Монахова А.А.

СТАТИСТИКА

краткий курс лекций
для студентов заочного отделения
Смольного института

Санкт-Петербург
2011

СОДЕРЖАНИЕ

Определение, предмет, методы и задачи статистики как науки ... 3
Основные понятия статистики ... 4
Этапы статистического исследования ... 5
статистические данные и статистическое наблюдение ... 5
точность наблюдения ... 7
этапы статистического исследования ... 8
Выборочный метод ... 9
Описательный анализ: Построение рядов распределения ... 10
атрибутивные ряды распределения ... 10
вариационные ряды распределения ... 10
дискретный вариационный ряд ... 11
интервальный вариационный ряд ... 11
Описательный анализ: Описательные статистики ... 13
средняя арифметическая ... 13
медиана ... 14
квартили, квартильный размах ... 15
стандартное отклонение ... 16
дисперсия ... 17
Описательный анализ: Нормальное распределение ... 18
проверка на нормальность ... 19
Схема выбора адекватных описательных статистик ... 23
Анализ влияния факторов ... 24
коэффициенты корреляции ... 24
коэффициент корреляции Пирсона ... 25
коэффициент корреляции Спирмена ... 28

ОПРЕДЕЛЕНИЕ, ПРЕДМЕТ, МЕТОДЫ И ЗАДАЧИ СТАТИСТИКИ КАК НАУКИ

Термин «статистика» происходит от латинского *status* – государство. Термин употребляется в нескольких значениях:

1) как отрасль общественных наук;
2) как совокупность данных о каком-либо явлении и процессе;
3) как некоторая обобщающая характеристика явления.

Принято считать, что термин «*статистика*» введен в науку немецким профессором физиологии и права Готфридом Ахенвалем (1719-1772 гг.) в 1746 г., когда он начал читать студентам новую учебную дисциплину, которую и назвал статистикой.

Статистика – *это общественная наука, которая изучает количественную сторону качественно определенных массовых социально-экономических явлений и процессов, их структуру и распределение, размещение в пространстве, движение во времени, выявляя действующие количественные зависимости, тенденции и закономерности, причем в конкретных условиях места и времени***.**

Например, публикуемые сведения о числе родившихся или умерших, данные о числе браков и разводов и т.д. – все это количественно определенные показатели различного качества массовых социально-экономических процессов. В этом смысле статистика знает все.

Но когда сообщается, что месячная норма потребления табачных изделий на душу населения составляет 8 пачек сигарет, или, что в месяц на душу населения в РФ приходится 10 литров выпитых спиртных напитков, то многие, примерив эти данные на себя, скажут, что приводимые факты далеки от истины.

Основное предназначение статистики – осуществлять научное прогнозирование и способствовать принятию правильных решений на основе исследования тенденций социально-экономического развития общества.

Предмет статистики изучение количественной стороны *массовых* общественных явлений в конкретных условиях *места* и *времени*. К отдельно взятому субъекту или явлению вне определенного места и времени ее выводы могут быть неприемлемыми. Следует твердо помнить, что статистические показатели имеют свои строго определенные условия и границы применения.

Изучение количественной стороны экономических явлений должно обязательно сопровождаться учетом их качественных особенностей.

Под ***методами статистики*** понимается система приемов, способов и правил, направленных на изучение количественных закономерностей, проявляющихся в структуре, динамике и взаимосвязях социально-экономических явлений.

Статистика связана со следующими областями:

- ✓ *экономической теорией*;
- ✓ *оперативно-техническим и бухгалтерским учетом*. Оперативно-технический учет фиксирует на предприятиях отдельные технико-экономические факты, например, простои оборудования по той или иной причине. Бухгалтерский учет характеризует кругооборот средств предприятия, источники их образования и направления использования.
- ✓ *математикой, теорией вероятностей*

Задачи статистики как науки:

I. Разработка системы показателей процессов общественного развития и методов их измерения:
   1) создание способов расчета и взаимной увязки показателей динамики различных сторон социально-экономических процессов;
   2) выявление тенденций развития процессов;
   3) разработка прогнозов.

II. Разработка методов расчета и сравнительного анализа экономики и социального развития различных стран, международных экономических связей.

ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

Каждая научная область имеет ряд основных понятий. В базовый аппарат статистики входят следующие термины: совокупность, вариация, признак, закономерность.

Статистическая совокупность *– это множество единиц, обладающих массовостью, однородностью, определенной целостностью, взаимозависимостью состояния отдельных единиц и наличием вариации.* Статистическая совокупность является объектом исследования статистики как науки.

Например, в качестве объектов статистического исследования, т. е. статистических совокупностей, может выступать множество коммерческих банков, зарегистрированных на территории РФ, множество акционерных обществ, множество граждан какой-либо страны и т. д.

Статистическая совокупность состоит из реально существующих материальных объектов. Может сложиться представление, что статистическая совокупность обязательно должна состоять из большего количества единиц, но в зависимости от задачи исследования ее численность может быть достаточно малой и составлять иногда 8-10 единиц.

Каждый отдельно взятый элемент статистической совокупности называется *единицей статистической совокупности*. Для того, чтобы множество единиц образовало статистическую совокупность они должны обладать общими свойствами, которые в статистике именуются признаками.

Признак *– общее свойство, характерная черта или иная особенность единиц совокупности, которые могут быть наблюдаемы или измерены.*

<u>По форме внешнего выражения</u> признаки делятся на:

- ✓ качественные (атрибутивные)
  - **номинальные** – переменные, которые не могут быть описаны количественно. Часто значениям номинальной переменной присваиваются коды (переменная категоризируется). Например, пол (1 – М, 2 – Ж), номер отдела (1 – отдел маркетинга, 2 – отдел сбыта, 3 – отдел рекламы) и т.п. Причем не важно, какому значению какой код будет присвоен: 1 – М, 2 – Ж или 1 – Ж, 2 – М. Также значения переменной могут представлять собой наименования (т. е. являются не категоризированными). Например, ФИО, код товара, номер по порядку и т.п. Для номинальных признаков не имеет смысла искать среднее.
  - **порядковые** – переменные, значения которых представляют собой некий рейтинг, имеют разные смысловые градации. Например, отношение к товару (1 – не покупаю, 2 – редко покупаю, 3 – чаще покупаю, чем нет, 4 – часто покупаю, 5 – постоянно покупаю), уровень образования (1 – без в/о, 2 - неполное высшее, 3 – высшее), возрастная группа (1 – до 20 лет, 2 – от 21 года до 30 лет, 3 – от 31 года до 40 лет и т.д.) и т.п. Т.е. значения переменной упорядочены, и между уровнями шкалы есть качественные неоднородные скачки.
- ✓ **количественные**

  В такой шкале представляются переменные, которые могут быть описаны количественно, т.е. измерены, взвешены, посчитаны. Значения таких переменных либо имеют определенные единицы измерения (натуральные, стоимостные, трудовые), либо представлены в виде процентов или коэффициентов.

  Например, возраст, заработная плата, объем продаж, количество человек, человеко-часы, производительность труда и т.п.

  Для количественных признаков всегда имеет смысл находить количественные характеристики (сумма, среднее и т.п.).

Другие классификации признаков представлены в Табл. 1.

Табл. 1. *Классификация признаков*

Классификация по:	Виды признаков
содержательности	существенные, несущественные, первичные, вторичные
принадлежности	индивидуальные, общие
направлению	прямые, косвенные
причинности	причины (факторные), следствия (результативные)
управляемости	управляемые, неуправляемые
степени детерминированности	детерминированные, стохастические
наблюдаемости	наблюдаемые, ненаблюдаемые
измеряемости	непосредственно измеряемые, косвенно измеряемые, условно измеряемые
времени	статические, динамические, периодические

Единицы совокупности наряду с общими для всех единиц признаками, обусловливающими качественную определенность совокупности, также обладают индивидуальными особенностями и различиями, отличающими их друг от друга, т. е. существует *вариация признаков*. Она обусловлена различным сочетанием условий, которые определяют развитие единиц совокупности.

Вариация *– многообразие, изменяемость значения признака у отдельных единиц статистической совокупности.*

Вариация признака может быть:

- ✓ прерывной (дискретной)

  Признак имеет только вполне определенные значения, между которыми не может быть промежуточных. Например, количество детей в семье.

- ✓ непрерывной

  Признак имеет любые значения с определенной степенью точности. Например, рост, вес, возраст.

Статистическая закономерность *– это форма проявления причинной связи, выражающаяся в последовательности, регулярности, повторяемости событий с достаточно высокой степенью вероятности, если причины (условия), порождающие события, не изменяются или изменяются незначительно*.

Статистическая закономерность – форма закономерной связи явлений, при которой вытекающее из нее предсказание носит вероятностный характер.

Статистическая закономерность предопределяет типичное распределение единиц статистической совокупности на определенный момент времени под воздействием всех факторов. Статистическая закономерность, не определяя положение каждого случая, устанавливает общее распределение в данных условиях времени и места.

ЭТАПЫ СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

СТАТИСТИЧЕСКИЕ ДАННЫЕ И СТАТИСТИЧЕСКОЕ НАБЛЮДЕНИЕ

Статистическое наблюдение *– это массовый, планомерный, научно организованный сбор статистических данных о социально-экономических процессах.*

Статистические данные *– это совокупность количественных характеристик социально-экономических явлений и процессов, полученных в результате статистического наблюдения, их обработки или соответствующих расчетов.*

Основными свойствами статистических данных являются массовость и стабильность. Первое свойство диктуется самим определением статистики как науки, а второе – говорит о том, что однажды собранная информация остается неизменной, что свидетельствует о ее способности к устареванию.

Классификация статистического наблюдения

1. По форме.
   1. *Отчетность*

      Это такая форма статистического наблюдения, при которой предприятия в определенные сроки в установленном виде представляют в статистические органы необходимые данные, скрепленные подписями лиц, ответственных за достоверность сообщаемых сведений.
   2. *Специально организованное статистическое наблюдение*

      Проводится в форме переписей и всякого рода обследований.
2. По виду с учетом полноты охвата единиц совокупности.
   1. *Сплошное*

      Это статистическое наблюдение, при котором обследованию подлежат все без исключения единицы совокупности.
   2. *Несплошное*

      Это статистическое наблюдение, при котором обследованию подлежат не все единицы совокупности.
      1) *наблюдение основного массива*

         Это статистическое наблюдение, при котором обследованию подвергается часть совокупности с преобладающей долей изучаемого признака.
      2) *выборочное*

         Это статистическое наблюдение, при котором обследованию подвергается некоторая часть совокупности, отобранная особым образом, а результаты обследования распространяются на всю совокупность.
      3) *монографическое*

         Это статистическое наблюдение с целью всестороннего обследования единицы типичного общественного явления.
3. По виду фактов по времени.
   1. *Непрерывное* (*текущее*, *постоянное*)

      Это статистическое наблюдение, при котором установление и обследование фактов происходит по мере их возникновения.
   2. *Прерывное* (*дискретное*)

      Это статистическое наблюдение, при котором установление и обследование фактов производится либо регулярно через определенные промежутки времени (*периодическое* статистическое наблюдение), либо по мере надобности (*единовременное* статистическое наблюдение).
4. По способу учета факторов.
   1. *Непосредственный учет*

      Это статистическое наблюдение, при котором необходимые сведения получают путем подсчета, измерения и взвешивания единиц совокупности.
   2. *Документальный учет*

      Это статистическое наблюдение, при котором все необходимые сведения получают на основе различной документации.
   3. *Опрос*

      Это статистическое наблюдение, при котором необходимые сведения регистраторы (работники статистических органов) получают у опрашиваемого.
      1) *Экспедиционный опрос*

         Регистраторы сами заполняют опросный бланк со слов опрашиваемого.

2) *Корреспондентский опрос*

   Опрашиваемый самостоятельно заполняет опросный бланк и возвращает его на добровольных началах.

3) *Саморегистрация*

   Опросные бланки заполняет сам опрашиваемый, а регистратор только проводит инструктаж, контроль правильности заполнения бланков при их получении.

ТОЧНОСТЬ НАБЛЮДЕНИЯ

Точность статистического наблюдения *– это степень соответствия величины какого-либо показателя (значение какого-либо признака), определенной по материалам статистического наблюдения, действительной его величине.*

Ошибка наблюдения – это расхождение между расчетным и действительным значениями изучаемых величин.

Точность данных – это основное требование, предъявляемое к статистическому наблюдению. Чтобы избежать ошибок наблюдения, предупредить, выявить и исправить их, необходимо:

- ✓ обеспечить качественное обучение персонала, который будет проводить наблюдение;
- ✓ организовать специальные частичные или сплошные контрольные проверки правильности заполнения статистических формуляров. *Статистический формуляр* – это документ единого образца, содержащий программу и результаты наблюдения.
- ✓ провести логический и арифметический контроль полученных данных после окончания сбора информации.

<u>Типы ошибок</u>:

- ➢ ***Случайные ошибки*** *- это результат действия различных случайных факторов* (например, цифры переставлены местами, перепутаны соседние строки или графы при заполнении статистического формуляра). Такие ошибки имеют разную направленность: они могут и повышать, и понижать значения показателей. При достаточно большой обследуемой совокупности эти ошибки взаимно погашаются.
- ➢ ***Систематические ошибки*** регистрации всегда *имеют одинаковую тенденцию либо к увеличению, либо к уменьшению значения показателей по каждой единице наблюдения, и поэтому величина показателя по совокупности в целом будет включать в себя накопленную ошибку*. Примером статистической ошибки регистрации при проведении социологических опросов населения может служить округление возраста населения, как правило, на цифрах, оканчивающихся на 5 и 0. Многие опрашиваемые, например, вместо 48-49 и 51-52 лет говорят, что им 50 лет.

Следующим этапом контроля точности информации является *арифметический контроль*. Он основывается на использовании количественных связей между значениями различных показателей.

Например, если среди собранных данных имеются сведения о численности промышленно-производственного персонала, выработке товарной продукции в среднем на одного работающего и стоимости товарной продукции, то произведение первых двух показателей должно дать значение третьего показателя.

Если арифметический контроль покажет, что данная зависимость не выполняется, это будет свидетельствовать о недостоверности собранных данных. Поэтому в программу статистического наблюдения целесообразно включать показатели, которые дают возможность провести арифметический контроль.

Логический контроль, так же как и арифметический, основывается на знании взаимосвязей между показателями, но не количественных, а логических. Например, человек в возрасте 6 лет не может иметь высшего образования. Поэтому если в бланке переписи имеются одновременно обе записи, то это показывает, что одна из них не соответствует действительности.

Обычно для исправления ошибок, выявленных в ходе логического контроля, требуется повторно обратиться к источнику сведений.

ЭТАПЫ СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

I. Статистическое наблюдение

На этом этапе:

1. Формируются цели и задачи наблюдения.
2. Определяются объект (совокупность общественных явлений или процессов) и единица наблюдения.
3. Разрабатываются программы исследования.
4. Определяются способы и методы исследования.
5. Составляется организационный план проведения исследования.

В результате статистического наблюдения получают первичный статистический материал.

II. Первичная обработка, сводка и группировка результатов наблюдения

Для получения обобщающих характеристик исследуемого процесса или явления первичный статистический материал необходимо специальным образом обработать. Для этого применяют статистическую сводку и группировку.

Статистическая сводка *– это научное обобщение первичного статистического материала с помощью итоговых подсчетов, выполняемых по определенной системе*. Сводка может осуществляться либо ручным, либо машинным способом.

Статистическая группировка *– это разделение единиц изучаемого общественного явления на однородные группы по существенным для него признакам.*

В зависимости от числа и вида признаков, решаемых задач и исходных данных группировки подразделяются на:

- *простые* (группы образованы по какому-либо одному признаку) и *комбинационные* (группы образуются на основе двух или более признаков, взятых в сочетании друг с другом);
- *группировка на основе качественных признаков*, которые характеризуют качество явления и не имеют количественного выражения (число групп определяется характером признака), и *группировка на основе количественных признаков*, вариация которых выражается числами (число групп определяется конкретными задачами того или иного исследования);
- *типологические* (выделение социально-экономических типов явлений), *структурные* (определение структуры однородных совокупностей) и *аналитические* (раскрытие связей и зависимостей между отдельными признаками общественных явлений);
- *многомерные* (цель – разделение совокупности на качественно однородные группы по большому числу признаков одновременно и определение на их основе связи и влияния факторных признаков на отклик);
- *вторичная группировка*. Если группировка первичного статистического материала не удовлетворяет целям исследования, то прибегают к вторичной группировке. Новых группы можно образовывать либо изменением интервалов первичной группировки, либо выделением определенной доли единиц совокупности.

Результаты статистической группировки и сводки, как привило, излагаются в виде *статистических таблиц* и *рядов распределения* (будут рассмотрены ниже).

III. Анализ полученных сводных материалов.

Выделяют следующие основные этапы анализа:

1. Констатация фактов и их оценка.
2. Установление характерных черт и причин явления.
3. Сопоставление явления с другими, принятыми за базу сравнения – нормативными, плановыми и прочими явлениями.
4. Формулирование гипотез, выводов и предположений.
5. Статистическая проверка выдвинутых гипотез.

Характерным для статистических методов на этой стадии является применение обобщающих показателей: абсолютных, относительных, средних величин и индексных систем. Некоторые общие черты формирования обобщающих показателей устанавливаются посредством измерения их вариации.

Закономерности причинно-следственных связей общественных процессов и явлений устанавливаются с помощью корреляционно-регрессионного анализа, а также методов многомерного статистического анализа. Взаимосвязи явлений также изучаются с помощью статистических группировок, параллельных рядов, взаимосвязанных индексов и т. д.

Широкое применение в статистике находят графические методы, позволяющие в наглядной форме представлять результаты статистических исследований.

Все эти этапы связаны между собой, отсутствие одного из них ведет к разрыву целостности статистического исследования.

ВЫБОРОЧНЫЙ МЕТОД

Как правило, в статистике используется выборочное статистическое наблюдение, т. е. обследованию подвергается некоторая часть статистической совокупности, отобранная особым образом, а результаты обследования распространяются на всю совокупность. Данные, полученные в результате такого наблюдения, называются *выборочной совокупностью* или *выборкой*.

Способы формирования выборочной совокупности представлены на Рис. 1.

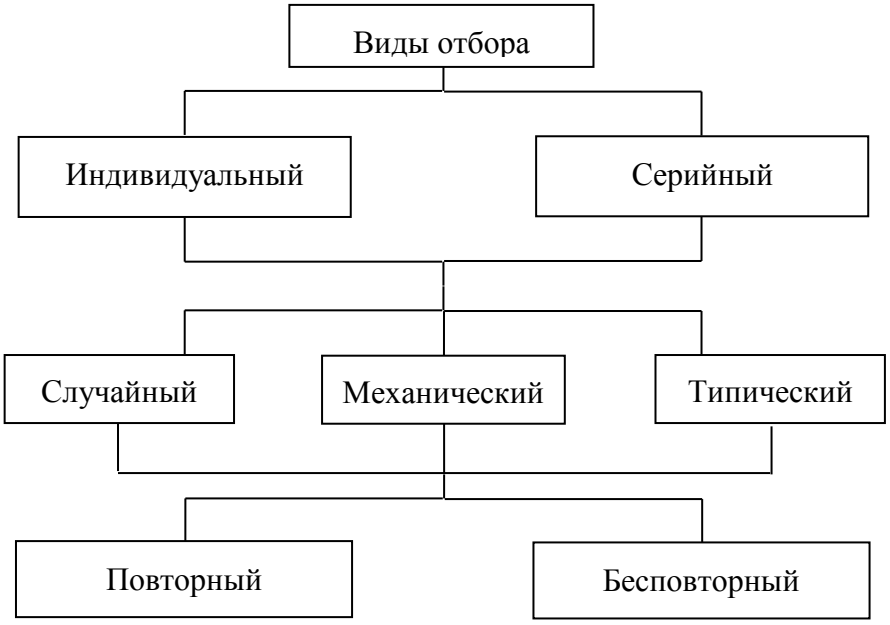


Рис. 1. *Способы формирования выборочной совокупности*

Выборка должна быть репрезентативной, т. е. достаточной по объему и представительной по изучаемому признаку.

Объем выборки (N) – это количество значений, полученных в результате выборочного статистического наблюдения.

Выборка обычно записывается в виде таблицы, в ячейках которой – ***варианты*** (результаты наблюдений, другими словами, конкретные значения признака).

ОПИСАТЕЛЬНЫЙ АНАЛИЗ: ПОСТРОЕНИЕ РЯДОВ РАСПРЕДЕЛЕНИЯ

Ряд распределения *– это упорядоченное распределение выборки на группы по определенному варьирующему признаку*. Ряд распределения принято оформлять в виде таблиц.

АТРИБУТИВНЫЕ РЯДЫ РАСПРЕДЕЛЕНИЯ

Это ряды распределения, построенные **по номинальным признакам**. В качестве примера рассмотрим распределение респондентов согласно их занятости (Табл. 2).

Табл. 2. *Распределение респондентов по занятости*

№	Категория занятости работника	Количество респондентов
1	Полная занятость	747
2	Частичная занятость	161
3	Безработный	32
4	Временно не работает	51
5	Пенсионер(ка)	231
6	Учащийся	42
7	Домохозяйка(ин)	200
8	Нет ответа	36
	Всего	1500

Графически атрибутивные ряды распределения можно представить в виде столбчатой или круговой диаграммы (Рис. 2, Рис. 3).

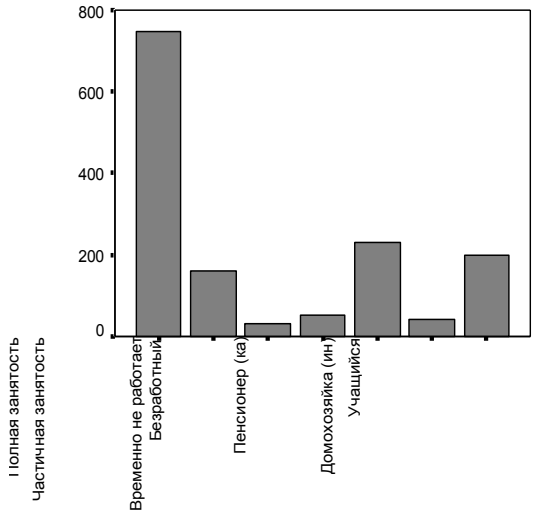


Рис. 2. *Столбчатый график*

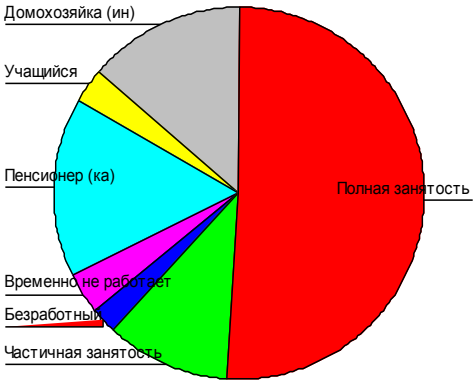


Рис. 3. *Круговая диаграмма*

ВАРИАЦИОННЫЕ РЯДЫ РАСПРЕДЕЛЕНИЯ

Это ряды распределения, построенные **по порядковым и количественным признакам**. Такой ряд представляет собой характеристику вариантов признака их частотами.

Вариантами считаются отдельные значения признака, которые он принимает в вариационном ряду, т. е. конкретное значение варьирующего признака.

Частота (ν) показывает, сколько раз встречается то или иное значение признака. Сумма всех частот – объем выборки: $\sum \nu_i = N$.

Иногда также вычисляют ***частости*** – частоты, выраженные в долях единицы или в процентах к

итогу. Соответственно сумма частостей равна 1 или 100%. Вычисляется по формуле: $w_i = \frac{\nu_i}{\sum \nu_i}$.

Накопленные частоты (***накопленные частости***) – это сумма предыдущих частот (частостей).

В зависимости от характера вариации признака различают дискретные и интервальные вариационные ряды.

дискретный вариационный ряд

Характеризует распределение единиц совокупности **по порядковому признаку**, величина которого принимает только целые значения.

Построим дискретный вариационный ряд на примере данных Табл. 3.

Табл. 3. *Данные о тарифных разрядах 50 рабочих одного из цехов завода*

3	5	6	3	2	4	3	5	5	5
4	3	2	3	4	5	4	2	4	4
5	3	4	5	4	3	3	6	2	2
4	6	3	4	4	5	4	5	3	3
2	6	3	4	5	3	4	4	5	5

Для того, чтобы показать распределение рабочих по тарифному разряду, построим вариационный ряд (Табл. 4).

Табл. 4. *Дискретный вариационный ряд*

Тарифный разряд (варианта x_i)	Численность рабочих (частота ν_i)
2	6
3	13
4	15
5	12
6	4
Всего:	50

Для визуализации ряда используется столбчатый график (Рис. 4).

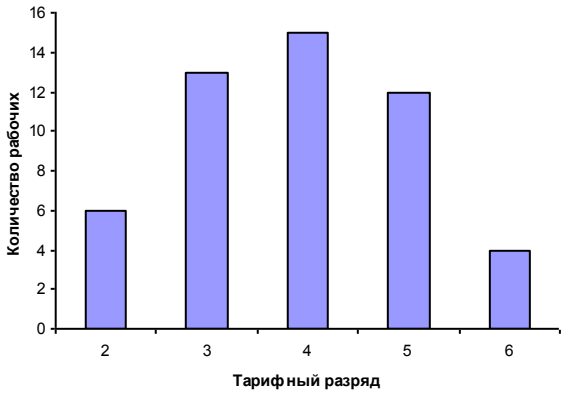


Рис. 4. *Столбчатый график*

интервальный вариационный ряд

Характеризует распределение единиц совокупности **по количественному признаку**, величина которого может принимать в определенных пределах любые значения, отличающиеся друг от друга на сколь угодно малую величину.

После определения исследуемого признака, необходимо решить вопрос о количестве групп (интервалов), на которые надо разбить выборку.

Для определения оптимального числа групп используется ***формула Стерджесса***: $n = 1 + 3,322 \cdot \lg N$, где n – число групп, N – объем выборки. Наличие «пустых» интервалов, в которые не попала ни одна из единиц выборки, свидетельствует о том, что группировка построена неправильно.

После определения числа групп необходимо определить интервалы. *Интервал* – это значения варьирующего признака, лежащие в определенных границах. Каждый интервал имеет верхнюю и нижнюю границы и свою длину (разность между верхней и нижней границами). Чаще всего строят группировку с равными интервалами. ***Длина интервала*** вычисляется по формуле: $h = \frac{R}{n-1}$, где $R = x_{\max} - x_{\min}$ - размах выборки, $x_{\max}, x_{\min}$ - максимальное и минимальное значения признака.

Первый интервал группировки обычно $\left(x_{\min} - \frac{h}{2}; x_{\min} + \frac{h}{2}\right)$, т.е. $x_{\min}$ является серединой первого интервала.

Построим вариационный ряд по данным Табл. 5.

Табл. 5. *Данные об объемах продаж колбасы «XX» по месяцам, млн. р.*

9,4	8,0	6,3	10,0	15,0	8,2	7,3	9,2	5,8	8,7
5,2	13,2	8,1	7,5	11,8	14,6	8,5	7,8	10,5	6,0
5,1	6,8	8,3	7,7	7,9	9,0	10,1	8,0	12,0	14,0
8,2	9,8	13,5	12,4	5,5	7,9	9,2	10,8	12,1	12,4
12,9	12,6	6,7	9,7	8,3	10,8	15,0	7,0	13,0	9,5

Число интервалов группировки: $n = 1 + 3,322 \cdot \lg N = 1 + 3,322 \cdot \lg 50 \approx 6,64 \approx 7$.

Длина интервала: $h = \frac{R}{n-1} = \frac{15-5,1}{6} \approx 1,65 \approx 1,7$.

Вариационный ряд представлен в Табл. 6.

Табл. 6. *Интервальный вариационный ряд*

№	Интервалы		Середина интервала x_i	Частота ν_i	Накопленные частоты
	нижняя граница	верхняя граница			
1	4,25	5,95	5,1	4	4
2	5,95	7,65	6,8	7	11
3	7,65	9,35	8,5	16	27
4	9,35	11,05	10,2	9	36
5	11,05	12,75	11,9	6	42
6	12,75	14,45	13,6	5	47
7	14,45	16,15	15,3	3	50
Сумма	—	—	—	50	—

Для изображения интервального вариационного ряда используется ***гистограмма***. На оси абсцисс откладываются величины интервалов, а частоты изображаются прямоугольниками, построенными на соответствующих интервалах. Высота столбцов гистограммы должна быть пропорциональна частотам.

Гистограмма для рассмотренного выше примера представлена на Рис. 5.

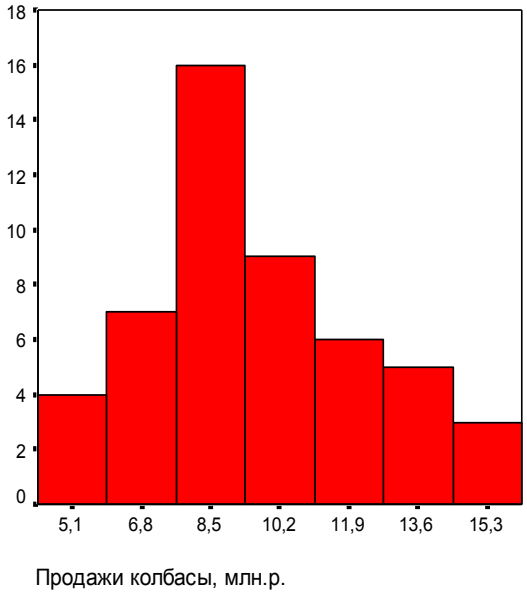


Рис. 5. *Гистограмма по продажам*

ОПИСАТЕЛЬНЫЙ АНАЛИЗ: ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ

СРЕДНЯЯ АРИФМЕТИЧЕСКАЯ

Наиболее распространенной формой статистических показателей является ***средняя величина****, которая представляет собой обобщенную количественную характеристику признака в выборке в конкретных условиях места и времени*. Показатель в форме средней величины *выражает типичные черты и дает обобщенную характеристику однотипных явлений* по одному из варьирующих признаков.

Значения признака отдельных единиц совокупности могут колебаться в ту или иную сторону под влиянием как основных, так и случайных факторов. Сущность средней заключается в том, что в ней взаимопогашаются отклонения значений признака отдельных единиц выборки, обусловленные действием случайных факторов, и учитываются изменения, вызванные действием основных факторов. Это позволяет средней отражать типичный уровень признака и абстрагироваться от индивидуальных особенностей, присущих отдельным единицам.

<u>Свойства средней:</u>

1. Произведение средней на сумму частот равно сумме произведений отдельных вариантов на соответствующие им частоты: $\bar{x}\sum \nu_i = \sum x_i \nu_i$.
2. Сумма отклонений индивидуальных значений признака от средней равна нулю: $\sum (x_i - \bar{x})\nu_i = 0$.

Средняя арифметическая невзвешенная (простая) используется в тех случаях, когда расчет осуществляется по несгруппированным данным:

$$\bar{x} = \frac{\sum x_i}{N} \text{ (здесь } x_i \text{ – каждое значение выборки).}$$

Рассчитаем средний стаж работы по данным Табл. 7.

Табл. 7. *Стаж работы рабочих завода*

Табельный номер рабочего	1	2	3	4	5	6	7
Стаж работы (лет)	10	3	5	12	11	7	9

С учетом имеющихся данных получим: $\bar{x} = \frac{10+3+5+12+11+7+9}{7} \approx 8{,}1$ года.

Средняя арифметическая взвешенная используется при расчете по сгруппированным данным или вариационным рядам:

$$\bar{x} = \frac{\sum x_i \nu_i}{\sum \nu_i}$$ (здесь x_i – середины интервалов группировки).

Найдем средние продажи колбасы в месяц по данным Табл. 5.

Табл. 8. *Расчетная таблица*

№	Интервалы		Середина интервала x_i	Частота ν_i	$x_i \nu_i$
	нижняя граница	верхняя граница			
1	4,25	5,95	5,1	4	20,4
2	5,95	7,65	6,8	7	47,6
3	7,65	9,35	8,5	16	136,0
4	9,35	11,05	10,2	9	91,8
5	11,05	12,75	11,9	6	71,4
6	12,75	14,45	13,6	5	68,0
7	14,45	16,15	15,3	3	45,9
Сумма	—	—	—	50	481,1

Средние продажи за месяц составили $\bar{x} = \frac{481,1}{50} = 9,62$ млн.р.

МЕДИАНА

Медиана Me - это значение признака, приходящееся на середину упорядоченной совокупности.

Пусть имеются следующие данные о рабочих бригады, состоящей из 9 человек:

Номер рабочего	1	2	3	4	5	6	7	8	9
Тарифный разряд	4	3	4	5	3	3	6	2	6

Для определения медианы необходимо провести ранжирование тарифных разрядов рабочих:

2 3 3 3 (4) 4 5 6 6

Центральным в этом упорядоченном ряду является пятое слева значение (рабочий 4-го разряда), следовательно, данное значение признака и будет медианным. Интерпретация: половина бригады имеет разряд ниже 4-го, а вторая половина – выше 4-го.

Если ранжированный ряд содержит четное число значений, то медиана определяется как средняя их двух центральных значений.

Нахождение моды и медианы для интервального вариационного ряда (по данным Табл. 5):

№	Интервалы		Середина интервала x_i	Частота ν_i	Накопленная частота
	нижняя граница	верхняя граница			
1	4,25	5,95	5,1	4	4
2	5,95	7,65	6,8	7	11
3	7,65	9,35	8,5	16	27
4	9,35	11,05	10,2	9	36
5	11,05	12,75	11,9	6	42
6	12,75	14,45	13,6	5	47
7	14,45	16,15	15,3	3	50
Сумма	—	—	—	50	—

Медиана рассчитывается по формуле:

$$Me = x_{Me} + h \cdot \frac{\frac{1}{2} \cdot \sum \nu_i - S_{Me-1}}{\nu_{Me}},$$

где x_{Me} – нижняя граница медианного интервала, т.е. интервала, накопленная частота которого первая превышает половину общей суммы частот;

h – длина медианного интервала;

S_{Me-1} – накопленная частота интервала, предшествующего медианному;

ν_{Me} – частота медианного интервала.

Медианным будет 3-ий интервал, т.к. его накопленная частота 27 первая превышает половину общей суммы частот (50/2=25).

$$Me = 7,65 + 1,7 \cdot \frac{\frac{1}{2} \cdot 50 - 11}{16} = 9,14 \text{ млн.р.}$$

Медиана показывает, что половину месяцев наблюдений продажи были менее 9,14 млн.р., а половину – больше.

КВАРТИЛИ, КВАРТИЛЬНЫЙ РАЗМАХ

Квартили – это значения признака, делящие ранжированную совокупность на четыре равновеликие части.

Различают:

- нижнюю квартиль $Q_н$, отделяющую ¼ часть совокупности с наименьшими значениями признака;
- среднюю квартиль – медиану;
- верхнюю квартиль $Q_в$, отделяющую ¼ часть совокупности с наибольшими значениями признака.

Для расчета квартилей по интервальному вариационному ряду используются формулы:

$$Q_н = x_{Q_н} + h \cdot \frac{\frac{1}{4} \cdot \sum \nu_i - S_{Q_н - 1}}{\nu_{Q_н}}, \quad Q_в = x_{Q_в} + h \cdot \frac{\frac{3}{4} \cdot \sum \nu_i - S_{Q_в - 1}}{\nu_{Q_в}},$$

где $x_{Q_н}$ – нижняя граница интервала, содержащего нижнюю квартиль (интервал определяется по накопленной частоте, первой, превышающей ¼ выборки);

$x_{Q_в}$ – нижняя граница интервала, содержащего верхнюю квартиль (интервал определяется по накопленной частоте, первой, превышающей ¾ выборки);

h – длина интервала;

$S_{Q_н - 1}$ – накопленная частота интервала, предшествующего интервалу, содержащему нижнюю квартиль;

$S_{Q_в - 1}$ – накопленная частота интервала, предшествующего интервалу, содержащему верхнюю квартиль;

$\nu_{Q_н}$, $\nu_{Q_в}$ – частота интервала, содержащего нижнюю и верхнюю квартили соответственно.

Квартильный размах – разница между верхней и нижней квартилями: $H = Q_в - Q_н$

Квартильный размах охватывает 50% значений выборки.

Для данных Табл. 5:

№	Интервалы		Середина интервала x_i	Частота ν_i	Накопленная частота
	нижняя граница	верхняя граница			
1	4,25	5,95	5,1	4	4
2	5,95	7,65	6,8	7	11
3	7,65	9,35	8,5	16	27
4	9,35	11,05	10,2	9	36
5	11,05	12,75	11,9	6	42
6	12,75	14,45	13,6	5	47
7	14,45	16,15	15,3	3	50
Сумма	—	—	—	50	—

Интервал, содержащий нижнюю квартиль – 3-ий, т.к. накопленная частота (27) первая превышает ¼ общей сумы частот (12,5).

Интервал, содержащий верхнюю квартиль – 5-ый, т.к. накопленная частота (42) первая превышает ¾ общей сумы частот (37,5).

$$Q_н = 7,65 + 1,7 \cdot \frac{\frac{1}{4} \cdot 50 - 11}{16} = 7,81; \; Q_в = 11,05 + 1,7 \cdot \frac{\frac{3}{4} \cdot 50 - 36}{6} = 11,48 .$$

Квартильный размах – полоса, шириной $11,48 - 7,81 = 3,67$ содержит 50% значений выборки.

СТАНДАРТНОЕ ОТКЛОНЕНИЕ

Стандартное отклонение оценивает разброс наблюдаемых значений вокруг их средней.

Вычисляется по формулам:

- невзвешенное: $\sigma = \sqrt{\frac{\sum x_i - \bar{x}^2}{N-1}}$, где x_i – каждое значение выборки;
- взвешенное: $\sigma = \sqrt{\frac{\sum x_i - \bar{x}^2 \nu_i}{N-1}}$, где x_i – середина интервала группировки.

Выражается в тех же единицах измерения, что и количественный признак.

В пределах $\bar{x} \pm \sigma$ находится 68,3% вариант. Более того, в пределах $\bar{x} \pm 2\sigma$ - 95,4% вариант, в пределах $\bar{x} \pm 3\sigma$ - 99,7%. Это положение называется ***правилом «трех сигм»***[*]. Таким образом, по показателю σ можно оценить полосу, в которой происходит вариация «основной массы» (95,4%) значений признака. Как правило, в оставшиеся 5% входят выбросы – нетипичные значения.

[*] Правило «трех сигм» имеет место в условиях нормального распределения, которое будет рассмотрено ниже.

Оценим разброс значений по данным Табл. 5.

№	Интервалы		Середина интервала x_i	Частота ν_i	$x_i - \bar{x}$	$x_i - \bar{x}^2$	$x_i - \bar{x}^2 \cdot \nu_i$
	нижняя граница	верхняя граница					
1	4,25	5,95	5,1	4	-4,52	20,43	81,72
2	5,95	7,65	6,8	7	-2,82	7,95	55,67
3	7,65	9,35	8,5	16	-1,12	1,25	20,07
4	9,35	11,05	10,2	9	0,58	0,34	3,03
5	11,05	12,75	11,9	6	2,28	5,20	31,19
6	12,75	14,45	13,6	5	3,98	15,84	79,20
7	14,45	16,15	15,3	3	5,68	32,26	96,79
Сумма:	—	—	—	50	—	—	367,67

$\sigma = \sqrt{\frac{\sum x_i - \bar{x}^2 \nu_i}{N-1}} = \sqrt{\frac{367,67}{50-1}} = 2,74$ млн.р., т.е. 95% продаж варьируются в интервале $\bar{x} \pm 2 \cdot \sigma = 9,42 \pm 2 \cdot 2,74$, т.е. от 3,94 до 14,9 млн.р.

ДИСПЕРСИЯ

Представляет собой стандартное отклонение в квадрате.

- невзвешенная: $\sigma^2 = \frac{\sum x_i - \bar{x}^2}{N-1}$;
- взвешенная: $\sigma^2 = \frac{\sum x_i - \bar{x}^2 \nu_i}{N-1}$.

Дисперсия обладает одним существенным недостатком: ее значение выражается в квадратных единицах. Но, несмотря на это, она обладает рядом замечательных свойств, использование которых получило широкое распространение в различных областях знаний, на их основе были разработаны новые методы исследования, а также новые показатели количественной характеристики большого класса явлений. В частности, дисперсии можно складывать и вычитать, в отличие от стандартных отклонений.

ОПИСАТЕЛЬНЫЙ АНАЛИЗ: НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Для правильного выбора адекватных описательных статистик и метода статистического анализа необходимо определить, как распределены количественные данные. В математической статистике выделяют большое число видов распределений. На практике часто ограничиваются делением распределений на два класса: нормальные и ненормальные распределения.

Графиком нормального распределения (Рис. 6) является симметричная колоколообразная кривая, которая задается уравнением:

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

Кривая нормального распределения полностью определяется средней арифметической и стандартным отклонением – $N\ \bar{x}, \sigma$.

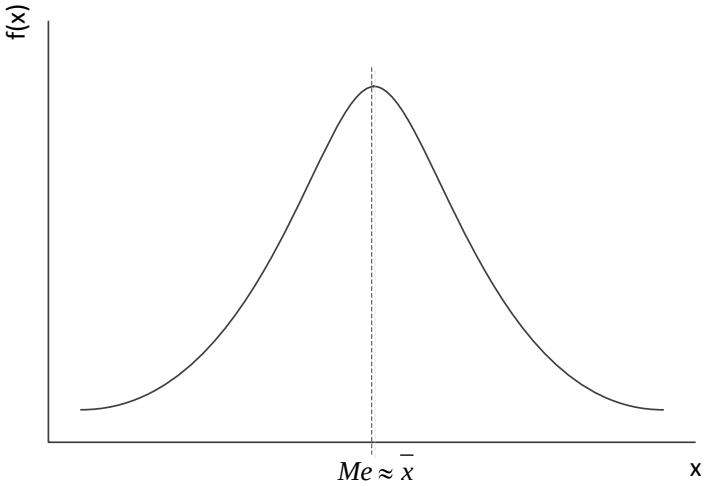


Рис. 6. *График нормального распределения*

Для нормального распределения значения средней и медианы примерно равны.

Для ненормальных распределений соотношение средней и медианы выглядит следующим образом (Рис. 7):

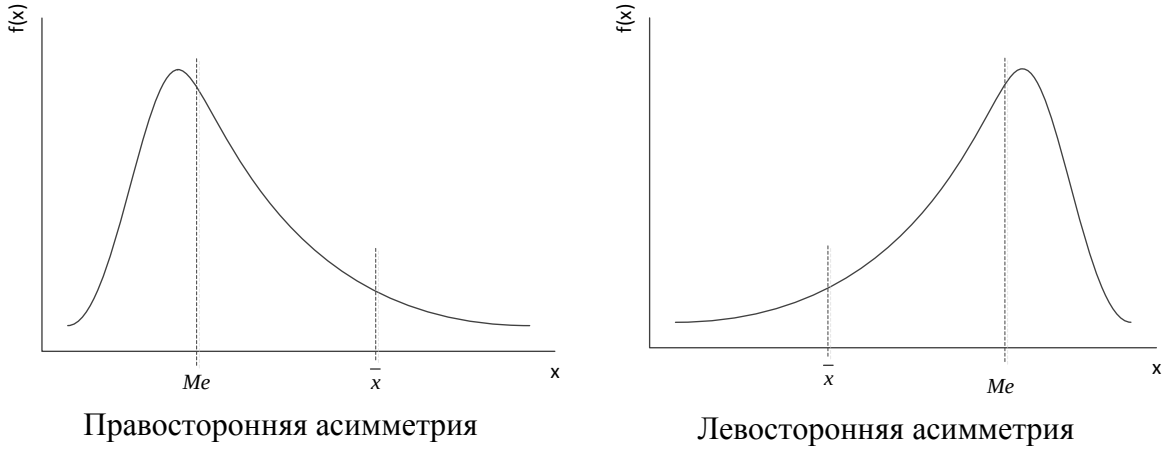


Рис. 7. *Среднее и медиана для ненормальных распределений*

Очевидно, что интерпретация среднего значения для ненормального распределения не даст ответов на поставленные вопросы. Поэтому для ненормальных распределений интерпретируют медиану.

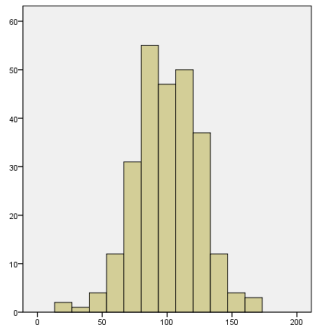
ПРОВЕРКА НА НОРМАЛЬНОСТЬ

Для проверки количественных данных на нормальность используются следующие методы:

1. **Графические**

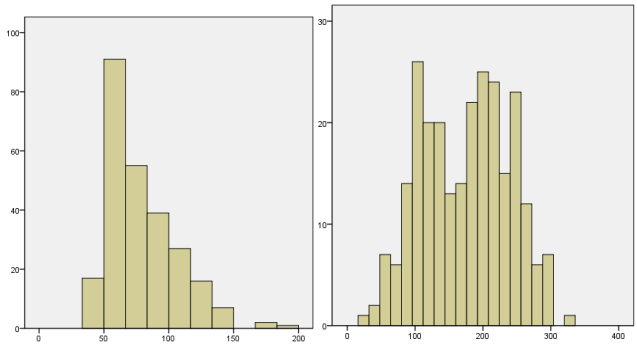
1.1. *Гистограмма*

Нормальное распределение



Симметричная гистограмма

Ненормальное распределение



Асимметричная гистограмма

Для данных из Табл. 5 гистограмма была построена ранее, она имела вид (Рис. 8):

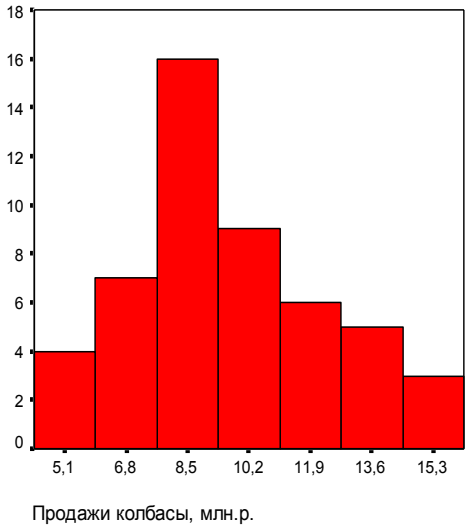


Рис. 8. *Гистограмма по продажам*

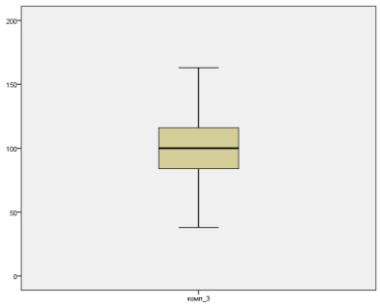
Гистограмму можно считать симметричной.

1.2. *Ящик с усами*

Идея построения диаграммы ящик с усами представлена на Рис. 9.

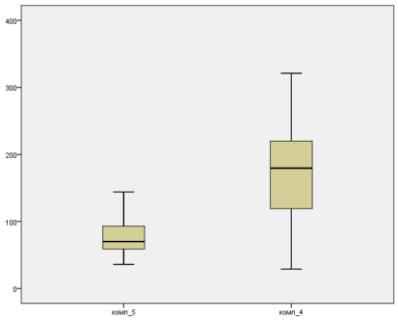
Рис. 9. *Идея построения ящика с усами*

Нормальное распределение



Симметричный ящик

Ненормальное распределение



Асимметричный ящик

Построим ящик с усами для данных из Табл. 5. Ранее было определено:

Максимальное значение	15
Верхняя квартиль	11,48
Медиана	9,14
Нижняя квартиль	7,81
Минимальное значение	5,1

Ящик будет иметь вид (Рис. 10):

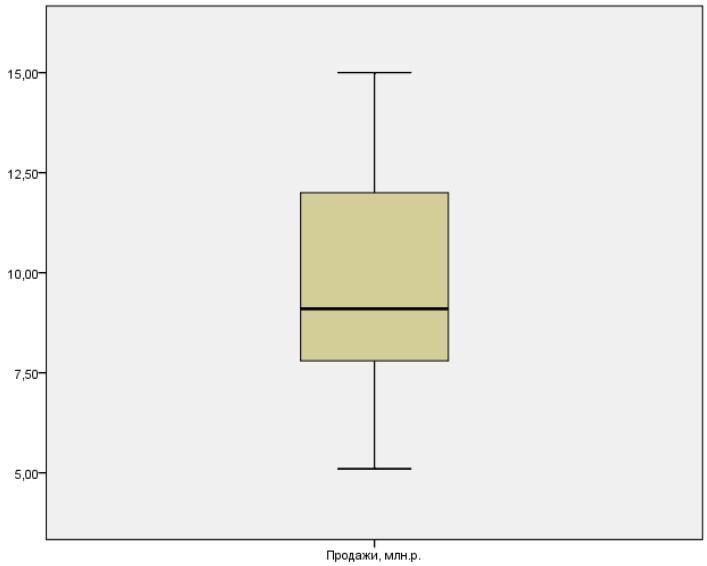


Рис. 10. *Ящик с усами по продажам*

Медиана расположена не симметрично относительно ящика. Окончательные выводы о нормальности распределения будут сделаны ниже после использования аналитического теста.

2. **Аналитические**

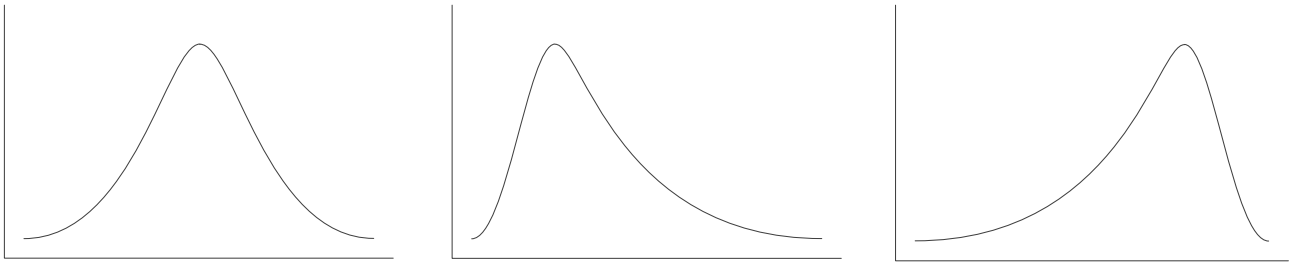
Для аналитической проверки на нормальность существует огромное число тестов, например, критерий Хи-квадрат, критерий Колмогорова, Шапиро-Уилка и другие. Рассмотрим ***критерий Жарка-Бера (Jarque-Bera)***. Идея этого критерия заключается в том, что по данным выборки оценивается скошенность (асимметрия) и «вытянутость» фактического распределения и сравнивается с нормальным.

За оценку асимметрии распределения отвечает ***коэффициент асимметрии***:

$$Sk = \frac{1}{N} \cdot \sum \left[\nu_i \cdot \left(\frac{x_i - \bar{x}}{\tilde{\sigma}} \right)^3 \right]$$ (здесь x_i – середины интервалов группировки, $\tilde{\sigma} = \sigma \cdot \sqrt{\frac{N-1}{N}}$).

В пакете Microsoft Excel за вычисление коэффициента асимметрии отвечает функция **=СКОС(исходная выборка)**.

Нормальное распределение Ненормальное распределение



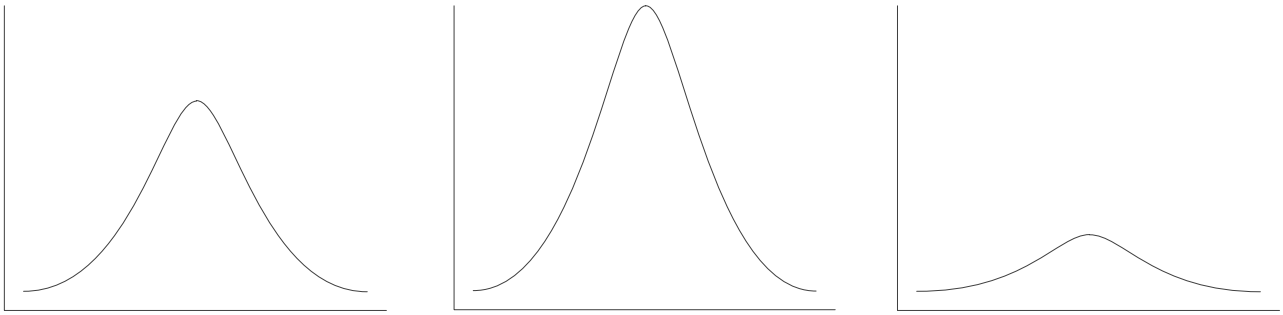
Симметричное, $Sk = 0$ | Правосторонняя асимметрия, $Sk > 0$ | Левосторонняя асимметрия, $Sk < 0$

За оценку «вытянутости» распределения отвечает ***эксцесс***:

$$K = \frac{1}{N} \cdot \sum \left[\nu_i \cdot \left(\frac{x_i - \bar{x}}{\tilde{\sigma}} \right)^4 \right]$$ (здесь x_i – середины интервалов группировки, $\tilde{\sigma} = \sigma \cdot \sqrt{\frac{N-1}{N}}$).

В пакете Microsoft Excel за вычисление эксцесса отвечает выражение =**ЭКСЦЕСС(исходная выборка)+3**.

Нормальное распределение Ненормальное распределение



$K = 3$ | $K > 3$ | $K < 3$

Алгоритм критерия Жарка-Бера:

1) Выдвинуть гипотезу H_0 о нормальном распределении выборки.
2) Вычислить фактическое значение критерия по формуле:

$$J - B = \frac{N}{6} \cdot \left(Sk^2 + \frac{K-3}{4}^2 \right).$$

3) Определить табличное значение критерия на основе специальных таблиц критических значений Пирсона на уровне значимости 0,05 и числе степеней свободы 2. Это значение равно 5,991.
   Уровень значимости – это вероятность ошибиться, утверждая, что распределение ненормальное. Общепринятым является вероятность ошибки не превышающая 5%.
   Число степеней свободы, в данном случае, отвечает за количество параметров в формуле критерия: там участвуют асимметрия и эксцесс.
4) Если $J - B > 5,991$, то гипотеза H_0 о нормальном распределении выборки отклоняется, т.е. распределение не является нормальным.
   Если $J - B < 5,991$, то гипотеза H_0 о нормальном распределении выборки принимается, т.е. распределение является нормальным.

Важно! Если объем выборки $N < 30$, то считают, что данные распределены ненормально. Но если известно, что выборка извлечена из нормально распределенной совокупности, то тогда можно говорить о нормальности.

Применим критерий Жарка-Бера к данным Табл. 5. Для начала вычислим асимметрию и эксцесс.

Будем производить вычисления по действиям:

1. $\tilde{\sigma} = \sigma \cdot \sqrt{\frac{N-1}{N}} = 2{,}74 \cdot \sqrt{\frac{50-1}{50}} = 2{,}71$
2. Составим расчетную таблицу для вычисления асимметрии и эксцесса:

№	Середина интервала x_i	Частота ν_i	$x_i - \bar{x}$	$\frac{x_i - \bar{x}}{\tilde{\sigma}}$	$\nu_i \cdot \left(\frac{x_i - \bar{x}}{\tilde{\sigma}}\right)^3$	$\nu_i \cdot \left(\frac{x_i - \bar{x}}{\tilde{\sigma}}\right)^4$
1	5,1	4	-4,52	-1,67	-18,56	30,96
2	6,8	7	-2,82	-1,04	-7,89	8,21
3	8,5	16	-1,12	-0,41	-1,13	0,47
4	10,2	9	0,58	0,21	0,09	0,02
5	11,9	6	2,28	0,84	3,57	3,01
6	13,6	5	3,98	1,47	15,84	23,26
7	15,3	3	5,68	2,10	27,62	57,89
Сумма:	—	50	—	—	19,54	123,82

$$Sk = \frac{1}{N} \cdot \sum \left[\nu_i \cdot \left(\frac{x_i - \bar{x}}{\tilde{\sigma}} \right)^3 \right] = \frac{1}{50} \cdot 19{,}54 = 0{,}391,$$

$$K = \frac{1}{N} \cdot \sum \left[\nu_i \cdot \left(\frac{x_i - \bar{x}}{\tilde{\sigma}} \right)^4 \right] = \frac{1}{50} \cdot 123{,}82 = 2{,}476.$$

3. Выдвинем гипотезу H_0 о нормальном распределении выборки.
4. Вычислим фактическое значение критерия Жарка-Бера:

$$J - B = \frac{N}{6} \cdot \left(Sk^2 + \frac{K-3}{4}^2 \right) = \frac{50}{6} \cdot \left(0{,}391^2 + \frac{2{,}476-3}{4}^2 \right) = 1{,}846.$$

5. $J - B = 1{,}846 < 5{,}991$, значит гипотеза H_0 о нормальном распределении выборки принимается, т.е. распределение является нормальным.

СХЕМА ВЫБОРА АДЕКВАТНЫХ ОПИСАТЕЛЬНЫХ СТАТИСТИК

АНАЛИЗ ВЛИЯНИЯ ФАКТОРОВ

Фактор (x) → [влияет на] → Отклик (y)

Общая идея методов: определить, влияет ли фактор на отклик. Вопросы о качественном характере этого влияния («каким образом влияет?») статистикой не рассматриваются.

В зависимости от типа фактора и отклика и нормальности распределения будут использоваться разные методы определения влияния (Табл. 9).

Табл. 9. *Выбор метода анализа влияния факторов*

Фактор	Отклик	Метод
Количественный (нормальный)	Количественный (нормальный)	**Корреляция Пирсона**
Количественный (ненормальный)	Количественный (ненормальный)	**Корреляция Спирмена**
Качественный (2 значения)	Количественный (нормальный)	**Т-критерий Стьюдента**
Качественный (2 значения)	Количественный (ненормальный)	**U-критерий Манна-Уитни**
Качественный (3 и более значений)	Количественный (нормальный)	**Дисперсионный анализ**
Качественный (3 и более значений)	Количественный (ненормальный)	**Н-критерий Краскела-Уоллиса**
Качественный	Качественный	**Критерий Хи-квадрат**

Ограничимся рассмотрением корреляции Пирсона и Спирмена.

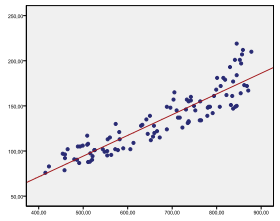
КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ

Рассмотрим влияние **количественного фактора** на **количественный отклик**.

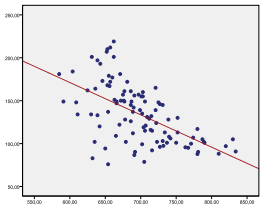
Корреляция – мера линейной связи между двумя количественными признаками.

Перед проведением расчетов полезно построить *диаграмму рассеяния* – график, на котором по горизонтальной оси отмечаются значения фактора (x), а по вертикальной – отклика (y). Расположение точек подскажет и о силе связи, и о ее характере.

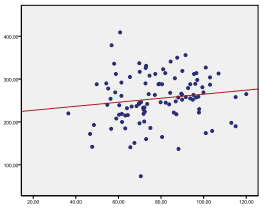
Сильная положительная корреляция



Умеренная отрицательная корреляция



Отсутствие корреляции



Свойства коэффициента корреляции:

1) изменяется в пределах от -1 до 1;
2) если коэффициент корреляции менее 0,3, то значимая статистическая связь отсутствует;
3) если коэффициент корреляции равен 1 или -1, то связь не корреляционная, а функциональная (полная);
4) если коэффициент больше 0, то связь называется прямой, если меньше 0 – то обратной;

5) сила связи может быть охарактеризована, например, следующим образом:

Теснота связи	Абсолютное значение коэффициента корреляции
отсутствует	Менее 0,3
слабая	0,3 – 0,5
умеренная	0,5 – 0,7
высокая	0,7 и более

В зависимости от нормальности распределения фактора и отклика используют следующие коэффициенты корреляции (Рис. 11):

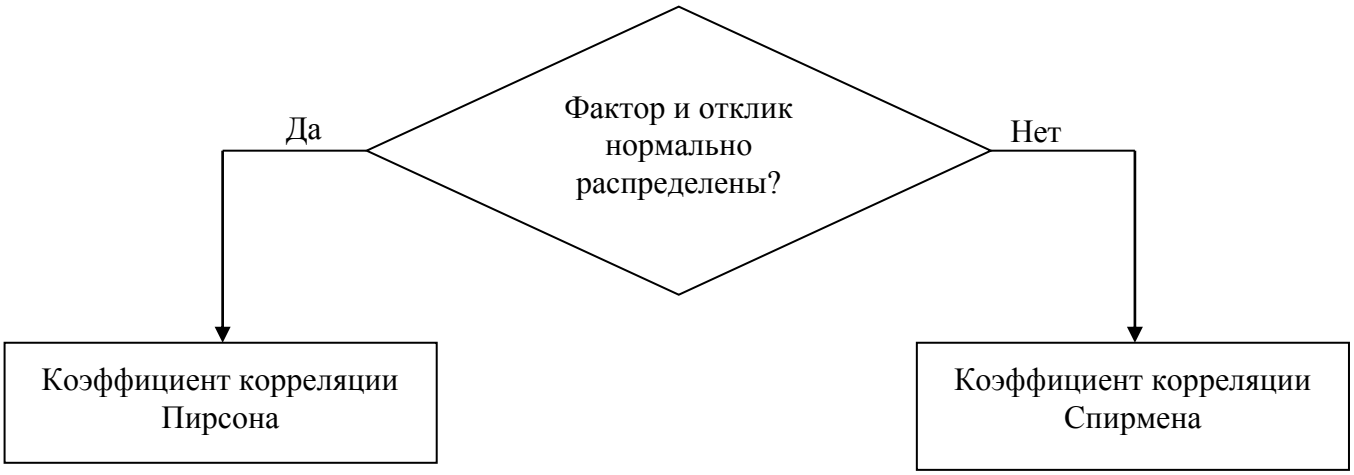


Рис. 11. *Схема использования коэффициентов корреляции*

коэффициент корреляции Пирсона

Особенности использования

1) число наблюдений фактора и отклика должно быть равно (обозначим N);
2) фактор и отклик должны иметь распределение, близкое к нормальному (или извлечены из нормально распределенных выборок бóльшего объема);
3) формула для расчета (пусть x – фактор, а y – отклик):

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y}, \text{ где}$$

$\overline{x}$ – среднее по фактору;

$\overline{y}$ – среднее по отклику;

$\overline{xy}$ – среднее произведений фактора на отклик;

σ_x – стандартное отклонение для фактора: $\sigma_x = \sqrt{\frac{\sum x_i - \overline{x}^2}{N-1}}$;

σ_y – стандартное отклонение для отклика: $\sigma_y = \sqrt{\frac{\sum y_i - \overline{y}^2}{N-1}}$.

Рассмотрим на примере Табл. 5. Ранее было доказано, что выборка не противоречит нормальному закону распределения. Разделим исходные данные на две части по 25 наблюдений (Табл. 10) и выясним, влияет ли фактор (x) на отклик (y).

Табл. 10. *Исходные данные для вычисления коэффициента корреляции Пирсона*

№	x	y
1	9,4	8,2
2	5,2	14,6
3	5,1	9
4	8,2	7,9
5	12,9	10,8
6	8	7,3
7	13,2	8,5
8	6,8	10,1
9	9,8	9,2
10	12,6	15
11	6,3	9,2
12	8,1	7,8
13	8,3	8
14	13,5	10,8
15	6,7	7
16	10	5,8
17	7,5	10,5
18	7,7	12
19	12,4	12,1
20	9,7	13
21	15	8,7
22	11,8	6
23	7,9	14
24	5,5	12,4
25	8,3	9,5

Для начала построим диаграмму рассеяния (Рис. 12).

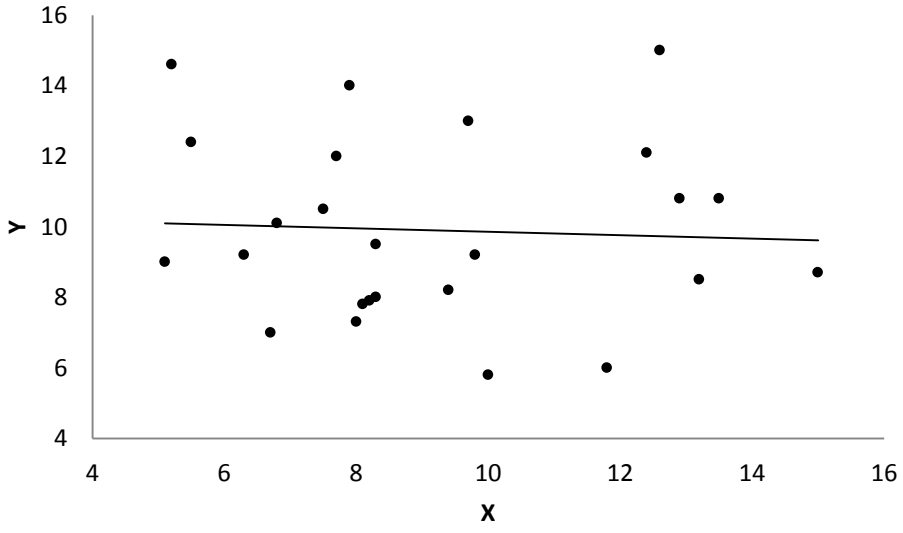


Рис. 12. *Диаграмма рассеяния*

По диаграмме рассеяния можно предположить отсутствие значимой связи. Проверим это с помо-

щью коэффициента корреляции Пирсона. Составим расчетную таблицу (Табл. 11).

Табл. 11. *Расчетная таблица для вычисления коэффициента корреляции Пирсона*

№	x	y	xy	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	9,4	8,2	77,08	0,04	2,88
2	5,2	14,6	75,92	15,97	22,13
3	5,1	9	45,9	16,78	0,80
4	8,2	7,9	64,78	0,99	3,98
5	12,9	10,8	139,32	13,72	0,82
6	8	7,3	58,4	1,43	6,74
7	13,2	8,5	112,2	16,03	1,95
8	6,8	10,1	68,68	5,74	0,04
9	9,8	9,2	90,16	0,36	0,48
10	12,6	15	189	11,59	26,05
11	6,3	9,2	57,96	8,39	0,48
12	8,1	7,8	63,18	1,20	4,39
13	8,3	8	66,4	0,80	3,59
14	13,5	10,8	145,8	18,52	0,82
15	6,7	7	46,9	6,23	8,39
16	10	5,8	58	0,65	16,78
17	7,5	10,5	78,75	2,88	0,36
18	7,7	12	92,4	2,24	4,43
19	12,4	12,1	150,04	10,27	4,86
20	9,7	13	126,1	0,25	9,63
21	15	8,7	130,5	33,69	1,43
22	11,8	6	70,8	6,78	15,18
23	7,9	14	110,6	1,68	16,84
24	5,5	12,4	68,2	13,66	6,27
25	8,3	9,5	78,85	0,80	0,16
Сумма	229,9	247,4	2265,92	190,69	159,49

Средние значения:

$$\bar{x} = \frac{\sum x_i}{N} = \frac{229{,}9}{25} = 9{,}20\,,\ \bar{y} = \frac{\sum y_i}{N} = \frac{247{,}4}{25} = 9{,}90\,,\ \overline{xy} = \frac{\sum x_i y_i}{N} = \frac{2265{,}92}{25} = 90{,}64\,.$$

Стандартные отклонения:

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{190{,}69}{25-1}} = 2{,}82\,,\ \sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N-1}} = \sqrt{\frac{159{,}49}{25-1}} = 2{,}58\,.$$

Корреляция Пирсона:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x \cdot \sigma_y} = \frac{90{,}64 - 9{,}20\cdot 9{,}90}{2{,}82\cdot 2{,}58} = -0{,}06\,.$$

Полученный коэффициент корреляции по абсолютному значению не превышает 0,3, значит, существенной статистической связи между фактором и откликом не наблюдается.

коэффициент корреляции Спирмена

Особенности использования

1) фактор и отклик могут иметь ненормальное распределение;
2) число наблюдений фактора и отклика должно быть равно (обозначим N);
3) формула для расчета:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)},$$ где d_i^2 - квадрат разности рангов фактора и отклика.

Замечание о ранжировании

Процедура ранжирования представляет собой упорядочивание значений по возрастанию. Ниже рассмотрены примеры ранжирования, в том числе и выставление средних рангов.

Выборка 1	Ранги
5	3
3	2
2	1
9	4
20	5

Выборка 2	Ранги
5	4
3	1
4	2,5
4	2,5
9	5

Выборка 3	Ранги
4	3
2	1
4	3
4	3
10	5

Быстро осуществить ранжирование можно с помощью пакета Microsoft Excel 2010 используя функцию **=РАНГ.СР( )**.

Рассмотрим расчет корреляции Спирмена на примере. Пусть требуется выяснить, влияют ли затраты на рекламу на объем продаж. Исходные данные представлены в Табл. 12.

Табл. 12. *Влияние рекламы на продажи*

Месяц	Затраты на рекламу, р	Продажи, тыс.р
Январь	2510	376
Февраль	2588	377
Март	2452	334
Апрель	2476	347
Май	1956	292
Июнь	2252	320
Июль	1808	271
Август	2356	353
Сентябрь	2660	451
Октябрь	2571	456
Ноябрь	2432	408
Декабрь	2376	440

Так как количество данных менее 30, будем считать, что признаки имеют ненормальное распределение.

Для начала построим диаграмму рассеяния (Рис. 13).

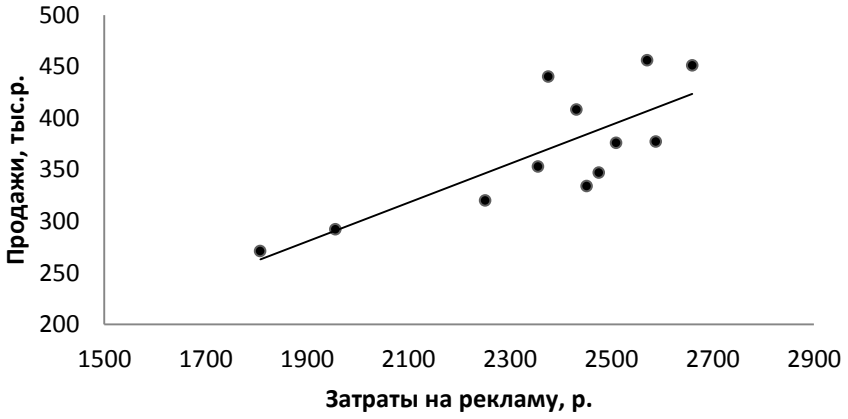


Рис. 13. *Диаграмма рассеяния*

По диаграмме рассеяния можно предположить наличие значимой корреляции. Вычислим коэффициент корреляции Спирмена. Для этого составим расчетную таблицу (Табл. 13).

Табл. 13. *Расчетная таблица для вычисления коэффициента корреляции Спирмена*

Месяц	Затраты на рекламу, р, (x)	Продажи, тыс.р (y)	Ранг x	Ранг y	d	d^2
Январь	2510	376	9	7	2	4
Февраль	2588	377	11	8	3	9
Март	2452	334	7	4	3	9
Апрель	2476	347	8	5	3	9
Май	1956	292	2	2	0	0
Июнь	2252	320	3	3	0	0
Июль	1808	271	1	1	0	0
Август	2356	353	4	6	-2	4
Сентябрь	2660	451	12	11	1	1
Октябрь	2571	456	10	12	-2	4
Ноябрь	2432	408	6	9	-3	9
Декабрь	2376	440	5	10	-5	25
Сумма	—	—	—	—	—	74

Коэффициент корреляции Спирмена: $r_S = 1 - \frac{6 \cdot \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} = 1 - \frac{6 \cdot 74}{12 \cdot (12^2 - 1)} = 0{,}741$.

Полученный коэффициент корреляции свидетельствует о прямой высокой связи между затратами на рекламу и продажами.