Лекция 1. Типы данных и выборочные характеристики

Курбацкий А. Н.

ВШГА МГУ

10 февраля 2020

Содержание

Два ключевых понятия

Исследуя некоторое множество объектов зачастую мы не имеем возможности получить о нём всю информацию. Нам приходится работать только с некоторым его подмножеством, которое, как правило, невелико.

Два ключевых понятия

Исследуя некоторое множество объектов зачастую мы не имеем возможности получить о нём всю информацию. Нам приходится работать только с некоторым его подмножеством, которое, как правило, невелико.

Определение

Генеральная совокупность (population) – *вся интересующая исследователя совокупность изучаемых объектов.*

Два ключевых понятия

Исследуя некоторое множество объектов зачастую мы не имеем возможности получить о нём всю информацию. Нам приходится работать только с некоторым его подмножеством, которое, как правило, невелико.

Определение

Генеральная совокупность (population) – *вся интересующая исследователя совокупность изучаемых объектов.*

Определение

Выборка, выборочная совокупность (sample) – *некоторая часть генеральной совокупности, отбираемая специальным образом и исследуемая с целью получения выводов о генеральной совокупности.*

Два ключевых понятия

Исследуя некоторое множество объектов зачастую мы не имеем возможности получить о нём всю информацию. Нам приходится работать только с некоторым его подмножеством, которое, как правило, невелико.

Определение

Генеральная совокупность (population) – *вся интересующая исследователя совокупность изучаемых объектов.*

Определение

Выборка, выборочная совокупность (sample) – *некоторая часть генеральной совокупности, отбираемая специальным образом и исследуемая с целью получения выводов о генеральной совокупности.*

В математичской статистике под выборкой $(x_1, \ldots, x_n)$ объёма n из распределения D называется набор из n независимых и одинаково распределённых случайных величин, имеющих распределение D.

Типы выборок

Отбор объектов для исследования должен быть осуществлён так, чтобы мы имели представление о всей генеральной совокупности в миниатюре.

Важно!

Говорят, что выборка должна быть **представительной** *или* **репрезентативной**.

Добиться этого можно грамотным отбором данных. Выделим некоторые типы.

- простой случайный отбор;
- механический отбор;
- стратифицированный отбор;
- серийный.

Типы выборок

Отбор объектов для исследования должен быть осуществлён так, чтобы мы имели представление о всей генеральной совокупности в миниатюре.

Важно!

Говорят, что выборка должна быть **представительной** *или* **репрезентативной**.

Добиться этого можно грамотным отбором данных. Выделим некоторые типы.

- простой случайный отбор;
- механический отбор;
- стратифицированный отбор;
- серийный.

Важно!

Неправильный отбор является причиной многих ошибок и неверных выводов!

Типы данных

Данные измерений бывают двух типов: дискретные и непрерывные.

Определение

Дискретные данные *представляют собой отдельные значения признака, общее число которых конечно либо если бесконечно, то является счётным.*

Типы данных

Данные измерений бывают двух типов: дискретные и непрерывные.

Определение

Дискретные данные *представляют собой отдельные значения признака, общее число которых конечно либо если бесконечно, то является счётным.*

Определение

Непрерывные данные *могут принимать любое значение в некотором интервале числовой прямой.*

Шкалы

Этим типам данных в свою очередь соответсвуют несколько шкал, которые зависят уже от природы исходных данных. Перечислим основные их виды.

- Номинальная шкала[1],
- порядковая шкала,
- интервальная шкала,
- относительная шкала.

[1] в частности, бинарная (дихотомическая) шкала

Шкалы

Этим типам данных в свою очередь соответсвуют несколько шкал, которые зависят уже от природы исходных данных. Перечислим основные их виды.

- Номинальная шкала[1],
- порядковая шкала,
- интервальная шкала,
- относительная шкала.

Замечание

В эконометрике данные дополнительно разбиваются в зависимости от их структуры.

[1] в частности, бинарная (дихотомическая) шкала

Шкалы для дискретных данных

Определение

Номинальная шкала состоит из названий или категорий для сортировки или классификации объектов по некоторому признаку.

Шкалы для дискретных данных

Определение

Номинальная шкала состоит из названий или категорий для сортировки или классификации объектов по некоторому признаку.

Пример

Примерами номинальной шкалы служат семейное положение, профессия, страна проживания, оператор связи.

Номинальная шкала, которая состоит из двух категорий, называется дихотомической или **бинарной**.

Шкалы для дискретных данных

Определение

Номинальная шкала состоит из названий или категорий для сортировки или классификации объектов по некоторому признаку.

Пример

Примерами номинальной шкалы служат семейное положение, профессия, страна проживания, оператор связи.

Номинальная шкала, которая состоит из двух категорий, называется дихотомической или **бинарной**.

Определение

Порядковая шкала означает, что числа присваиваются объектам, чтобы обозначить относительные позиции объектов.

Пример

Воинское звание, учёная степень, итоговые места спортсменов.

Шкалы для непрерывных данных

Определение

Интервальная шкала позволяет указать количественное значение измеряемого признака и находить разницу между двумя величинами. Недостатком служит отсутствие абсолютного нуля в качестве точки отсчета.

Шкалы для непрерывных данных

Определение

Интервальная шкала позволяет указать количественное значение измеряемого признака и находить разницу между двумя величинами. Недостатком служит отсутствие абсолютного нуля в качестве точки отсчета.

Шкала времени, например, может быть разделена на годы, каждый год разделен на дни, дни на часы и далее.

Определение

Относительная шкала обладает абсолютным нулем в качестве точки отсчета.

Для данных этой шкалы осмысленными являются все операции, включая вычитание и деление.

Содержание

Вариационный ряд

Описательная статистика занимается начальным анализом данных.

Первым шагом в анализе данных для нас будет их упорядочивание и разбиение на группы.

Определение

Упорядоченные по возрастанию значения выборки называются **вариационным рядом (set of order statistic)**.

Вариационный ряд

Описательная статистика занимается начальным анализом данных.

Первым шагом в анализе данных для нас будет их упорядочивание и разбиение на группы.

Определение

Упорядоченные по возрастанию значения выборки называются **вариационным рядом (set of order statistic)**.

Группы, на которые разбивается множество значений будем называть **интервалами группировки** .

Важно!

Пусть мы упорядочили наши n наблюдений $x_1, \ldots, x_n$. *Они лежат в некотором интервале, который мы разбиваем еще на m интервалов. Последние и называются интервалами группировки. Их длины обозначим через* $\Delta_1, \ldots, \Delta_m$, *а середины интервалов группировки - через* $c_1, \ldots, c_m$.

Ранжирование

Что делать, когда признаки объектов наблюдения не являются количественными или их численные значения указывают только на порядок?

Ранжирование

Что делать, когда признаки объектов наблюдения не являются количественными или их численные значения указывают только на порядок?

Определение

Рангом наблюдения называется порядковый номер наблюдения в вариационном ряду. Если значения наблюдаемых величин повторяются, то каждому из этих значений (наблюдений), присваивается одинаковый ранг, равный среднему арифметическому номеров занимаемых мест.

Ранжирование

Что делать, когда признаки объектов наблюдения не являются количественными или их численные значения указывают только на порядок?

Определение

Рангом наблюдения называется порядковый номер наблюдения в вариационном ряду. Если значения наблюдаемых величин повторяются, то каждому из этих значений (наблюдений), присваивается одинаковый ранг, равный среднему арифметическому номеров занимаемых мест.

Переход от самих наблюдений к последовательности их рангов называется *ранжированием* .

Графическое представление данных

- Графические изображения дают возможность сразу получить представление о поведении и распределении данных.

Графическое представление данных

- Графические изображения дают возможность сразу получить представление о поведении и распределении данных.

- Базовыми графическими инструментами представления данных являются гистограммы, полигоны и кумуляты (накопительные гистограммы). Рассмотрим их по порядку.

Гистограмма

Графическое изображение числа наблюдений n_i выборки, соответствующих каждому интервалу, называется **гистограммой** выборки.

Важно!

По горизонатльной оси откладываются значения наблюдаемой величины, по вертикальной – частота их появления.

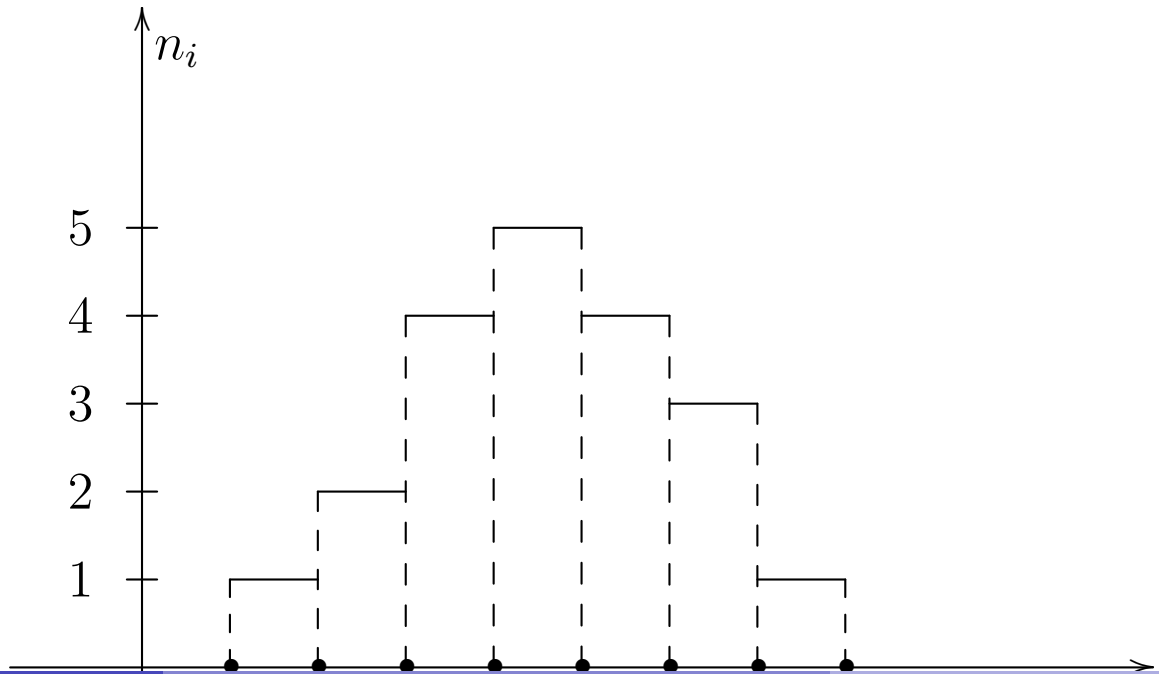
Гистограмма

Графическое изображение числа наблюдений n_i выборки, соответствующих каждому интервалу, называется **гистограммой** выборки.

Важно!

По горизонатльной оси откладываются значения наблюдаемой величины, по вертикальной – частота их появления.

Изобразим это на графике:

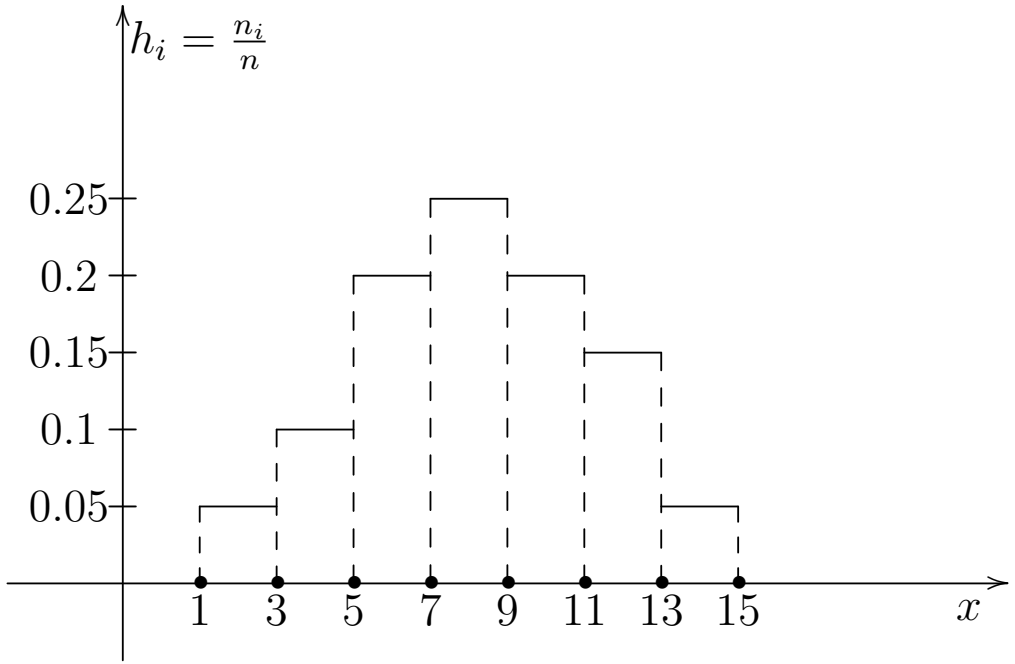


Гистограмма частот

Это графическое изображение зависимости частоты $h_i = \frac{n_i}{n}$ попадания элементов выборки от соответствующего интервала.

Гистограмма частот

Это графическое изображение зависимости частоты $h_i = \frac{n_i}{n}$ попадания элементов выборки от соответствующего интервала.



Важно!

Такую гистограмму ещё называют **гистограммой относительных частот** *. Отличие гистограммы относительных частот от гистограммы состоит в том, что на оси y вместо количества наблюдений на данном интервале отмечены их доли (или процент) от общего числа.*

Гистограмма частот

Чтобы каждый раз не думать о том, какой длины выбирать интервал группировки, можно пользоваться формулой Стерджеса $m \approx 1 + \log_2 n$. Длина каждого интервала будет равна $\Delta = \frac{x_{\max} - x_{\min}}{m}$.

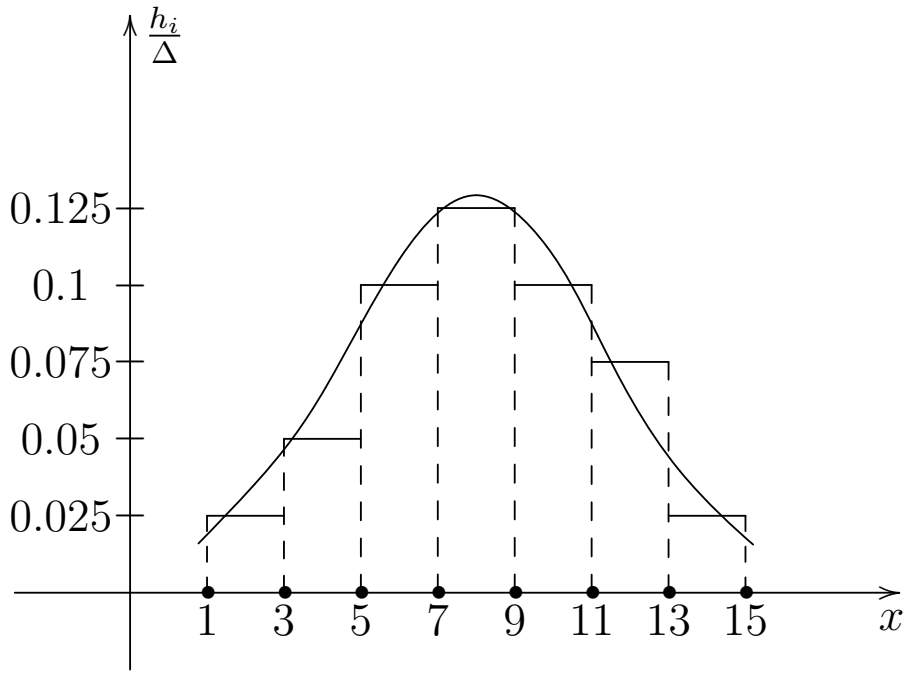
Гистограмма частот

Чтобы каждый раз не думать о том, какой длины выбирать интервал группировки, можно пользоваться формулой Стерджеса $m \approx 1 + \log_2 n$. Длина каждого интервала будет равна $\Delta = \frac{x_{\max} - x_{\min}}{m}$.

Можно избавиться от влияния размера интервала группировки, поделив частоты h_j на соответствующие длины Δ_j. В таком случае площадь фигуры под гистограммой становится равной единице и поэтому её можно назвать эмпирической функцией плотности.

Гистограмма частот(график)

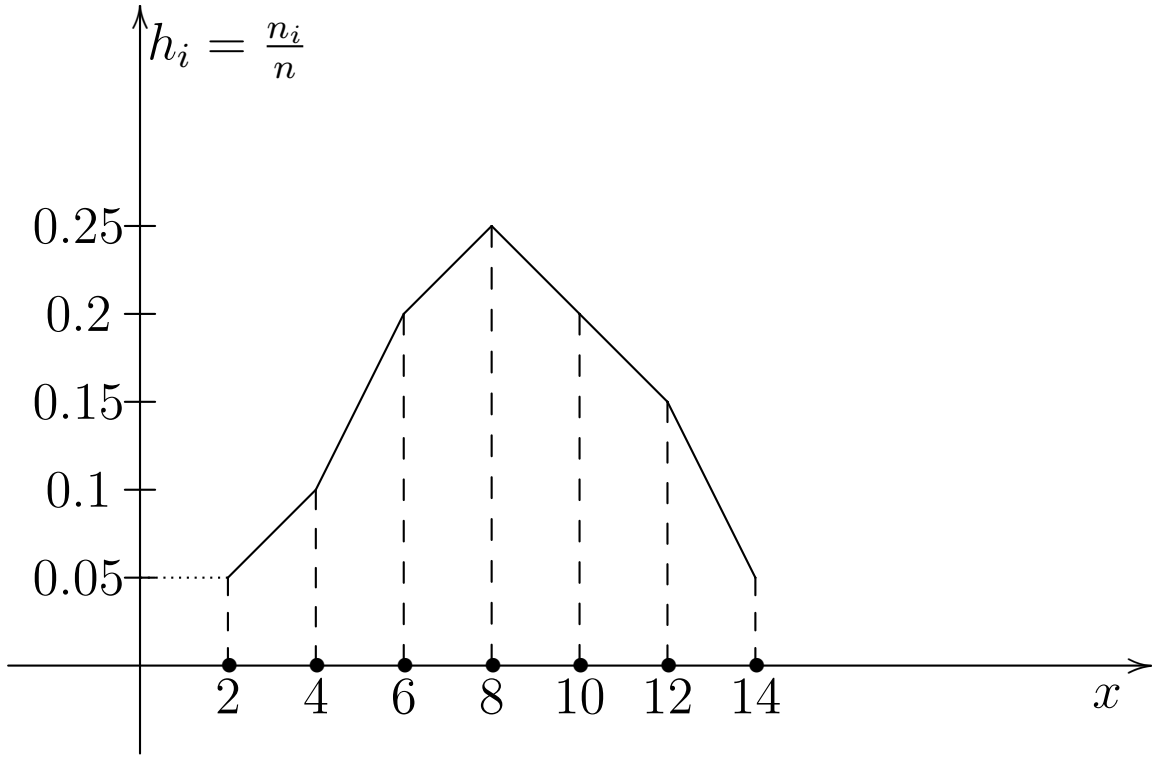
Изобразим это на графике вместе со сглаженной гистограммой, которую также часто рисуют, чтобы лучше представлять, какому непрерывному распределению приблизительно соответствует распределение относительных частот[2]:



[2]Если плотность распределения элементов выборки является непрерывной функцией и количество k интервалов группировки стремится к бесконечности таким образом, что $\frac{k}{n} \to 0$, то имеет место сходимость по вероятности гистограммы к плотности в каждой точке.

Полигон

Несколько иное графическое представление данных дает *полигон* . Полигон строится в виде области, ограниченной линией, которая проходит через точки $(c_i; h_i)$, где c_i - середина интервала, а h_i- частота.

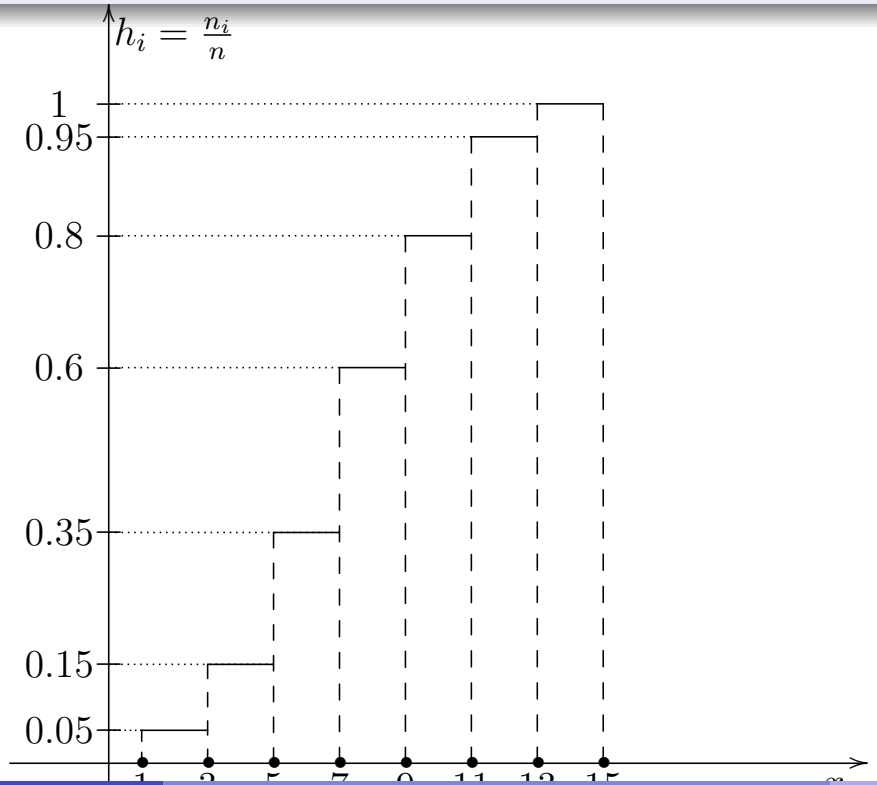


Накопительная гистограмма

И ещё одно ключевое графическое изображение данных - это *кумулята* или *накопительная гистограмма* .

Определение

Графическое изображение зависимости накопленных частот $\omega_i = \sum_{j=1}^{i} h_j$ *называется* **кумулятой** *выборки.*



Эмпирическая функция распределения

Определение

Эмпирической функцией распределения случайной величины, построенной по выборке $x_1, \ldots, x_n$, называется функция $F_n(x)$, которая равна доле таких значений x_i, для которых $x_i \leq x$.

То есть $F_n(x) = n_x/n$, где n_x - число наблюдений меньших или равных x, а n - объем выборки.

Эмпирическая функция распределения

Определение

Эмпирической функцией распределения случайной величины, построенной по выборке $x_1, \ldots, x_n$, называется функция $F_n(x)$, которая равна доле таких значений x_i, для которых $x_i \leq x$.

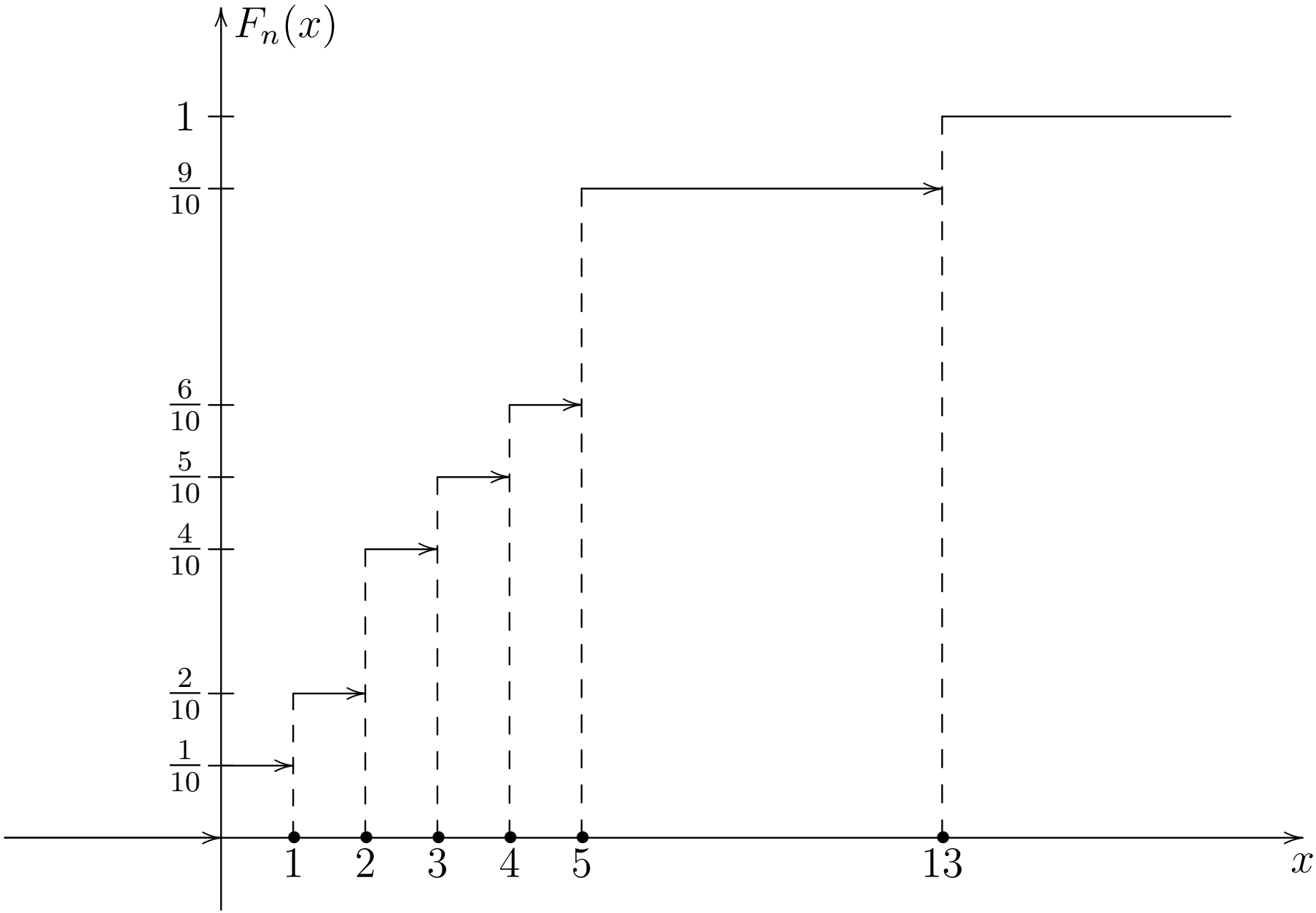
То есть $F_n(x) = n_x/n$, где n_x - число наблюдений меньших или равных x, а n - объем выборки.

Теорема

С ростом объема выборки эмпирическая функция распределения приближается к теоретической функции распределения, более точно $\lim\limits_{n \to +\infty} P(\sup |F_n(x) - F(x)| = 0) = 1.$

Этот замечательный факт доставляет нам теорема Гливенко-Кантелли.

Пример


$F_n(x)$
1
$\frac{9}{10}$
$\frac{6}{10}$
$\frac{5}{10}$
$\frac{4}{10}$
$\frac{2}{10}$
$\frac{1}{10}$
1
2
3
4
5
13
x

Свойства ЭФР

Свойства эмпирической функции распределения аналогичны свойствам произвольной функции распределения:

1. $0 \leq F_n(x) \leq 1$.
2. $F_n(x)$ - неубывающая функция.
3. $F_n(x)$ непрерывна справа.
4. $F_n(x) = 0$ при $x < x_{min}$ и $F_n(x) = 1$ при $x \geq x_{max}$.

Содержание

Мода

Наша текущая задача состоит в выборе одного числа, которое можно было бы назвать центральным значением для набора данных.

Мода

Наша текущая задача состоит в выборе одного числа, которое можно было бы назвать центральным значением для набора данных.

Определение

Мода Mo – наиболее часто встречающееся значение в выборке.

Мода

Наша текущая задача состоит в выборе одного числа, которое можно было бы назвать центральным значением для набора данных.

Определение

Мода Mo – наиболее часто встречающееся значение в выборке.

Мода может быть не одна!

Мода

Наша текущая задача состоит в выборе одного числа, которое можно было бы назвать центральным значением для набора данных.

Определение

Мода Mo – наиболее часто встречающееся значение в выборке.

Мода может быть не одна!
В выборке $1, 3, 4, -1, 2, 3, 5, 4$ есть две моды 3 и 4. В таком случае распределение будет называться *бимодальным*.

Пример

В результате независимых наблюдений случайной величины были получены следующие ее значения: мегафон, билайн, теле2, теле2, мтс, мтс, теле2. Укажите количество мод данной выборки.

Медиана

Еще одна характеристика среднего - это **медиана** (оценка медианы), которая определяется как значение, которое делит упорядоченную выборку пополам по количеству наблюдений.

Медиана

Еще одна характеристика среднего - это **медиана** (оценка медианы), которая определяется как значение, которое делит упорядоченную выборку пополам по количеству наблюдений.

Важно!

Для нечетного числа наблюдений медиана есть просто центральное наблюдение $x_{(n+1)/2}$. *Для четного числа наблюдений медиана - это среднее арифметическое двух соседних центральных наблюдений* $x_{\frac{n}{2}}$ *и* $x_{\frac{n}{2}+1}$.

Пример

Рассмотрим выборку $1, 0, 3, 6, -1, 2, 7, 5, 4$.

Медиана

Еще одна характеристика среднего - это **медиана** (оценка медианы), которая определяется как значение, которое делит упорядоченную выборку пополам по количеству наблюдений.

Важно!

Для нечетного числа наблюдений медиана есть просто центральное наблюдение $x_{(n+1)/2}$. *Для четного числа наблюдений медиана - это среднее арифметическое двух соседних центральных наблюдений* $x_{\frac{n}{2}}$ *и* $x_{\frac{n}{2}+1}$.

Пример

Рассмотрим выборку $1, 0, 3, 6, -1, 2, 7, 5, 4$.
Выпишем её вариационный ряд $-1, 0, 1, 2, 3, 4, 5, 6, 7$.

Медиана

Еще одна характеристика среднего - это **медиана** (оценка медианы), которая определяется как значение, которое делит упорядоченную выборку пополам по количеству наблюдений.

Важно!

Для нечетного числа наблюдений медиана есть просто центральное наблюдение $x_{(n+1)/2}$. *Для четного числа наблюдений медиана - это среднее арифметическое двух соседних центральных наблюдений* $x_{\frac{n}{2}}$ *и* $x_{\frac{n}{2}+1}$.

Пример

Рассмотрим выборку $1, 0, 3, 6, -1, 2, 7, 5, 4$.
Выпишем её вариационный ряд $-1, 0, 1, 2, 3, 4, 5, 6, 7$.
Объем выборки равен 9, поэтому медиана - это просто центральный (пятый) элемент в выборке $Me = x_{(9+1)/2} = x_5 = 3$.

Среднее

Наиболее распространённой характеристикой безусловного математического ожидания при работе с числовыми данными является **среднее арифметическое**.

Определение

Среднее значение выборки объема n вычисляется по формуле:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

[3] как и мода с медианой

Среднее

Наиболее распространённой характеристикой безусловного математического ожидания при работе с числовыми данными является **среднее арифметическое**.

Определение

Среднее значение выборки объема n вычисляется по формуле:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

[3] как и мода с медианой

Среднее

Наиболее распространённой характеристикой безусловного математического ожидания при работе с числовыми данными является **среднее арифметическое**.

Определение

Среднее значение выборки объема n вычисляется по формуле:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Среднее значение[3], сами по себе малоценны в качестве информации о выборке. Примером может служить средняя температура по больнице. Необходимы и характеристики разброса данных.

[3] как и мода с медианой

Содержание

Размах

Простейшей мерой разброса является **размах** (range).
Размах - это разность между минимальным и максимальным значениями выборки, то есть $x_{\max} - x_{\min}$.

Размах

Простейшей мерой разброса является **размах** (range).
Размах - это разность между минимальным и максимальным значениями выборки, то есть $x_{\max} - x_{\min}$.

Пример

В результате независимых наблюдений случайной величины были получены следующие ее значения: -1, 2, 4, 6, 5, 7, 1, 4, 0, 2. Чему равен размах?

Решение

Минимальный элемент равен -1, а максимальный равен 7. Значит, размах равен $7 - (-1) = 8$.

Размах

Простейшей мерой разброса является **размах** (range).
Размах - это разность между минимальным и максимальным значениями выборки, то есть $x_{\max} - x_{\min}$.

Пример

В результате независимых наблюдений случайной величины были получены следующие ее значения: -1, 2, 4, 6, 5, 7, 1, 4, 0, 2. Чему равен размах?

Решение

Минимальный элемент равен -1, а максимальный равен 7. Значит, размах равен $7 - (-1) = 8$.

Чтобы ввести ещё одну меру разброса нам потребуется определить понятие выборочной квантили.

Выборочная квантиль

Определение

Выборочной квантилью x_p называется решение уравнения $F_n(x) = p$, где $F_n(x)$ - это эмпирическая функция распределения.

Смысл квантили состоит в том, что левее точки x_p лежит приблизительно $100p\%$ наблюдений.

Выборочная квантиль

Определение

Выборочной квантилью x_p называется решение уравнения $F_n(x) = p$, где $F_n(x)$ - это эмпирическая функция распределения.

Смысл квантили состоит в том, что левее точки x_p лежит приблизительно $100p\%$ наблюдений.

Наиболее используемыми в описательной статистике являются

- квантиль $x_{0.5}$, называемая медианой;
- квантиль $x_{0.25}$, называемая нижней квартилью;
- квантиль $x_{0.75}$, называемая верхней квартилью;
- квантили $x_{0.1}$, $x_{0.2}$, $x_{0.3}$, $x_{0.4}$, $x_{0.5}$, $x_{0.6}$, $x_{0.7}$, $x_{0.8}$, $x_{0.9}$, называемые децилями.

Выборочная квантиль

Определение

Выборочной квантилью x_p называется решение уравнения $F_n(x) = p$, где $F_n(x)$ - это эмпирическая функция распределения.

Смысл квантили состоит в том, что левее точки x_p лежит приблизительно $100p\%$ наблюдений.

Наиболее используемыми в описательной статистике являются

- квантиль $x_{0.5}$, называемая медианой;
- квантиль $x_{0.25}$, называемая нижней квартилью;
- квантиль $x_{0.75}$, называемая верхней квартилью;
- квантили $x_{0.1}$, $x_{0.2}$, $x_{0.3}$, $x_{0.4}$, $x_{0.5}$, $x_{0.6}$, $x_{0.7}$, $x_{0.8}$, $x_{0.9}$, называемые децилями.

А ещё есть перцентили - это квантили $x_{0.01}$, $x_{0.02}$,. . .,$x_{0.99}$.

Выборочная квантиль

Уравнение $F_n(x) = p$ не всегда однозначно разрешимо! Поэтому ...

Важно!

Выборочная квантиль порядка p $(0 < p < 1)$ *равна* $X_{([pn]+1)}$.

При ручном счёте часто используют другие формулы[4]! Например, медиану мы уже ввели и не так, как здесь.

[4] А в MS Excel функция КВАРТИЛЬ часто может давать совсем не то, что получается по нашему правилу! Как так?

Выборочная квантиль

Уравнение $F_n(x) = p$ не всегда однозначно разрешимо! Поэтому ...

Важно!

Выборочная квантиль порядка p $(0 < p < 1)$ *равна* $X_{([pn]+1)}$.

При ручном счёте часто используют другие формулы[4]! Например, медиану мы уже ввели и не так, как здесь.

С квартилями при ручном счёте будем поступать следующим образом:

- сначала находится медиана, которая разбивает выборку на две равные подвыборки;
- для каждой из подвыборок ищем медианы и называем их верхней и нижней квартилью.

[4] А в MS Excel функция КВАРТИЛЬ часто может давать совсем не то, что получается по нашему правилу! Как так?

Выборочная квантиль

Уравнение $F_n(x) = p$ не всегда однозначно разрешимо! Поэтому ...

Важно!

Выборочная квантиль порядка p $(0 < p < 1)$ *равна* $X_{([pn]+1)}$.

При ручном счёте часто используют другие формулы[4]! Например, медиану мы уже ввели и не так, как здесь.

С квартилями при ручном счёте будем поступать следующим образом:

- сначала находится медиана, которая разбивает выборку на две равные подвыборки;
- для каждой из подвыборок ищем медианы и называем их верхней и нижней квартилью.

Замечание

Если выборка **нечётная**, то медиана включается в нижнюю и верхнюю подвыборки. Данными не разбрасываемся!

[4] А в MS Excel функция КВАРТИЛЬ часто может давать совсем не то, что получается по нашему правилу! Как так?

Межквартильный размах

Ещё одна мера вариации данных называется **межквартильным размахом**.

Определение

Межквартильный размах d - это разность между верхней и нижней квартилями, то есть $d = Q_{0.75} - Q_{0.25}$. Иногда используется обозначение IR (interquartile range).

Межквартильный размах

Ещё одна мера вариации данных называется **межквартильным размахом**.

Определение

Межквартильный размах d - это разность между верхней и нижней квартилями, то есть $d = Q_{0.75} - Q_{0.25}$. Иногда используется обозначение IR (interquartile range).

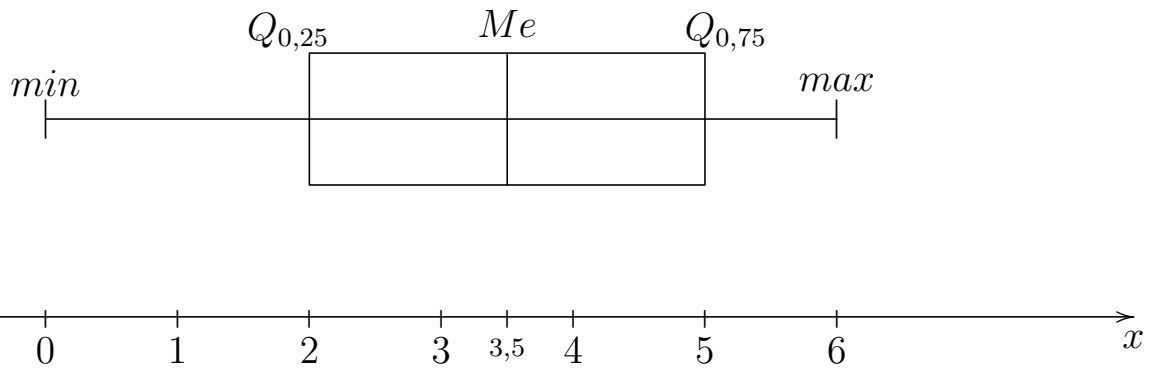
В отличие от размаха, который полностью игнорирует распределение данных между минимальным и максимальным элементами, межквартильный размах показывает, где расположены 50% центральных данных. Крайние же значения выпадают из обозрения.

Коробчатая диаграмма (boxplot)

Коробчатая диаграмма представляет собой необычный рисунок, так называемый, "ящик с усами"[5]:

- отрезок прямой от минимального до максимального значения;
- ящик, в котором заключены 50% наблюдений между нижней и верхней квартилью, с отмеченной медианой;
- иногда особо выделяют выбросы, то есть такие значения $x \notin [Q_{0.25} - 1.5d; Q_{0.75} + 1.5d]$.

Коробчатая диаграмма



[5] В описательной статистике можно встретить и другие диаграммы, например, точечные диаграммы (dot plot) и стебель с листьями (stem and leaf plot).

Дисперсия и стандартное отклонение

Когда речь идет о так называемых параметрических методах статистики, то на первый план среди различных мер разброса данных выходят выборочные *дисперсия* и *стандартное отклонение*.

Дисперсия и стандартное отклонение

Когда речь идет о так называемых параметрических методах статистики, то на первый план среди различных мер разброса данных выходят выборочные *дисперсия* и *стандартное отклонение*.

Определение

Выборочная дисперсия вычисляется по формуле

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

а выборочное стандартное отклонение - это корень из дисперсии.

Дисперсия и стандартное отклонение

Когда речь идет о так называемых параметрических методах статистики, то на первый план среди различных мер разброса данных выходят выборочные *дисперсия* и *стандартное отклонение*.

Определение

Выборочная дисперсия вычисляется по формуле

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

а выборочное стандартное отклонение - это корень из дисперсии.

Зачем извлекать корень, может, лучше прологарифмировать? Почему бы нам не взять просто отклонения от среднего $\sum(x_i - \bar{x})$ или модули отклонений $\sum|x_i - \bar{x}|$? А почему в знаменателе $n-1$, а не n?

Сгруппированные данные

Если среди значений x_i выборки имеется только k различных(то есть каждое из k значений a_j повторяется n_j раз), то обозначим частоту значения a_j через $f_j = \frac{n_j}{n}$. Тогда формулы для среднего и дисперсии могут быть записаны в виде:

Определение

Формулы среднего и дисперсии для сгруппированных данных

$$\bar{x} = \sum_{j=1}^{k} f_j a_j.$$

$$s^2 = \frac{n}{n-1} \sum_{j=1}^{k} f_j (a_j - \bar{x})^2.$$

Асимметрия и эксцесс

Это две характеристики, которыми часто руководствуются, чтобы делать вывод о соответствии данных некоторому распределению.

Определение

Коэффициент асимметрии *характеризует симметричность в распределении наблюдений и равен* $As = \dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$.

Асимметрия и эксцесс

Это две характеристики, которыми часто руководствуются, чтобы делать вывод о соответствии данных некоторому распределению.

Определение

Коэффициент асимметрии *характеризует симметричность в распределении наблюдений и равен* $As = \dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}$.

Определение

Коэффициент эксцесса *характеризует вероятность появления больших (по модулю) значений и равен* $Kurt = \dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\right)^4}$.

То есть это оценки для третьего и четвёртого центральных нормированных моментов. Есть и другие формулы для их оценивания!

Интерпретация

- Наличие симметрии характеризуется близостью коэффициента асимметрии к нулю.

Интерпретация

- Наличие симметрии характеризуется близостью коэффициента асимметрии к нулю.
- Эксцесс характеризует островершинность распределения, а также частоту появления значений, которые удалены от среднего, то есть насколько много наблюдений находится в "хвостах" распределения.

Интерпретация

- Наличие симметрии характеризуется близостью коэффициента асимметрии к нулю.
- Эксцесс характеризует островершинность распределения, а также частоту появления значений, которые удалены от среднего, то есть насколько много наблюдений находится в "хвостах" распределения.

Важно!

Часто хочется проверить данные на нормальность. Как это сделать? Для нормального распределения коэффициент асимметрии равен нулю, а эксцесс - трем.

Если эксцесс сильно отличается от трёх, то говорят о наличии "тяжёлых хвостов".

Интерпретация

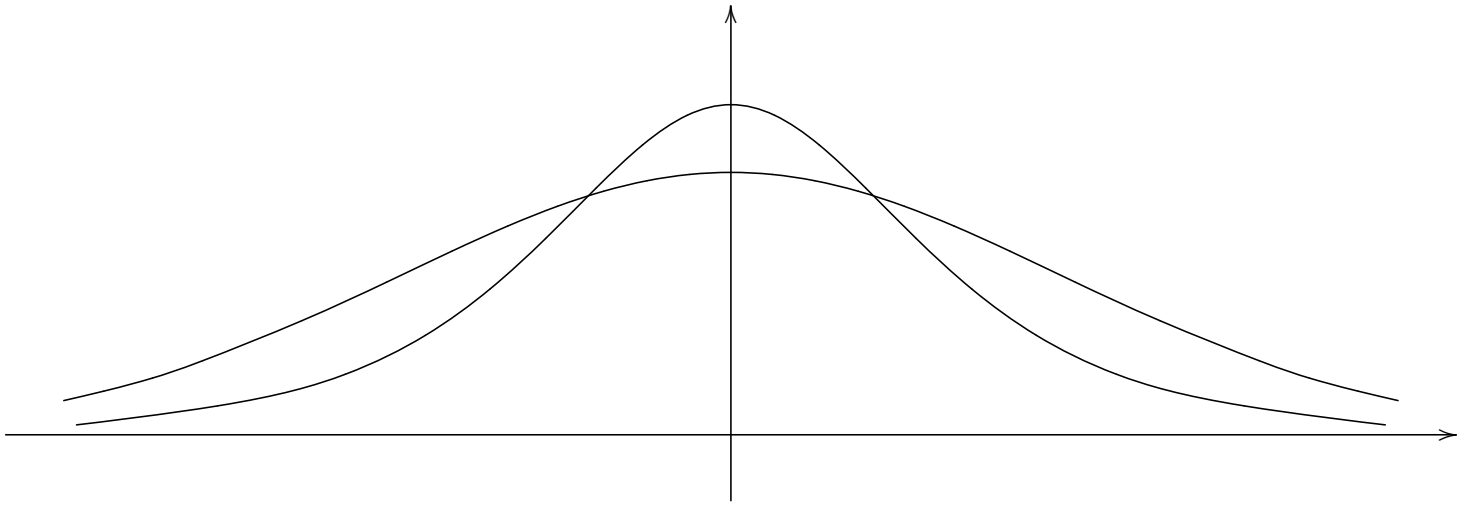
- Наличие симметрии характеризуется близостью коэффициента асимметрии к нулю.
- Эксцесс характеризует островершинность распределения, а также частоту появления значений, которые удалены от среднего, то есть насколько много наблюдений находится в "хвостах" распределения.

Важно!

Часто хочется проверить данные на нормальность. Как это сделать? Для нормального распределения коэффициент асимметрии равен нулю, а эксцесс - трем.

Если эксцесс сильно отличается от трёх, то говорят о наличии "тяжёлых хвостов".
Далее мы узнаем и о других способах проверки на соответствие распределения данных некоторому известному распределению.

ХВОСТЫ



У кого больше хвосты, у того больше вероятность оказаться далеко от МГУ.

Содержание

Где и что почитать?

Тема: Генеральная и выборочная совокупности. Случайные выборки. Виды выборок. Эмпирическая функция распределения. Выборочные характеристики.([И-М], §9-10; [Ф,Л], глава 10).


Ивашев-Мусатов О. С., Теория вероятностей и математическая статистика: учеб. пособие. - 2-е изд., перераб. и доп. - М.: ФИМА, 2003. - 224 с.


Фадеева Л. Н., Лебедев А. В., Теория вероятностей и математическая статистика: учебное пособие. - 2-е изд., перераб. и доп. - М.: Эксмо, 2010. - 496 с. – (Новое экономическое образование).