Федеральное агентство по образованию

Государственное образовательное учреждение высшего профессионального образования

Восточно - Сибирский государственный технологический университет

(ГОУВПО ВСГТУ)

А. Е. Бордоева

Описательная статистика и проверка статистических гипотез средствами EXCEL

в примерах

Издательство ВСГТУ
Улан - Удэ
2009

Бордюева А. Е.
Описательная статистика и проверка статистических гипотез средствами Excel. Учебно – методическое пособие. Улан-Удэ: Изд – во ВСГТУ, 2009. – 60 с.

Учебно – методическое пособие содержит теоретический и практический материал по применению методов теории вероятностей и математической статистики для обработки экспериментальных данных в среде табличного процессора Excel.

В пособии имеется минимально необходимый теоретический материал из теории вероятностей и математической статистики, а также дана методика применения соответствующих статистических функций Excel.

На примерах показана методика определения и анализа параметров описательной статистики и законов распределения случайных величин, построения доверительных интервалов для статистических параметров на основе экспериментальных данных средствами Excel. Также рассматриваются критерии значимости и методы проверки статистических гипотез. Отдельной главой даны статистические методы контроля качества продукции.

В пособии приведены варианты индивидуальных заданий для практического освоения рассмотренной методики.

Работа является первой частью разработок автора. Вторая часть будет посвящена построению и анализу уравнений связи средствами Excel.

Предназначено для студентов всех специальностей, изучающих курс «Статистическое моделирование», а также тем, кто хочет получить первоначальные навыки применения методов математической статистики в своих исследованиях с применением средств табличного процессора EXCEL.

Рецензент: Е.Г.Чимитова, зав.каф. информатики и ВТ БФ СГУ телекоммуникаций и информатики, доц. к.п.н.

Печатается по решению редакционно – издательского совета ВСГТУ

ПРЕДИСЛОВИЕ

Все явления, процессы в природе, экономике, человеческом обществе и т.д. взаимосвязаны между собой. Но такого рода связи являются приближенными и носят случайный характер, так как невозможно предусмотреть и проанализировать влияние всех факторов. Поэтому вид таких зависимостей и их достоверность устанавливается с применением методов теории вероятностей и математической статистики, которые дают возможность специалисту в любой области построить математическую модель изучаемого процесса и оценить с той или иной вероятностью достоверность полученных параметров.

Применение статистических методов требует вычислений большого объема, что невозможно выполнить без средств автоматизации – будь это калькуляторы или мощные компьютеры.

В настоящее время существуют множество программных средств реализации статистических методов. Каждая программа дает возможность получить статистические параметры изучаемого ряда экспериментальных данных и оценить их достоверность. Одни из них удобны для тех, кто только начинает использовать статистические методы при анализе изучаемого процесса, другие программы предназначены для опытных исследователей.

Предлагаемое учебное пособие ориентировано на пользователей, которые хотят получить первоначальные навыки применения методов статистической обработки данных. Поэтому в качестве программного продукта выбран табличный процессор *Microsoft Excel* и его стандартные статистические функции. Применение этих функций доступно любому пользователю, имеющему навыки работы с этим приложением.

Глава 1. Некоторые понятия математической статистики

Математическая статистика – это раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей. Статистические данные представляют собой информацию, полученную в результате обследования большого числа объектов или явлений; следовательно, математическая статистика имеет дело с массовыми явлениями. Для анализа и обработки используются математические методы, которые называются *статистическими методами.*

Современная математическая статистика подразделяется на две обширные области:

- **описательная статистика** – это методы описания статистических данных с помощью числовых параметров, представление данных в форме таблиц, распределений и прочее;
- **аналитическая статистика** – это теория статистических выводов. Ее предметом является обработка данных, полученных в ходе эксперимента, и формулировка выводов. Применяется в различных областях человеческой деятельности и базируется на математическом аппарате теории вероятностей.

Основным методом статистического исследования является *выборочный метод.*

1.1. Выборочный метод

Методы математической статистики позволяют осуществить сбор, анализ и обработку данных, полученных в результате обследования большого числа объектов или явлений. Так как на практике чаще всего невозможно сплошное обследование изучаемых объектов, то применяют выборочное обследование, т.е. используют **выборочный метод**. Отобранные данные образуют статистическую совокупность, а отбор осуществляется по определенному *качественному* или *количественному* признаку. Например, если имеется *партия деталей*, то *качественным признаком* может служить *стандартность* детали, а *количественным* – контролируемый *размер* детали.

При этом исходная совокупность называется **генеральной совокупностью**, а отобранные из нее для исследования данные – **выборочной совокупностью** (*выборкой*). Важнейшая характеристика выборки – ее **объем** (*количество отобранных элементов).*

Анализируемые признаки объекта варьируют под воздействием большого числа различных факторов, лишь небольшую часть которых можно контролировать и предсказать их влияние. Поэтому наблюдаемые значения результатов обследования носят *случайный характер.*

Величина называется *случайной,* если в результате испытания принимает одно и только одно возможное значение, наперед неизвестное, и зависящее от случайных причин, которые заранее не могут быть учтены.

Случайная дискретная величина принимает отдельные возможные значения с определенной вероятностью. Число возможных значений может быть *конечным* или *бесконечным. Например,* для оценки качества отобраны *100 лампочек.* Возможность появления бракованной лампочки есть *случайная величина,* которая может иметь значения *0, 1, 2……100.*

Случайная непрерывная величина может принимать все значения из некоторого конечного или бесконечного промежутка. *Например,* расстояние, которое пролетит ядро (копье), брошенное спортсменом, *случайная величина.* Возможные значения принадлежат промежутку *(a, b).*

1.2. Статистическое распределение выборки

Пусть из генеральной совокупности извлечена выборка. Отдельные числовые значения варьирующего признака называются **вариантами** и обозначаются X_i. При этом если

значение X_1 наблюдалось n_1 раз, X_2 - n_2 раз и т. д., то величины n_i называются **частотами** и $\Sigma n_i=n$, где n – объем выборки; $W_i = \frac{n_i}{n}$ - относительная частота (вероятность появления случайной величины).

Статистическим распределением выборки называется перечень вариант и соответствующих им частот или относительных частот.

Если варианты расположить в возрастающем порядке, то такая последовательность называется ранжированным **вариационным рядом.**

Пример: даны варианты некоторого признака и их частоты – *распределение частот выборки:*

X_i	2	6	12
n_i	3	10	7

Объем выборки $n = 3 + 10 + 7 = 20$.

Относительные частоты:

$$W_1 = \frac{3}{20} = 0{,}15;\ W_2 = \frac{10}{20} = 0{,}50;\ W_3 = \frac{7}{20} = 0{,}35;$$

Распределение относительных частот

X_i	2	6	12
W_i	0,15	0,5	0,35

Эти величины показывают вероятность появления соответствующей варианты и их *сумма равна 1.*

1.3. Определение параметров описательной статистики

Числовые характеристики выборки называются *параметрами описательной статистики.* К ним относятся:

1. **Выборочная средняя величина** – среднее арифметическое значение признака выборочной совокупности. Вычисляется по **формуле:**

$$x_B = \frac{x_1 + x_2 + \ldots + x_n}{n},$$ где n - объем выборки.

2. **Выборочная дисперсия** – это среднее арифметическое значение квадратов отклонений наблюдаемых значений признака от их среднего значения $x_в$

Расчетная формула: $$D_B = \frac{\sum (x_i - x_B)^2}{n}$$

Дисперсия характеризует рассеяние наблюдаемых значений признака вокруг среднего значения.

3. Выборочное среднеквадратическое отклонение (стандартное отклонение):

$$\sigma_B = \sqrt{D_B}.$$

Эта величина также характеризует рассеяние значений признака около средней, только в отличие от дисперсии ее размерность совпадает с размерностью исследуемого признака, а у дисперсии размерность равна квадрату размерности признака.

4. **Медиана** – варианта, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант **нечетно**, т.е. $n=2k+1$, то $m_e=x_{k+1}$; при **четном** числе вариант $n=2k$, медиана $m_e = \frac{x_k + x_{k+1}}{2}$

Федеральное агентство по образованию

Государственное образовательное учреждение высшего профессионального образования

Восточно - Сибирский государственный технологический университет

(ГОУВПО ВСГТУ)

А. Е. Бордоева

Описательная статистика и проверка статистических гипотез средствами EXCEL

в примерах

Издательство ВСГТУ
Улан - Удэ
2009

Пример:

a) дан вариационный ряд 2, 3, 5, 6, 7, где n=5. Тогда медиана равна 5;

b) у вариационного ряда 2, 3, 5, 6, 7, 9 объем выборки n=6, а медиана $m_e = \frac{5+6}{2} = 5{,}5$.

5. **Мода** - варианта, которая имеет наибольшую частоту.

Пример:

Варианты	1	4	7	9
Частоты	5	1	20	6

Мода равна 7 - наиболее часто встречающаяся варианта.

6. **Размах вариации** - это разность между значениями максимальной и минимальной вариант, т.е.

$$R = x_{max} - x_{min}.$$

Это простейшая характеристика рассеяния вариационного ряда. Используется обычно при небольшом объеме выборки.

7. **Коэффициент вариации** - это отношение выборочного стандартного отклонения к выборочной средней величине. Измеряется в процентах, т.е. $V = \frac{\sigma_в}{x_в} * 100\%$.

Этот коэффициент служит для сравнения величин рассеяния по отношению к выборочной средней для двух вариационных рядов: тот ряд имеет большее рассеяние, у которого коэффициент вариации выше. Это безразмерная величина, поэтому пригодна для сравнения рассеяния рядов, варианты которых имеют разную размерность, т. к. в этом случае применение стандартного отклонения неудобно. Практически коэффициент вариации применяется в основном для сравнения выборок из однотипных генеральных совокупностей.

8. **Коэффициент эксцесса** - величина, которая характеризует относительную **остроконечность** или **сглаженность** распределения по сравнению с нормальным распределением. Если его значение **больше нуля**, то график распределения **остроконечен**, иначе – **сглажен**.

9. **Коэффициент асимметрии** – величина, которая указывает на симметричность распределения относительно математического ожидания. Если *значение коэффициента асимметрии равно нулю*, то распределение симметрично относительно математического ожидания (средней величины). При положительном значении коэффициента преобладают положительные отклонения от математического ожидания *(положительная асимметрия)*, а при отрицательном значении – отрицательные отклонения (*отрицательная асимметрия).*

1.4. Применение стандартных функций EXCEL

Стандартные функции **Excel** для вычисления параметров описательной статистики образуют категорию **Статистические** и вычисляются с помощью инструмента **Мастер функций**. Каждая функция имеет следующий **синтаксис**:

=ИМЯ (аргумент), где

ИМЯ - стандартное имя по умолчанию;

аргумент - список операндов, перечисленные через точку с запятой и представляющие собой числовые константы, адреса ячеек или блока ячеек.

Таблица 1.1. Статистические функции

№	Синтаксис функции	Выполняемое действие
1	СРЗНАЧ (аргумент)	Среднее арифметическое значение
2	ДИСП (аргумент)	Выборочная дисперсия
3	ДИСПР (аргумент)	Дисперсия генеральной совокупности
4	КВАДРОТКЛ (аргумент)	Сумма квадратов отклонений значений ряда от выборочной средней
5	МЕДИАНА (аргумент)	Медиана выборки
6	МОДА (аргумент)	Мода выборки
7	СТАНДОТКЛОН(аргумент)	Стандартное отклонение выборки
8	МАКС (аргумент)	Максимальное значение
9	МИН (аргумент)	Минимальное значение
10	ЭКСЦЕСС (аргумент)	Эксцесс выборки
11	СКОС (аргумент)	Асимметрия выборки

1.5. Оценка генеральных параметров

Так как эксперимент для всей генеральной совокупности нереализуем или неоправдан, то на основании данных, полученных по выборке, делается вывод относительно всей генеральной совокупности. Для этого используются методы теории статистических выводов, которые делятся на два класса: *оценка параметров* и *проверка гипотез*.

Задача *оценки генеральных параметров* состоит в получении наилучших в определенном смысле оценок параметров распределения генеральной совокупности на основе выборочных данных.

Проверка гипотез - это методы для проверки предположений о распределении и параметрах распределения генеральной совокупности, выдвигаемые до получения выборочных данных.

Чтобы по выборке можно было делать выводы о свойствах всей генеральной совокупности, она должна быть *представительной* (репрезентативной). Это обеспечивается лишь

тогда, когда выборка носит случайный характер. Для обеспечения случайной выборки на практике применяются разные способы отбора: *случайный бесповторный отбор, серийный отбор* и т.п.

1.6. Ошибки выборочного наблюдения

Информация, полученная в результате любого статистического наблюдения, имеет расхождение с реальной действительностью. Такое расхождение называют *ошибками статистического наблюдения.*

Причины возникновения этих ошибок различны, но они не имеют ничего общего с ошибками измерения. Например, *статистическая ошибка* или ошибка репрезентативности - *отклонение выборочных параметров от оценок генеральных параметров* - возникает потому, что не все объекты генеральной совокупности представлены в выборке. Такая ошибка является *случайной* и уменьшается при достаточно большом объеме выборки.

Различают среднюю *(стандартную)* и *предельную* ошибку выборки.

Стандартная ошибка - расхождение между средней величиной выборочной и генеральной совокупностей, не превышающей значения генерального стандартного отклонения. Величина этой ошибки вычисляется по разным формулам в зависимости от типа выборки. Так, для случайного отбора при достаточно большом объеме выборки можно применить формулу: $\mu = \frac{\sigma}{\sqrt{n}}$, где σ - стандартное отклонение генеральной совокупности; n - объем выборки ($n>30$).

Эта величина показывает, какая ошибка допускается в среднем, если использовать вместо генерального среднего его выборочную оценку.

Примечание: для малой выборки размера от 5 до 30 единиц *стандартная ошибка* вычисляется по формуле:

$$\mu = \frac{\sigma}{\sqrt{n-1}}.$$

Предельная ошибка - это максимально возможное расхождение между средними величинами выборки и генеральной совокупности.

Формула для вычисления: $\Delta = t\mu$, где t - заданный коэффициент доверия.

С учетом формулы для стандартной ошибки предельная ошибка вычисляется так:

$$\Delta = \frac{t\sigma}{\sqrt{n}}.$$

Так при t = 1 величина $\Delta = \mu$, которая гарантируется с вероятностью 0,683, что означает следующее: в 683 выборках из 1000 подобных максимальная ошибка выборки не превысит значения $\pm 1\mu$.

При t =2 с вероятностью 0,954 она не выйдет за пределы $\pm 2\mu$. На практике, как правило, максимальный предел ошибок достаточен в пределах $\pm 3\mu$, что соответствует вероятности 0,997.

Значения вероятностей и t определяются из таблицы значений функции распределения вероятностей нормального закона – *функции Ф(t).*

Примечание: для выборок большого объема выборочную и генеральную дисперсии можно считать равными. Поэтому в формуле для стандартной ошибки σ можно считать выборочным стандартным отклонением.

1.7. Типы оценок генеральных параметров

Статистические ошибки применяются для оценки генеральных параметров на основе соответствующих параметров

выборки. Оценки подразделяются на два типа: *точечные и интервальные.*

Точечной оценкой числовых параметров генеральной совокупности являются значения этих параметров, полученные по выборке. Так, оценкой генерального среднего является *выборочное среднее*, а оценкой генеральной дисперсии - *выборочная дисперсия.* Так, для генеральной средней величины *точечная оценка* на основе выборочной средней выглядит так: $X_{ср} \pm \mu$.

Нужно отметить, что точечную оценку можно применять для выборок большого объема. При выборке малого объема точечная оценка значительно отличается от оцениваемого генерального параметра.

Интервальная оценка определяет границы интервала, между которыми с большой вероятностью находятся истинные значения параметров.

Вероятности, признанные достаточными для того, чтобы уверенно судить о генеральных параметрах на основании выборочных характеристик, называются *доверительными.*

Интервал, в котором с заданной доверительной вероятностью находится оцениваемый генеральный параметр, называется *доверительным интервалом.*

Обычно в качестве доверительных вероятностей выбирают значения *0.95, 0.99, 0.999* (их выражают в %). Они соответствуют *95%, 99%, 99.9%.* Выбор той или иной вероятности производится исследователем исходя из практических соображений (чаще всего используется вероятность *0.95*).

При анализе говорят о *100(1 - α) -* процентном доверительном интервале, где *(1 - α) - доверительная вероятность*, а α - некоторое малое число *(0.05; 0.01; 0.001)*, задающее вероятность того, что оцениваемый генеральный параметр выходит за границы доверительного интервала.

Доверительный интервал для генеральной средней величины определяется следующим образом:

$$X_{выб} - \frac{t\sigma}{\sqrt{n}} \le X_{ген} \le X_{выб} + \frac{t\sigma}{\sqrt{n}},$$

где $X_{ген}$ - средняя величина генеральной совокупности; $X_{выб}$ - средняя величина выборочной совокупности; t - параметр *Стьюдента*, который определяется из таблицы распределения *Стьюдента* при заданной доверительной вероятности; σ - стандартное отклонение выборки; n - объем выборки.

Как известно, $\Delta = \frac{t\sigma}{\sqrt{n}}$ - *предельная ошибка выборки.*

Смысл формулы заключается в следующем: *с заданной вероятностью можно утверждать, что значение генеральной средней можно ожидать в пределах от* $X_{выб} - \Delta$ *до* $X_{выб} + \Delta$, *т.е. доверительный интервал* $(X_{выб} - \Delta, X_{выб} + \Delta)$ *с заданной вероятностью заключает в себе генеральную среднюю величину.*

Для определения предельной ошибки $\Delta = \frac{t\sigma}{\sqrt{n}}$ в среде Excel применяется стандартная статистическая функция, синтаксис которой следующий:

=ДОВЕРИТ(*α; станд - откл; размер*), где

α - некоторое малое число *(0.05; 0.01; 0.001)*, задающее вероятность того, что оцениваемый генеральный параметр выходит за границы доверительного интервала. Тогда доверительная вероятность равна (1 - α). *Например,* при α =0,05 доверительная вероятность равна 0,95 или уровень надежности равен 95%;

станд - откл - стандартное отклонение генеральной совокупности, которое при больших выборках заменяется выборочным стандартным отклонением;

размер – объем выборки.

1.8. Лабораторная работа № 1. Определение параметров описательной статистики. Оценка генеральных параметров.

Задание 1. Определить параметры описательной статистики по данным *объема поступления* продукции за *10* дней.

Таблица 1.2. Исходные данные

Дни	1	2	3	4	5	6	7	8	9	10
Объем	300	280	400	350	530	350	480	250	330	440

Параметры описательной статистики вычисляются с помощью стандартных функций EXCEL из категории Статистические. Результаты выведены в таблице 1.3.

Таблица 1.3. Параметры описательной статистики

Среднее	Максимум	Минимум	Дисперсия	Стандартное отклонение	Медиана	Мода	Эксцесс	Размах вариации	Асимметрия
371	530	250	8143,33	90,24	350	350	-0,67	280	0,51

Примечания.

- стандартное отклонение указывает на *меру рассеяния* данных около средней величины, т.е. в среднем величины поставок отличаются от среднего объема на 90 единиц;
- медиана - это *середина* ряда;
- мода - наиболее *часто встречающаяся варианта.*
- эксцесс имеет *отрицательное* значение. Следовательно, график распределения сглажен по отношению к нормальному распределению;
- размах вариации показывает *разброс* данных ряда;
- значение асимметрии 0,51 указывает на *положительную асимметрию.*

Задание 2: на основе вычисленного значения выборочной средней величины *задания 1* дать *точечную* и *интервальную* оценки генеральной средней величины.

Для *точечной оценки* вычисляется стандартная ошибка по формуле для малой выборки: $\mu = \frac{\sigma}{\sqrt{n-1}}$.

Таблица 1.4. Точечная оценка

Точечная оценка					
Выборочная средняя	Стандартное отклонение	Объем выборки	Стандартная ошибка	Нижняя оценка	Верхняя оценка
371	90,24	10	30,08	340,92	401,08

Для интервальной оценки вычисляется предельная ошибка $\Delta = \frac{t\sigma}{\sqrt{n}}$ с применением статистической функции ДОВЕРИТ при разных значениях уровня значимости.

Таблица 1.5. Интервальная оценка

Интервальная оценка						
Выборочная средняя	Стандартное отклонение	Объем выборки	Параметр α	Предельная ошибка	Нижняя граница	Верхняя граница
371	90,24	10	0,01	73,50	297,50	444,50
			0,05	55,93	315,07	426,93
			0,1	46,94	324,06	417,94
			0,03	61,93	309,07	432,93
			0,046	56,94	314,06	427,94

1.9. Задания для самостоятельного выполнения

В каждом из заданий необходимо:

- подобрать исходные данные;
- определить параметры описательной статистики;
- дать точечную оценку средней величины генеральной совокупности;
- указать доверительный интервал для генеральной средней величины.

Примечание: указанный в заданиях объем генеральной совокупности будет использован для следующей лабораторной работы.

Варианты заданий:

1. *Исходные данные* - значения температуры воздуха за 10 дней. Генеральная совокупность - данные за месяц.

2. *Исходные данные* - объем продаж за 10 дней. Генеральная совокупность - данные за месяц.

3. *Исходные данные* - выработка 12 рабочих. Генеральная совокупность - данные 25 рабочих.

4. *Исходные данные* - значения прибыли за полугодие. Генеральная совокупность - данные за год.

5. *Исходные данные* - товарооборот за 15 дней. Генеральная совокупность - данные за месяц.

6. *Исходные данные* - показания прибора в 20 измерениях. Генеральная совокупность - 30 измерений.

7. *Исходные данные*- значения роста 15 человек. Генеральная совокупность – 25 человек.

8. *Исходные данные* – значения веса 12 человек. Генеральная совокупность – 25 человек.

9. *Исходные данные* – стоимость акций за 15 дней. Генеральная совокупность - данные за месяц.

10. *Исходные данные* – количество деталей, выточенных токарем за день в течение 10 дней. Генеральная совокупность – данные за 20 дней.

11. *Исходные данные* – годовой процент автомобилей импортного производства в России с 1990 по 1999 гг. Генеральная совокупность – данные с 1985 по 2003 годы.

12. *Исходные данные* – влажность зерна в % за 15 дней. Генеральная совокупность – данные за месяц.

13. *Исходные данные* - размер 16 выточенных деталей на станке – автомате. Генеральная совокупность – 30 деталей.

14. *Исходные данные* – вес 20 упаковок с макаронными изделиями, изготовляемые автоматической линией. Генеральная совокупность – 35 упаковок.

15. *Исходные данные* – температура технологической установки за 15 дней. Генеральная совокупность – данные за месяц.

16. *Исходные данные* – сумма издержек торговой фирмы за 15 дней. Генеральная совокупность – данные за месяц.

ГЛАВА 2. Вариационные ряды

2.1. Понятие вариационного ряда

Необработанные (первичные) экспериментальные данные представлены в виде неупорядоченного набора чисел, записанных исследователем в порядке их поступления. Этот набор данных трудно обозрим, и сделать по ним какие – то выводы невозможно. Поэтому первичные данные нуждаются в обработке и упорядочении в виде вариационного ряда, который может быть *безынтервальным* и *интервальным*.

Вариационный ряд – это двойной числовой ряд, показывающий, каким образом числовые значения изучаемого признака связаны с их повторяемостью в выборке.

В *безынтервальном вариационном ряду* частоты распределяются непосредственно по значениям варьирующего признака. Для построения такого ряда нужно варианты расположить в порядке возрастания или убывания и подсчитать количество повторений каждого значения (частота). Безынтервальный ряд применяется в тех случаях, когда исследуемый признак варьирует дискретно и слабо.

В *интервальном вариационном ряду* частоты распределяются по интервалам группировки. Такой ряд строится в том случае, если изучаемый признак варьирует непрерывно или тогда, когда дискретно варьирующие признаки меняются в широких пределах.

2.2. Построение безынтервального вариационного ряда

Как уже было сказано, для построения такого ряда необходимо варианты расположить в порядке возрастания или убывания и подсчитать количество повторений каждого значения (частота).

Задание 1. Дан выборочный ряд из **30** значений, т.е. **n = 30**.
Построить безынтервальный вариационный ряд.

Таблица 2.1. Исходные данные.

№	Исходный ряд	Ранжированный ряд
1	5	5
2	16	5
3	18	6
4	19	7
5	14	7
6	12	8
7	22	8
8	23	9
9	25	11
10	20	12
11	32	13
12	17	14
13	34	14
14	25	14
15	14	14
16	14	16
17	17	17
18	8	17
19	5	18
20	11	19
21	13	20
22	6	21
23	7	22
24	9	23
25	14	23
26	7	25
27	21	25
28	28	28
29	23	32
30	8	34

Примечание: для ранжирования необходимо произвести сортировку исходного ряда по возрастанию.

Для построения безынтервального вариационного ряда нужно перечислить неповторяющиеся значения вариант, их частоты и частости. Полученный ряд дан в *таблице 2.2.*

Таблица 2.2. Безынтервальный ряд

Варианты	Частота	Частость
5	2	0,067
6	1	0,033
7	2	0,067
8	2	0,067
9	1	0,033
11	1	0,033
12	1	0,033
13	1	0,033
14	4	0,133
16	1	0,033
17	2	0,067
18	1	0,033
19	1	0,033
20	1	0,033
21	1	0,033
22	1	0,033
23	2	0,067
25	2	0,067
28	1	0,033
32	1	0,033
34	1	0,033
n	**30**	**1**

Задание на самостоятельное выполнение: определить параметры описательной статистики.

2.3. Построение интервального вариационного ряда

Построение интервального вариационного ряда начинается с группировки.

Группировка – процесс систематизации или упорядочения данных для удобства обработки. Существует множество методов группировки, но чаще всего она сводится к представлению данных в виде *статистических таблиц*.

В такой таблице данные распределяются по группам, каждая из которых содержит некоторый диапазон значений изучаемого признака. Ключевым моментом при этом является определение *числа интервалов* и *ширины* каждого из них. Обычно предпочтительны интервалы *одинаковой ширины*. Выбор числа интервалов зависит от цели исследования, объема выборки и степени варьирования признака в выборке. Однако приближенно число интервалов можно оценить по *формуле Стерджеса:* $k = 1+3{,}32\ lg(n)$, где n – объем выборки или выбрать из *следующей таблицы:*

Таблица 2.3. Выбор числа интервалов группировки

Объем выборки, n	Число интервалов, k
25 – 40	5 - 6
40 – 60	6 – 8
60 – 100	7 – 10
100 – 200	8 – 12
Больше 200	10 - 15

После выбора числа интервалов в зависимости от объема выборки определяется *ширина интервала* по формуле:

$h = \frac{x_{max} - x_{min}}{k}$, *где* x_{max} - максимальная варианта;

x_{min} – минимальная варианта; k - число интервалов.

Далее определяются границы интервалов группировки. Нижняя граница первого интервала равна минимальной варианте или в некоторых случаях *определяется по формуле:*

$$\mathbf{x_{n1} = x_{min} - \frac{h}{2}}$$

Нижняя граница второго интервала будет: $\mathbf{x_{n2} = x_{n1} + h}$.

Она же будет одновременно верхней границей предыдущего интервала (первого интервала). Аналогично определяются границы остальных интервалов.

После определения границ интервалов необходимо заполнить статистическую таблицу, в которой нужно в начале указать *границы интервалов* и *частоты интервалов* – количество вариант, относящихся к данному интервалу (n_i), при этом $\mathbf{\Sigma\, n_i = n}$.

Задание 2. По данным *задания 1* построить интервальный вариационный ряд.

Для этого в первую очередь определяем количество интервалов по *формуле Стерджеса:* k = 1+3,32 lg(n) и шаг интервала по формуле

$h = \frac{x_{max} - x_{min}}{k}$, *где* x_{max} - максимальная варианта;

x_{min} – минимальная варианта; k - число интервалов.

$x_{max} = 34;\ x_{min} = 5;\ k = 6;\ h = 5.$

Для определения частот применяется стандартная статистическая функция Excel со следующим синтаксисом:

ЧАСТОТА(массив _ вариант; массив_ границ интервала), где *массив _ вариант* – диапазон с исходным рядом; *массив_ границ интервала* – диапазон с границами за исключением нижней и верхней границ интервала.

Построенный интервальный вариационный ряд дан в *таблице 2.4.*

Таблица 2.4. Интервальный ряд

Интервалы	Границы интервалов	Частоты	Частости
5 - 10	10	8	0,27
10 - 15	15	7	0,23
15- 20	20	6	0,20
20 - 25	25	6	0,20
25 - 30	30	1	0,03
30 - 35		2	0,07
	n	30	1

Задание 3. Определить точечную и интервальную оценки средней величины

Таблица 2.5. Точечная оценка средней величины

Xср	ско	Стандартная ошибка	Нижняя оценка	Верхняя оценка
16,2	7,9	1,44	14,76	17,64

Таблица 2.6. Интервальная оценка средней величины

Xср	ско	Предельная ошибка	Нижняя граница	Верхняя граница
16,2	7,9	2,83	13,37	19,03

Глава 3. Закон распределения случайных величин

3.1. Формы представления закона распределения

Случайная величина – это величина, принимающая одно из возможных значений в результате испытаний. Это значение заранее не известно.

Случайная дискретная величина задается отдельными значениями; **случайная непрерывная величина** принимает все значения из некоторого промежутка. Значения случайной величины образуют *статистический ряд.*

Для анализа такого ряда можно использовать числовые характеристики описательной статистики: *среднее значение, стандартное отклонение, моду, медиану, размах* и т.д., как было указано ранее. Каждая из этих величин определенным образом характеризует данный ряд.

Наиболее полной характеристикой случайной величины является закон ее распределения. **Закон распределения** случайной дискретной величины задает соответствие между ее возможными значениями и их вероятностями. Закон распределения имеет две **формы представления:**

- функция распределения - $F(x)=P(x<=a)$, определяет вероятность появления случайной величины; принимает значения от 0 до 1 (*интегральная функция*);
- плотность распределения - $f(x)=F^{1}(x)$, т.е. первая производная от функции распределения (*дифференциальная функция*). Показывает вероятность попадания случайной величины в заданный интервал, т.е. какие значения случайной величины наиболее вероятны.

При решении практических задач встречаются различные законы распределения случайной величины (*равномерный, нормальный, показательный, биномиальный и т.п.).*

3.2. Нормальный закон распределения

Распределение вероятностей случайной величины называется ***нормальным,*** если оно описывается дифференциальной функцией следующего вида:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

В этой формуле: a – математическое ожидание (средняя величина); σ - стандартное отклонение.

Если $a=0$; $\sigma=1$, то такое нормальное распределение называется **нормированным** для величины $U=\frac{x-a}{\sigma}$.

Чаще всего эмпирические данные имеют **нормальный закон распределения**. График плотности нормального распределения называют **нормальной кривой** (*кривой Гаусса*), вид которой зависит от параметров распределения *следующим образом*:

- *изменение величины a не меняет форму кривой, а приводит к сдвигу вдоль* ***оси X: вправо,*** *если значение* ***a*** *возрастает;* ***влево,*** *если значение* ***a*** *убывает;*
- *изменение* σ *меняет* ***форму кривой****: с возрастанием* σ *ордината кривой уменьшается, а сама кривая становится более пологой, т. е. сжимается к оси X; при убывании* σ *кривая становится более* ***островершинной.***

Нужно отметить, что не существует распределений эмпирических данных, которые были бы в точности нормальными, поскольку нормальная случайная величина находится в пределах от - ∞ до + ∞, чего не бывает на практике. Однако нормальное распределение очень часто хорошо подходит как приближение. К тому же многие распределения при $n\to\infty$ переходят в нормальное распределение.

Для нормального распределения такие параметры описательной статистики как *средняя величина, мода* и *медиана* имеют равные значения. Следовательно, если закон эмпирического ряда неизвестен, но значения указанных величин близки между собой, то можно предположить, что изучаемый показатель описывается моделью *нормального распределения.* Также для нормального распределения такие параметры как *асимметрия* и *эксцесс* равны *нулю.* Следовательно, если эти величины близки к нулю, то можно принять для описания нормальный закон.

Теоретически случайная величина может изменяться от - ∞ до + ∞. Однако с точностью до долей процента можно указать ее пределы изменения по закону *трех сигм,* который формулируется следующим образом:

если случайная величина распределена нормально, то абсолютная величина ее отклонения от математического ожидания не превосходит утроенного стандартного отклонения, т.е.: a - 3*σ<=X<=a + 3*σ - пределы изменения случайной величины.

На практике его применяют так: *если распределение случайной величины неизвестно, но закон трех сигм выполняется, то предполагается, что она имеет нормальное распределение.*

3.3. Определение форм нормального распределения средствами EXCEL

Поставленная задача сводится к определению функции распределения F(x) и плотности распределения f(x). Для этого используется стандартная статистическая функция следующего вида:

=НОРМРАСП(*x; среднее значение; стандартное отклонение; интегральная*), где x - данное значение случайной величины, для которой определяются *F(x)* и *f(x)*;
интегральная - определяет форму распределения, для которой определяется значение.

Данная функция используется следующим образом:

- для определения значений функции распределения *F(x)* функция имеет следующий вид:
  =НОРМРАСП(*x; среднее значение; стандартное отклонение; истина*);
- для определения значений плотности распределения *f(x)* используется функция: =НОРМРАСП(*x; среднее значение; стандартное отклонение; ложь*).

Алгоритм определения функции распределения и плотности распределения.

1. Определить ***a*** и σ - *среднее значение,* и *стандартное отклонение.*

2. Найти пределы изменения случайной величины, используя закон трех сигм (*генеральная совокупность*):

$$x_{min}=a-3*\sigma; \quad x_{max}=a+3*\sigma.$$

3. Вычислить шаг изменения случайной величины для определения интервала изменения генеральной совокупности:

$h=\frac{x_{max}-x_{min}}{n}$, где n - количество значений в интервале *(задается исследователем как объем генеральной совокупности).*

4. Определить значения случайной величины в найденном интервале через вычисленный шаг (от x_{min} до x_{max} с шагом h) – генеральная совокупность.

5. Вычислить значения плотности распределения f(x) с помощью функции НОРМРАСП с параметром "*интегральная = ложь*".

6. Вычислить значение функции распределения F(x) с помощью функции НОРМРАСП с параметром "*интегральная = истина*".

7. Построить график плотности распределения с помощью мастера диаграмм. Тип диаграммы – График.

8. Построить график функции распределения с помощью мастера диаграмм. Тип диаграммы – График.

9. Построить совместные графики плотности и функции распределения с помощью мастера диаграмм. Тип диаграммы - *Нестандартная, График с двумя осями.*

3.4. Лабораторная работа № 2. Определение форм нормального распределения

Задание 1: определить значения функций f(x) и F(x) по данным поступления продукции в течение *10 дней* (данные из работы 1):

Таблица 3.1. Исходные данные

Дни	1	2	3	4	5	6	7	8	9	10
Объем	300	280	400	350	530	350	480	250	330	440

Выполняемые действия:

1. Определить *среднее значение* и *стандартное отклонение:* a= 371; σ= 90,24.

2. Определить *пределы изменения* случайной величины:

   $x_{min} = a - 3*\sigma = 100{,}28$; $x_{max} = a + 3*\sigma = 641{,}72$.

   При построении таблицы вычисленные значения округлить соответственно до 100 и 650.

3. Вычислить *шаг изменения* случайной величины для определения *интервала изменения*:

   $h = \frac{x_{max} - x_{min}}{n} = \frac{650 - 100}{15} = 36{,}6$ (*округлить до 40).*

   n = 15 – объем генеральной совокупности.

4. Вычислить значения объема в пределах от 100 до 650 с *шагом* 40 (генеральная совокупность).

5. Вычислить значения *плотности распределения* f(x) и *функции распределения* F(x) с помощью функции НОРМРАСП с соответствующими значениями параметра «*интегральная*».
   Для этого нужно вставить функцию для первого значения объема, а затем распространить вниз по столбцу для остальных значений.
   Так как значения средней величины и стандартного отклонения остаются неизменными при распространении формулы, то ссылка на их адреса должна быть абсолютной, т.е. с применением знака *доллара.*

   Полученные результаты даны в таблице 2.2.

Таблица 3.2. Результаты расчета

Объем	f(x)	F(x)
100	0.00005	0.0013
140	0.00017	0.0052
180	0.00047	0.0171
220	0.00109	0.0471
260	0.00207	0.1093
300	0.00324	0.2157
340	0.00417	0.3656
380	0.00440	0.5397
420	0.00381	0.7064
460	0.00272	0.838
500	0.00159	0.9236
540	0.00077	0.9695
580	0.00030	0.9897
620	0.00010	0.9971
660	0.00003	0.9993

6. Построить график *плотности распределения.* Для этого выделить столбец со значениями f(x) вместе с заголовком, вызвать **Мастер диаграмм** и выбрать тип диаграммы – **График**. Для вывода значений **объема по оси X** необходимо во **втором окне** диалога мастера диаграмм нажать на вкладыш **Ряд** и в поле **Подписи по оси X** указать мышью значения объема без заголовка (**Рис. 3.1.**).

7. Построить аналогичным образом график функции распределения F(x) по данным таблицы 3.2. График дан на **Рис. 3.2.**

8. Построить совместный график плотности и функции распределения. Для этого нужно выделить значения f(x) и

F(x) вместе с заголовком, вызвать **Мастер диаграмм** и во вкладыше **Нестандартные** выбрать тип диаграммы – **График (2 оси)**. Как указано выше вывести значения объема на **оси X** (Рис. 3.3.).

Рис. 3.1. График плотности распределения

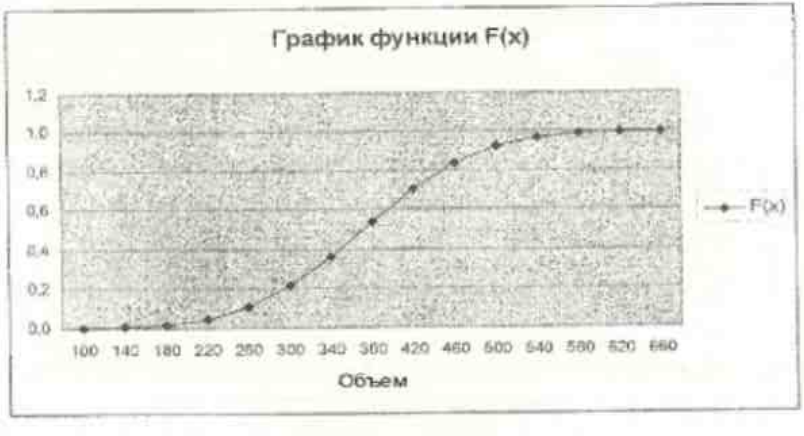

Рис. 3.2. График функции распределения

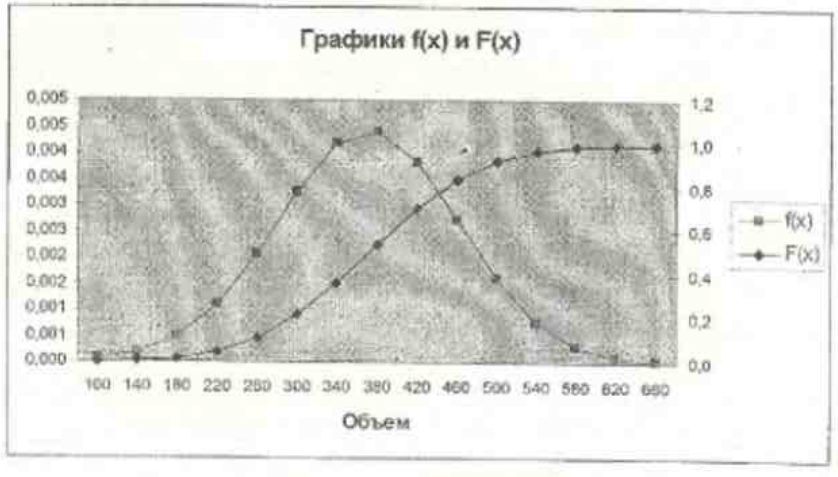

Рис. 3.3. Совместный график f(x) и F(x)

Примечание: для чисел по осям уменьшен размер шрифта и количество знаков после запятой. Для этого нужно вызвать контекстное меню около каждой оси и в строке **Формат оси** во вкладках **Число** и **Шрифт** установить нужные размеры.

3.5. Применение функции распределения и плотности распределения для решения задач

С помощью полученного графика можно провести следующий анализ:

- график плотности распределения f(x) показывает вероятность появления каждого значения случайной величины. **Например**, по графику видно, что наиболее вероятно появление значения объема равного 371, т.к. наиболее вероятным значением случайной величины является ее **математическое ожидание** (вероятность **равна** 1);

более вероятны появление значений объема в интервале от 250 до 500 (вероятность их появления от 0,4 до 1);

- график функции распределения F(x) служит для определения вероятности появления значения случайной величины. С его помощью можно решать две задачи:

1) *прямая задача* - какова вероятность того, что значение случайной величины X будет не менее заданного значения. Например, из графика видно, что вероятность того, что значение объема будет не менее 370 равна примерно 0,5;
2) *обратная задача* - чему должна быть равна случайная величина X, чтобы вероятность ее появления равнялась бы заданному значению. Например, по графику можно определить, что вероятность 0,8 справедлива для значений X<=450.

Точнее эти вычисления можно выполнить с помощью статистических функций *Excel* при решении следующих задач.

1. Прямая задача.

Определение *вероятности* появления заданной случайной величины. Для этого используется *функция вида*:
=НОРМРАСП(*x; среднее значение; стандартное отклонение; истина).*

2. Обратная задача.

Определение значений случайной величины при заданной *вероятности*. Для этого используется *функция вида*:
=НОРМОБР (*вероятность; среднее; стандартное отклонение),*
где *вероятность* - заданное число.

3. Вероятность попадания в интервал. Задача звучит так: *определить вероятность того, что значения X входят в заданный*

интервал от X_1 до X_2. Для решения используется следующая формула: $P(X_1<=X<=X_2) = F(X_2) - F(X_1)$.
Значения $F(X_1)$ и $F(X_2)$ определяются с помощью функции =НОРМРАСП(*x; среднее значение; стандартное отклонение; истина).*

Задание 1: Какова *вероятность* появления значений объема, равных: *400 250 320 370.*
Например, для значения объема *370* функция выглядит так: =НОРМРАСП (*370; 371; 90.24; истина*) и дает результат *0,496.*

Результаты расчета сведены в следующую таблицу.

Прямая задача	
Объем	Вероятность
400	0,626
250	0,089
320	0,286
370	0,496

Задание 2. Какое значение *объема* соответствует *вероятностям*: 0,34; 0,65; 0,15; 0,8.
Например, для вероятности *0,8* функция выглядит так: =НОРМОБР(0,8; 371; 90,24) и дает результат 447.

Вычисленные значения сведены в таблицу.

Обратная задача	
Вероятность	Объем
0,34	333,78
0,65	405,77
0,15	277,47
0,8	447

Задание 3. Определить *вероятности* попадания величин объема в заданный интервал.

Например, вероятность вхождения значения объема в интервал от **300** до **500** определяется так:

$$P(300 <= X <= 500) = F(500) - F(300) = 0,9 - 0,2 = 0,7.$$

Значения ***вероятностей*** *0,9* и *0,2* определены с помощью выражения *=НОРМРАСП(500, 371; 90,24; истина) – =НОРМРАСП(300, 371; 90,24; истина).*

Результаты расчета в таблице.

Вероятность вхождения в интервал		
Начало интервала	Конец интервала	Вероятность
300	500	0,7
130	250	0,1
440	580	0,2
260	610	0,9

3.6. Задания на самостоятельное выполнение

В каждом из заданий главы 1 необходимо:

- построить и проанализировать графики плотности распределения и функции распределения;
- применить функцию распределения и плотность распределения для решения задач по определению:
  1) *вероятности появления заданного значения случайной величины (прямая задача);*
  2) *значения случайной величины по заданной вероятности (обратная задача);*
  3) *вероятности попадания случайной величины в заданный интервал.*

ГЛАВА 4. КРИТЕРИИ ЗНАЧИМОСТИ И ПРОВЕРКА ГИПОТЕЗ

4.1. Введение

Часто на практике исследователи сталкиваются со случаями, когда неизвестен закон распределения изучаемого показателя или не известны параметры имеющегося закона распределения. Рассматриваемые методы применяются в тех случаях, когда предстоит проверить какие–то теоретические предположения о неизвестном законе распределения случайной величины или о значениях параметра, т.е.:

1. О *законе распределения* генеральной совокупности, из которой сделана выборка и для которой выдвинута гипотеза об эмпирической модели распределения, чаще всего нормальной формы.
2. О *параметрах* генеральной совокупности с известным распределением, чаще всего нормального распределения.

Статистической гипотезой называют гипотезу о виде неизвестного распределения или о параметрах известных распределений.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Поэтому *различают:*

1. Нулевая (основная) гипотеза – выдвинутая гипотеза H_0.
2. Конкурирующая (альтернативная) гипотеза – гипотеза H_1.

Например, *нулевая гипотеза*: математическое ожидание нормальной величины: $a=10$, а *конкурирующая гипотеза* - $a \neq 10$.

Коротко это записывают так: H_0: $a=10$; H_1: $a \neq 10$.

Гипотезы проверяют с помощью **статистического критерия** – некоторой случайной величины (*обозначим ее К*), для которой известен точно или приближенно закон распределения.

Каждый критерий разбивает все множество возможных значений на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается. Это подмножество считается **критической областью**. Другое подмножество содержит значения критерия, при которых нулевая гипотеза принимается. Это область **принятия гипотезы**.

Следовательно, основной *принцип проверки гипотезы такой*: если наблюдаемое значение критерия принадлежит критической области, то гипотеза отвергается; в противном случае – принимается.

Критическими точками (границами) $K_{кр}$ называются точки, отделяющие критическую область от области принятия гипотезы.

Правосторонней критической областью называется область, которая определена неравенством: $K > K_{кр}$, где $K_{кр} > 0$.

Левосторонней называют критическую область, для которой $K < K_{кр}$, где $K_{кр} < 0$.

Односторонней критической областью называют правостороннюю или левостороннюю области (*отклонения в одну сторону от критического значения*).

При **двухстороннем** критерии допускаются отклонения в обе стороны от критического значения.

4.2. Критерий значимости

Методы, которые для каждой выборки определяют принять или отвергнуть выдвинутую гипотезу, называют **критерием значимости**.

Процедура проверки гипотезы сводится к тому, что по выборочным данным вычисляется значение критерия (некоторой величины, имеющей известное стандартное распределение), а затем оно сравнивается с критическим значением

критерия, взятым из соответствующих таблиц. Если вычисленное значение критерия не превосходит критического значения, то нулевая гипотеза принимается на заданном уровне значимости α. В этом случае говорят: наблюдаемое различие *незначимо*, а в противном случае – *значимо*, т.е. вычисленное значение критерия превосходит критическое значение и нулевая гипотеза отвергается

Наблюдаемые значения критерия вычисляются по заданным формулам в зависимости от сравниваемых *параметров (средняя величина, дисперсия и т.п.)*. Уровень значимости обычно на практике равен одному из указанных значений: α=0,05; 0,01; 0,001.

Критерии значимости подразделяются на *два типа*:

1. *Критерии*, которые служат для проверки гипотез *о параметрах распределения* генеральной совокупности (чаще всего нормального распределения). Они называются *параметрическими* критериями.
2. *Критерии согласия* служат для проверки гипотез о согласии распределения генеральной совокупности, из которой сделана выборка, с принятой гипотезой о теоретической модели распределения (чаще всего о нормальном распределении).

Проверку параметрических гипотез можно применить во многих случаях (распределение считается *нормальным*):

- *сравнение выборочной и генеральной средней величин;*
- *сравнение средних величин двух независимых выборок;*
- *сравнение двух выборочных дисперсий;*
- *сравнение выборочных средних величин связанных выборок (до испытания; после испытания).*

В принципе алгоритм проверки гипотез во всех случаях примерно одинаков. Поэтому рассмотрим порядок выполняе-

мых действий для сравнения средних величин выборочной и генеральной совокупностей.

4.3. Алгоритм проверки гипотезы сравнения выборочной и генеральной средних величин

Рассматриваемый алгоритм подходит для малых выборок *(n<30)* и применяет *t – критерий Стьюдента*, который основан на предположении о нормальности распределения генеральной совокупности. Результаты справедливы и при небольших отклонениях от нормального распределения.

Предполагаем, что имеется выборка из нормально распределенной совокупности с параметрами: *a – математическое ожидание* (средняя величина) и *S – стандартное отклонение.*

Выдвигаемая *нулевая гипотеза* H_0: $a=a_0$, где a_0 – предполагаемое значение генеральной средней величины.

Альтернативная гипотеза H_1: $a \neq a_0$.

Так как при сравнении допускаются отклонения в обе стороны от a_0, т.е. $a<a_0$ или $a>a_0$, то применяется *двухсторонний* критерий.

Выполняемые действия:

1. *Формулировка выдвигаемых гипотез и выбор уровня значимости α.*
2. *Отбор малой выборки объема n (n<30).*
3. *Определение выборочной средней величины $X_{ср}$ и стандартного отклонения выборки S.*
4. *Определение наблюдаемого значения критерия по формуле:* $t_{набл} = \frac{|X_{ср} - a_0|}{S}\sqrt{n}$.

   Величина t имеет t – распределение Стьюдента с (n-1) степенями свободы.

5. *По таблице определяют критическое значение $t_{кр}$ при уровне значимости α и числе степеней свободы υ=n-1 для двухстороннего критерия.*
6. *Если $t_{набл} < t_{кр}$, то нулевая гипотеза принимается на заданном уровне значимости, т.е. различие между средними величинами не значимо. В противном случае гипотеза H_0 отвергается, т.е. различие значимо.*

Примечание: при больших объемах выборки (n>=30) *t – распределение Стьюдента* переходит в нормированное нормальное распределение. В этом случае применяют *U – критерий*, для которого наблюдаемое значение вычисляется по формуле: $U_{набл} = \frac{|X_{ср} - a_0|}{S}\sqrt{n}$.

Как видим, формула такая же, что и для вычисления $t_{набл}$, но критическое значение определяется из другой таблицы.

Например, при уровне значимости $\alpha=0,01$ критические значения равны: $t_{кр}=2,756$; $U_{кр}=2,58$ при $n=30$.

Разница не значительная, но *при n<30* это различие существенно, поэтому при малых выборках используют критерий Стьюдента.

Пример:

На контрольных испытаниях при $n=16$ было определено, что средний срок службы ламп *292 часа*. Предполагая, что срок службы ламп распределен нормально со стандартным отклонением *20 часов*, проверить на уровне значимости *0,1* нулевую гипотезу H_0: $a_0=300$ *ч* против конкурирующей гипотезы H_1: $a_0 \neq 300$ *ч*.

Решение:

Дано: $X_{ср}=292$; $S=20$; $\alpha=0,1$; $n=16$

1. Определим наблюдаемое значение критерия по формуле для малой выборки:

$$t_{набл} = \frac{|X_{ср} - a_0|}{S}\sqrt{n} = \frac{|292-300|}{20}\sqrt{16} = 1,6.$$

2. По таблице при уровне значимости $\alpha = 0{,}1$ (уровень достоверности $1-\alpha = 0{,}9$) определяем $t_{кр} = 1{,}75$.
3. Так как $t_{набл} < t_{кр}$, то нулевая гипотеза не отвергается. Следовательно, средняя величина генеральной совокупности равна **300 часов** при уровне достоверности **90%**.

4.4. Применение функций Excel при проверке гипотез

При проверке гипотез применяются следующие функции *Excel* из категории ***Статистические.***

1. Определение *критического значения* распределения Стьюдента при малых выборках:
   =СТЬЮДРАСПОБР (*вероятность; степень_свободы*),
   где *вероятность* – уровень значимости α (0,1; 0,01; 0,05*);*
   степень_свободы = n -1.
   В выше рассмотренном примере для определения критического значения указанная функция будет выглядеть так: = СТЬЮДРАСПОБР(0,1; 15) = 1,75.

2. Определение *критического значения* при n>30 для нормированного нормального распределения:
   =НОРМОБР (*вероятность; среднее; станд_отклонение*), *где*
   вероятность – уровень значимости α (*0,1; 0,01; 0,05;*)
   среднее – среднее значение выборки;
   станд_ отклонение – стандартное отклонение выборки.

3. *Функция с синтаксисом:*
   = СТЬЮДРАСП (*X; степ_ свободы; хвосты*), где
   X – случайная величина, значение которой проверяется на достоверность; *степ_ свободы* – число степеней свободы (n – 1*);* *хвосты* – число возвращаемых хвостов. Если *хвосты = 1*, то функция СТЬЮДРАСП возвращает одностороннее распределение. Если *хвосты = 2*, то данная функция возвращает двухсторонне распределение.

Рассматриваемая функция возвращает α - *вероятность* того, что значение величины *X не достоверно*, а (1- α) - вероятность достоверности величины *X*.

4.5. Лабораторная работа № 3. Проверка гипотез

Задание 1. Определить, значимо ли отличие значения выборочной средней от предполагаемого значения генеральной средней, т.е. является ли выборочная средняя величина надежной оценкой средней величины генеральной совокупности. Для решения задачи используем данные из *примера 2 главы 1*. В качестве генеральной совокупности используем значения объема *из таблицы 1.6.*

Таблица 4.1. Генеральная совокупность

№	X
1	100
2	140
3	180
4	220
5	260
6	300
7	340
8	380
9	420
10	460
11	500
12	540
13	580
14	620
15	660
Сумма	5700

Выполняемые действия:

1. Из *таблицы 1.3. главы 1* известно, что выборочная *средняя величина* a=371; *стандартное отклонение* S = 90,24; *объем выборки* n=10.

2. Определить генеральную среднюю величину по данным **таблицы 4.1.**: $X_{ген} = \frac{5700}{15} = 380$.

3. Выдвигаемые гипотезы:
   нулевая гипотеза H_0: $X_{ген} = 380$;
   альтернативная гипотеза H_1: $X_{ген} \neq 380$;

4. Определить *наблюдаемое значение критерия* - величины, которая подчиняется распределению Стьюдента, т.к. выборка является малой.

$$t_{набл} = \frac{|a - x_{сов}|}{S}\sqrt{n} = \frac{|371 - 380|}{90{,}24}\sqrt{10} = 0{,}315$$

5. Определить критическое значение $t_{кр}$ с помощью функции:

= СТЬЮДРАСПОБР(*вероятность; степень_свободы*), где
вероятность – уровень значимости – 0,05;
степень_свободы - (n – 1) = 9.

Тогда $t_{кр}$ = СТЬЮДРАСПОБР(*0,05; 9*)=2,26 (*Рис. 4.1.*).

Рис. 4.1. Вставка функции СТЬЮДРАСПОБР

6. Так как $t_{набл} < t_{кр}$, то нулевая гипотеза принимается на уровне значимости *0,05*, т.е. с вероятностью *0,95* можно принять, что разница между средними величинами выборки и генеральной совокупности не значима, и что выборочная средняя является надежной оценкой генеральной средней.

7. Доверительные границы для генеральной средней величины: $a - \Delta \le x_{ген} \le a + \Delta$, где $\Delta = \frac{t\,s}{\sqrt{n}}$ - стандартная ошибка средней величины при уровне значимости α.

Значение Δ определяется с помощью функции:

=ДОВЕРИТ (*α; станд - откл; размер*), где
α - уровень значимости, равный 0,05;
станд - откл - стандартное отклонение выборки;
размер - объем выборки.

В нашем случае функция выглядит так:

=ДОВЕРИТ (*0,05; 90,24; 10*) = 56

Следовательно, *доверительные границы* следующие:

$371 - 56 < X_{ген} < 371 + 56$ или $315 < X_{ген} < 427$

4.6. Проверка гипотезы о законе распределения

Допустим, случайная эмпирическая величина имеет неизвестный закон распределения. Проверяем гипотезу о том, что неизвестный закон распределения является *нормальным законом* или близок к нему.

Гипотеза H_0: неизвестный закон является нормальным.
Гипотеза H_1: неизвестный закон не является нормальным.

Алгоритм проверки гипотезы:

I. Для нормального закона распределения характерны следующие *особенности его параметров.*

1. *Значения средней величины, моды и медианы равны между собой, т.е.* $X_{ср} = M_o = M_e$.

2. *Значения коэффициентов эксцесса и асимметрии равны нулю, т.е.* $E_k = A_s = 0$.

Для *малых* выборок (*при n<30*) необходимо дополнительно вычислить и проанализировать *следующие величины:*

a) *несмещенные оценки* для коэффициентов *асимметрии и эксцесса:*

$$G_1 = \frac{\sqrt{n-1}}{n-2} A_s;\quad G_2 = \frac{n-1}{(n-2)(n-3)}((n+1)E_k + 6);$$

b) *среднеквадратические отклонения* для вычисленных оценок:

$$S_{G_1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}};\quad S_{G_2} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}};$$

c) проверяются следующие *условия:*

$$|G_1| <= 3\,S_{G_1};\quad |G_2| <= 5\,S_{G_2}.$$

Если *условия выполняются,* то предполагаемый закон случайной величины близок к *нормальному закону* распределения.

II. Для *элементов* выборки при нормальном законе распределения выполняются следующие *условия:*

1. Отклонения всех значений элементов выборки от значения выборочной средней величины должны быть меньше $\pm 3\sigma$ (99,7 % отклонений).

2. Примерно 2/3 всех отклонений (68,3 %) от среднего значения должны быть меньше $\pm \sigma$.

3. Половина всех отклонений от среднего значения должны быть меньше $\pm 0{,}657\sigma$.

Задание 2.

Проверить гипотезу о том, что выборочные данные *объема поступления* продукции за *10* дней имеют нормальный закон распределения (*пример* рассмотрен в первой главе, где подсчитаны *параметры описательной статистики).*

Таблица 4.2. Исходные данные

Дни	1	2	3	4	5	6	7	8	9	10
Объем	300	280	400	350	530	350	480	250	330	440

Выполняемые действия:

I. Анализ значений *параметров описательной статистики:*

1. Средняя величина, мода и медиана имеют следующие значения: $X_{ср} = 371;\quad M_o = 350;\quad M_e = 350.$
Так как значения этих величин близки между собой, то можно предположить, что закон распределения рассматриваемого ряда близок к нормальному закону.

2. Величина коэффициента эксцесса $E_k = -0,67$ - приближается *к нулю*, причем *отрицательное* значение указывает на *пологость кривой* распределения. Величина коэффициента асимметрии $A_s = 0,51$ также приближается к *нулю*. *Положительное* значение указывает на *правостороннюю асимметрию*.

Факт близости к *нулю* значений этих коэффициентов дает право предположить о *нормальности закона* распределения рассматриваемого ряда.

Так как мы имеем *малую выборку (n=10)*, то необходимо вычислить и проанализировать следующие *величины*:

a) *несмещенные оценки* асимметрии и эксцесса:

$$G_1 = \frac{\sqrt{10-1}}{10-2} * 0,51 = 0,19125;$$

$$G_2 = \frac{10-1}{(10-2)(10-3)}((10+1)*(-0,67)+6) = 0,22;$$

b) *среднеквадратические отклонения оценок:*

$$S_{G_1} = \sqrt{\frac{6*10*(10-1)}{(10-2)(10+1)(10+3)}} = 0,687;$$

$$S_{G_2} = \sqrt{\frac{24*10*(10-1)^2}{(10-3)(10-2)(10+3)(10+5)}} = 1,33;$$

c) *проверяем условие:*

$|0,19125| <= 3*0,687$ или $0,19125 <= 2,06$;

$|0,22| <= 5*1,33$ или $0,22 <= 6,65$.

Так как условия выполняются, то закон распределения исследуемого ряда можно считать *нормальным*.

II. Анализ *отклонений* значений выборки от средней величины:

1. Определяем отклонения значений выборки от средней величины (*таблица 4,3).*

Таблица 4.3. Исходные данные и отклонения

№	X	ABS(X-Xcp)	Сравнение с 3δ	Сравнение с δ	Сравнение с 0,657δ
1	300	71	ИСТИНА	ИСТИНА	ЛОЖЬ
2	280	91	ИСТИНА	ЛОЖЬ	ЛОЖЬ
3	400	29	ИСТИНА	ИСТИНА	ИСТИНА
4	350	21	ИСТИНА	ИСТИНА	ИСТИНА
5	530	159	ИСТИНА	ЛОЖЬ	ЛОЖЬ
6	350	21	ИСТИНА	ИСТИНА	ИСТИНА
7	480	109	ИСТИНА	ЛОЖЬ	ЛОЖЬ
8	250	121	ИСТИНА	ЛОЖЬ	ЛОЖЬ
9	330	41	ИСТИНА	ИСТИНА	ИСТИНА
10	440	69	ИСТИНА	ИСТИНА	ЛОЖЬ
Количество:			**10**	**6**	**4**
%			**100%**	**60%**	**40%**

2. Для анализа необходимо вычислить значения 3σ, $0,657\sigma$, где значение $\sigma = 90,24$; тогда $3\sigma = 270,72$; $0,657\sigma = 59,3$.

Сравнение вычисленных значений и значений отклонений из *таблицы 4.3.* приводит к следующим выводам:

a) *все отклонения меньше* $\pm 3\sigma$;

b) *шесть значений отклонений (60%) меньше* $\pm \sigma$;

c) *четыре значения отклонений (40%) меньше* $\pm 0,657\sigma$.

Следовательно, можно приближенно принять для анализируемого ряда *нормальный закон распределения.*

Примечания:

1. Столбцы для сравнений получены вставкой логической функции ЕСЛИ, в которой первым аргументом является сравнение разности с соответствующим значением, *например* **ABS(X-Xср) < 3σ** и т.д. Второй и третий аргументы соответственно равны значениям **«истина»** и «ложь».

2. Количество подсчитано с помощью статистической функции СЧЕТЕСЛИ с условием **«истина»**.

4.7. Задания для самостоятельного выполнения

1. Проверить гипотезу о значимости генеральной средней для заданий главы 1.

2. Проверить гипотезу о законе распределения данных заданий главы 1.

Глава 5. Статистический контроль качества продукции

5.1. Понятие контроля качества продукции

Контроль качества продукции – это система мероприятий, обеспечивающих экономичное производство товаров и услуг, качество которых соответствует требованиям потребителей.

Статистический контроль – это контроль с применением статистических методов на основе выборки. Существуют две основные задачи статистического контроля.

1. *Статистическое регулирование качества продукции.*
2. *Статистический приемочный контроль.*

Первая задача позволяет с помощью регулярных отборов небольших по объему проб предупреждать появление брака в процессе производства.

Вторая задача служит для определения доли брака в уже изготовленной и представленной к сдаче продукции.

5.2. Статистическое регулирование качества продукции Контрольные диаграммы

Методы статистическое регулирование качества продукции служат для текущего предупредительного контроля качества продукции в процессе производства. Часто применяемым методом является *метод контрольных диаграмм.* Существует несколько типов контрольных диаграмм: карта средних величин; карта медиан; карта стандартных отклонений; комбинированные карты и т.д.

Для контроля на такую карту наносятся следующие линии:

1. *Контрольные линии* – средняя величина; медиана; максимум; минимум и т.п.
2. *Контролируемые значения* показателя качества, которые при отсутствии брака не должны выходить за пределы контрольных линий.

5.2.1. Контрольная диаграмма средних величин

На контрольную диаграмму средних величин наносятся в виде линий значения следующих величин.

1. **Верхняя и нижняя границы допуска.**

Для определения этих величин необходимо построить доверительный интервал для средней величины с вероятностью 0,997 по следующим формулам:

$$X_{max} = \overline{X} + 3\frac{\sigma}{\sqrt{n}}; X_{min} = \overline{X} - 3\frac{\sigma}{\sqrt{n}}$$

2. Предупреждающие границы допуска.

Для этого нужно определить доверительный интервал для средней величины с вероятностью 0,954:

$$A_1 = \overline{X} + 2\frac{\sigma}{\sqrt{n}}; A_2 = \overline{X} - 2\frac{\sigma}{\sqrt{n}}$$

Таким образом построенные линии являются предупреждающими границами и называются *«статистическое сито»*.

3. Затем на диаграмму для контроля наносятся значения выборочных средних величин (средние величины проб). Ес-

ли эти значения не выходят за пределы A_1 и A_2, то система работает устойчиво. Если же какие – то значения вышли за пределы A_1 или A_2, то в системе произошел сбой и может возникнуть брак, следовательно, нужно принять меры по наладке системы. Если же значения контролируемого признака вышли за пределы X_{max} или X_{min}, то уже продукция является бракованной.

Пример. Малое предприятие по выпуску полуфабрикатов производит тефтели в упаковке общим весом 400 грамм. Исследования показали, что допускается стандартное отклонение в 0,17 грамм на упаковку. В процессе производства через каждый час в течение *семичасового* рабочего дня осуществляется контрольный замер из 5 упаковок (выборка из 5 упаковок). Результаты замера даны в следующей таблице:

Таблица 5.1. Результаты замера

Номер выборки	1	2	3	4	5	6	7
Вес упаковки	400,1	400	400,2	399,8	399,9	400	400,1
	400,1	399,7	399,6	400,1	400	400,2	400,2
	399,8	399,7	400	400,2	400,3	400	399,9
	400,2	400,1	400	399,9	399,7	399,8	400
	399,8	399,9	398,9	399,6	399,8	399,8	399,9
Среднее значение	400	399,88	399,74	399,92	399,94	399,6	400,02

Построить контрольную диаграмму протекания процесса. Оценить качество упаковки продукции.

Решение задачи.

1. Средний вес упаковки – $\overline{X} = 400$ г.
   Стандартное отклонение - $\sigma = 0,17$ г.

2. Определяем *верхнюю* и *нижнюю* границы допуска веса с вероятностью 0,997 по формулам:

$$X_{max} = \overline{X} + 3\frac{\sigma}{\sqrt{n}}; \qquad X_{min} = \overline{X} - 3\frac{\sigma}{\sqrt{n}}$$

Предельную ошибку допуска $3\frac{\sigma}{\sqrt{n}}$ вычисляем с помощью стандартной функции **ДОВЕРИТ (0,003; 0,17; 5) = 0,23.**

Тогда верхняя граница допуска X_{max}= 400 + 0,23 = 400,23, а нижняя граница допуска X_{min} = 400 – 0,23 = 399, 77.

3. *Предупреждающие границы* («статистическое сито») определяются с вероятностью 0,954 по формулам:

$$A_1 = \overline{X} + 2\frac{\sigma}{\sqrt{n}}; \quad A_2 = \overline{X} - 2\frac{\sigma}{\sqrt{n}}.$$

Предельную ошибку допуска $2\frac{\sigma}{\sqrt{n}}$ вычисляем с помощью стандартной функции **ДОВЕРИТ (0,046; 0,17; 5) = 0,15.**

Тогда A_1 = 400 + 0,15 = 400,15; A_2 = 400 – 0,15 = 399,85.

4. Строим контрольную диаграмму с помощью **Мастера диаграмм**. Тип диаграммы **График**. Для этого необходимо нанести на диаграмму линии их значений верхней и нижней границ допуска, а также значения предупреждающих границ.
Значения средних величин каждой выборки заносятся в поле диаграммы.

Для построения диаграммы исходные данные представим в виде следующей таблицы:

Таблица 5.2. Исходные данные

№	X_{max}	X_{min}	A_1	A_2	$X_{ср}$	$X_{срвыб}$
1	400,23	399,77	400,15	399,85	400	400
2	400,23	399,77	400,15	399,85	400	399,88
3	400,23	399,77	400,15	399,85	400	399,74
4	400,23	399,77	400,15	399,85	400	399,92
5	400,23	399,77	400,15	399,85	400	399,94
6	400,23	399,77	400,15	399,85	400	399,96
7	400,23	399,77	400,15	399,85	400	400,2

Для построения диаграммы необходимо выделить всю таблицу с заголовками кроме столбца с порядковыми номерами, вызвать **Мастер диаграмм** и выбрать тип **График**. Далее с помощью вкладыша **Ряд** расположить под осью X значения выборочных средних величин и задать заголовки. Построенная контрольная диаграмма имеет вид:

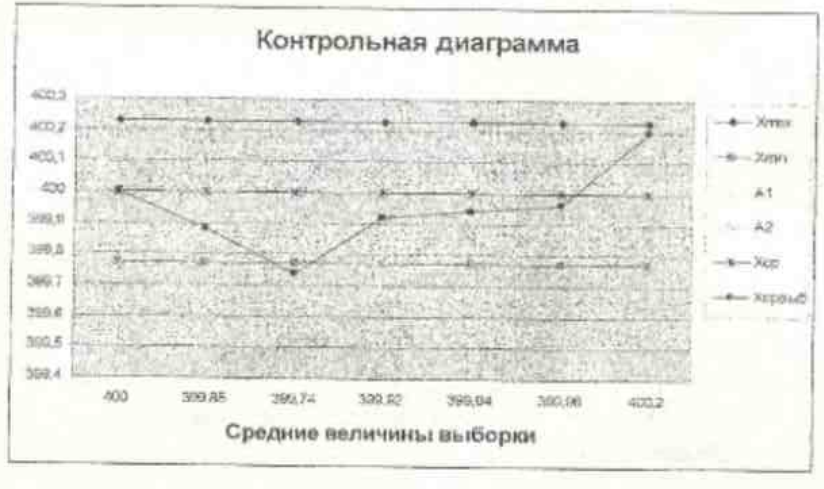

Рис. 5.1. Контрольная диаграмма средних величин.

Анализ построенной контрольной диаграммы дает основание сделать вывод о том, что качество расфасовки в целом устойчивое. Но появление средних величин, выходящих за пределы контрольных линий в третий и седьмой часы работы говорят о том, что процесс может иметь сбой и нужно принять соответствующие меры.

5.2.2. Контрольная диаграмма медиан

Контролируемыми точками на такой диаграмме являются значения **медиан**, а контрольными линиями – следующие границы.

1. Статистические границы, которые вычисляются с вероятности **0,999** по формулам:

$$A_1 = \overline{X} + 3{,}5\frac{\sigma}{\sqrt{n}}; \quad A_2 = \overline{X} - 3{,}5\frac{\sigma}{\sqrt{n}}.$$

2. Линии технического допуска, вычисляемые по формулам:

$$B_1 = \overline{X} + 2{,}5\sigma; \quad B_2 = \overline{X} - 2{,}5\sigma.$$

3. Максимальное и минимальное значения контролируемой величины согласно ГОСТ .

Для построения рассматриваемой диаграммы необходимо вычислить выше данные значения границ и медиан, занести в такую же таблицу, как и в предыдущем случае и построить диаграмму в виде графика.

Результаты диаграммы анализируются так же, как и в случае средних величин.

СПИСОК ЛИТЕРАТУРЫ

1. Гмурман В.Е. Теория вероятностей и математическая статистика. - М.: Высшая школа, 1972.
2. Основы математической статистики/Под ред. В.С. Иванова.- М.: 1990.
3. Теория статистики с основами теории вероятностей/Под ред. И.И. Елисеевой. – М.: ИНИТИ – ДИАН, 2001.
4. Кремер Н.Ш. Теория вероятностей и математическая статистика. – М.: ИНИТИ – ДИАН, 2002.
5. Гранберг Д. Статистическое моделирование и прогнозирование. – М.: Финансы и статистика, 1990.
6. Четыркин Е.М. Статистические методы прогнозирования. – М.: Статистика, 1977.
7. Основы экономического и социального прогнозирования. – М.: Высшая школа, 1995.
8. Бордоева А.Е. Табличный процессор EXCEL.- Улан – Удэ: ВСГТУ, 2002.
9. Орехова Р.А. Моделирование экономических процессов. – Улан – Удэ: ВСГТУ, 2000.

ОГЛАВЛЕНИЕ

Бордоева Анна Евдокимовна

Описательная статистика и проверка гипотез средствами EXCEL

Учебно – методическое пособие

Подписано в печать 13.05.2009 г. Формат 60x84 1/16.
Усл.п.л. 3,49. Тираж 100 экз. Заказ № 178.

Издательство ВСГТУ.
670013, г. Улан – Удэ, ул. Ключевская, 40, в.