Описательная статистика в анализе данных

Стати́стика — отрасль знаний, **наука**, в которой излагаются **общие вопросы сбора, измерения, мониторинга, анализа** массовых статистических (количественных или качественных) данных и их **сравнение**; **изучение количественной стороны** массовых явлений в числовой форме.

Слово «статистика» происходит от латинского **status** — состояние и положение дел.

В науку термин «статистика» ввёл немецкий учёный Готфрид Ахенвалль в 1746 году, предложив заменить название курса «Государствоведение», преподававшегося в университетах Германии, на «Статистику», положив тем самым начало развитию статистики как науки и учебной дисциплины.

Несмотря на это, статистический учёт вёлся намного раньше: проводились переписи населения в Древнем Китае, осуществлялось сравнение военного потенциала государств, вёлся учёт имущества граждан в Древнем Риме и тому подобное.

Статистика разрабатывает специальную методологию исследования и обработки материалов: массовые статистические наблюдения, метод группировок, средних величин, индексов, балансовый метод, метод графических изображений, кластерный, дискриминантный, факторный и компонентный анализы, оптимизацию и другие методы анализа статистических данных.

https://habr.com/ru/company/piter/blog/320772/

ph_piter 30 января 2017 в 10:37

Разница между статистикой и наукой о данных

Блог компании Издательский дом «Питер» , Data Mining *, Алгоритмы *, Big Data *, R *

Мнение: Статистикам весь тренд, связанный с наукой о данных, кажется слегка высокомерным. . . эта сфера деятельности весьма пересекается с той работой, которой статистики занимаются уже не одно десятилетие.

“Думаю, data-scientist – распиаренный синоним для «специалист по статистике»” – заявил *Нейт Сильвер** в 2013 году на лекции в Joint Statistical Meeting.

Брэд Шлумич специалист по data science в Twitch: “**Статистика – важнейшая составляющая науки о данных**. У нас в Twitch команда data science обладает тремя компетенциями: статистика, программирование и понимание продукта. **Мы никогда не взяли бы на работу человека, слабо ориентирующегося в статистике**.”

“Некоторые считают, что наука о данных – это всего лишь прикладная статистика, но мы – определенно не просто статистики. . . Гораздо эффективнее работать, если все одинаково понимают смысл продукта, решают, какие параметры важнее, понимают с точки зрения программиста, как реализовать трекинг, и с точки зрения статистика – как делать анализ. Не понимая, как люди будут пользоваться продуктом, и каковы цели компании, можно исказить весь анализ данных. Задача data scientist'a – держать в голове сразу всю эту информацию и знать, к каким данным обратиться, чтобы ответить на любой нечетко определенный вопрос.

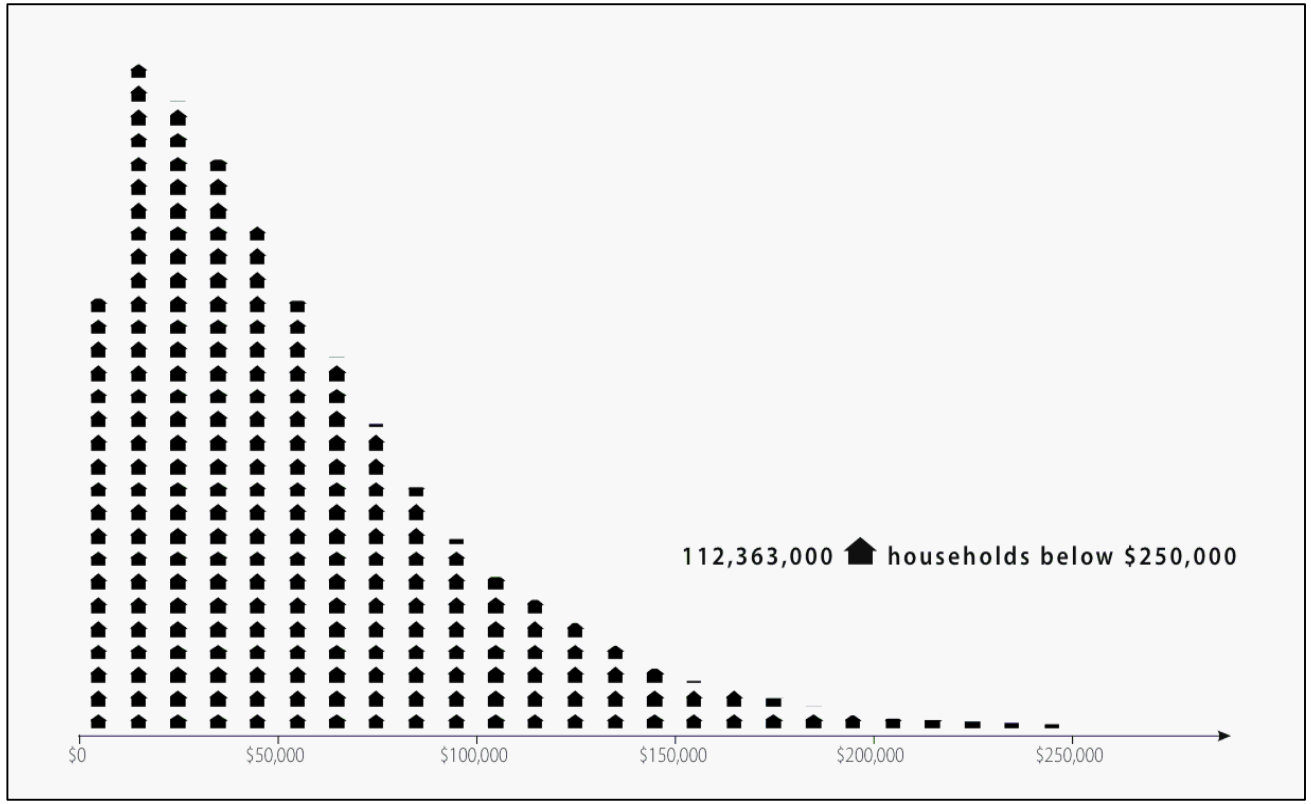
** Нейт Сильвер (Nate Silver) - тот самый человек, который верно спрогнозировал итоги голосования на президентских выборах 2008 года в 49 из 50 штатов США. В 2012 году у него получилось уже 50 из 50.*

Базовые знания статистики крайне полезны в повседневной жизни.

Например, в 2005 году британские СМИ писали о том, что средний уровень дохода населения снизился на 0,2 % по сравнению с предыдущим годом. Некоторые политики даже использовали этот факт, критикуя действующее правительство.

Однако, важно понимать, что среднее арифметическое — хороший показатель, когда наш признак имеет симметричное распределение (богатых столько же, сколько бедных). Реальное же распределение доходов имеет скорее следующий вид:

Распределение имеет явно выраженную асимметрию: очень состоятельных людей заметно меньше, чем представителей среднего класса. Это приводит к тому, что **в данном случае банкротство одного из миллионеров может значительно повлиять на этот показатель.**
Гораздо **информативнее использовать значение медианы для описания таких данных.** И, как ни удивительно, медиана дохода в 2005 году в Великобритании, в отличие от среднего значения, продолжила свой рост.



Крылатая фраза: *Существует три вида лжи: ложь, наглая ложь и статистика*

Таблица 2.3. Количество выпадений каждой цифры (Kansas Pick 3 Lottery, 15 марта 1997 года)

Цифра	Количество выпадений
0	485
1	468
2	513
3	491
4	484
5	480
6	487
7	482
8	475
9	474

Рис. 2.1. Столбиковая диаграмма, показывающая количество выпадений каждой цифры

Таблица 2.4. Процент выпадений каждой цифры

Цифра	Количество выпадений	Процент выпадений
0	485	10,0% = 485/4 839
1	468	9,7% = 486/4 839
2	513	10,6% = 513/4 839
3	491	10,1% = 491/4 839
4	484	10,0% = 484/4 839
5	480	9,9% = 480/4 839
6	487	10,0% = 487/4 839
7	482	10,0% = 482/4 839
8	475	9,8% = 475/4 839
9	474	9,8% = 474/4 839

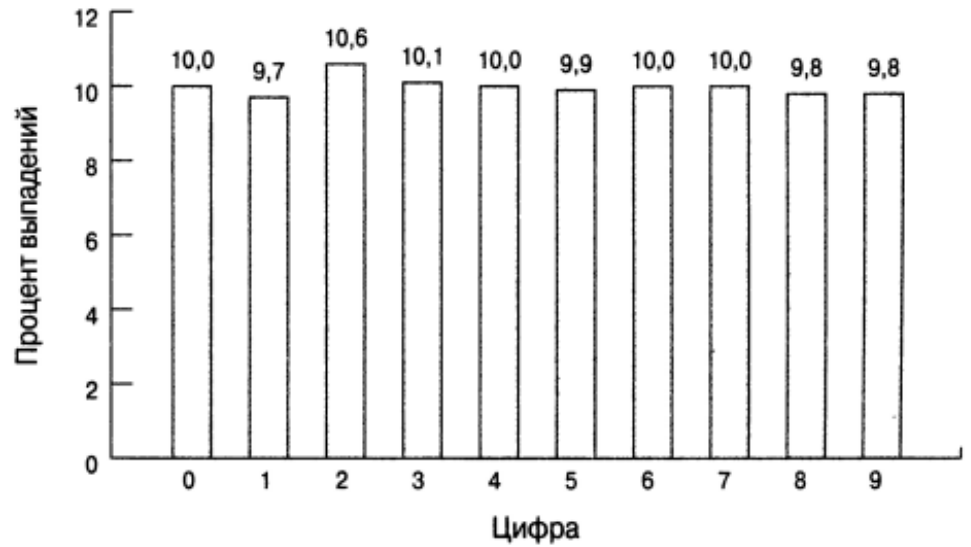
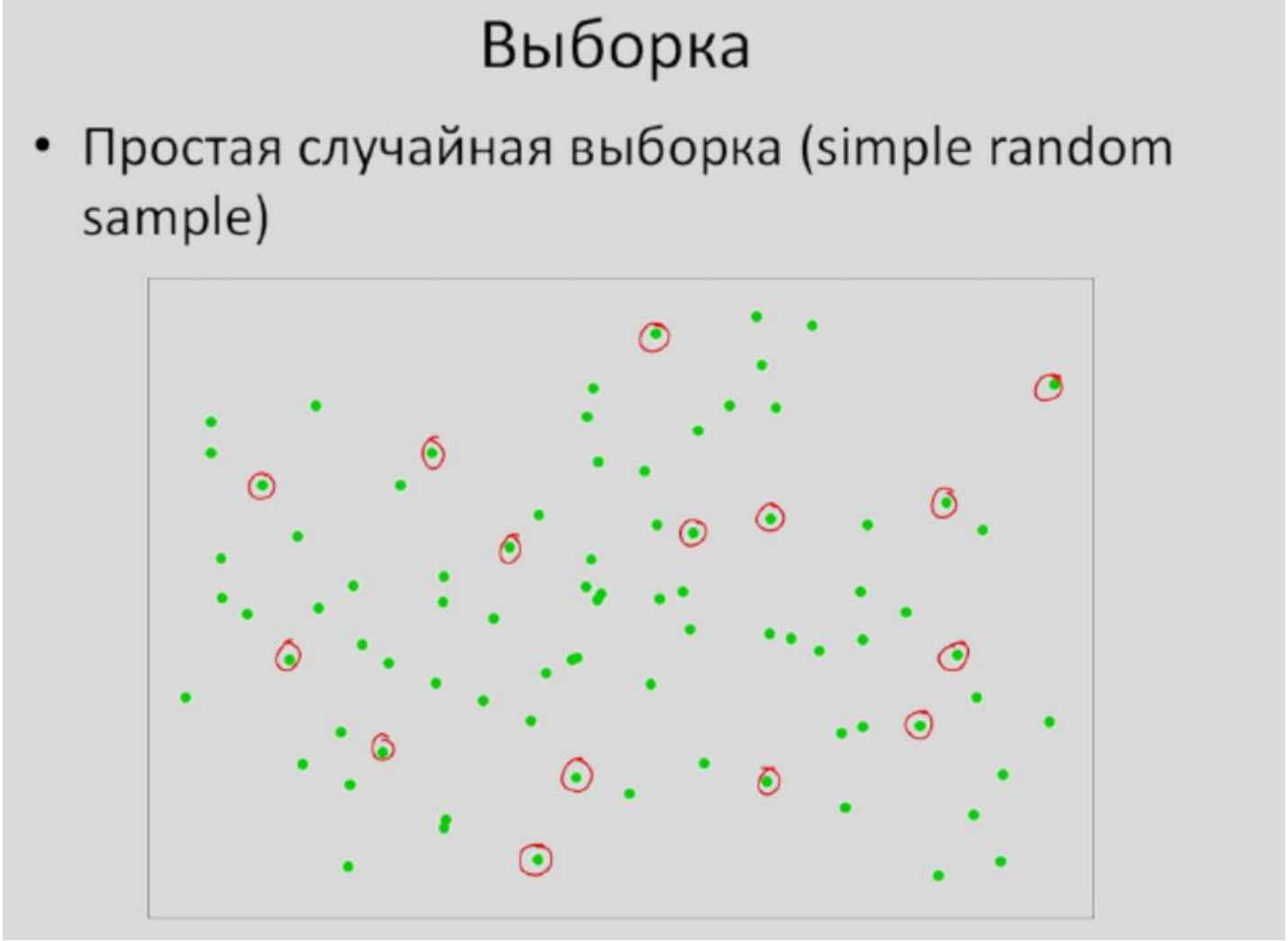


Рис. 2.2. Столбиковая диаграмма, показывающая процентное отношение количества выпадений каждой цифры

Курс – «Основы статистики»

https://stepik.org/

Конспект курса
https://github.com/dgokondra/stepik_notebooks/tree/master/basis_of_statistics



Генеральная совокупность

Суммарная численность объектов наблюдения, обладающих определенным набором признаков, ограниченная в пространстве и времени.

Примеры генеральных совокупностей (социология, медицина, биологические исследования . . .)

Выборка (Выборочная совокупность)

Часть объектов из генеральной совокупности, отобранных для изучения, с тем чтобы сделать заключение о всей генеральной совокупности. Для того чтобы заключение, полученное путем изучения выборки, можно было распространить на всю генеральную совокупность, выборка должна обладать свойством репрезентативности.

Репрезентативность выборки

Свойство выборки корректно отражать генеральную совокупность. Одна и та же выборка может быть репрезентативной и нерепрезентативной для разных генеральных совокупностей.

Примеры:
Выборка, целиком состоящая из горожан, владеющих автомобилем, не репрезентирует все население города.
Выборка только из женщин не репрезентирует все население.
Выборка только из учеников одной школы не репрезентирует всех школьников страны.

Количественные переменные:

- Дискретные данные являются значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.
  Пример дискретных данных. Продолжительность маршрута троллейбуса (количество вариантов продолжительности конечно): 10, 15, 25 мин.
- Непрерывные данные - данные, значения которых могут принимать какое угодно значение в некотором интервале. Измерение непрерывных данных предполагает большую точность.
  Пример непрерывных данных: температура, высота, вес, длина и т.д.

Качественные (номинативные) переменные

Такие переменные используются для разделения наших испытуемых или наблюдений на группы.

Например, мы можем сказать, что все участники эксперимента женского пола будут обозначены цифрой 1, а все участники мужского пола - цифрой 2 соответственно.

Таким образом, в случае номинативных переменных за цифрами не стоит никакого математического смысла. В данном случае цифры используются как маркеры различных смысловых групп, в отличие от количественных переменных.

Ранговые переменные

Представьте, что у нас есть информация о марафонском забеге: кто прибежал в каком порядке. Мы можем сказать, что испытуемый с рангом 1 быстрее, выше, сильнее испытуемого с рангом 5. Но вот насколько или во сколько он опережает этого испытуемого мы сказать не можем. Единственной возможной математической операцией является сравнение - кто быстрее, а кто медленнее.

Диспе́рсия случа́йной величины́ — мера разброса значений случайной величины относительно её математического ожидания.

Математи́ческое ожида́ние — среднее (взвешенное по вероятностям возможных значений) значение случайной величины. На практике математическое ожидание обычно оценивается как среднее арифметическое наблюдаемых значений случайной величины (выборочное среднее, среднее по выборке).

Среднеквадрати́чное отклоне́ние — наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания (аналога среднего арифметического с бесконечным числом исходов). Обычно он означает квадратный корень из дисперсии случайной величины, но иногда может означать тот или иной вариант оценки этого значения.

В статистике принято два обозначения: **sigma — для генеральной совокупности и sd (с англ. standard deviation — стандартное отклонение) — для выборки**.

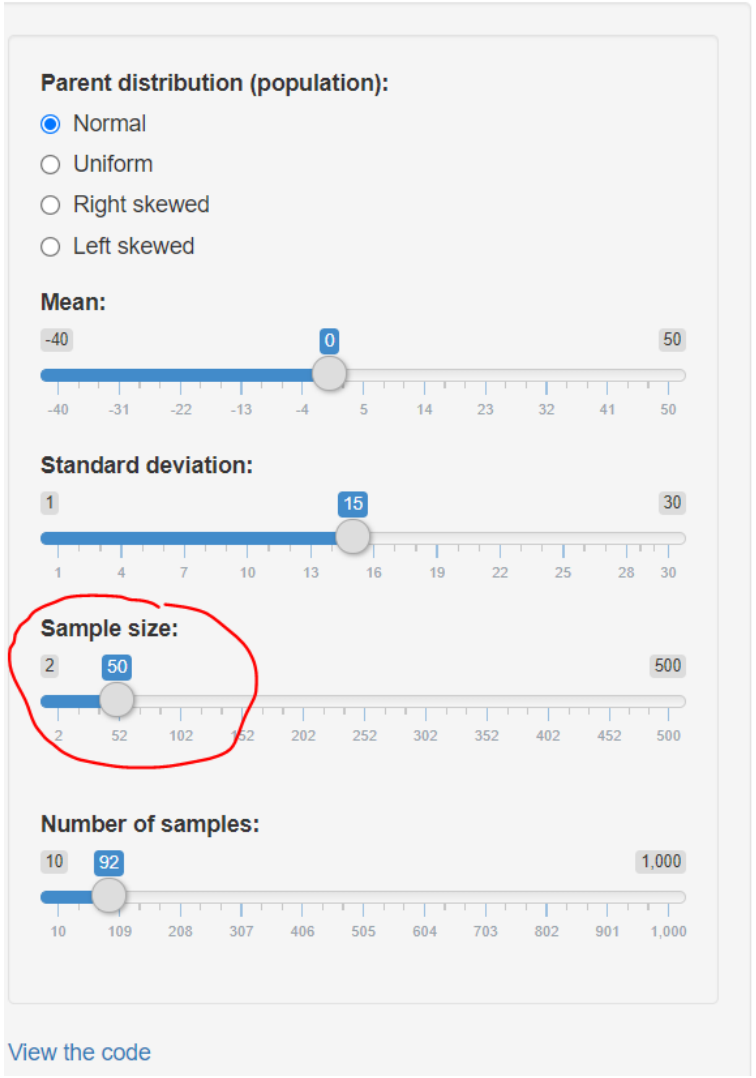
Допустим, у нас есть массив возрастов всех людей, живущих на улице.

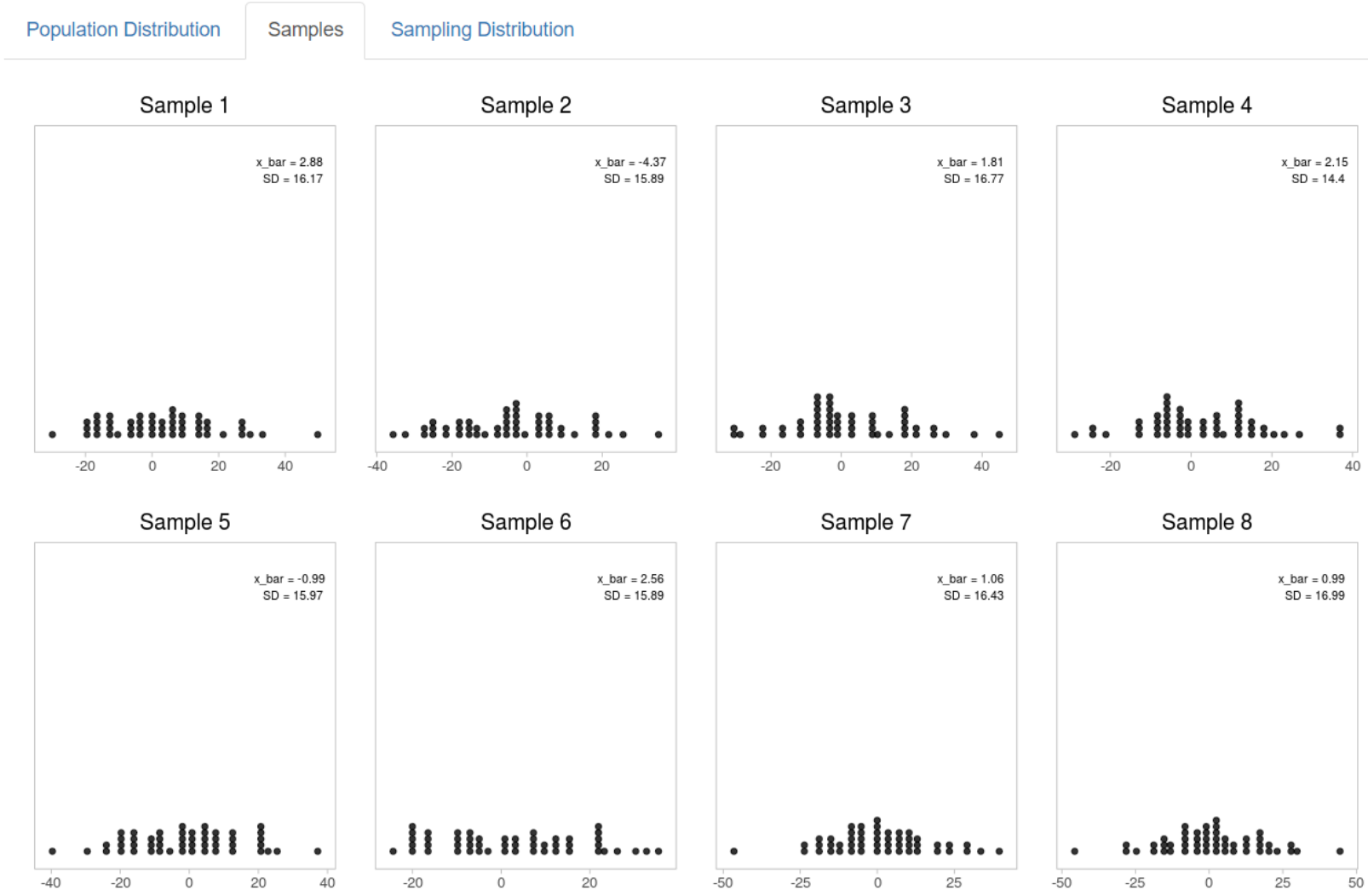
ages = [5,31,43,48,50,41,7,11,15,39,80,82,32,2,8,6,25,36,27,61,31]

Что такое 75-й **процентиль**? Ответ - 43, что означает, что 75% людей моложе 43 лет.

Central Limit Theorem for Means

https://gallery.shinyapps.io/CLT_mean/





Parent distribution (population):

- (•) Normal
- ( ) Uniform
- ( ) Right skewed
- ( ) Left skewed

Mean:

-40 [0] 50

-40 -31 -22 -13 -4 5 14 23 32 41 50

Standard deviation:

1 [15] 30

1 4 7 10 13 16 19 22 25 28 30

Sample size:

2 [50] 500

2 52 102 152 202 252 302 352 402 452 500

Number of samples:

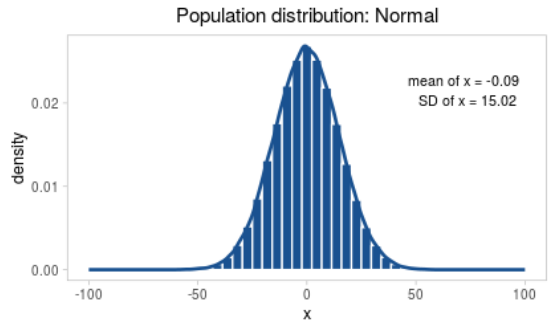
10 [92] 1,000

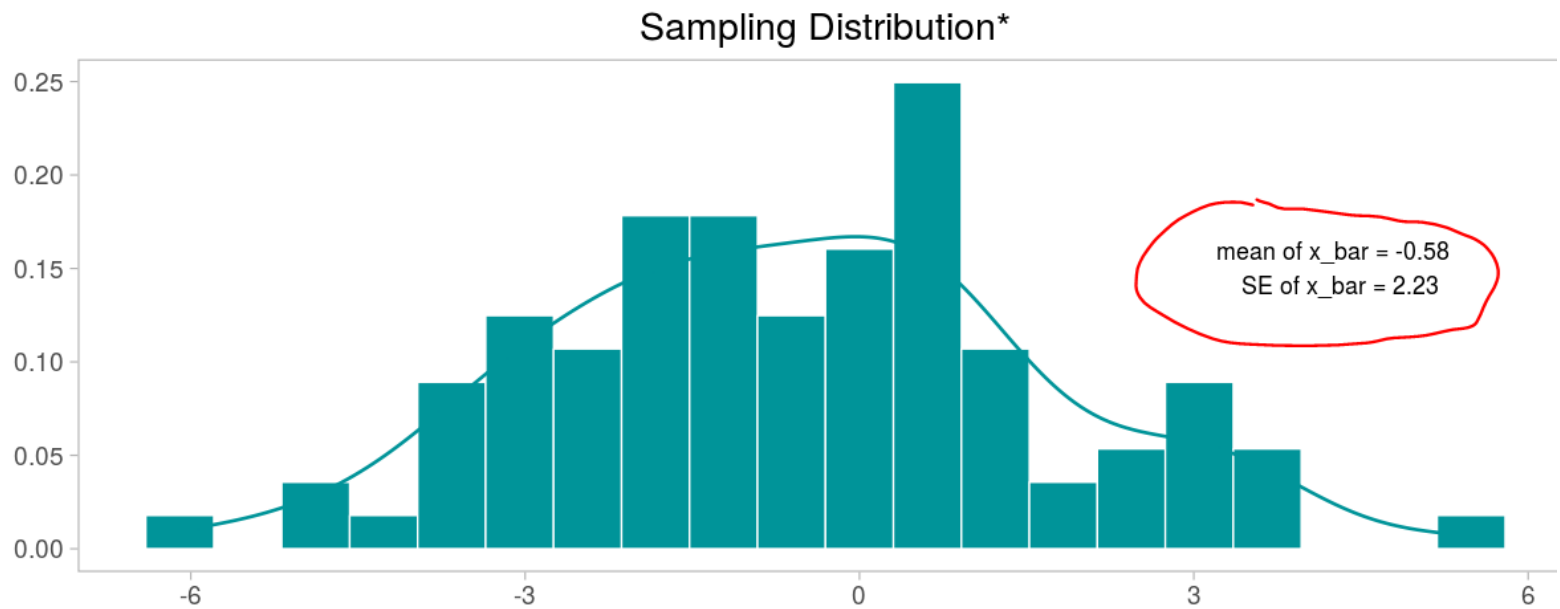
10 109 208 307 406 505 604 703 802 901 1,000

View the code

Population Distribution | Samples | Sampling Distribution

According to the Central Limit Theorem (CLT), the distribution of sample means (the sampling distribution) should be nearly normal. The mean of the sampling distribution should be approximately equal to the population mean (-0.09) and the standard error (the standard deviation of sample means) should be approximately equal to the SD of the population divided by square root of sample size (15.02/sqrt(50) = 2.12). Below is our sampling distribution graph. To help compare, population distribution plot is also displayed on the right.





Parent distribution (population):

- (•) Normal
- ( ) Uniform
- ( ) Right skewed
- ( ) Left skewed

Mean:

-40 | 0 | 50

-40 -31 -22 -13 -4 5 14 23 32 41 50

Standard deviation:

1 | 15 | 30

1 4 7 10 13 16 19 22 25 28 30

Sample size:

2 | 200 | 500

2 52 102 152 202 252 302 352 402 452 500

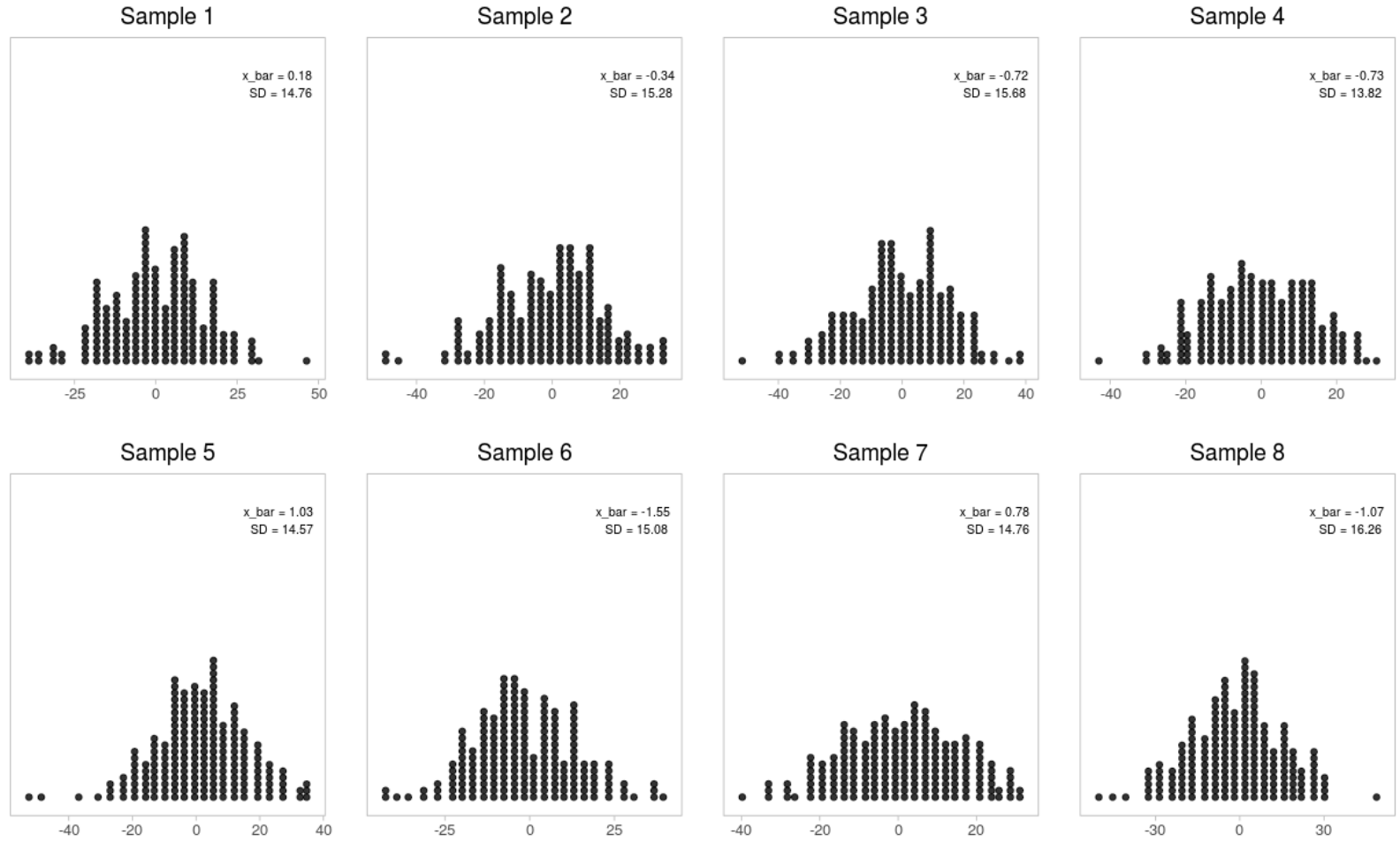
Number of samples:

10 | 92 | 1,000

10 109 208 307 406 505 604 703 802 901 1,000

View the code

Population Distribution | Samples | Sampling Distribution



Parent distribution (population):

- (•) Normal
- ( ) Uniform
- ( ) Right skewed
- ( ) Left skewed

Mean:

-40 [0] 50

-40 -31 -22 -13 -4 5 14 23 32 41 50

Standard deviation:

1 [15] 30

1 4 7 10 13 16 19 22 25 28 30

Sample size:

2 [200] 500

2 52 102 152 202 252 302 352 402 452 500

Number of samples:

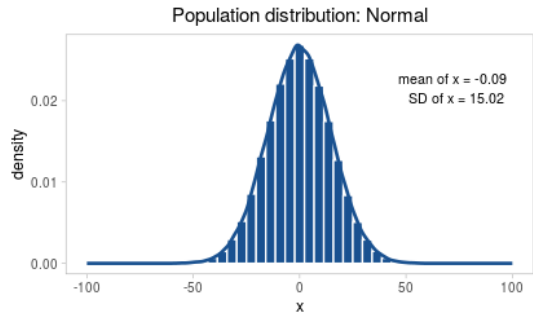
10 [92] 1,000

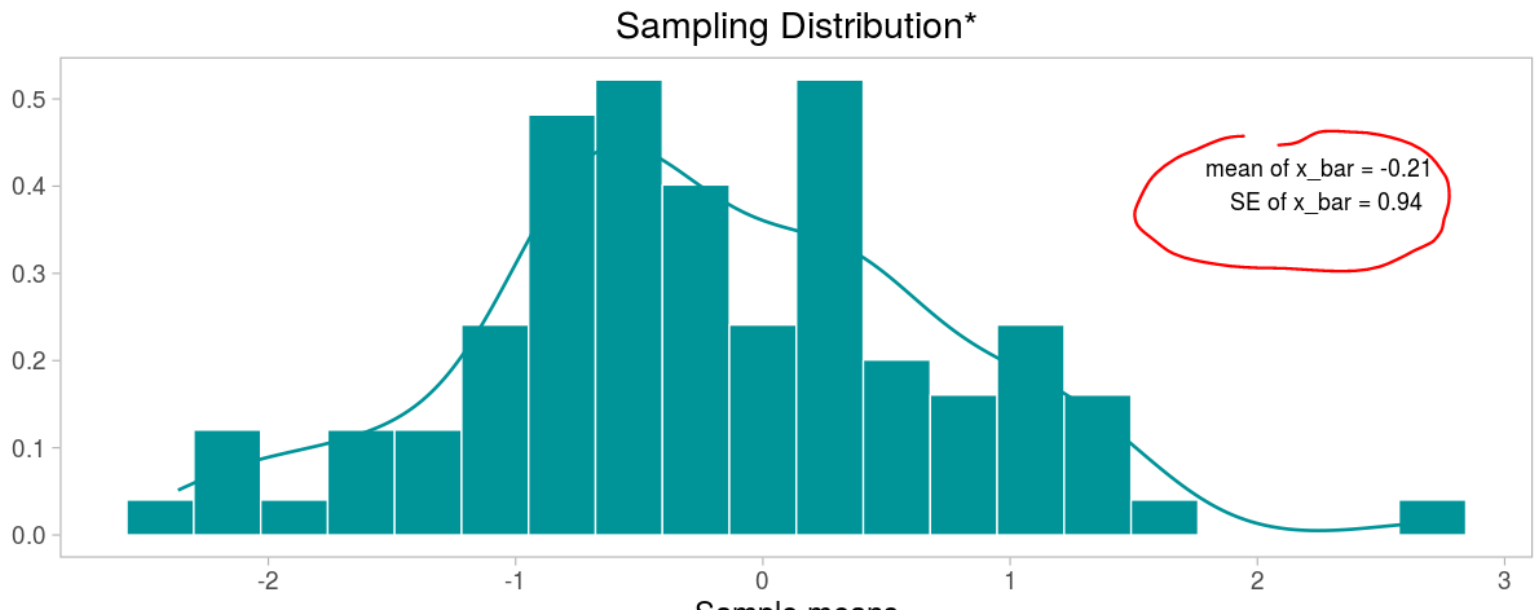
10 109 208 307 406 505 604 703 802 901 1,000

View the code

Population Distribution | Samples | Sampling Distribution

According to the Central Limit Theorem (CLT), the distribution of sample means (the sampling distribution) should be nearly normal. The mean of the sampling distribution should be approximately equal to the population mean (-0.09) and the standard error (the standard deviation of sample means) should be approximately equal to the SD of the population divided by square root of sample size (15.02/sqrt(200) = 1.06). Below is our sampling distribution graph. To help compare, population distribution plot is also displayed on the right.





In [1]:
```
import pandas as pd
```

In [2]:
```
titanic = pd.read_csv("data/titanic.csv")
```

Как рассчитать сводную статистику?

Агрегированная статистика

Какой средний возраст пассажиров Титаника?

In [6]:
```
titanic["Age"].mean()
```

Out[6]:
```
29.69911764705882
```

Каков средний возраст и стоимость билета пассажиров «Титаника»?

In [7]:
```
titanic[["Age", "Fare"]].median()
```

Out[7]:
```
Age     28.0000
Fare    14.4542
dtype: float64
```

Функция **pandas.DataFrame.describe** рассчитывает параметры описательной статистики

In [8]:
```
titanic[["Age", "Fare"]].describe()
```

Out[8]:

	Age	Fare
count	714.000000	891.000000
mean	29.699118	32.204208
std	14.526497	49.693429
min	0.420000	0.000000
25%	20.125000	7.910400
50%	28.000000	14.454200
75%	38.000000	31.000000
max	80.000000	512.329200

Вместо предопределенной статистики можно определить конкретные комбинации агрегированной статистики для заданных столбцов с помощью **DataFrame.agg()метода**:

In [10]:
```
titanic.agg(
    {
        "Age": ["min", "max", "median"],
        "Fare": ["min", "max", "median", "mean"],
    }
)
```

Out[10]:

	Age	Fare
min	0.42	0.000000
max	80.00	512.329200
median	28.00	14.454200

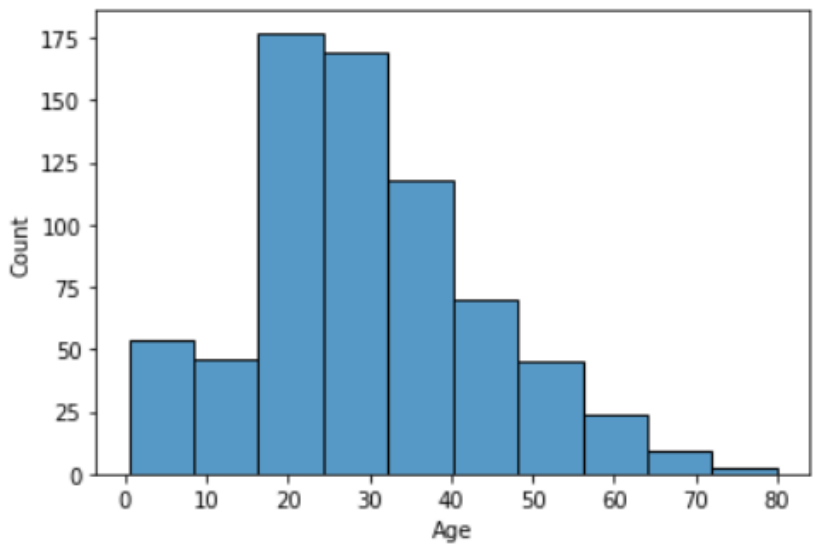
Процентиль — мера, в которой процентное значение общих значений равно этой мере или меньше ее.

In [3]:
```
titanic["Age"].describe(percentiles=[0.05, 0.25, 0.75, 0.95])
```

Out[3]:
```
count    714.000000
mean      29.699118
std       14.526497
min        0.420000
5%         4.000000
25%       20.125000
50%       28.000000
75%       38.000000
95%       56.000000
max       80.000000
Name: Age, dtype: float64
```

In [4]:
```
import seaborn as sns
sns.histplot(data=titanic["Age"],bins=10)
```

Out[4]:
```
<AxesSubplot:xlabel='Age', ylabel='Count'>
```



Группировки

Каков средний возраст пассажиров Титаника мужчинами и женщинами?

In [11]:
```
titanic[["Sex", "Age"]].groupby("Sex").mean()
```

Out[11]:

Sex	Age
female	27.915709
male	30.726645

Если не указывать столбцы, то mean-метод применяется к каждому столбцу, содержащему числовые данные:

In [12]:
```
titanic.groupby("Sex").mean()
```

Out[12]:

Sex	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
female	431.028662	0.742038	2.159236	27.915709	0.694268	0.649682	44.479818
male	454.147314	0.188908	2.389948	30.726645	0.429809	0.235702	25.523893

Группировки

Какова средняя цена билета для каждой комбинации пола и класса салона?

In [16]:
```
titanic.groupby(["Sex", "Pclass"])["Fare"].mean()
```

Out[16]:
```
Sex     Pclass
female  1         106.125798
        2          21.970121
        3          16.118810
male    1          67.226127
        2          19.741782
        3          12.661633
Name: Fare, dtype: float64
```

Расчет количества записей по категориям

Метод **value_counts()** подсчитывает количество записей для каждой категории в колонке.

Какое количество пассажиров в салоне каждого класса?

In [17]:
```
titanic["Pclass"].value_counts()
```

Out[17]:
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

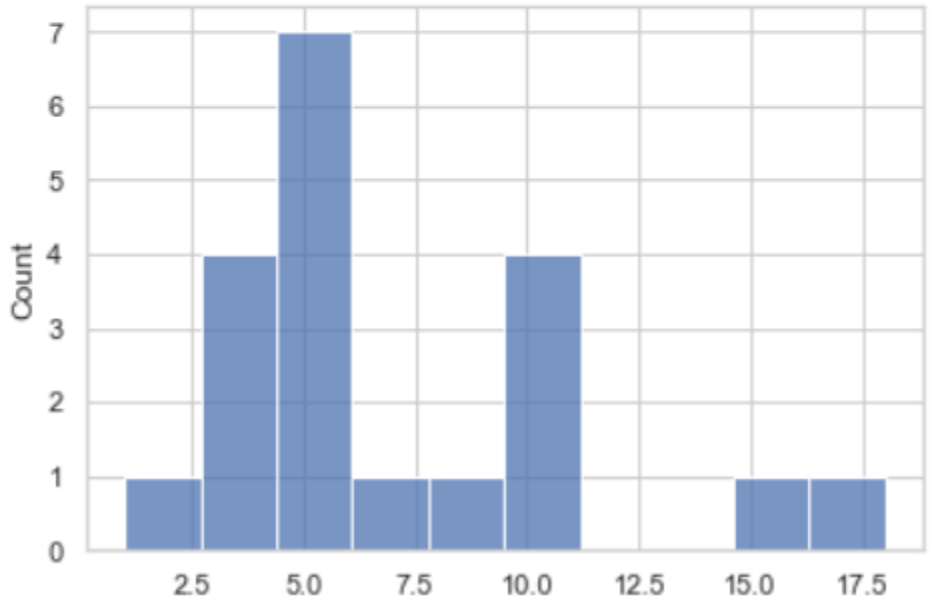
Гистограмма

In [35]:
```
s = pd.Series([1,3,5,11,10,3,6,5,6,6,7,8, 4,6,15,18,6,4,11,10])
```

Seaborn — это библиотека для создания статистических графиков на Python. Она основывается на matplotlib и тесно взаимодействует со структурами данных pandas.

In [47]:
```
import seaborn as sns
sns.histplot(data=s,bins=10)
```

Out[47]:
```
<AxesSubplot:ylabel='Count'>
```



```
In [36]: s.describe()

Out[36]: count    20.000000
         mean      7.250000
         std       4.191156
         min       1.000000
         25%       4.750000
         50%       6.000000
         75%      10.000000
         max      18.000000
         dtype: float64

In [34]: s.median()

Out[34]: 6.0

In [48]: sns.boxplot(x=s)

Out[48]: <AxesSubplot:>
```

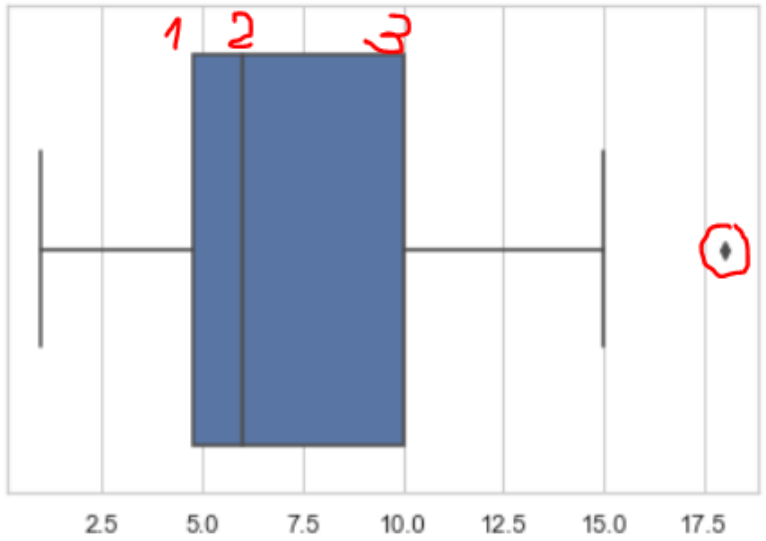


График box-plot

Центром ящика является медиана наших данных или второй квартиль, верхняя граница = 3-й квартиль, а нижняя граница = 1-й квартиль.

Почему некоторые точки на графике отображены отдельно?

Если мы посчитаем разность между 3-м и 1-м квартилем - это межквартильный размах (мера изменчивости).

Чем выше межквартильный размах, тем больше вариативность нашего признака.

Отложим мысленно 1,5 межквартильного размаха вверх и вниз от 1-го и 3-го квартилей. Те значения признака, которые последними принадлежат этому промежутку и будут границами усов.

Точки, которые превосходят полтора межквартильного размаха - наносятся на график отдельно.