Измерительные шкалы

- *Измерение* – это приписывание объекту числа по определенному правилу.
- Это правило устанавливает соответствие между измеряемым свойством и его значением.

Измерительные шкалы

(С. Стивенс, 1951 год)

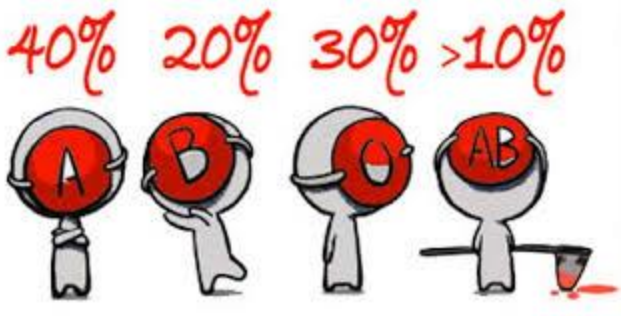
Неметрические:

- Номинативная шкала (шкала наименований)
- Ранговая (порядковая) шкала

Метрические:

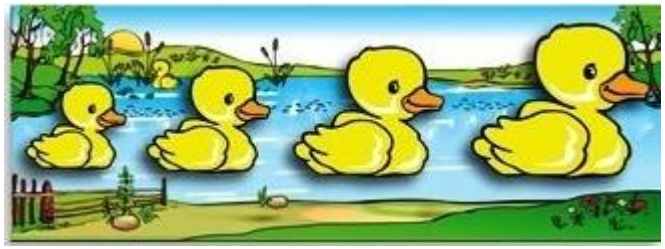
- Интервальная шкала
- Абсолютная шкала (шкала отношений)

Номинативная шкала (шкала наименований)



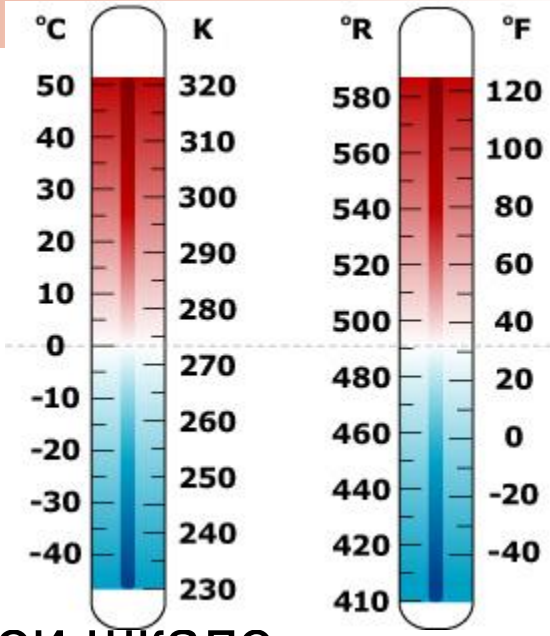
- Неметрическая.
- Измерение состоит в присвоении признаку определенного обозначения или символа.
- Процедура измерения состоит в классификации объектов.
- При сравнении различных значений между собой можно только сказать, что они разные, но упорядочивать, сравнивать по степени выраженности признака нельзя.
- Широко используются, но для них необходимы специальные процедуры обработки данных.

Ранговая (порядковая) шкала



- Неметрическая.
- Измерение предполагает приписывание объектам чисел в зависимости от степени выраженности измеряемого свойства.
- Измерение по этой шкале расчленяет всю совокупность измеренных признаков на такие множества, которые связаны между собой отношениями типа «больше — меньше», «выше — ниже», «сильнее - слабее» и пр.
- Не позволяет делать заключение "на сколько больше" или "на сколько меньше".
- Размер интервала между категориями не может быть выражен количественно.

Интервальная шкала



- Метрическая
- Измерение отражает не только различия в уровне выраженности признака, но и то, на сколько больше или меньше выражен этот признак.
- Равным разностям между числами в этои шкале соответствуют равные разности в уровне выраженности измеренного признака.
- Главное понятие этой шкалы – интервал, являющийся долей или частью измеряемого свойства между двумя соседними позициями на шкале. Размер интервала – величина фиксированная и постоянная на всех участках шкалы.
- Нет естественной точки отсчета (нуль условен и не означает отсутствие измеряемого свойства).

Абсолютная шкала (шкала отношений)

- Метрическая
- Установлена нулевая точка, соответствующая полному отсутствию выраженности измеряемого признака.
- В силу абсолютности нулевой точки можно сказать не только о том, насколько больше или меньше выражено свойство, но и о том, во сколько раз больше или меньше оно выражено.
- Наиболее информативна, допускает различные математические операции и использование разнообразных статистических методов.

Сила шкал

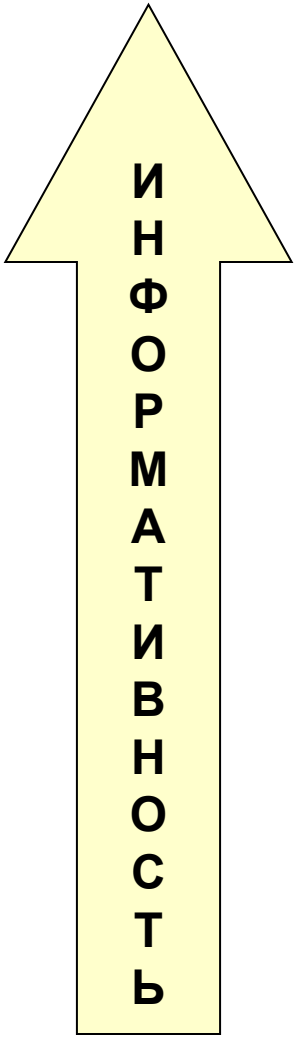

Шкала наименований
Ранговая шкала
Шкала интервалов
Шкала отношений

Типы данных (применительно к статистической обработке)

- Качественные
  - Номинативные
  - Ранговые (порядковые, полуколичественные)
- Количественные
  - Дискретные
  - Непрерывные

Информативность шкал данных

- Непрерывные
- Дискретные
- Ранговые
- Номинативные



Преобразование данных

- количественные
- ранговые
- номинативные

Описательная статистика

Генеральная совокупность

Всё множество объектов, обладающее изучаемым признаком.

Генеральная совокупность

Выборка

несколько элементов
из генеральной совокупности

Генеральная совокупность

Отбор

Репрезантативная выборка

Анализ

Выводы о генеральной совокупности

- Характеристики, которые базируются на данных массовых наблюдений, называют **обобщающими показателями** или **числовыми характеристиками**.
- Эти показатели характеризуют значения признака, его вариацию.
- Их вычисляют с помощью вариант и соответствующих частот (относительных частот).

Описательная статистика нужна для:

- «Сжатия» и концентрирования информации
- Первичного анализа полученной информации
- Представления и сравнения результатов

- *Ценность описательной статистики* в том, что она дает сжатую и концентрированную характеристику изучаемого явления.
- **Например:**
- На некотором предприятии работает 1500 человек.
- Бухгалтерская ведомость на зарплату довольно большая.
- Информация о том, что средняя месячная зарплата работников этого предприятия составляет 8200 рублей, дает определенное, хотя и неполное представление об уровне заработной платы на этом предприятии.

Что характеризует?

- Центр распределения
- Разброс значений
- Форму кривой

§1. Меры центральной тенденции

- Важнейшие среди обобщающих показателей - **средние величины**, т. е. такие значения признака, вокруг которых группируются отдельные наблюдаемые значения элементов.
- Отсюда и название - **меры центральной тенденции**.

- В зависимости от характера задачи пользуются тем или иным видом средней величины.
- К ним принадлежат
  - *среднее арифметическое* (*выборочная средняя*)
  - *мода*
  - *медиана*
- Обобщающие показатели только тогда объективно будут соответствовать своему назначению, если применяются к однородным совокупностям.

1.1. Среднее арифметическое (выборочная средняя)

- *Выборочная средняя* – это средняя арифметическая всех вариант в выборке.
- Обозначается $\overline{X_B}$
- Вычисляется по формуле:

$$\overline{X_B} = \frac{1}{n}\left(x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots + x_k \cdot n_k\right) = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

(для группированной выборки)

$$\overline{X_B} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(для негруппированной выборки).

- Выборочная средняя характеризует среднюю варианту признака.

1.1. Среднее арифметическое (выборочная средняя)

- Сущность среднего арифметического состоит в следующем: если каждое наблюдение заменить средним, то общая сумма не изменится.
- Среднее можно интерпретировать еще и так: если все наблюдения будут равны между собой, а сумма наблюдений останется неизменной, то каждое наблюдение будет равно среднему.
- Поскольку среднее сохраняет неизменной сумму при равномерном распределении значений, то оно наиболее полезно в качестве обобщающего показателя при отсутствии резко выделяющихся наблюдений, или как их называют, выбросов, т. е. когда набор данных представляет собой более менее однородную группу.

1.1. Среднее арифметическое (выборочная средняя)

- Еще одно свойство выборочной средней состоит в том, что сумма расстояний от среднего арифметического до объектов, имеющих большее значение, равна сумме расстояний до объектов, имеющих меньшее значение.
- Можно ее использовать только для шкал, где вычисление расстояний между объектами имеет смысл, то есть **для числовых шкал**.

1.1. Среднее арифметическое (выборочная средняя)

- **Например:**
- Рассмотрим среднюю месячную зарплату работников некоторого предприятия. Пусть, например, в фирме работает 20 человек, зарплата 19 из них составляет 10 000 рублей, а зарплата 10-го, руководителя, - 1 000 000 рублей.

- Тогда средняя зарплата одного работника на этой фирме будет равна

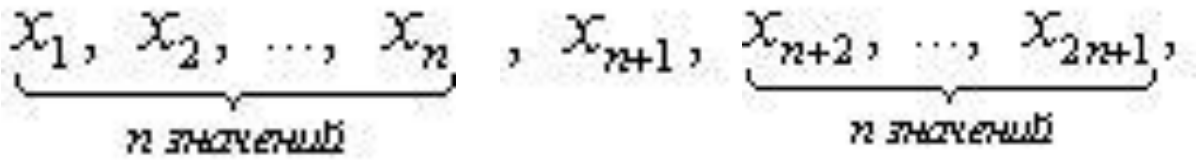
$$\overline{X} = \frac{1000000 + 10000 \cdot 19}{20} = 59500$$

1.2. Медиана

- **Медиана** (обозначается *Md* или *Me*) — это значение, которое делит упорядоченное множество данных пополам, так что одна половина значений оказывается больше медианы, а другая — меньше.
- При нахождении медианы дискретного вариационного ряда следует различать два случая:
  - объем совокупности нечетный;
  - объем совокупности четный.

1.2. Медиана

- Если объем совокупности нечетный и равен $2n + 1$, и варианты размещены в порядке возрастания их значений, то $Me = x_{n+1}$.



1.2. Медиана

- Если количество элементов четное и равно $2n$, то нет варианты, которая бы делила совокупность на две равные по объему части.

$$\underbrace{x_1,\ x_2,\ \dots,\ x_n}_{n\ \text{значений}}\quad \underbrace{x_{n+1},\ \dots,\ x_{2n}}_{n\ \text{значений}}.$$

- В качестве медианы условно берется полусумма вариант, находящихся в середине вариационного ряда:

$$Me = \frac{x_n + x_{n+1}}{2}.$$

1.2. Медиана

- Ранее рассматривался пример с зарплатой работников некоторой фирмы, в которой работает 20 человек, зарплата 19 из них составляет 10 000 рублей, а зарплата 20-го, руководителя, - 1 000 000 рублей.
- Средняя зарплата одного работника на этой фирме составляет 59 500 рублей.
- Медиана данной совокупности равна 10 000 рублей. Она лучше характеризует совокупность, состоящую из размеров зарплат работников фирмы.

1.2. Медиана

- Вычисление медианы имеет следующие преимущества:
  - она мало чувствительна к выбросам
  - ее возможно вычислять не только для метрических данных, но и для данных, измеренных в ранговой шкале

1.3. Мода

- **Мода** - это такое значение признака, которое встречается наиболее часто. В случае дискретных рядов вычислить моду нетрудно. Достаточно найти варианту, которая имеет наибольшую частоту или относительную частоту, это и будет мода.
- Обозначается символом ***Mo***.
- Если все значения в группе встречаются одинаково часто, то мода отсутствует.
- **Например:** в группе (1, 1, 2, 2, 13, 13) моды нет.

1.3. Мода

- Когда два соседних значения имеют одинаковые частоты и они больше частоты любого другого значения, мода есть среднее этих двух значений.
- **Например:** в группе (1, 2, 2, 5, 5, 5, 6, 6, 6, 9, 9, 10) мода равна 5,5.

- Если два несмежных значения в группе имеют равные частоты и они больше частот любого другого значения, то существуют две моды. В этом случае говорят, что группа оценок является *бимодальной*.
- **Например:** в группе (1,4,4,4,7,7,9,9,9,10) модами являются 4 и 9.

1.4. Ограничения при работе с мерами центральной тенденции

- Следует всегда помнить, что меры центральной тенденции отражают адекватно реальную ситуацию, только если мы имеем дело с однородной совокупностью:
  - для выборок, имеющих более чем одну моду, любая мера центральной тенденции, включая среднее, будет недостаточно хороша. Центральной тенденции в таком распределении просто не существует;
  - две одинаковые меры центральной тенденции можно сравнивать, только если они имеют относительно одинаковые распределения. Нельзя сказать, что средние в ряде 20, 20, 20 и 2, 18, 40 равны;
  - нельзя с уверенностью сказать, что среднее показывает нам «типичный» случай, если не знать кривой распределения;
  - Моду незачем вычислять, когда частоты всех наблюдаемых значений почти равны.

1.5. Ограничения при работе с мерами центральной тенденции

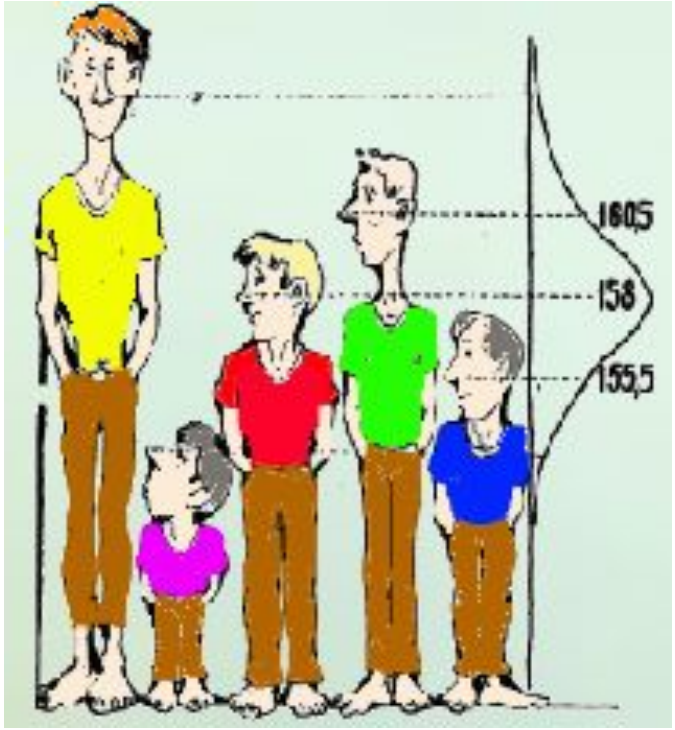
- Выбирая меру центральной тенденции, нужно руководствоваться знанием ее свойств, общей формой распределения и, наконец, здравым смыслом.

- Однажды пятеро мужчин сидели рядом на скамейке парка. Двое были бродягами, имущество которых выражалось в 25 центах. Третий был рабочим, чей счет в банке и другое имущество составляли 2000 долларов. Четвертый владел 15 000 долларов в различных формах. Пятый же был мультимиллионером с чистым доходом 5 000 000 долларов.

- Мода – 25 центов.
- Медиана – 2 000 долларов.
- Среднее – 1 003 400,10 долларов.

§2. Меры изменчивости

- Чтобы определить, насколько хорошо та или иная мера центральной тенденции выражает «типичного» представителя совокупности, следует воспользоваться какой-либо мерой изменчивости, разброса.
- К мерам разброса относят
  - размах,
  - квартильный размах,
  - дисперсию,
  - среднеквадратическое и стандартное отклонение,
  - коэффициент вариации.



2.1. Размах

- *Размах* просто измеряет на числовой шкале расстояние, в пределах которого изменяются оценки.
- Поскольку существуют несколько иные определения размаха, то надо разграничить два его типа: *включающий* и *исключающий*.

2.1. Размах

- *Исключающий размах* — это разность максимального и минимального значений в выборке.
- **Например:**
  - исключающий размах значений 0, 2, 3, 5, 8 равен 8 – 0 = 8.
  - Значения: –0,2; 0,4; 0,8; 1,6 имеют исключающий размах, равный

    1,6 – (–0,2) = 1,8.

2.1. Размах

- *Включающий размах* — это разность между *естественной верхней границей* интервала, содержащего максимальное значение, и *естественной нижней границей* интервала, включающего минимальное значение.
- **Например**,
  - рост пяти мальчиков измеряется с точностью до ближайшего сантиметра.
  - Получены следующие значения: 150, 155, 157, 165, 168 см.
  - Фактический рост самого низкого мальчика находится где-то между 149,5 и 150 см и действительная нижняя граница равна 149,5 см.
  - Верхняя граница интервала, содержащего максимальное значение, составляет 168,5 см.
  - Таким образом, включающий размах равен разности 168,5 – 149,5 = 19, которая на единицу больше, чем 168 – 150.

2.2. Квартильный размах

- **Кванти́ль** в математической статистике - такое число, что заданная случайная величина не превышает его с фиксированной вероятностью.
  - 0,25-квантиль называется **первым (или нижним) квартилем**;
  - 0,5-квантиль называется **медианой** или **вторым квартилем**;
  - 0,75-квантиль называется **третьим (или верхним) квартилем**.
- Таким образом, квартили – это значения признака, делящие ранжированную совокупность на четыре равновеликие части.

2.2. Квартильный размах

- **Квартильный размах** – это интервал, в котором вокруг медианы сосредоточилось 50% значений.
- Он равен разности значений верхнего и нижнего квартиля.
- Термин был впервые использован Гальтоном в 1882 г. Это единственная мера вариации для порядковых шкал

2.3. Дисперсия

- *Размах* представляет собой меру рассеяния, разброса, неоднородности или изменчивости.
- Эта величина возрастает с ростом рассеяния и уменьшением однородности.
- Так же как и для моды и медианы, в ходе вычисления этой меры не учитывается каждое отдельное значение.
- Поэтому необходима другая мера, при вычислении которой, как и для среднего, используется каждая оценка. Такая мера изменчивости называется *дисперсией*.

- *Выборочная дисперсия* – это средняя арифметическая квадратов отклонений вариант от выборочной средней
- Обозначается D_B
- Вычисляется по формуле:

$$D_B = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{X_B}\right)^2 n_i$$

(для группированной выборки)

$$D_B = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X_B}\right)^2$$

(для негруппированной выборки).

- Выборочная дисперсия описывает разброс вариант относительно выборочной средней и характеризует точность измерений.
- Выборочная дисперсия всегда положительна.

2.3. Дисперсия

- *Исправленная выборочная дисперсия*:

$$S^2 = \frac{n}{n-1} D_B$$

- Чаще всего вычисляют сразу исправленную дисперсию по формуле:

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}.$$

2.3. Дисперсия

- Ценность дисперсии заключается в том, что, являясь мерой варьирования числовых значений признака вокруг его среднего значения, она измеряет внутреннюю изменчивость значений признака, зависящую от разностей между наблюдениями.
- Преимущество дисперсии перед другими показателями вариации состоит также и в том, что она разлагается на составные компоненты, позволяя тем самым оценивать влияние различных факторов на величину учитываемого признака.

2.4. Среднеквадратическое и стандартное отклонение

- Мерой изменчивости, тесно связанной с дисперсией, является стандартное отклонение.
- *Среднеквадратическое* (*стандартное отклонение),* обозначаемое σ_x (или S_x), определяется как положительное значение квадратного корня из дисперсии (исправленной дисперсии).
- Для определения S_x надо сначала найти исправленную дисперсию, а затем вычислить квадратный корень из нее.

2.5. Коэффициент вариации *Cv*

- Дисперсия и среднее отклонение применимы и для сравнительной оценки одноимённых средних величин.
- В практике же довольно часто приходится сравнивать изменчивость признаков, выраженных разными единицами.
- В таких случаях используют не абсолютные, а относительные показатели вариации.
- Дисперсия и среднее отклонение как величины, выражаемые теми же единицами, что и характеризуемый ими признак, для оценки изменчивости разноимённых величин непригодны.

2.5. Коэффициент вариации *Cv*

- Одним из относительных показателей вариации является коэффициент вариации.
- Этот показатель представляет собой среднее квадратическое отклонение (среднее отклонение), выраженное в процентах от величины среднего значения:

$$Cv = \frac{S_x}{\overline{x}} \cdot 100 \ \ \%.$$

2.5. Коэффициент вариации *Cv*

- Различные признаки характеризуются различными коэффициентами вариации.
- Но в отношении одного и того же признака значение этого показателя *Cv* остаётся более или менее устойчивым и при симметричных распределениях обычно не превышает 50 %.
- При сильно асимметричных рядах распределения коэффициент вариации может достигать 100 % и даже выше.
- Варьирование считается
  - слабым, если *Cv* не превосходит 10 %,
  - средним, когда *Cv* составляет 11—25 %,
  - значительным при *Cv* > 25 %.

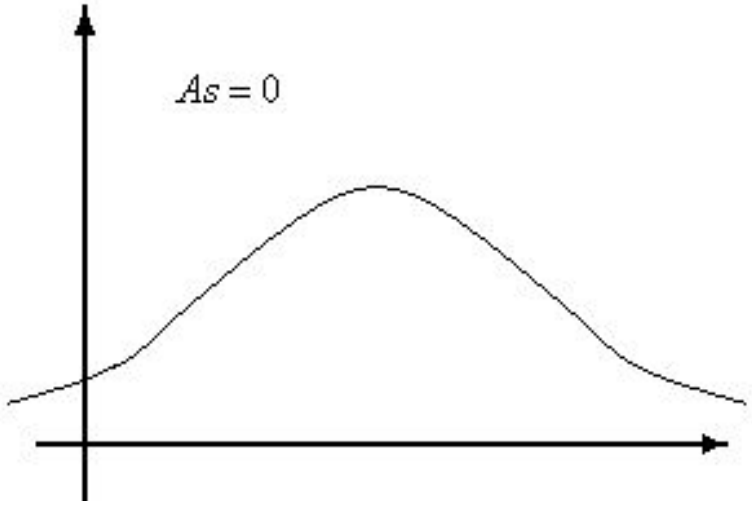
§3. Показатели формы кривой распределения

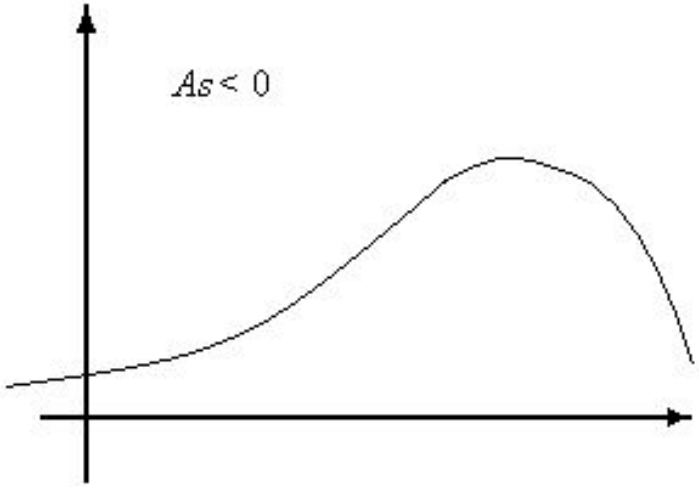
3.1. Асимметрия

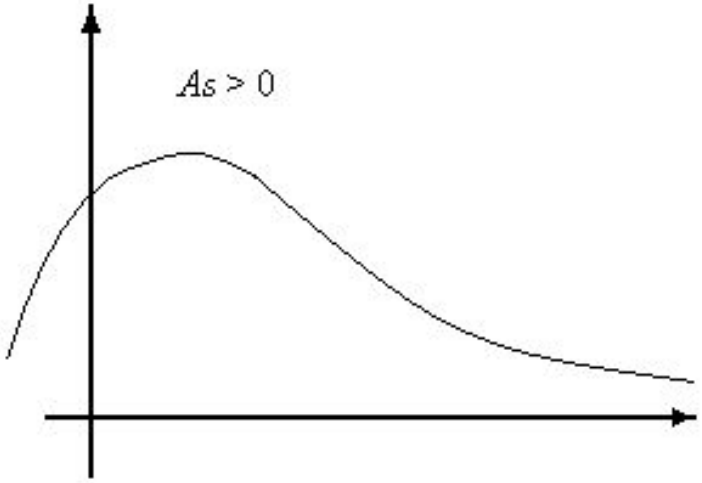
- Одно из наиболее важных свойств распределения частот – степень асимметрии.
- Практически точно симметричные полигоны частот и гистограммы почти никогда не встречаются.
- Степень асимметрии распределения частот для выборки называется его *асимметрией*.
- Легко выявить и распознать асимметрию, если рассматривать полигон частот или гистограмму, но это не всегда возможно или удобно.
- Поэтому изобретены различные обобщенные статистические характеристики, оценивающие вид и степень асимметрии группы наблюдений.
- Наилучшая мера асимметрии (*As*) для группы данных выражается формулой

$$As = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3 / n}{S_x^3}.$$

3.1. Асимметрия





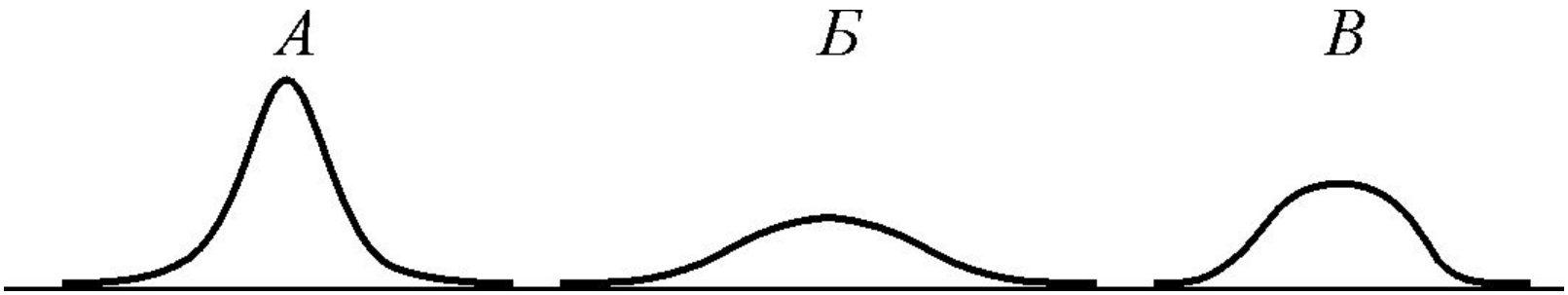


3.2. Эксцесс

- Статистики описывают три свойства или особенности выборок:
  - центральную тенденцию
  - изменчивость
  - симметрию
- Четвертое свойство завершает набор особенностей распределений, представляющих интерес при анализе данных.
- Иногда важно получить представление о том, являются ли полигон частот или гистограмма островершинными или плоскими.
- *Эксцесс* — греческое слово, обозначающее свойство «остроконечности» кривой. (Карл Пирсон формализовал понятие «эксцесс» в статистике и предложил метод его оценки.)

2.8. Эксцесс

- Первая (*А*) является совсем острой: подобная кривая называется *островершинной*.
- Вторая (*Б*) — сравнительно плоская: такие кривые называются *плосковершинными*.
- «Островершинность», или степень эксцесса, третьей кривой (*В*) представляет собой норму, по отношению к которой измеряется эксцесс других кривых.
- Третья кривая — нормальная кривая, которая будет обсуждаться в соответствующей главе; принято говорить, что она является *средневершинной*.



2.8. Эксцесс

- Понятие «эксцесс» применимо лишь к унимодальным распределениям и относится к крутизне кривой в окрестности единственной моды.
- Если распределение имеет две моды, то принято говорить об эксцессе кривой в окрестности каждой моды.
- Обычная мера эксцесса (*Ex*) определяется следующей формулой:

$$Ex = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^4 / n}{s_x^4} - 3.$$

Спасибо за внимание!