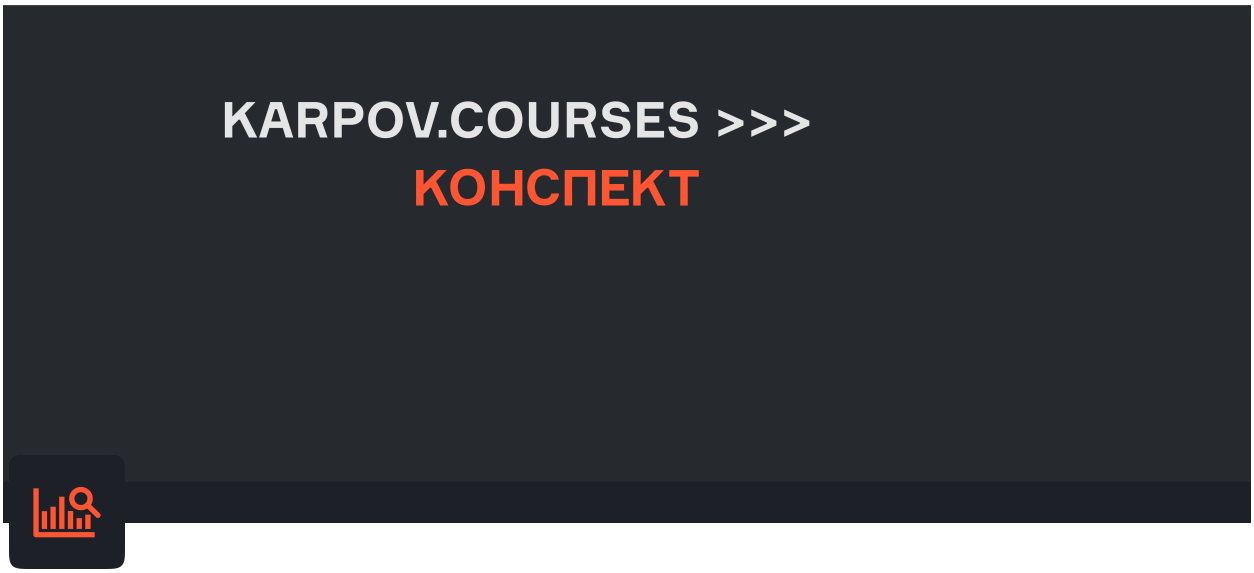


> Конспект > 1 урок > СТАТИСТИКА

> Оглавление

> Генеральная совокупность

Генеральная совокупность – совокупность всех объектов (единиц), относительно которых предполагается делать выводы при изучении конкретной задачи.

Пример: уровень тестостерона всех мужчин, участвующих в Олимпийских играх

Зачастую очень сложно исследовать все объекты, поэтому из генеральной совокупности берут **выборки**.

Важной характеристикой выборки является её **репрезентативность**. Под этим термином понимается соответствие характеристик выборки характеристикам генеральной совокупности в целом.

Виды выборок

1. Вероятностные выборки – при создании таких выборок мы предполагаем, что генеральная совокупность достаточно однородна и все её элементы одинаково доступны.

Простая случайная выборка (simple random sample) – случайный набор объектов из генеральной совокупности. Пример: 100 мужчин, участвующих в Олимпийских играх

Стратифицированная выборка (stratified sample) – перед тем, как случайным образом отобрать объекты из генеральной совокупности, мы разбиваем её на несколько страт (групп).

Пример: мужчины 18-25 лет, 36-31, 32-36 и так далее.

Потом уже из этих групп случайно набираем по N человек.

Групповая выборка (cluster sample) – также сначала делим генеральную совокупность на кластеры, только считаем, что они между собой схожи.

Пример: рост жителей Санкт-Петербурга. Мы делим их на районы (Адмиралтейский, Василеостровский и т.д.), а потом случайно набираем людей из нескольких случайно выбранных районов для исследования.

2. Невероятностные выборки – отбор в такой выборке осуществляется не по принципам случайности, а по субъективным критериям – доступности объектов, типичности или равного представительства. Такие выборки часто встречаются в социологических исследованиях, однако данные, полученные на них обладают меньшей достоверностью и лучше их обходить стороной.

Метод снежного кома - у каждого респондента, начиная с первого, просятся контакты его друзей, коллег, знакомых, которые подходили бы под условия отбора и могли бы принять участие в исследовании. Основная проблема такой выборки -

то, что затрагивается не случайная группа лиц, а лица, связанные общими интересами, хобби и т.д.

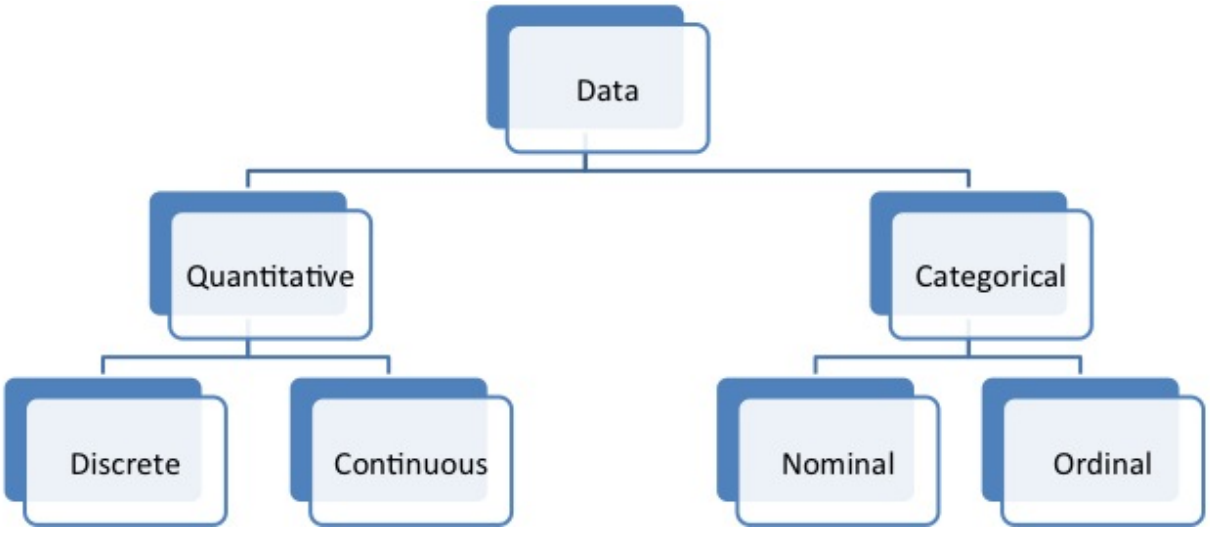
Стихийная выборка - производится опрос наиболее доступных респондентов. Размер и состав стихийных выборок заранее не известен и определяется только одним параметром - активностью респондентов.

Пример: опрос, проведенный в газете или журнале, большинство интернет-опросов

Выборка типичных случаев происходит отбор отдельных единиц генеральной совокупности, которые обладают типичным значением признака (часто это среднее значение). При этом возникает проблема выбора признака и определения его типичного значения.

> Типы переменных

Два основных типа переменных: количественные и качественные.



Количественные – измеренные значения некоторого признака.

- непрерывные – могут принимать любое значение на определенном промежутке. *Пример:* рост человека
- дискретные – могут принимать определенные значения. *Пример:* число детей в семье (целые неотрицательные числа, то есть 3.5 ребенка быть не может)

Качественные (номинативные / категориальные) – делят наши объекты на группы. *Пример:* кодировка пола человека (0 - мужчина, 1 женщина)

Также переменные могут быть **ранговыми**, например, когда мы смотрим на результаты марафона: 1 – прибежал первым, 2 – вторым и так далее. Мы не знаем, насколько различаются результаты участников между собой, но знаем их порядок.

Важно отметить, что некоторые переменные, в зависимости от того, в какой шкале они представлены, могут относиться к разным категориям. Одним из таких примеров является переменная возраста.

Количественная непрерывная - возраст, измеренный в днях/месяцах/годах

Ранговая переменная - возраст разбит на группы (очень часто встречается в анкетировании) - от 14-17 лет; 18-29 лет и так далее.

> Меры центральной тенденции

Мода (mode) – значение измеряемого признака, которое встречается максимально часто. Мод может быть несколько.

```
import pandas as pd
df.column_1.mode()  # pandas wayfrom scipy import stats
stats.mode(df.column_1)  # scipy way
```

Медиана (median) – значение признака, которое делит упорядоченное множество данных пополам. Берем множество значений признака, сортируем и берем центральное значение.

Это легко сделать, когда количество наблюдений нечетное.

Пример: 1 2 3 4 5.

Когда количество наблюдений четное, то точка, делящая упорядоченное множество пополам, окажется между числами. Тогда нужно брать среднее от значений, окружающих эту точку.

Пример: 1 2 3 **.** 4 5 6

$$median = \frac{3+4}{2} = 3.523 + 4 = 3.5$$

```
import pandas as pd
df.column_1.median()  # pandas wayimport numpy as np
np.median(df.column_1)  # numpy way
```

Среднее (mean, среднее арифметическое) – сумма всех значений измеренного признака, деленная на количество измеренных значений.

$$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$\overline{X}$ - среднее выборки

M - среднее генеральной совокупности

Свойства среднего значения:

$$M_{x+c} = M_x + c$$

$$M_{x*c} = M_x * c$$

$$\sum(x_i - M_x) = 0$$

← сумма всех отклонений от среднего равняется нулю

Когда не стоит использовать среднее значение, а лучше брать моду или медиану:

- явная ассиметрия
- заметные выбросы
- несколько мод

Пример: в группе студентов 10 человек ростом 150, 150, 155, 155, 160, 165, 165, 170, 170, 175 см.

- Средний рост: 161.5
- Медиана: 162.5

После в группу перевелись Майкл и Джордан, оба ростом 198 см.

- Средний рост теперь: 182.8
- Медиана: 165

Как посчитать в питоне:

```
import pandas as pd
df.column_1.mean()  # pandas way

import numpy as np
np.mean(df.column_1)  # numpy way
```

> Меры изменчивости

Размах (range) – разность между максимальным и минимальным значением из распределения

$$R = X_{max} - X_{min}$$

Минусы: размах характеризует распределение, используя только 2 значения. Так что если в данных появится аутлаер (выброс), то данные сильно изменятся (вспомним пример про группу студентов, к которым перевелись Майкл и Джордан). Поэтому лучше использовать каждое значение для оценки изменчивости.

```
import numpy as np
np.percentile(df.A, [0, 100])
```

Дисперсия (variance) – средний квадрат отклонений индивидуальных значений признака от их средней величины.

$$D = \frac{\sum(x_i - \overline{X})^2}{n}$$

- для генеральной совокупности

$$D = \frac{\sum(x_i - \overline{X})^2}{n - 1}$$

- для выборки

Как посчитать в python:

```
import pandas as pd
df.A.var()  # pandas way

import numpy as np
np.var(df.A)  # numpy way
```

Среднеквадратическое отклонение – квадратный корень из дисперсии.

$$\sigma = \sqrt{D} = \sqrt{\frac{\sum(x_i - \overline{X})^2}{n}}$$

- для генеральной совокупности

$$sd(standarddeviation) = \sqrt{\frac{\sum(x_i - \overline{X})^2}{n-1}}$$

- для выборки

Показывает реальную среднюю разницу каждого значения и среднего в выборке. Дисперсия же отражает квадрат этой разницы.

Как посчитать в python:

```
import pandas as pd
df.A.std()  # pandas way
import numpy as np
np.std(df.A)  # numpy way
```

Свойства дисперсии и стандартного отклонения:

$$D_{x+c} = D_x$$

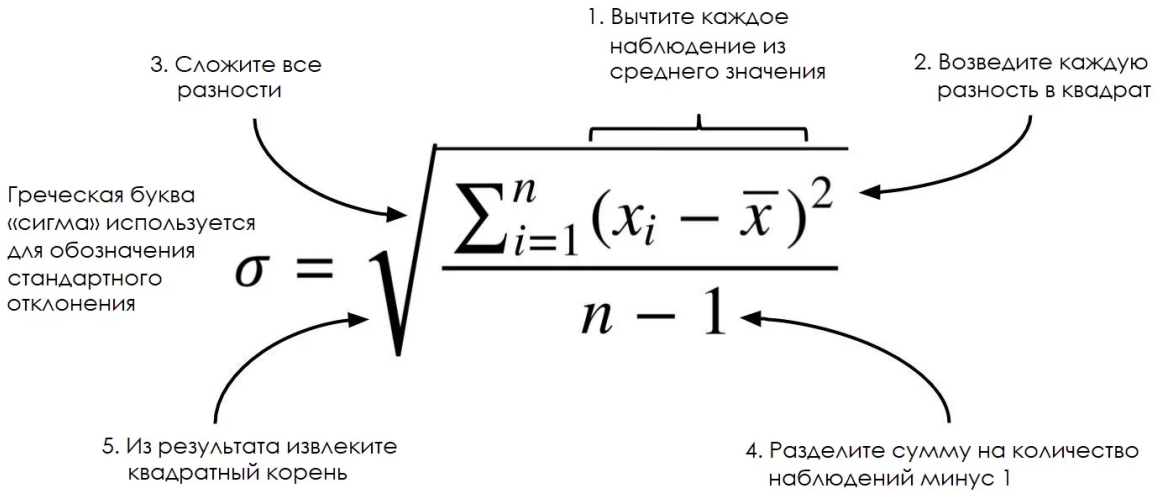
$$sd_{x+c} = sd_x$$

$$D_{x*c} = D_x * c^2$$

$$sd_{x*c} = sd_x * c$$

Можно на этой картинке запомнить как считать, и что представляет собой среднеквадратичное отклонение.

Правда, на картинке есть неточность: σ - это стандартное отклонение в генеральной совокупности, а формула здесь для выборки. Правильнее было бы написать s_x.



> Квантили распределения

Квантили распределения – значения признака, делящие распределение на некоторое число равных частей.

Квартили – три точки (значения признака), которые делят упорядоченное множество данных на 4 равных части.

Как посчитать в python?

`quantile` – метод для поиска определённых перцентилей. Принимает число от 0 до 1, обозначающее перцентиль в виде доли:

- 0 – 0-ой перцентиль
- 0.1 – 10-ый перцентиль

- 0.75 – 75-ый перцентиль (он же 3-ий квартиль)

```
import pandas as pd

df.quantile(q=0.75)
```

Также в `q` можно передать список всех желаемых перцентилей

```
df.quantile(q=[0.5, 0.7])
```

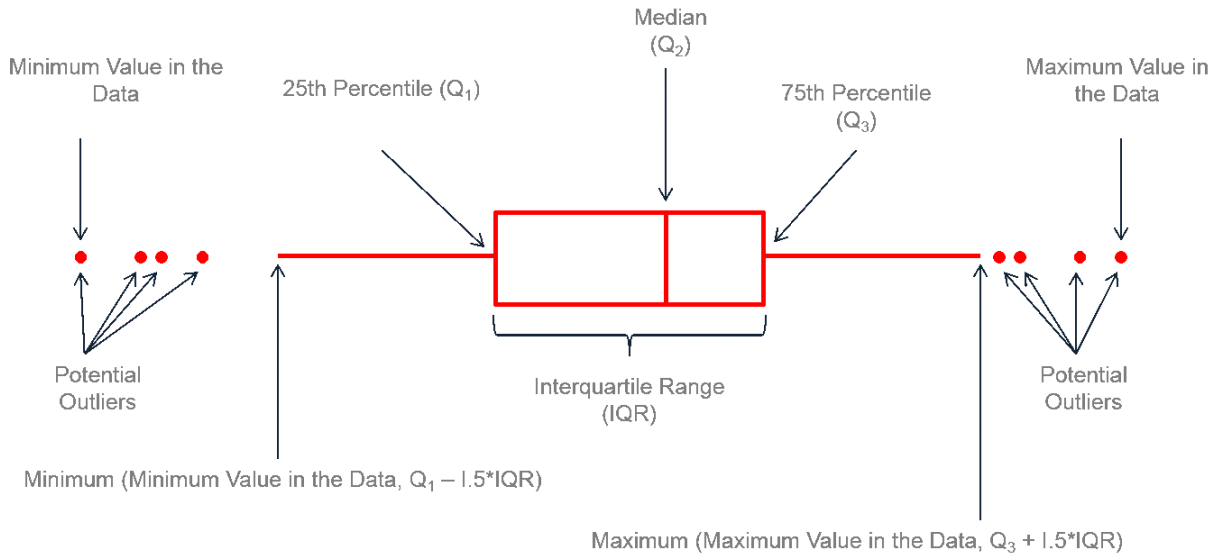
> Боксплот (график)

Межквартильный размах (IQR) – разница между Q_1 и Q_3. Чем больше межквартильный размах – тем шире "ящик".

Усы боксплота = $1.5 * IQR$ (полтора межквартильных расстояния).

Значения, лежащие за усами, обозначаются жирными точками.

```
import seaborn as sns
sns.boxplot(df.A)
```



Однако боксплот может редуцировать информацию из-за того, что распределение может бимодальным или полимодальным. Поэтому лучше на боксплоте отражать ещё и наблюдения из выборки в виде точек.

```
import seaborn as sns

ax = sns.boxplot(df.A)
ax = sns.swarmplot(df.A)
```