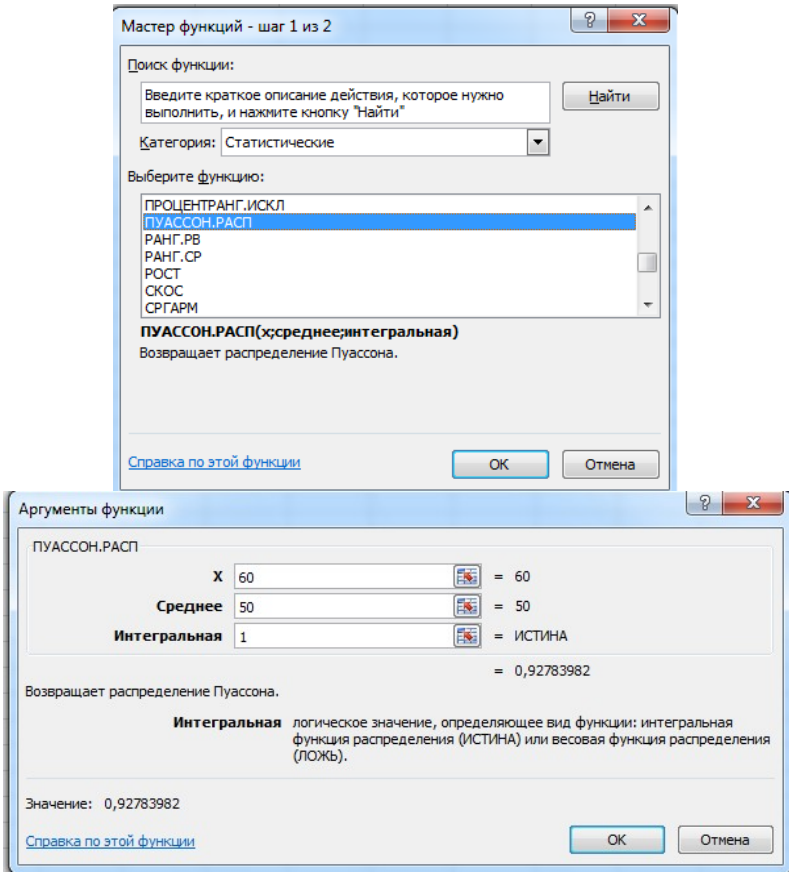


Рис. 1.7. Встроенная функция ПУАССОН.РАСП .

ЗАДАНИЕ

1. Введите в таблицу значения аргумента x в диапазоне от 0 до 40 с шагом 1.
2. Вычислите вероятности того, что успех в серии из 40 испытаний произойдет ровно x раз(x от 0 до 40) при λ= 10; λ= 20; λ= 30.
3. Используя мастер диаграмм, постройте соответствующие графики распределения.
4. Отредактируйте графики в соответствии с образцом оформления (рис.1.8).

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Вычислить значение функции нормального распределения с математическим ожиданием 12 и стандартным отклонением 2 при x= 8.
2. Построить кривую нормального распределения с математическим ожиданием 12 и стандартным отклонением2.
3. Построить кривую экспоненциального распределения с параметром λ= 0,001.

4. Какова вероятность, что при10 подбрасываниях монеты герб выпадет ровно два раза? Воспользоваться встроенной функцией биномиального распределения.

5. Предприятие отпустило поставщику партию из1000 изделий. Вероятность повреждения в пути составляет0,002. Какова вероятность, что поставщик получит пять изделий дефектными? Воспользоваться встроенной функцией распределения Пуассона.

6. Во многих статистических расчетах используется бетараспределение. Ознакомиться по справке с встроенной функцией БЕТА.РАСП.

7. Вычислить значение плотности бетараспределения с параметрами $\alpha = 5$ и $\beta = 3$ при $x = 6$.

8. В условиях предыдущего примера построить кривую бетараспределения.

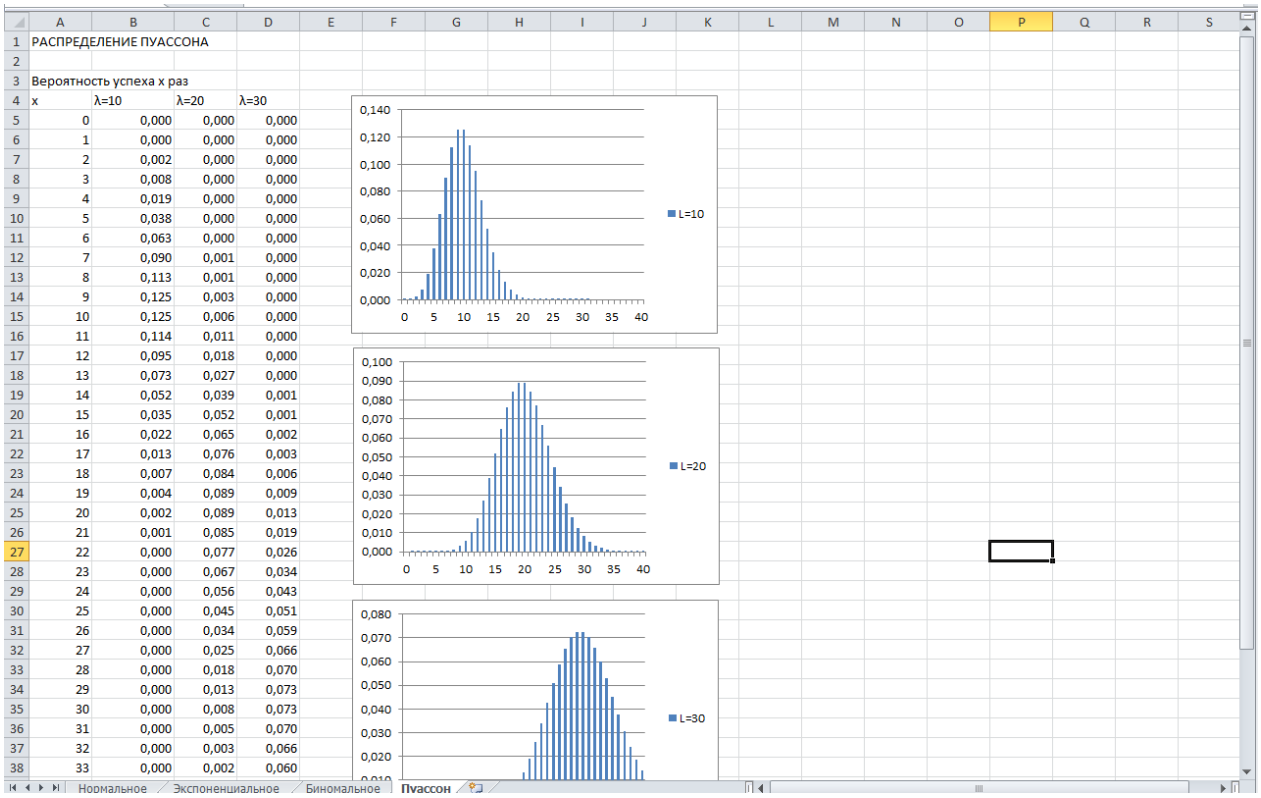


Рис. 1.8. Образец оформления рабочего листа «Распределение Пуассона»

Лабораторная работа №3
Поиск параметров распределений случайных величин

Результаты наблюдений в выборке записываются в порядке их регистрации x1, x2, … , xn; n– объем выборки. Вариационным называется ряд, составленный из элементов выборки в порядке их возрастания: $x^{(1)} \le x^{(2)} \le \ldots \le x^{(n)}$. При этом минимальный элемент выборки $x_{min}= x^{(1)}$, максимальный элемент $x_{max}= x^{(n)}$. Разность между максимальным и минимальным элементами выборки называется размахом:

$$R= x_{max} - x_{min}. \quad (2.1)$$

При достаточно большом объеме выборки данные группируют– разбивают на интервалы, как правило, одинаковой длины. Количество интервалов k выбирается в зависимости от объема выборки, обычно от 8 до 20 интервалов. Иногда используется эмпирическая формула

$$k= 1 + 3{,}32 \lg n. \quad (2.2)$$

Ширина интервала

$$w= R / k. \quad (2.3)$$

Количество n_i элементов выборки, попавших в i-й интервал (i= 1, 2, …, k), называется частотой. Результаты расчета сводят в таблицу частот, в которой показывают границы интервалов, середины z_i каждого интервала, частоты, относительные частоты n_i / n, накопленные относительные частоты $\sum n_i / n$, а также относительные частоты, деленные на длину интервала n_i / w_n. Эти данные используются для графического представления выборки.

Выборочным распределением называется распределение дискретной случайной величины, принимающей значения x_1, x_2, … , x_n с вероятностями 1/n. График выборочной функции распределения F*(x) строится по значениям накопленных относительных частот. Можно показать, что при большом объеме выборки выборочная функция распределения является приближенной оценкой функции распределения F(x) генеральной совокупности.

Гистограмма частот строится по значениям ni/wn и является приближенной оценкой плотности распределения f(x) генеральной совокупности.

Часто для простоты на гистограмме откладывают значения абсолютных частот n_i . При этом меняется только масштаб по оси ординат.

Гистограмма позволяет визуально представить характер распределения изучаемой величины: оценить его симметричность, положение центра, рассеяние, проверить, является ли распределение унимодальным или имеется несколько вершин, сравнить положение центра распределения с требуемым математическим ожиданием(если оно задано), а рассеяние - с границами допуска.

Характерные типы гистограмм показаны на рис.1

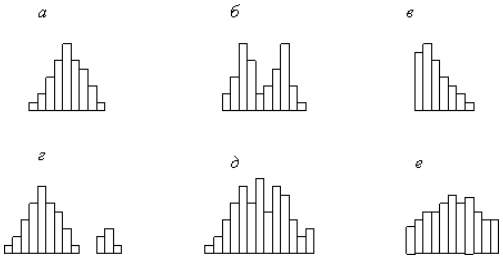


Рис. 1. Характерные типы гистограмм.

На рис.1 а) показан обычный тип гистограммы с двусторонней симметрией, что указывает на стабильность процесса.

На рис1 б) в распределении имеется два пика (двугорбая гистограмма). Такая гистограмма получается при объединении двух распределений, например, в случае двух видов сырья, изменения настройки процесса или объединения в одну партию изделий, обработанных на двух разных станках. Требуется расслоение продукции.

На рис.1 в) показана гистограмма с обрывом. Такое распределение получается, когда невозможно получить значение ниже (или выше) некоторой величины. Подобное распределение имеет место также, когда из партии исключены все изделия с показателем ниже (и/или выше) нормы, т.е. изначально это была партия с большим количеством дефектных изделий. Такое же распределение получается, когда измерительные приборы были неисправны.

На рис.1 г) показана гистограмма с островком. Получается при ошибках в измерениях, или когда некоторое количество дефектных изделий перемешано с доброкачественными.

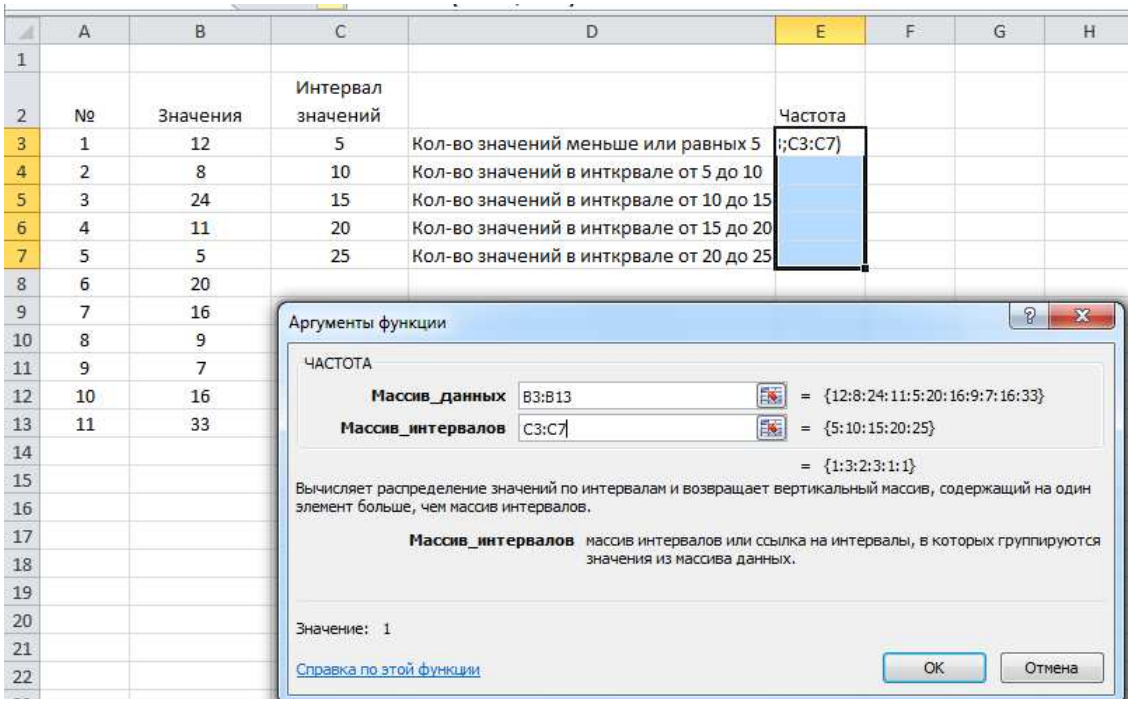
На рис.1 д) показана гистограмма «гребёнка». Получается, когда ширина интервала не кратна единице измерения или при ошибках оператора.

На рис. 1е) показана гистограмма в форме плато. Получается, когда объединяются несколько распределений при небольшой разнице средних значений. В этом случае требуется расслоение.

При анализе характера распределения иногда полезна стратификация данных. Если одно и то же изделие изготавливается разными рабочими, часто имеет смысл проанализировать работу каждого из них отдельно: провести стратификацию, или расслоение, по квалификации рабочих. При использовании материала из разных партий иногда уточнить природу дефекта можно, если анализировать эти партии раздельно.

В производстве для стратификации удобен метод, называемый5М(по первым буквам английских наименований): необходимо провести стратификацию данных по квалификации работников(men), по используемому оборудованию(machine), по материалам(material), по технологии изготовления(method), по методам и средствам измерения(measure).

Для построения гистограммы вExcel необходимо ввести в таблицу результаты наблюдений и подготовить столбец рассчитанных значений границ интервалов. Для подсчета частот используется функция массива ЧАСТОТА из списка Статистических функций (рис. 2.1), которая возвращает частотное распределение в виде массива из одного столбца. Функция служит для подсчета количества значений в интервале значений, например, для подсчета количества результатов тестирования, попадающих в интервалы результатов. Поскольку данная функция возвращает массив, ее необходимо вводить как формулу массива.



	A	B	C	D	E
1					
2	№	Значения	Интервал значений		Частота
3	1	12	5	Кол-во значений меньше или равных 5	1
4	2	8	10	Кол-во значений в инткрвале от 5 до 10	3
5	3	24	15	Кол-во значений в инткрвале от 10 до 15	2
6	4	11	20	Кол-во значений в инткрвале от 15 до 20	3
7	5	5	25	Кол-во значений в инткрвале от 20 до 25	1
8	6	20			
9	7	16			
10	8	9			
11	9	7			
12	10	16			
13	11	33			

Рис. 2.1. Встроенная функция массива ЧАСТОТА

Синтаксис:
ЧАСТОТА(массив_данных, массив_интервалов)

- Массив_данных — обязательный аргумент. Массив или ссылка на множество значений, для которых вычисляются частоты. Если аргумент "массив_данных" не содержит значений, функция ЧАСТОТА возвращает массив нулей.
- Массив_интервалов — обязательный аргумент. Массив или ссылка на множество интервалов, в которые группируются значения аргумента "массив_данных". Если аргумент "массив_интервалов" не содержит значений, функция ЧАСТОТА возвращает количество элементов в аргументе "массив_данных".

Пример:

В приведенной на рис. 2.1 таблице в столбце В содержится список неотсортированных измерений. В столбце С содержится верхний предел для интервалов, на которые требуется разделить данные из столбца В. В соответствии с пределом в ячейке С3 функция ЧАСТОТА возвращает количество значений, которые меньше либо равны 5. Предел в ячейкеС4 равен10, поэтому функция ЧАСТОТА возвращает количество значений, которые больше5 или больше либо равны10.

ЗАДАНИЕ

1. Введите в один столбец результаты измерений, выполненных на двух станках А и Б:

№	1	2	3	4	5	6	7	8	9	10	11
Значение	9,94	9,74	10,05	10,12	10,1	10,1	9,56	9,95	10,22	9,78	9,86
Станок	А	А	А	А	А	А	А	А	А	А	А
№	12	13	14	15	16	17	18	19	20	21	22
Значение	9,66	9,63	9,8	9,85	9,58	9,89	9,92	10,03	9,93	9,93	9,93
Станок	А	А	А	А	А	А	А	А	А	А	А
№	23	24	25	26	27	28	29	30	31	32	33
Значение	10,21	9,98	9,96	9,9	10,22	10,17	10,48	9,87	11,04	10,89	11,72
Станок	А	А	А	А	А	А	А	А	Б	Б	Б
№	34	35	36	37	38	39	40	41	42	43	44
Значение	10,76	11,35	11	11	10,63	10,88	10,92	11,11	10,76	11,01	10,84
Станок	Б	Б	Б	Б	Б	Б	Б	Б	Б	Б	Б
№	45	46	47	48	49	50	51	52	53	54	55
Значение	10,84	10,93	11,13	10,5	11,34	10,51	10,42	10,67	10,98	10,65	11,07
Станок	А	А	А	А	А	Б	Б	Б	Б	Б	Б

2. Найдите максимальное и минимальное значения, используя встроенные статистические функции МАКС и МИН.

3. Вычислите размах выборки.

4. Определите количество интервалов и найдите их ширину.

5. Подготовьте массив классов: в качестве первого значения введите найденное минимальное значение, последующие значения– с шагом, равным ширине интервала.

6. Вычислите частоты, используя функцию массива ЧАСТОТА. Для этого выделите ячейки, в которых требуется рассчитать частоты. Нажмите знак равно или знак *f(x)* и выберите функцию ЧАСТОТА. Введите необходимые диапазоны и нажмите сочетание

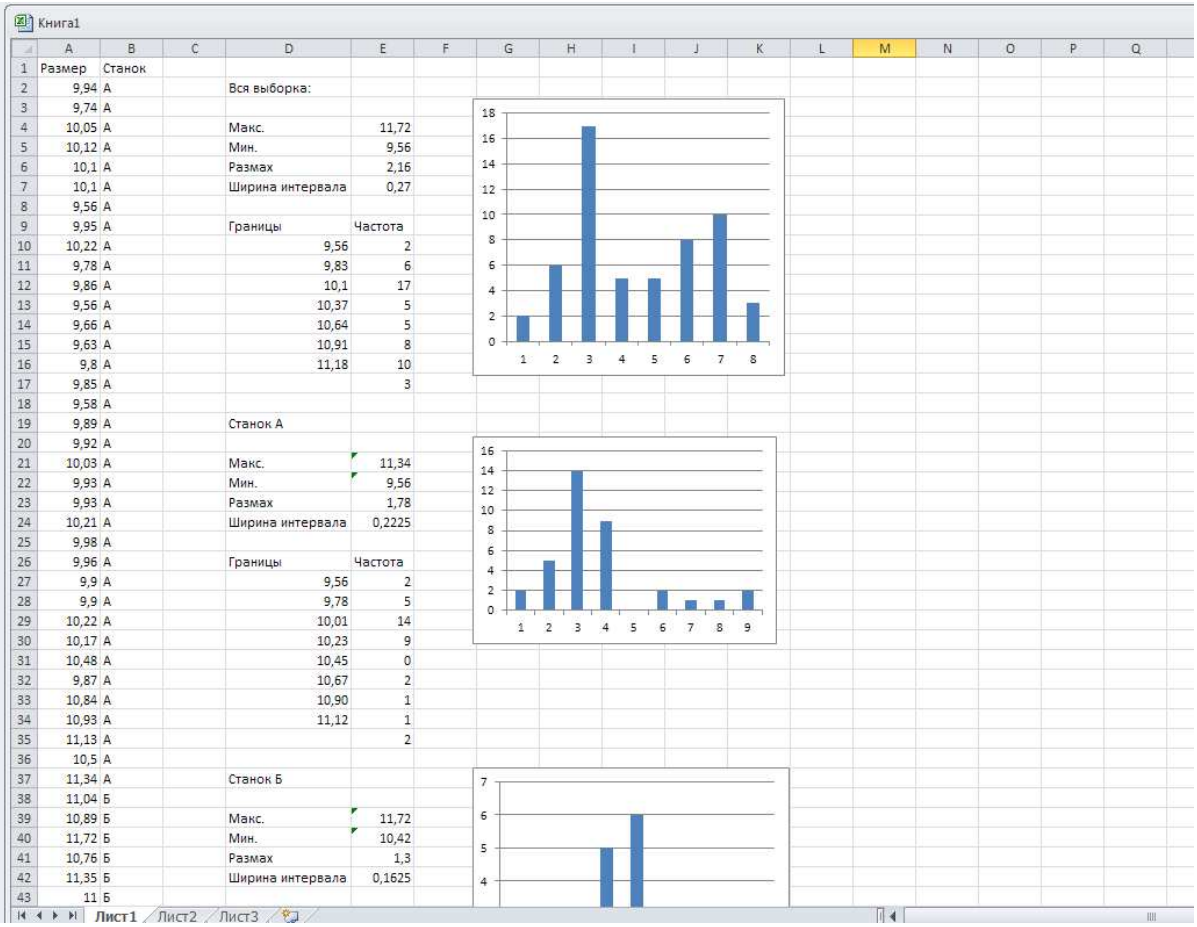
клавиш CTRL+SHIFT+ENTER. Если нажать просто ОК, т.е. формула не будет введена как формула массива, то отобразится только один результат в самой первой ячейке.
7. С помощью мастера диаграмм постройте гистограмму частот.
8. Отредактируйте графики в соответствии с образцом оформления (рис. 2.2).
9. Стратифицируйте гистограмму по станкам: проведите расчеты по пунктам 2 – 8 отдельно для станка А и для станка Б.
10. Сформулируйте выводы по результатам рассмотрения трех построенных гистограмм.

2.1 Числовые характеристики выборки

Выборочное среднее (математическое ожидание выборки)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad (2.4)$$

выборочная мода Mo*– элемент выборки, встречающийся с наибольшей частотой (для унимодального – одновершинного распределения); выборочная медиана Me* – число, которое делит вариационный ряд на две части, содержащие одинаковое количество элементов; если объем выборки нечетен n = 2t + 1, то Me* = $x^{(t+1)}$; при n = 2t Me* = $0{,}5(x^{(t)}+x^{(t+1)})$;



Размер	Станок			
9,94	А	Вся выборка:		
9,74	А			
10,05	А	Макс.	11,72	
10,12	А	Мин.	9,56	
10,1	А	Размах	2,16	
10,1	А	Ширина интервала	0,27	
9,56	А			
9,95	А	Границы	Частота	
10,22	А	9,56	2	
9,78	А	9,83	6	
9,86	А	10,1	17	
9,56	А	10,37	5	
9,66	А	10,64	5	
9,63	А	10,91	8	
9,8	А	11,18	10	
9,85	А		3	
9,58	А			
9,89	А	Станок А		
9,92	А			
10,03	А	Макс.	11,34	
9,93	А	Мин.	9,56	
9,93	А	Размах	1,78	
10,21	А	Ширина интервала	0,2225	
9,98	А			
9,96	А	Границы	Частота	
9,9	А	9,56	2	
9,9	А	9,78	5	
10,22	А	10,01	14	
10,17	А	10,23	9	
10,48	А	10,45	0	
9,87	А	10,67	2	
10,84	А	10,90	1	
10,93	А	11,12	1	
11,13	А		2	
10,5	А			
11,34	А	Станок Б		
11,04	Б			
10,89	Б	Макс.	11,72	
11,72	Б	Мин.	10,42	
10,76	Б	Размах	1,3	
11,35	Б	Ширина интервала	0,1625	
11	Б			

Рис. 2.2. Образец оформления рабочего листа «Гистограмма»

выборочная дисперсия

$$D_X^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2; \qquad (2.5)$$

несмещенная дисперсия

$$s^2 = \frac{n}{n-1} D_X^* = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2; \quad (2.6)$$

выборочное стандартное отклонение

$$\sigma_X^* = \sqrt{D_X^*}; \quad (2.7)$$

или

$$s = \sqrt{s^2} \quad (2.8)$$

выборочный коэффициент асимметрии

$$a_X^* = \frac{\mu_3}{\mu_2^{3/2}} \quad (2.9)$$

(здесь $\mu_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k$ – выборочный центральный момент k-го порядка);

выборочный коэффициент эксцесса

$$e_X^* = \frac{\mu_4}{\mu_2^2} - 3. \quad (2.10)$$

Для вычисления значений этих характеристик в Excel используются встроенные статистические функции:

СРЗНАЧ – среднее значение по формуле (2.4),

МОДА.ОДН – мода,

МЕДИАНА – медиана,

ДИСП.Г - дисперсия (2.5)

ДИСП.В– дисперсия (2.6),

СТАНДОТКЛОН.Г – стандартное отклонение (2.7),

СТАНДОТКЛОН.В – стандартное отклонение (2.8),

ЭКСЦЕСС – коэффициент эксцесса (2.10).

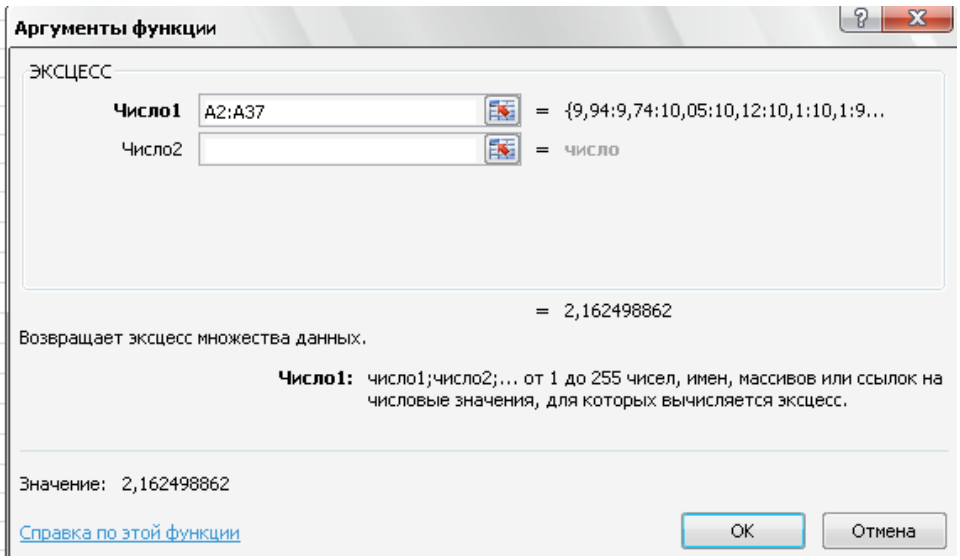


Рис. 2.3. Встроенная статистическая функция KURT

Синтаксис этих функций практически одинаков. Например, для функции ЭКСЦЕСС (рис. 2.3):

ЭКСЦЕСС(число1, [число2],...)

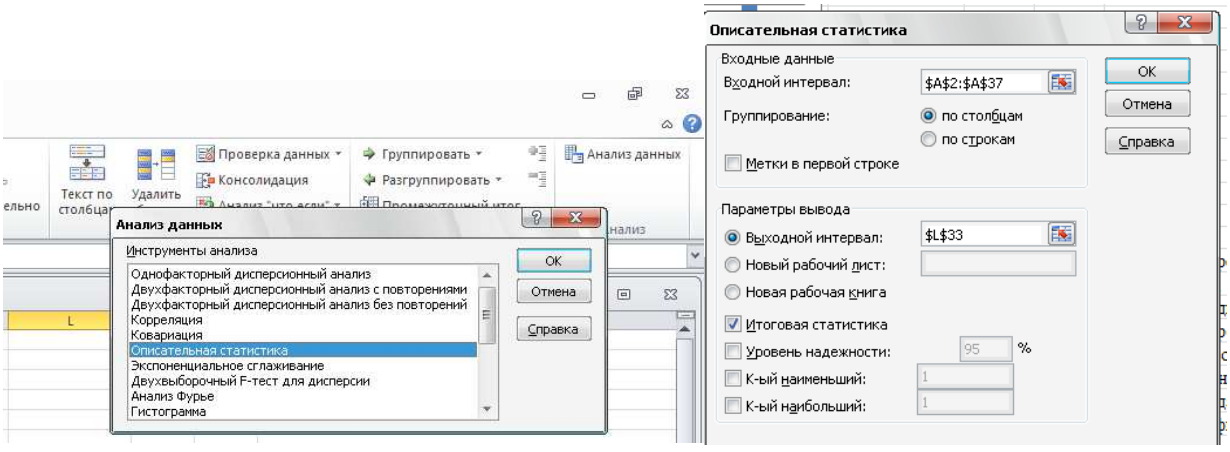
Число1, число2,... Аргумент "число1" является обязательным, последующие числа необязательные. От 1 до 255 аргументов, для которых вычисляется эксцесс. Вместо аргументов, разделенных запятой, можно использовать один массив или ссылку на массив.

Также все указанные характеристики можно вычислить с помощью пакета анализа данных (Описательная статистика). Пакет анализа представляет собой надстройку Microsoft Excel, т. е. программу, которая становится доступной при установке Microsoft Office или Excel. Однако чтобы использовать надстройку в Excel, необходимо сначала загрузить ее. Если вы работаете в Microsoft Office 2010:

1. Откройте вкладку Файл и выберите пункт Параметры.
2. Выберите команду Надстройки, а затем в поле Управление выберите пункт Надстройки Excel.
3. Нажмите кнопку Перейти.
4. В окне Доступные надстройки установите флажок Пакет анализа, а затем нажмите кнопку ОК.
5. После загрузки пакета анализа в группе Анализ на вкладке Данные становится доступной команда Анализ данных.

Если вы работаете в Microsoft Office 2003:

1. В меню Сервис выберите команду Надстройки.
2. В окне Список надстроек установите флажок рядом с элементом Пакет анализа, а затем нажмите кнопку ОК.
3. Если в списке отсутствует элемент Пакет анализа, нажмите кнопку Обзор, чтобы найти надстройку самостоятельно.
4. В случае появления сообщения о том, что пакет анализа не установлен на компьютере, нажмите кнопку Да, чтобы установить его.
5. Нажмите кнопку Сервис в строке меню. После загрузки пакета анализа в меню Сервис добавляется пункт Анализ данных.



Для того чтобы выполнить вычисления, введите в поле «Водной интервал» адреса ячеек, в которых записаны выборочные значения; пометьте «Выходной интервал» и введите в поле адрес первой ячейки, начиная с которой в листе Excel будет отображён результат; поставить галочку в окне «Итоговая статистика»

Лабораторная работа №4,5,6
Расчет статистических оценок прогнозов Построение и оценка парной регрессии Построение и оценка множественной регрессии

Основные понятия

Статистической гипотезой называют гипотезу о виде неизвестного распределения генеральной совокупности или о параметрах известных распределений.
Нулевой (основной) называют выдвинутую гипотезу H_0. **Конкурирующей (альтернативной)** называют гипотезу H_1, которая противоречит нулевой.
Пример. Пусть H_0 заключается в том, что математическое ожидание генеральной совокупности $a = 3$. Тогда возможные варианты H_1: а) $a \neq 3$; б) $a > 3$; в) $a < 3$.

Проверка гипотез

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное или приближенное распределение которой известно. Эту величину называют статистическим критерием K. Для проверки гипотезы по данным выборок вычисляют точные значения входящих в критерий величин и таким образом получают наблюдаемое значение критерия $K_{набл}$.

После выбора определенного критерия множество всех его возможных значений разбивают два непересекающихся подмножества: W, содержащее значения критерия, при которых H_0 принимается — область принятия гипотезы, и $\overline{W}$ — при которых нулевая гипотеза H_0 отвергается, — критическая область. При использовании любого критерия возможны ошибки следующих видов:

1) ошибка первого рода — принять гипотезу H_1 , когда верна H_0;
2) ошибка второго рода — принять гипотезу H_0 , когда верна H_1 .

Для каждой из них могут быть заданы соответствующие вероятности:

$\alpha = P\{K \in \overline{W} | H_0\}$, $\alpha' = P\{K \in W | H_1\}$.

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка второго рода)
	H_1	H_0 неверно отвергнута (Ошибка первого рода)	H_0 верно отвергнута

Вероятность ошибки первого рода называется критерием (или уровнем) значимости. Нулевую гипотезу отвергают, если для нее значение p ниже уровня значимости, т. е., если $p < \alpha$. Обычно назначают условное значение 0,05, тогда шанс допустить ошибку 1-го рода никогда не превысит выбранного уровня значимости, скажем $\alpha = 0{,}05$, так как нулевую гипотезу отвергают только тогда, когда $p < 0{,}05$. Если обнаружено, что $p > 0{,}05$,

то нулевую гипотезу не отвергнут и, следовательно, не допустят ошибки 1-го рода. Величина (1-β) называется *мощностью критерия*.

Односторонние и двусторонние критические области.

Точки, отделяющие критическую область от области принятия гипотезы называют критическими точками $k_{кр}$. Различают одностороннюю (правостороннюю и левостороннюю) и двустороннюю критические области:

1. Правосторонняя критическая область определяется неравенством $K > k_{кр}$, где $k_{кр} > 0$. Например, когда проверяется гипотеза о том, что среднее значение равно 10.
2. Левосторонняя критическая область $K < k_{кр}$, где $k_{кр} < 0$.
3. Двусторонняя критическая область $K < k_1$, $K > k_2$, где $k_2 > k_1$. В частности, если критические точки симметричны относительно нуля, то двусторонняя критическая область определяется как $|K| > k_{кр}$.

Для нахождения критической точки $k_{кр}$ правосторонней критической области при условии, что справедлива гипотеза H_0:

1) задаются достаточно малой вероятностью — уровнем значимости α (связанной с доверительной вероятностью β соотношением $\alpha = 1 - \beta$);
2) ищут $k_{кр}$, исходя из требования: $P(K > k_{кр} = \alpha)$;
3) по таблицам соответствующих критериев находят $k_{кр}$ и сравнивают с наблюдаемым значением критерия $K_{набл}$;
4) если $K_{набл} > k_{кр}$, то H_0 отвергают; если $K_{набл} < k_{кр}$, нет оснований отвергать H_0.

Отыскание левосторонней и двусторонней критических областей сводится к нахождению соответствующих критических точек:

$P(K < k_{кр}) = \alpha$ — требование для левосторонней критической области;

$P(K < k1) + P(K > k2) = \alpha$ — требование для двусторонней критической области.

Если выбрать симметричные относительно нуля точки: $-k_{кр}$ и $k_{кр}$ ($k_{кр} > 0$), то получим следующее соотношение для отыскания критических точек двусторонней области:

$$P(K > k_{кр}) = \frac{\alpha}{2}.$$

Для пересчета одностороннего значения уровня значимости в двустороннее значение можно воспользоваться простой формулой: $\alpha_{двустор} = 2 \cdot \alpha_{одностор}$.

Существенным моментом является то, что статистические методы не позволяют подтвердить гипотезу. Корректное утверждение, которое можно сформулировать в результате анализа, звучит, например: «проведенный анализ не позволяет с заданным уровнем значимости отвергнуть гипотезу».

Проверить гипотезу о равенстве среднего значения заданному можно двумя способами.

Первый способ:

Пусть генеральная совокупность ***X*** имеет нормальное распределение, и требуется проверить предположение о том, что ее математическое ожидание равно некоторому числу m_0. Рассмотрим две возможности.

1) Известна дисперсия σ^2 генеральной совокупности. Тогда по выборке объема ***n*** найдем выборочное среднее $\bar{x}$ и проверим нулевую гипотезу $\boldsymbol{H_0}$: $\boldsymbol{M(X)} = \boldsymbol{m_0}$. Критерием может служить величина

$$t = \frac{\bar{x} - m_0}{o / \sqrt{n}}$$

где t – квантиль нормального распределения. Если же дисперсия неизвестна, то используется статистика

$$t = \frac{\bar{x} - m_0}{s / \sqrt{n}}$$

имеющая распределение Стъюдента с (n–1) степенью свободы. t – квантиль распределения Стьюдента, s – среднее квадратическое отклонение выборки.

Найденное значение t сравнивается с критическим $t_{кр}$, которое определяется по таблице квантилей в зависимости от уровня значимости или вычисляется встроенными функциями Excel. При попадании выборочного значения в критическую область гипотеза отвергается.

Второй способ:

Для проверки гипотезы о равенстве среднего заданному значению может быть использована функция Z.ТЕСТ, которая вычисляет двухстороннюю вероятность значений z-теста при стандартном распределении.

Синтаксис:

Z.ТЕСТ (Данные; Число; Сигма)

Данные: массив данных.

Число: значение для теста.

Сигма: необязательное стандартное отклонение генеральной совокупности. Если аргумент не указан, используется стандартное отклонение выборки.

Пример:

=Z.ТЕСТ (A1:A50; 12) вычисляет вероятность того, что значение12 относится к стандартному распределению генеральной совокупности данных в A1:A50.

При использовании этой функции вычисляется вероятность того, что для генеральной совокупности справедлива гипотеза H0: m= m0. Используется двухсторонний критерий, то есть альтернативная гипотеза H1: m≠m0. Если эта вероятность меньше заданного уровня значимости, гипотеза отклоняется.

ЗАДАНИЕ №1

Шарики, изготовленные станком-автоматом, должны иметь диаметр 10 мм; проверить эту гипотезу по заданной выборке на уровне значимости 0.05, если:

А) Дисперсия известна и равна0,1 мм^2

В) Дисперсия неизвестна.

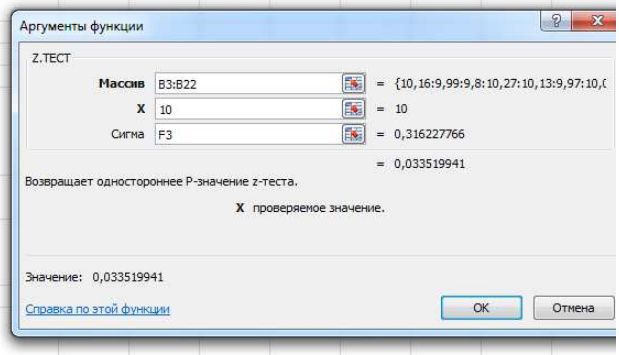
Результаты наблюдений приведены в таблице.

10,16	9,99	9,8	10,27	10,13	9,97	10,04	10,16	10,19	10,26
9,96	9,89	10,11	10,3	10,15	10,04	10,39	10,43	10,03	10,32

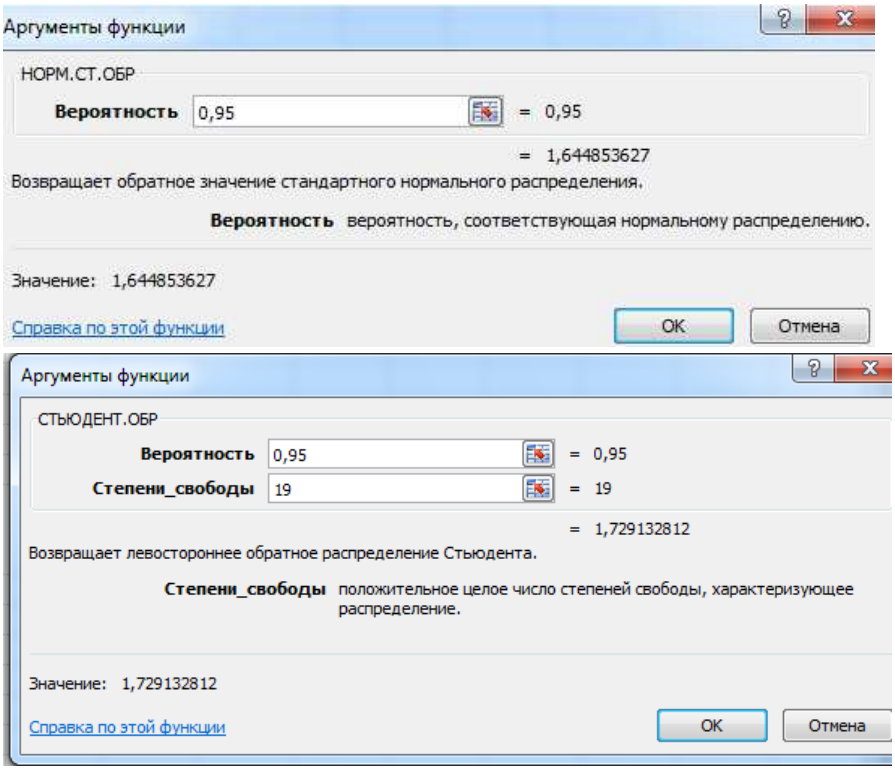
Воспользоваться стандартным алгоритмом проверки гипотез и встроенной функцией Z.ТЕСТ.

Пример выполнения задания:

	A	B	C	D	E	F
1					Проверяется гипотеза Н0 о том, что среднее выборки равно 10	
2	Результаты наблюдений				Дисперсия D	0,1
3	1	10,16			σ=√D	0,31623
4	2	9,99			среднее выборки	10,1295
5	3	9,8			среднее ген.совокупности	10
6	4	10,27				
7	5	10,13			***Стандартный метод***	
8	6	9,97				
9	7	10,04			Уровень значисости α (вероятность принять гипотезу Н1 , когда верна Н0)	0,05
10	8	10,16			Вероятность принятия гипотезы Н0 1-α	0,95
11	9	10,19			Квантиль нормального распределения для 0,95 tкр	1,64485
12	10	10,26				
13	11	9,96			Вычисляем действительную ошибку Δ=|Xср_ген-Xср_выб|	0,1295
14	12	9,89			С другой стороны Δ=μt, где μ=σ/√n, n - объем выборки, t - квантиль распределения	
15	13	10,11			Выразим значение t для данной ошибки tнаб=Δ√n/σ	1,83141
16	14	10,3				
17	15	10,15			***tнабл>tкр***	***гипотеза Н0 отвергается***
18	16	10,04				
19	17	10,39				
20	18	10,43			***Решение задачи с помошью функции Z.ТЕСТ***	
21	19	10,03				
22	20	10,32			Z тест	;10;F3)
23						
24					***Вероятность принятия гипотезы Н0 (Шарики, изготовленные станком-автоматом, должны иметь диаметр10 мм) 0,033 меньше 0,05. Гипотеза отклоняется***	



Для определения квантилей распределения воспользуйтесь функциями НОРМ.СТ.ОБР и СТЬЮДЕНТ.ОБР



	A	B	C	D	E	F	G	H	I	J	K
1					Проверяется гипотеза H0 о том, что среднее выборки равно 10						
2	Результаты наблюдений				среднее выборки	10,1295					
3	1	10,16			среднее ген.совокупности m0	10					
4	2	9,99			дисперсия выборки Dв	0,02801					
5	3	9,8			s=√Dв	0,16735					
6	4	10,27									
7	5	10,13			***Стандартный метод***						
8	6	9,97			вычислим квантиль распределения Стьюдента по формуле						
9	7	10,04			$u = \frac{\bar{x} - m_0}{s/\sqrt{n}}$						
10	8	10,16									
11	9	10,19									
12	10	10,26			tнаб	3,46072					
13	11	9,96			Найдем tкр квантиль распределения Стьюдента для вероятности 0,95 и n-1=19	1,72913					
14	12	9,89									
15	13	10,11			***tнабл>tкр***	***гипотеза H0 отвергается***					
16	14	10,3									
17	15	10,15									
18	16	10,04									
19	17	10,39									
20	18	10,43			***Решение задачи с аомошью функции Z.ТЕСТ***						
21	19	10,03									
22	20	10,32			Z тест	0,00027					
23											
24					***Вероятность принятия гипотезы H0 (Шарики, изготовленные станком-автоматом, должны иметь диаметр10 мм) меньше 0,05. Гипотеза отклоняется***						
25											

Проверка гипотез о равенстве дисперсий

При проверке гипотеза о равенстве дисперсий двух нормально распределенных совокупностей H0: $\sigma_1^2 = \sigma_2^2$ при неизвестных математических ожиданиях m1 и m2 используется статистика

$$F = \frac{s_1^2}{s_2^2} \qquad (1)$$

которая имеет F-распределение Фишера с числом степеней свободы (n1– 1) и (n2– 1);

здесь n1 и n2– объемы выборок, s_1^2 и s_2^2 – соответствующие несмещенные дисперсии;

при этом предполагается, что $s_1^2 > s_2^2$.

Для проверки этой гипотезы в Excel есть функция F.ТЕСТ, которая возвращает результат F-теста. Синтаксис функции :

F.ТЕСТ (Данные1; Данные2)

Данные1: первый массив записей.

Данные2: второй массив записей.

Пример:

=F.ТЕСТ (A1:A30; B1:B12) вычисляет различие дисперсий для двух множеств данных и возвращает вероятность того, что оба множества представляют собой выборку из общей совокупности.

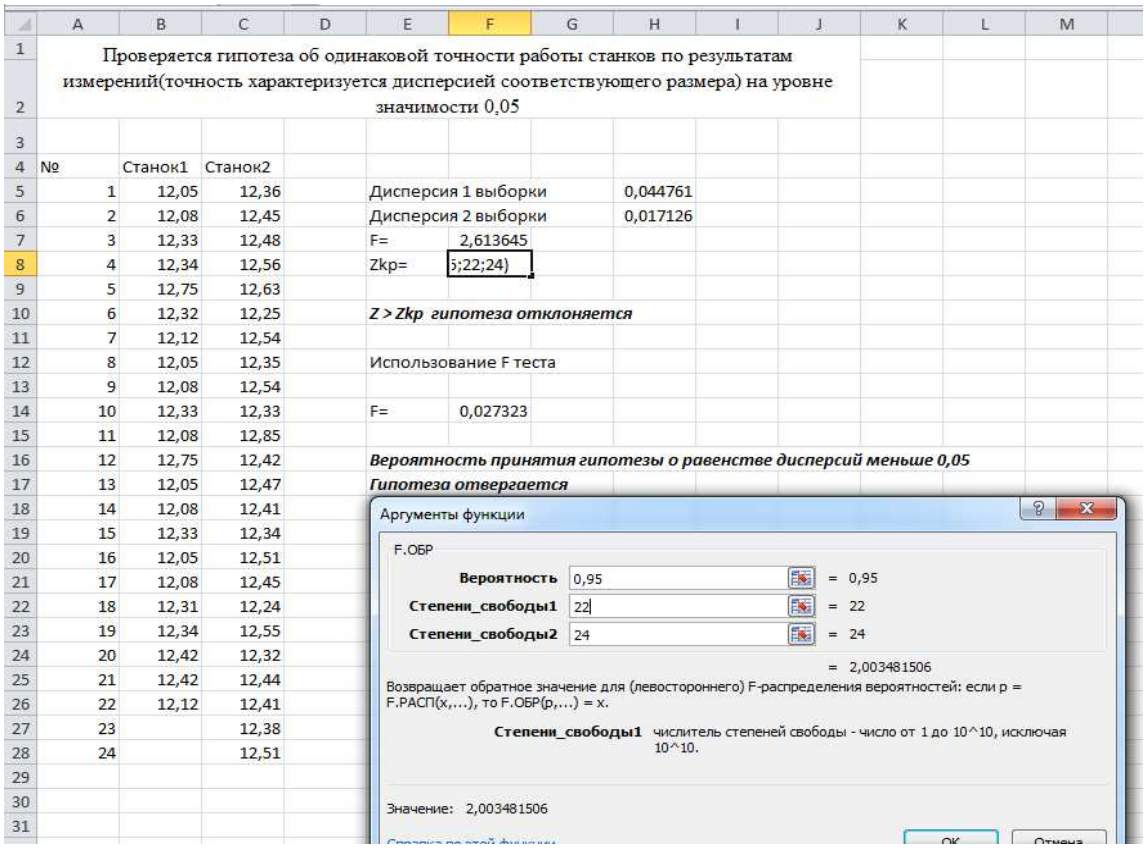
ЗАДАНИЕ№2

Проверить гипотезу об одинаковой точности работы станков по результатам измерений(точность характеризуется дисперсией соответствующего размера) на уровне

значимости 0,05 с использованием формулы (1) и функции F.ТЕСТ. Результаты измерений контролируемого параметра на двух станках приведены в таблице.

№	Станок1	Станок2	№	Станок1	Станок2
1	12,05	12,36	13	12,05	12,47
2	12,08	12,45	14	12,08	12,41
3	12,33	12,48	15	12,33	12,34
4	12,34	12,56	16	12,05	12,51
5	12,75	12,63	17	12,08	12,45
6	12,32	12,25	18	12,31	12,24
7	12,12	12,54	19	12,34	12,55
8	12,05	12,35	20	12,42	12,32
9	12,08	12,54	21	12,42	12,44
10	12,33	12,33	22	12,12	12,41
11	12,08	12,85	23		12,38
12	12,75	12,42	24		12,51

Пример выполнения задания



Проверка гипотез о виде распределения

Другой группой статистических гипотез являются гипотезы о проверке вида распределения: неизвестен вид распределения генеральной совокупности, и в частности, неизвестна функция распределения F(x).

Пусть $x_1,x_2,\ldots,x_n$ – выборка наблюдений случайной величины X. Проверяется гипотеза H_0 о том, что случайная величина X имеет функцию распределения F(x). Разобьем область

возможных значений X на r интервалов Δ_1, Δ_2,… Δ_r. Пусть n_i – число элементов выборки, принадлежащих интервалу $\Delta_i(i=1,\ldots,r)$; при малых значениях n_i интервалы объединяют таким образом, чтобы в каждом из них было $n_i \geq 5$. Используя предполагаемый закон распределения – с функцией F(x), с учетом оценок параметров этого закона, найденных по выборке, находят вероятности того, что значения X принадлежат интервалу Δ_i, то есть

$$p_i = P\{X \in \Delta_i\}, i = \overline{1,r}.$$

Статистика

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

имеет распределение χ^2 с числом степеней свободы (r – l – 1), где r– число интервалов, l – число неизвестных параметров распределения. Например, для нормального распределения l = 2 (неизвестные параметры m и σ). Считается, что гипотеза H_0 согласуется с опытом, если

$$\chi^2 < \chi^2_{1-\alpha}(r-l-1)$$

, где χ^2 – выборочное значение статистики, $\chi^2_{1-\alpha}(r-l-1)$ – квантиль порядка (1 – α) распределения χ^2 с числом степеней свободы(r – l – 1). Рассмотренный метод проверки гипотезы вида распределения назы-вается критерием хи-квадрат или критерием согласия Пирсона.

ЗАДАНИЕ №3

Дана выборка из100 наблюдений; определить числовые характеристики, построить гистограмму частот, проверить нормальность распределения по критерию хи-квадрат.

12,01	12	11,64	12,09	11,79
11,64	11,99	11,7	11,79	12,16
11,59	11,45	11,9	11,86	12,25
11,4	11,84	12,15	11,79	11,92
12,37	12	11,93	11,98	11,72
11,76	11,97	12,25	11,76	11,9
11,7	11,58	12,1	12,39	11,74
12,04	11,58	11,81	12,13	12,09
12,02	12,16	11,94	12,2	11,66
12,01	11,35	12,2	11,84	11,84
12,07	12,15	12,1	11,52	11,84
11,58	11,78	11,79	11,78	11,83
12,07	11,42	12,08	12,03	12,03
11,79	11,8	12,7	11,65	11,96
12,19	11,85	12,42	11,72	12,4
12,34	12,15	11,65	12,27	11,81
11,91	12,03	12,16	12,11	11,92

11,81	11,74	12,54	11,98	11,84
11,9	11,73	11,9	11,67	12,4
11,81	11,74	12,14	12,25	11,93

Для выполнения задания необходимо рассчитать и внести в таблицу следующие данные:

n_i – частота попаданий элементов выборки в i-й интервал;

x_i – верхняя граница i-го интервала;

$F(x_i)$ – значение функции нормального распределения;

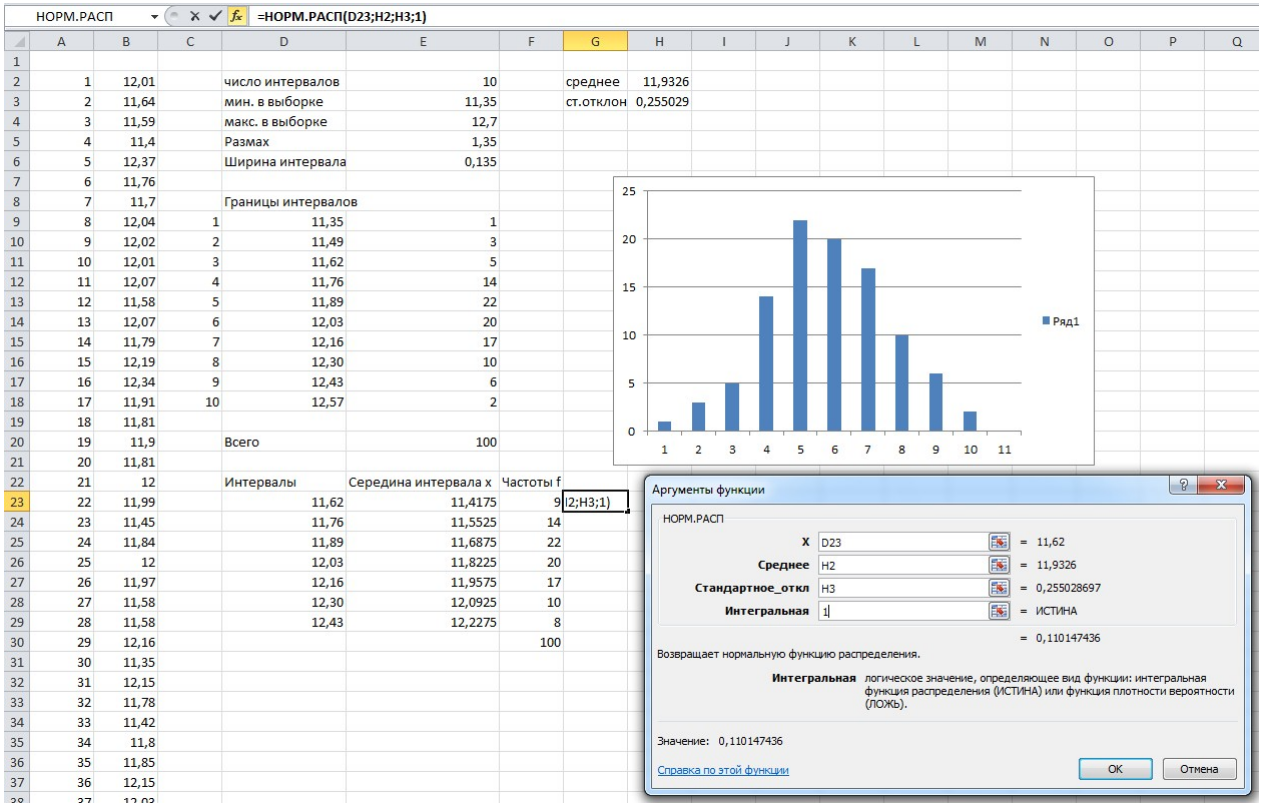
ΔF_i – теоретическое значение вероятности попадания случайной величины в i-й интервал

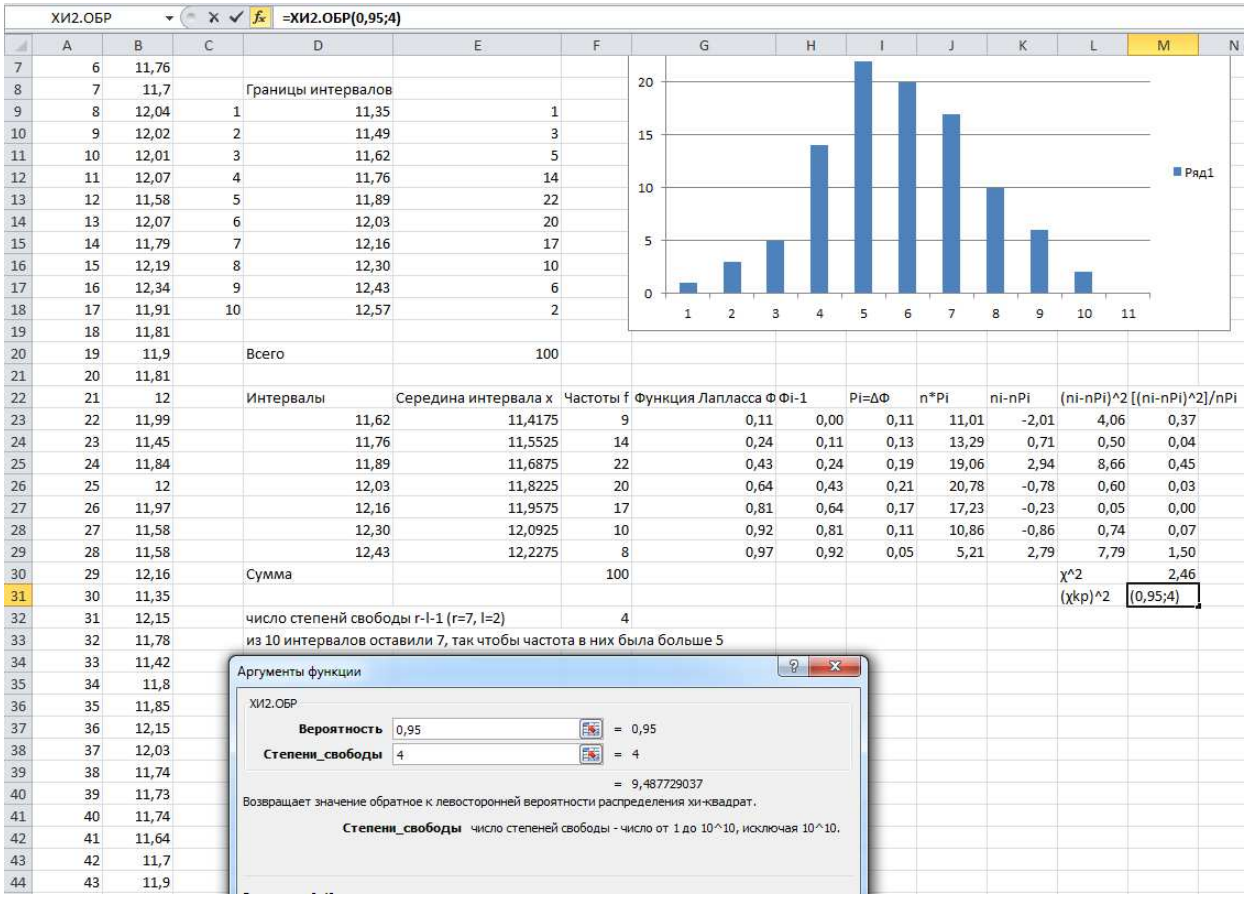
$$\Delta F_i = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_{i-1}}^{x_i} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right) = F_x(x_i) - F_x(x_{i-1}) =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{x_i} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right) - \frac{1}{\sigma\sqrt{2\pi}} \int_{\infty}^{x_{i-1}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right);$$

$F_i = \Delta F_i * n$ – теоретическая частота попадания случайной величины в i-й интервал;

$(n_i - F_i)^2/F_i$ – взвешенный квадрат отклонения.





Лабораторная работа №7
Подготовка эталонов распознавания печатных знаков

При выборочном приёмочном контроле по результатам контроля выборок принимается решение принять или отклонить партию продукции. При этом в случае контроля по альтернативному признаку единицы продукции делятся на годные и дефектные, а партия, поступающая на контроль, имеет входной уровень дефектности q. Входной уровень дефектности - это доля дефектных единиц продукции, которая заранее неизвестна, и её надо оценить по результатам контроля. Обычно при выборочном контроле партии разделяют на хорошие и плохие с помощью двух чисел – AQL (приёмочный уровень дефектности) и LQ (браковочный уровень дефектности). Партии считаются хорошими при q < AQL и плохими при q > LQ. При AQL < q < LQ качество партии считается ещё допустимым. Приёмочный уровень дефектности AQL – это предельно допустимое значение уровня дефектности в партии, изготовленной при нормальном ходе производства. Браковочный уровень качества LQ – это граница для отнесения продукции к браку.

При выборочном контроле по альтернативному признаку план контроля включает значения объёма выборки n и приёмочного числа c. Партия принимается, если число дефектных единиц продукции в выборке m < c.

Оперативной характеристикой плана контроля называется функция P(q), равная вероятности принять партию с долей дефектных единиц продукции q.

$$P(q) = \sum_{m=0}^{c} P_n(m)$$

где $P_n(m)$ – вероятность появления m дефектных единиц продукции в выборке объёмом n. Чаще всего оперативная характеристика отображается в виде графика.

P(q) = 1 - α при q = AQL
P(q) = β при q = LQ

Здесь α - риск поставщика, равный вероятности забраковать партию с q = AQL, β - риск потребителя, равный вероятности принять партию с q = LQ.

Пример 1. Для контроля качества партий из N = 20 изделий используют одноступенчатый выборочный план с параметрами n = 5 и с =1. Построить оперативную характеристику плана контроля.

Поскольку приёмочное число равно 1, то партия будет принята при числе дефектных изделий в выборке 0 или 1. Вероятность приёмки равна сумме вероятностей появления в выборке 0 или 1 дефектных изделий:

$$P(q) = \sum_{m=0}^{c} P_n(m) = P_5(0) + P_5(1)$$

Вероятности $P_5(0)$ и $P_5(1)$ можно найти, исходя из гипергеометрического распределения вероятностей

$$P(m) = \frac{C_D^m * C_{N-D}^{n-m}}{C_N^n}$$

$$C_N^n = \frac{N!}{n!(N-n)!}$$

Величина P(m) может быть рассчитана в программе Excel при помощи статистической функции ГИПЕРГЕОМЕТ. Диалоговое окно, открывающееся при выборе этой функции, имеет четыре строки для ввода данных:

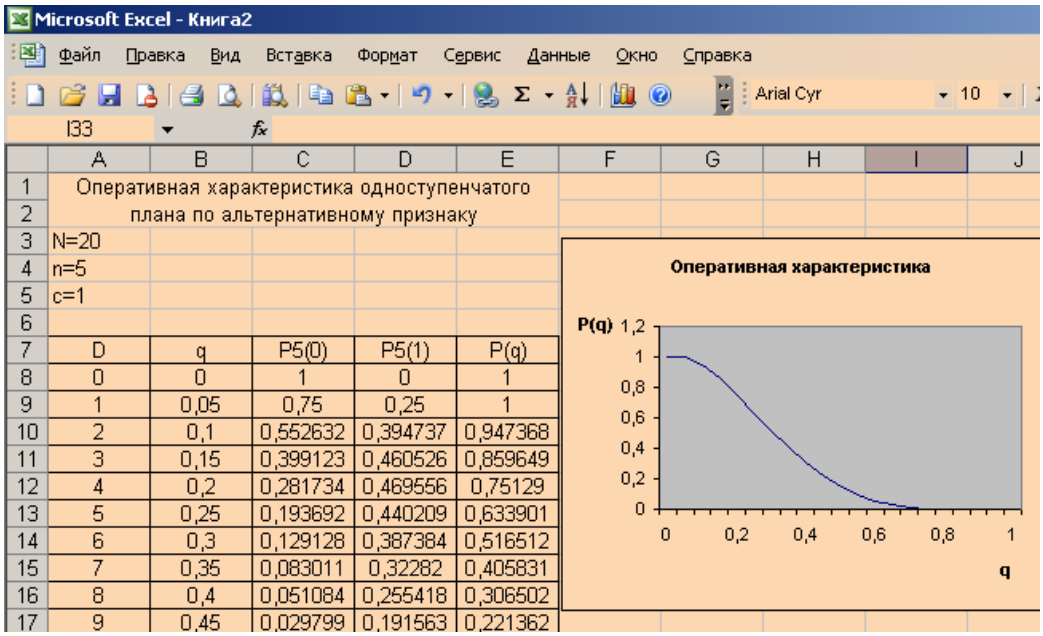
Число успехов в выборке: необходимо ввести количество успешных испытаний в выборке. При этом под количеством успешных испытаний понимается количество элементов выборки, обладающих определённым признаком, в нашем случае – количество дефектных изделий в выборке.
Размер выборки: Вводится размер выборки.
Число успехов в совокупности: надо ввести количество успешных испытаний в генеральной совокупности. В нашем случае это количество дефектных изделий в партии.
Размер ген совокупности: Вводится объём партии.
Таким образом, для построения оперативной характеристики потребуются столбцы с заголовками: D (количество дефектных изделий в партии), q, P5(0), P5(1), P(q). Эти заголовки вводим в ячейки A7:E7. В ячейки B3:B5 вводим исходные данные - значения объёма партии, объёма выборки и приёмочного числа.
В ячейки A8:A28 вводим возможные значения количества дефектных изделий в партии от 0 до 20. В ячейке B8 рассчитываем q при D = 0 по формуле =A8/B3, затем копируем эту формулу в диапазон B9:B28, предварительно указав в формуле абсолютную адресацию для объёма партии.
В ячейке C8 рассчитываем значение P5(0) для D = 0 по статистической формуле ГИПЕРГЕОМЕТ, и после указания абсолютной адресации в тех ячейках, где это необходимо, копируем формулу в диапазон C9:C28. При этом в диапазоне C24:C28

результатом расчёта является ошибка. Это связано с тем, что при D > 15 вероятность P5(0) = 0, но при расчёте вместо нуля получается очень маленькое число, которое слишком мало, чтобы его можно было представить в Excel. В эти ячейки следует с клавиатуры ввести значения 0.

Исходя из аналогичных соображений, в ячейке D8 рассчитываем значение P5(1) для D = 0 по статистической формуле ГИПЕРГЕОМЕТ (получится ошибка, поскольку для D = 0 P5(1) = 0), и после указания абсолютной адресации в тех ячейках, где это необходимо, копируем формулу из D8 в диапазон D9:D28. При этом в диапазоне D25:D28 результатом расчёта является ошибка. В ячейки D8 и D25:D28 с клавиатуры вводим 0.

Далее в ячейке E8 рассчитываем значение P(q) как сумму вероятностей P5(0) и P5(1). Формулу из ячейки E8 копируем в диапазон E9:E28.

По полученным данным строим оперативную характеристику. Результаты расчётов и построений показаны на рис. 1.



	A	B	C	D	E
1	Оперативная характеристика одноступенчатого				
2	плана по альтернативному признаку				
3	N=20				
4	n=5				
5	c=1				
6					
7	D	q	P5(0)	P5(1)	P(q)
8	0	0	1	0	1
9	1	0,05	0,75	0,25	1
10	2	0,1	0,552632	0,394737	0,947368
11	3	0,15	0,399123	0,460526	0,859649
12	4	0,2	0,281734	0,469556	0,75129
13	5	0,25	0,193692	0,440209	0,633901
14	6	0,3	0,129128	0,387384	0,516512
15	7	0,35	0,083011	0,32282	0,405831
16	8	0,4	0,051084	0,255418	0,306502
17	9	0,45	0,029799	0,191563	0,221362

Рис 1. Результаты расчёта и построения оперативной характеристики в примере 1.

Пример 2: Для контроля качества партий из 1000 изделий, с входным уровнем дефектности не более 0,08, используют одноступенчатый выборочный план с параметрами n = 50 и c =2. Построить оперативную характеристику плана контроля.

Если q < 0,1 и n < 0,1N, что обычно и имеет место в практике статистического контроля, то биномиальное распределение, как и гипергеометрическое, можно приближённо заменить ещё более простым для расчётов распределением Пуассона, в котором

$$P(m) = \frac{\lambda^m e^{\lambda}}{m!}$$

, где λ= nq – математическое ожидание числа дефектных изделий в выборке.

При распределении Пуассона величина P(m) может быть рассчитана в программе Excel при помощи статистической функции ПУАССОН. Диалоговое окно, открывающееся при выборе функции, имеет три строки для ввода данных:

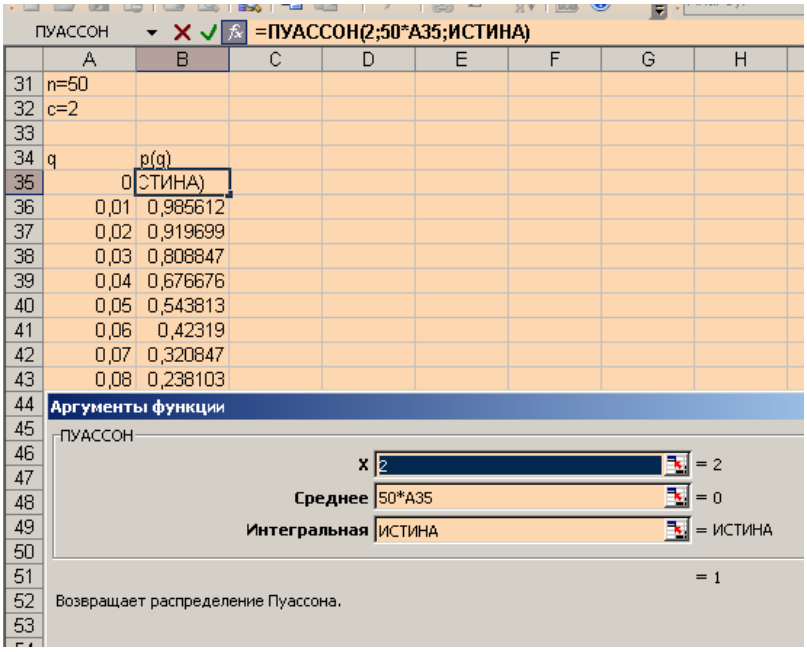
X: Количество событий, в нашем случае - количество дефектных изделий в выборке.

Среднее: Среднее ожидаемое численное значение, в нашем случае – параметр λ, т.е. математическое ожидание числа дефектных изделий в выборке.

Интегральная: Вводится истина, если рассчитывается значение интегральной функции распределения, и ложь, если рассчитывается значение дифференциальной функции распределения, т.е. в нашем случае – значение P(m).

Поскольку в статистической функции ПУАССОН возможно рассчитывать значения не только дифференциальной, но и интегральной функции распределения, то оперативная характеристика P(q) может быть рассчитана непосредственно. Для этого в третьей строке диалогового окна функции ПУАССОН следует вводить значение истина. При этом значение функции будет сразу же рассчитываться как P(q), т.е. как сумма вероятностей $P_n(m)$ при изменении m от 0 до приёмочного числа, значение которого вводится в первой строке диалогового окна. Поэтому понадобится всего два столбца расчётных значений: q и P(q). Соответствующие заголовки вводим в ячейки A6 и B6.

В диапазон A7:A15 вводим значения q от 0 до 0,08 с шагом 0,01. В ячейке B7 рассчитываем значение интегральной статистической функции ПУАССОН.



Затем, после установки в формуле ячейки B7 необходимой абсолютной адресации, копируем эту формулу в диапазон B8:B15. По полученным столбцам значений q и P(q) строим оперативную характеристику. Результаты расчётов и построений показаны на рис

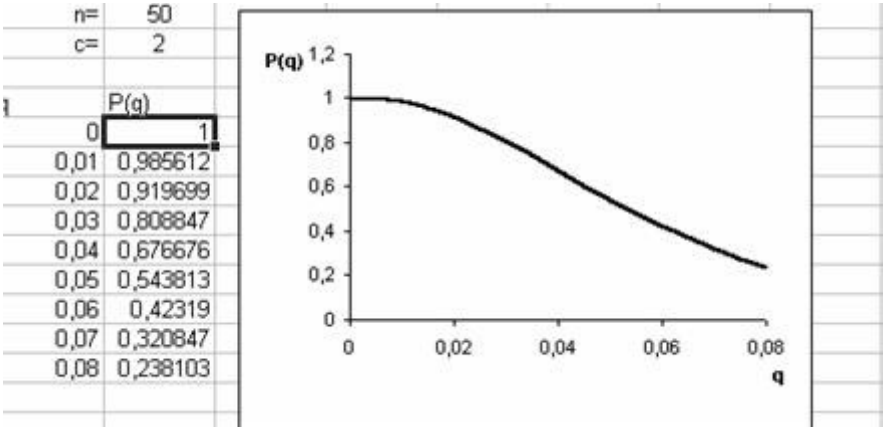


Рис 2. Результаты расчёта и построения оперативной характеристики

Задание

1. Выполнить расчёты и построения в соответствии с примером 1. Чему равны риски поставщика и потребителя при приёмочном уровне дефектности 0,1 и браковочном уровне дефектности 0,4?
2. Выполнить расчёты и построения в соответствии с примером 2.
3. Построить на одной диаграмме три оперативные характеристики планов одноступенчатого выборочного контроля с параметрами, указанными в табл. 1, учитывая, что n < 0,1N и q не превышает 0,4. Как изменяется вероятность приёмки партии при заданном входном уровне дефектности с увеличением объёма выборки? Как изменяется вероятность приёмки партии при заданном входном уровне дефектности с увеличением приёмочного числа?

Таблица 1.

План	Вариант 1		Вариант 2		Вариант 3		Вариант 4		Вариант 5	
	n	c	n	c	n	c	n	c	n	c
1	20	1	20	2	25	1	25	2	30	2
2	20	2	20	2	25	2	25	2	30	3
3	30	2	30	3	35	2	35	3	20	3
План	Вариант 6		Вариант 7		Вариант 8		Вариант 9		Вариант 10	
	n	c	n	c	n	c	n	c	n	c
1	15	2	15	1	30	1	15	2	25	1
2	15	1	15	2	30	2	15	3	25	2
3	20	1	20	2	20	2	25	3	30	2

Лабораторная работа №8
Визуализация данных с помощью диаграмм

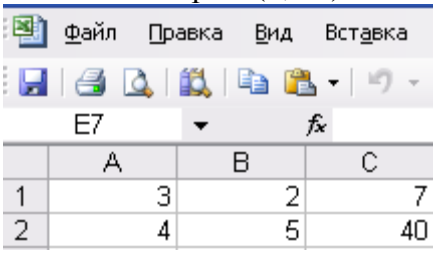
Решение n линейных уравнений с n неизвестными.

Если дана система линейных уравнений матричного вида $A * X = B$, то ее решение состоит в нахождении матрицы, обратной матрице A, которую необходимо умножить слева на вектор – столбец B.

Пример 1. Решить систему линейных уравнений

$$\begin{cases} 3x + 2y = 7 \\ 4x - 5y = 40 \end{cases}$$

1) Введите матрицу A в диапазон A1: B2. Вектор B=(7; 40) в диапазон C1:C2.



2) Найдем обратную матрицу (функция МОБР).
3) В результате обратная матрица выглядит следующим образом:

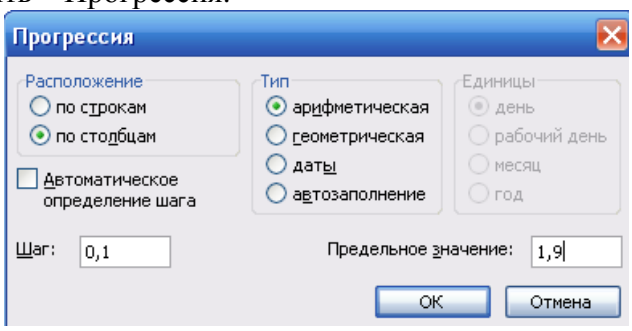
3	0,714286	-0,28571
4	-0,57143	0,428571

4) Умножением обратной матрицы на вектор B найдем вектор X (функция МУМНОЖ). Результат поместите в диапазон C3:C4.

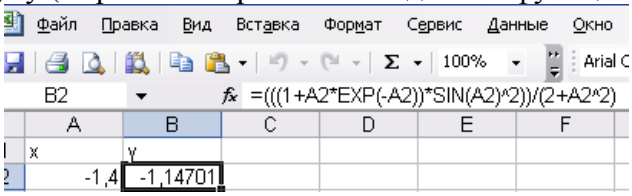
Построение графиков функций

Пример 1. Постройте график функции $y = \frac{1 + xe^{-x}}{2 + x^2} \sin^2 x, x \in [-1.4; 1.9]$

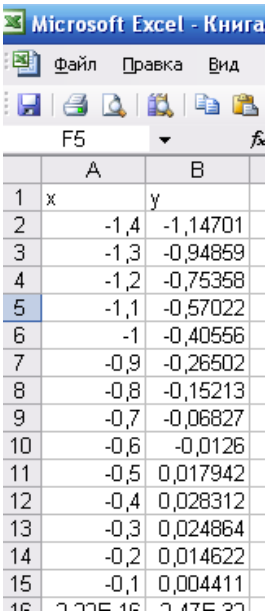
Для построения графика функции необходимо сначала построить таблицу значений функции при различных значениях аргумента, причем аргумент, согласно условию, изменяется с фиксированным шагом, например 0,1. Выбор этого шага обусловлен необходимостью более наглядного отображения значений функции на интервале табуляции. Создадим таблицу, представленную на рисунке. В ячейку A1 вводим начальное значение аргумента. Затем выполняем команду Правка – Заполнить – Прогрессия.



В результате диапазон ячеек заполнится значениями аргумента.
В ячейку B2 вводим формулу (выражение правой части данной функции).

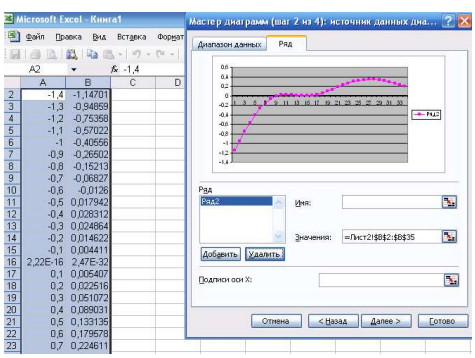


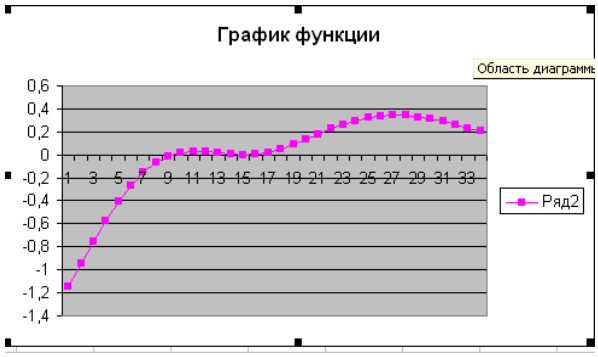
С помощью маркера автозаполнения найдем значения функции на остальном диапазоне.



Для построения графика функции выделим диапазон ячеек, содержащий таблицу значений функции и ее аргументы, и вызовем мастер диаграмм командой Вставка – Диаграмма. В появившемся окне необходимо выбрать закладку Стандартные и Тип – График. Затем нажать Далее и перейти на закладку ряд, в которой удалить в соответствующем поле Ряд 1. введем название диаграммы График Функции, нажмем Далее. В появившемся окне в группе Имеющемся

установить Лист 1, что предполагает размещение диаграммы на листе, на котором выполнялись все расчеты. Нажав кнопку готово, получим график функции.





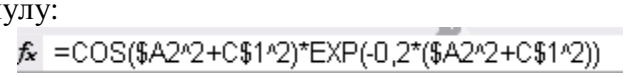
Построение поверхностей

Технологию построения поверхностей рассмотрим на примере.

Пример 1. Построить поверхность функции Z(x, y) с шагом 0,2 при -3<x<3 и -3<y<3

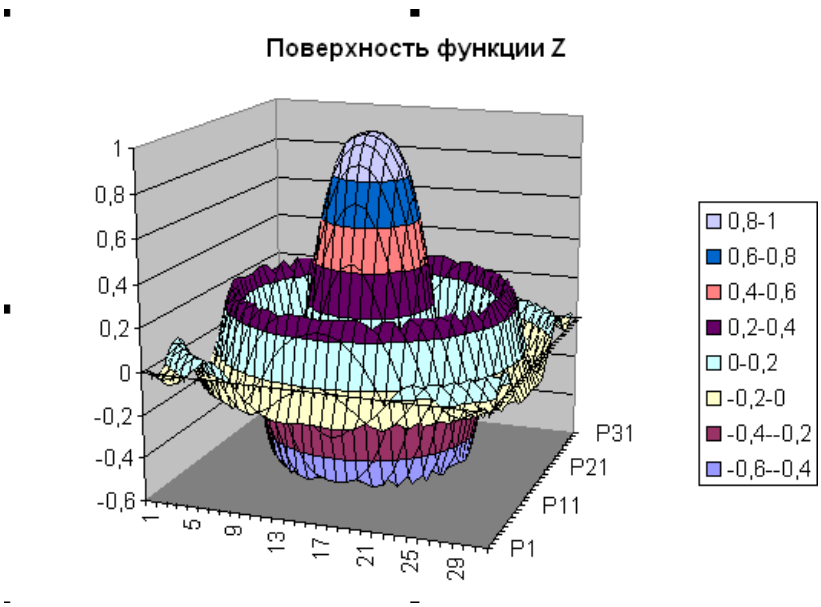
$$Z(x, y) = \cos(x^2 + y^2) \cdot e^{-0.2(x^2+y^2)}$$

Для этого в ячейку A2 введите значение -3 и заполните столбец A вниз значениями арифметической прогрессии с шагом 0,2 до значения 3. В ячейку B1 введите значение -3 и заполните строку 1 вправо значениями арифметической прогрессии с шагом 0,2 до значения 3.
В ячейку B2 введите формулу:



Растяните формулу этой ячейки вниз до значения x=3, т.е. до ячейки B32, затем размножьте эту формулу на весь массив C2: F32, т.е до значения y=3

Далее строим поверхность для заданного массива: надо выделить массив B2: AF32, вызвать Мастер диаграмм и выбрать в нем Поверхности, нажав Далее. Присвоить поверхности название, перейти на вкладку Оси и снять флажок ось y, нажать Далее и в появившемся окне выбрать построение поверхности на этом же листе, затем нажать готово. Полученную поверхность отформатировать в соответствии с рисунком



Задания для самостоятельной работы

1) Решите систему линейных уравнений

$$\begin{cases} 2x + y + 4z = 7 \\ 2x - y - 3z = -5 \\ 3x + 4y - 5z = -14 \end{cases}$$

2) Решите систему линейных уравнений

$$\begin{cases} 3x + 2y = 7 \\ 4x - 5y = 40 \end{cases}$$

3) Постройте графики функции.

а) $y = e^x + \sqrt[3]{x}$ Отрезок построения и шаг задайте самостоятельно.

б) $y = 2\sin 2\pi \cos 4\pi x$

в) $z = \cos^2 3\pi x - \cos \pi x \sin \pi x$

4) Постройте поверхность функции $y = \sin x - \sin y$ при -3<x<3, -3<y<3 с шагом 0,1

5) Постройте поверхность функции $N = 5x^2 \cos^2 y - 2e^y y^2$ при -1<x<1, -1<y<1 с шагом 0,1

Лабораторная работа № 9
Интервальные оценки

Цель работы: получение практических навыков по определению основных выборочных характеристик количественного признака генеральной совокупности.

Краткие теоретические сведения

Статистические оценки параметров распределения. Обычно в распоряжении исследователя имеются лишь выборочные данные. Если из теоретических соображений удалось установить, какое именно распределение имеет признак генеральной совокупности, то возникает задача оценки параметров, которыми определяется это распределение. Для описания случайных величин используются описательные статистики: минимум, максимум, среднее, дисперсия, стандартное отклонение, медиана, мода и т.д. Статистики дают общее представление о значениях, которые принимают случайные величины. Получаемые оценки могут носить точечный и интервальный характер.

Оценка называется *точечной*, если определяется одним числом; *интервальной* – если по данным выборки строится числовой интервал, внутри которого на основании заранее выбранной вероятности находится оцениваемый параметр.

Оценка должна быть близка к оцениваемому параметру. Близость характеризуется несмещенностью оценки, ее состоятельностью и эффективностью.

Несмещенность оценки означает отсутствие систематических погрешностей в наблюдаемых данных, для этого ее математическое ожидание должно быть равно оцениваемому параметру.

Состоятельность оценки заключается в том, что с ростом числа наблюдений дисперсия стремится к нулю.

Для исследуемого параметра оценка эффективна, если имеет минимальную дисперсию среди всех возможных оценок, построенных по данной выборке.

Пусть из генеральной совокупности извлечена выборка объема *n*. *Выборочное среднее* (m^*) – сумма значений переменной, делённая на *n* (число значений переменной)

$$m^* = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Выборочное среднее может быть посчитано по частотно-вариационному ряду

$$m^* = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n},$$

где k – количество вариантов в ряду, или по интервальному ряду

$$m^* = \frac{\sum_{i=1}^{k} x'_i \cdot n_i}{n},$$

где $x'i$ - середина i-го интервала, k - количество интервалов.

Среднее выборочное является несмещенной, состоятельной и эффективной оценкой математического ожидания генеральной совокупности, т.е. точечная оценка математического ожидания является доброкачественной

$$\tilde{x} = m^*.$$

Выборочная дисперсия (D)*- мера изменчивости случайной величины. Вычисляется по формуле:

$$D^* = \frac{\sum_{i=1}^{n} (x_i - m^*)^2}{n}.$$

Значение 0 означает отсутствие изменчивости, т.е. переменная постоянна. Выборочная дисперсия является смещенной оценкой дисперсии генеральной совокупности, поэтому доброкачественной оценкой генеральной дисперсии является исправленная выборочная дисперсия

$$\tilde{D} = D^* \frac{n}{n-1} = \frac{\sum_{i=1}^{n} (x_i - m^*)^2}{n-1}.$$

Выборочное стандартное отклонение (S) - корень квадратный из дисперсии. Более удобная характеристика, так как измерена в тех же единицах, что и исходная величина. Чем выше дисперсия и стандартное отклонение, тем сильнее разбросаны значения случайной величины относительно среднего. Для оценки среднего квадратичного отклонения генеральной совокупности применяют выборочное среднее квадратичное отклонение

$$S = \sqrt{D^*}$$

или исправленное среднее квадратичное отклонение

$$\tilde{S} = \sqrt{\tilde{D}} = \sqrt{\frac{n}{n-1} D^*}.$$

Для более подробного описания свойств распределения вводятся эмпирические начальные

$$\lambda^p = \frac{\sum_{i=1}^{n} x_i^p}{n}$$

и центральные

$$\mu^p = \frac{\sum_{i=1}^{n} (x_i - x^*)^p}{n}$$

моменты p-го порядка или их комбинаций. В частности, *коэффициент асимметрии* позволяет судить о симметричности выборочных данных

$$A = \left[\frac{n \cdot \mu^3}{(n-1) \cdot (n-2) \cdot S^3} \right]$$

Если коэффициент значительно отличается от 0, распределение является асимметричным. Показатель эксцесса служит мерой крутизны (заостренности) гистограммы по отношению к кривой нормального распределения (для нормально распределенной случайной величины $E=0$).

$$E = \frac{\left| n \cdot (n+1) \cdot \mu^4 - 3 \cdot \mu^2 \cdot \mu^2 \cdot (n-1) \right|}{\left| (n-1) \cdot (n-2) \cdot (n-3) \cdot S^4 \right|}.$$

Медиана – значение, которое разбивает выборку на две равные части. Половина наблюдений лежит выше медианы, и половина – ниже. В некоторых случаях, например, при описании доходов населения медиана более удобна, чем среднее.

Медиана дает общее представление о том, где сосредоточены значения переменной, иными словами, где находится ее центр. Сумма *абсолютных* расстояний между точками выборки и медианой *минимальна*. Медиана вычисляется следующим образом. Выборка упорядочивается в порядке возрастания. Если количество элементов в выборке определяется как *2m+1* (нечетно), то медиана выборки оценивается как $Me = x_{m+1}$. Если число наблюдений четно, то медиана оценивается как $Me = (x_m + x_{m+1})/2$.

Квантиль – число t_p, ниже которого находится *p*-я часть (доля) выборки.

Процентиль – значение квантили в процентах.

Мода – наиболее часто встречающееся выборочное значение, варианта, имеющая наибольшую частоту.

Доверительным интервалом для параметра θ называется интервал $(\theta^* - \delta,\ \theta + \delta)$, который с заданной надежностью β покрывает реальное значение параметра θ, здесь θ^* – оценка параметра, δ – точность оценки. Число $\beta = 1 - \alpha$ называется доверительной вероятностью, а значение α - уровнем значимости. В качестве β, как правило, выбираются значения, близкие к единице: 0,95; 0,99; 0,999.

Точечная оценка m^* даже, если она несмещенная, состоятельная, эффективная дает приближенное значение параметра генеральной совокупности и, особенно для выборок малого объема, отличается от истинного значения параметра, т.е. от m.

Представление о том, к каким ошибкам может привести замена параметра m на его точечную оценку m^* и с какой степенью уверенности можно ожидать, что эти ошибки не выйдут за известные пределы дает *мера достоверности* (или *интервальная оценка*).

В качестве меры достоверности принимают:

1) *доверительную вероятность* β (точный метод), с которой истинное значение параметра a будет находится в заданном относительно стат. оценки интервале;

2) *доверительный интервал* $I_\beta(m^* - \varepsilon; m^* + \varepsilon)$ (грубый метод) относительно статистической оценки, в который с заданной вероятностью β попадет истинное значение параметра m.

Понятие оценки меры достоверности. Назначим некоторую достаточно большую вероятность (β = 0,9; 0,95; 0,997) такую, что событие с этой вероятностью β можно считать практически достоверным.

Требуется найти доверительный интервал: $P(a_n^{(1)} < a < a_n^{(2)}) = \beta$,

где границы интервала $a_n^{(1)}; a_n^{(2)}$ – *доверительные границы*.

Интервальная оценка параметра a (доверительный интервал) - числовой интервал $I_\beta[a] = (a^{*(1)}_n; a^{*(2)}_n)$ относительно статистической оценки параметра, который с заданной вероятностью β накрывает реальное значение параметра a.

Чаще всего доверительный интервал выбирают *симметричным* относительно статистического параметра (см. рис. 2.1).

$I_\beta[a] = (a^{*(1)}_n; a^{*(2)}_n)$, $a^{*(1)}_n = a^*_n - \varepsilon;\ a^{*(2)}_n = a^*_n + \varepsilon$.

$P(|a^*_n - a| < \varepsilon) = \beta;\ P(a^*_n - \varepsilon < a < a^*_n + \varepsilon) = \beta;\ I_\beta[a] = (a^*_n - \varepsilon; a^*_n + \varepsilon)$.

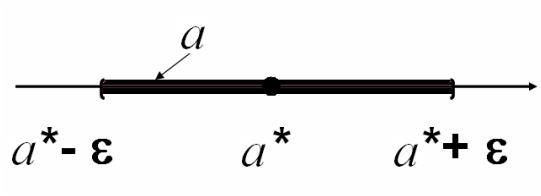


Рисунок 2.1 - Симметричный доверительный интервал

$\alpha = 1 - \beta$ - уровень значимости, вероятность того, что расхождения между параметром и его оценкой больше либо равно абсолютной величине доверительного интервала:

$P(|a - a_n^*| \geq \varepsilon) = \alpha$.

Чаще всего $\alpha = 0{,}05;\ 0{,}1$.

Доверительный интервал – числовой интервал значений параметра a ГС, которые не противоречат опытным данным или совместимы с опытными данными. Границы интервала и его величина получены по выборочным данным и поэтому случайны в отличие от самого параметра a.

Величина доверительного интервала ε существенно зависит:

- от объема выборки (с ростом n величина интервала уменьшается;
- от величины доверительной вероятности: чем больше доверительная вероятность β, тем больше ε.

Оценка доверительного интервала для математического ожидания. Пусть для параметра m генеральной совокупности получена доброкачественная оценка m^*. Нужно оценить полученную при этом ошибку «грубым» и «точным» методами. Определение I_β возможно, если известен закон распределения статистической оценки, который зависит от закона распределения самой СВ, и от конкретного значения параметра ГС.

«Грубый метод» используется при следующих допущениях:

- допущение нормальности закона распределения СВ;
- замена параметров этого закона их статистическими оценками.

Пусть имеется случайная величина X – описывающая ГС, с неизвестными параметрами $m,\ D$. Найти доверительный интервал для m, если задана доверительная вероятность и получены результаты эксперимента. Т.е., дано: $x_1, \ldots, x_n; \beta$. Найти: $I_\beta[m]: a = m_X;\ a^* = m^*$.

Известно, что статистическая оценка математического ожидания равна:

$$m^* = \tilde{m} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

В качестве оценки реального m по выборке принимается среднее арифметическое n независимых наблюденных значений.

x_i –некоторый экземпляр случайной величины X с параметрами m_X. Оценка m_X - это сумма n независимых одинаково распределенных СВ, тогда, по центральной предельной теореме при достаточно большом n закон распределения этой суммы близок к нормальному.

В практической статистике даже при относительно небольшом числе испытаний (от 10 до 20) считается, что закон распределения стремится к нормальному. Тогда, вероятность попадания в интервал для нормального закона равна:

$$P(a < X < b) = \Phi^*\left(\frac{b - m_X}{\sigma_X}\right) - \Phi^*\left(\frac{a - m_X}{\sigma_X}\right),$$

В симметричный интервал $\pm\varepsilon$ относительно m_X:

$$P(m_X - \varepsilon < X < m_X + \varepsilon) = \Phi\left(\frac{\varepsilon}{\sigma_X}\right) - 1 + \Phi\left(\frac{\varepsilon}{\sigma_X}\right) = 2\Phi\left(\frac{\varepsilon}{\sigma_X}\right) - 1 = 2\Phi^*\left(\frac{\varepsilon}{\sigma_X}\right).$$

Рассматриваемая СВ X - это оценка матожидания: $\tilde{m} = m^* \approx N(m, \frac{D}{n})$,

$$P(|\tilde{m} - m| < \varepsilon) = \beta = 2\Phi(\frac{\varepsilon}{\sigma_{m^*}}) - 1 \Rightarrow \Phi(\frac{\varepsilon}{\sigma_{m^*}}) = \frac{1+\beta}{2} \Rightarrow \varepsilon = \sigma_{m^*} \cdot \arg\Phi[\frac{1+\beta}{2}] =$$

$$= \sigma_{m^*} \cdot \arg\Phi^*[\frac{\beta}{2}], \sigma_{m^*} = \frac{\sigma_X}{\sqrt{n}} \Rightarrow \varepsilon = \frac{\sigma_X}{\sqrt{n}} \cdot \arg\Phi(\frac{1+\beta}{2}) = \frac{\sigma_X}{\sqrt{n}} \cdot \arg\Phi^*(\frac{\beta}{2}).$$

Величина доверительного интервала для матожидания равна (“грубый метод”):

$$u_\beta = \arg\Phi[\frac{1+\beta}{2}] = u_{1-\alpha/2},$$

где $u_\beta = u_{1-\alpha/2}$ - квантиль нормального распределения. Тогда

$$I_\beta[m] = [m^* - u_\beta \sqrt{\frac{D^*}{n}}; m^* + u_\beta \sqrt{\frac{D^*}{n}}].$$

Для примера 1 $m^* \approx 107{,}33$, $D^* \approx 250{,}64$, $u_{0,9} = 0{,}3289$, $u_{0,95} = 0{,}3340$, $u_{0,99} = 0{,}3389$. Тогда $I_{0,9}[m] = (81{,}82; 132{,}84)$, $I_{0,95}[m] = (81{,}42; 133{,}23)$, $I_{0,99}[m] = (81{,}04; 133{,}62)$.

Полученные с помощью «грубого» метода границы интервалов для математического ожидания, нанесем на полигон частот (см. рис. 2.2).

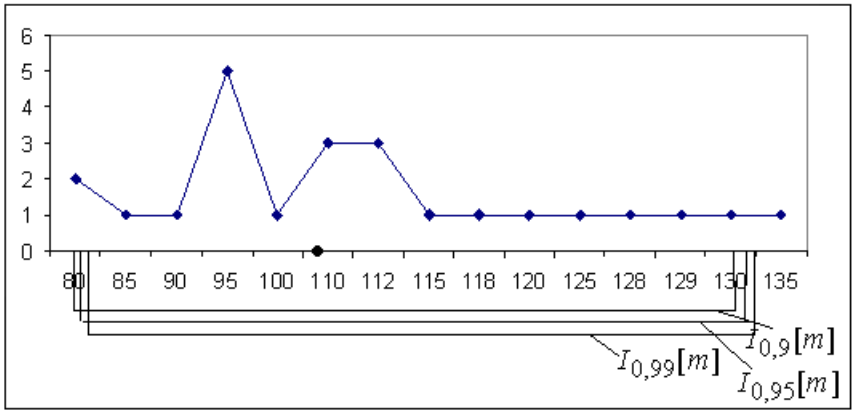


Рисунок 2.2 - Границы доверительных интервалов для мат. ожидания

“***Точный ” метод оценки достоверности матожидания.*** Если σ не известно, то используют σ^* и вместо нормального распределения t -распределение Стьюдента:

$$\varepsilon = \frac{\sigma^*}{\sqrt{n}} t_{1-\alpha/2}(n-1),$$

где $t_{1-\alpha/2}$ - квантиль t -распределения (табличное значение).

Доверительный интервал для D_X**.** Дана СВ X с нормальным законом распределения и неизвестными параметрами m и D**.** Произведено n независимых испытаний. Требуется по заданной доверительной вероятности найти доверительный интервал для D.

В качестве оценки D принимаем:

$$\tilde{D} = \frac{\sum_{i=1}^{n} (x_i - \tilde{m})^2}{n - 1}.$$

По аналогии с математическим ожиданием, ***оценка*** D ***грубым методом:***

$$I_\beta[D] = (\tilde{D} - \varepsilon; \tilde{D} + \varepsilon), \quad \varepsilon = u_\beta \cdot \sigma[\tilde{D}], \quad D[\tilde{D}] = \frac{\mu_4}{n} - \frac{(n-3)}{n(n-1)} \cdot D_X^2$$

Чтобы воспользоваться этими формулами вместо реальных D и μ_4 пользуются их оценками:

$$\mu_4^*[X] = \frac{\sum_{i=1}^{n} (x_i - m^*)^4}{n}$$

Нормальный закон:

$$D[\tilde{D}] = \frac{2}{n-1} \cdot D^2.$$

Равномерный:

$$D[\tilde{D}] = \frac{0{,}8n + 1{,}2}{n(n-1)} \cdot D^2.$$

Оценка D ***«точным» методом:*** если m известно, то

$$I_\beta^D = \left[\frac{nD^*}{\chi^2_{1-\alpha/2}(n)}; \frac{nD^*}{\chi^2_{\alpha/2}(n)} \right];$$

если m неизвестно, то берут m^*:

$$I_\beta^D = \left[\frac{nD^*}{\chi^2_{1-\alpha/2}(n-1)}; \frac{nD^*}{\chi^2_{\alpha/2}(n-1)} \right].$$

Рекомендуемая литература

Основная литература

1 Балдин, К.В. Эконометрика : учебное пособие / К.В. Балдин, О.Ф. Быстров, М.М. Соколов. - 2-е изд., перераб. и доп. - М. : Юнити-Дана, 2015. - 254 с. - Библиогр. в кн. - ISBN 5-238-00702-7 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=114533

2 Доррер, Г.А., Теория принятия решений: учебное пособие для студентов направления 230100.62 – Информатика и вычислительная техника. [Электронный ресурс] / Г.А. Доррер. – Красноярск : ФГБОУ ВПО «Сибирский государственный технологический университет», 2013. –

180 с. – Режим доступа : https://biblioclub.ru/index.php?page=book_view_red&book_id=428854 https://biblioclub.ru/index.php?page=book_view_red&book_id=208939

Дополнительная литература

1. Моделирование систем: Подходы и методы : учебное пособие / В.Н. Волкова, Г.В. Горелова, В.Н. Козлов и др. ; Министерство образования и науки Российской Федерации, Санкт-Петербургский государственный политехнический университет. - СПб. : Издательство Политехнического университета, 2013. - 568 с. : схем., ил., табл. - Библиогр. в кн. - ISBN 978-5-7422-4220-8 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=362986

2. Интеллектуальные системы : учебное пособие / А. Семенов, Н. Соловьев, Е. Чернопрудова, А. Цыганков ; Министерство образования и науки Российской Федерации, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Оренбургский государственный университет». - Оренбург : ОГУ, 2013. - 236 с. ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=259148

3. Иванова, В.В. Основы бизнес-информатики : учебник / В.В. Иванова, Т.А. Лёзина, А.А. Салтан ; Санкт-Петербургский государственный университет ; под ред. В.В. Ивановой. - СПб. : Издательство Санкт-Петербургского Государственного Университета, 2014. - 244 с. : табл., ил. - Библиогр. в кн. - ISBN 978-5-288-05538-6 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=458093

4. Чернышов, В.Н., Системный анализ и моделирование при разработке экспертных систем: учебное пособие. [Электронный ресурс] / В.Н. Чернышов, А.В. Чернышов. – Тамбов : Из-во Тамб. гос. техн. ун-та, 2012. – 128 с. – Режим доступа : https://biblioclub.ru/index.php?page=book_view_red&book_id=277638

5. Сурина, Е. Е. Методы анализа экономической информации [Текст] : учебно-методическое пособие / Е. Е. Сурина. - Орск : Изд-во ОГТИ (филиала) ОГУ, 2014. - 129 с. - ISBN 978-5-8424-0736-1. [Электронный ресурс]

Периодические издания

1. Журнал «Вестник компьютерных и информационных технологий »
2. Журнал «Информационные технологии и вычислительные системы»
3. Журнал «Стандарты и качество»
4. Журнал «Прикладная информатика»