Описательные статистики. Статистика вывода

Цель занятия

В результате обучения на этой неделе вы научитесь:

- отличать и понимать базовые статистические концепции;
- различать статистические распределения;
- считать точечные оценки и доверительные интервалы;
- тестировать ряд статистических гипотез.

План занятия

1. Типы переменных
2. Меры центральности
3. Меры “НЕцентральности”
4. Меры вариации
5. Ящик с усами
6. Вычисление ключевых метрик в Pandas
7. Асимметрия и эксцесс
8. Выборка и генеральная совокупность
9. Распределения
10. Оценки
11. Тестирование гипотез

Используемые термины

Генеральная совокупность (N) — это весь набор объектов, о которых мы хотим получить информацию.

Выборка (n) — подмножество объектов из генеральной совокупности.

Доверительный уровень — вероятность того, что реальный параметр лежит в границах полученного доверительного интервала: значение параметра ± ошибка выборки (Δ).

Стратификация — выделение подгрупп (страт) на основе важных признаков.

Случайная величина — это такая величина, для которой возможное значение в результате эксперимента зависит от такого большого количества разных факторов, что предсказать ее с ходу невозможно.

Плотность распределения — это взаимоотношение между значением величины и вероятностью того, что величина примет именно это значение (первая производная от ее функции распределения).

Точечная оценка параметра — число, оцениваемое на основе наблюдений, предположительно близкое к оцениваемому параметру.

Надежность — вероятность того, что оценка параметра принадлежит доверительному интервалу.

Конспект занятия

1. Типы переменных

Категориальные (неквантифицируемые характеристики) — например, имена людей:

```
categorical_var = ['Misha','Dasha','John']
```

Численные (квантифицируемые характеристики) — например, рост человека:

```
numeric_var = [187, 165, 163]
```

Каждый из типов переменных можно разделить на два подтипа.

Категориальные (Неквантифицируемые характеристики)

Номинальные (НЕТ естественного порядка категорий)

Порядковые (ЕСТЬ определенный порядок категорий)

A	90% – 100%
B	80% – 89%
C	70% – 79%
D	60% – 69%
F	0% – 59%

Пример номинальных переменных — знаки зодиака. 12 знаков зодиака ассоциируются с определенными характеристиками, но мы не можем установить среди них никакого порядка.

Пример порядковой переменной — оценки. Например, система оценивания из букв A, B, C, D, F. Буквам соответствует определенный процент выполнения некоторой работы.

Мы не можем сказать, насколько A отличается от B, B отличается от C. Но мы понимаем, что есть определенный логический порядок категорий, который является социальной конвенцией. Этот порядок не естественный в том смысле, что для некоторых людей удовлетворительная оценка C будет более предпочтительная отличной оценки A.

При определении порядковых переменных важно понимать, что этот порядок есть определенная договоренность.

В случае интервальной переменной в качестве примера можно привести доход в России. Мы можем определить нижнюю границу, а с определением верхней границы могут возникнуть проблемы. Она будет сводиться к количеству всех денег, которые существуют. На этом промежутке существует бесконечное количество значений, которые будут являться возможными для данной переменной.

В силу той или иной ограниченности инструментов измерителей мы не можем дробить наши значения до бесконечности. Поэтому в реальной жизни эмпирически интервальными переменными будем считать те переменные, для которых число реальных значений примерно бесконечно, то есть стремится к бесконечности.

Примером дискретной случайной величины служат оценки на каком-нибудь соревновании. В фигурном катании были оценки от 0.0 до 6.0 с шагом в 0.1. Мы можем перечислить все возможные значения оценок.

2. Меры центральности

Для понимания трендов в переменных необходим подсчет определенных метрик, или мер.

Для числовых переменных важной характеристикой являются различные **меры центральности** — меры, которые описывают «типичное» или центральное значение в распределении.

Обозначим некоторые такие меры:

- **Среднее** (среднее арифметическое):

$$x_{mean} = \frac{\sum x_i}{n}$$

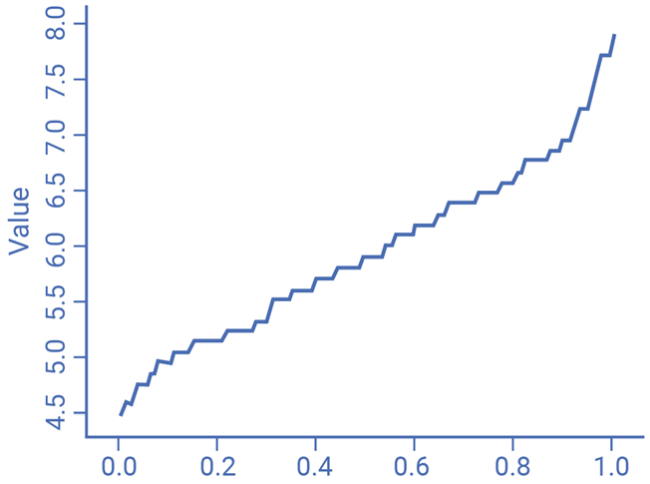
- **Медиана**:
  - n четное — значение элемента центрального ранга упорядоченного вектора чисел;
  - n нечетное — среднее арифметическое элементов двух центральных рангов вектора чисел.
- **Мода** — самое часто встречающееся значение в выборке.

Среднее зависимо от выбросов (нетипичных значений), медиана — нет. Вариант решения — **усеченное среднее**:

$$x_{mean_{trimmed}} = \frac{\sum_{i=p+1}^{n-p} x_i}{n-2p}$$

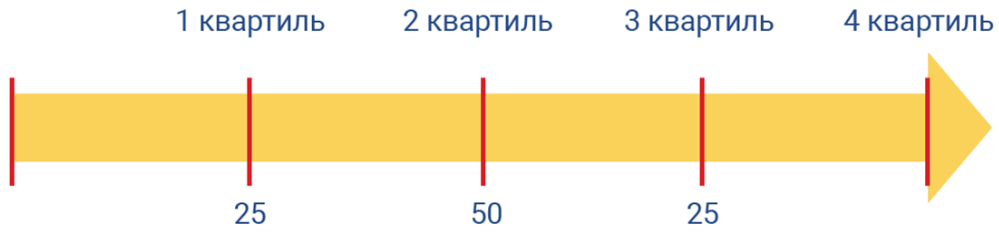
3. Меры «НЕцентральности»

Процентиль — значение упорядоченного вектора чисел, которое делит выборку в определенной пропорции. $i\%$ наблюдений имеют значения ниже i-го процентиля, $(100 - i)\%$ наблюдений — выше.



Частный случай процентиля — **квартили**:

- **Нижний квартиль** (25-й процентиль) — процентиль, разбивающий выборку в соотношении 25% к 75%.
- **Верхний квартиль** (75-й процентиль) — процентиль, разбивающий выборку в соотношении 75% к 25%.



4. Меры вариации

Меры вариации — меры, которые показывают, насколько наша переменная «богата» на различные значения.

Основные меры:

- Размах $x_{max} - x_{min}$.
- Межквартильный размах (IQR):

$$\hat{x}_{0.75} - \hat{x}_{0.25}.$$

- Выбросы:

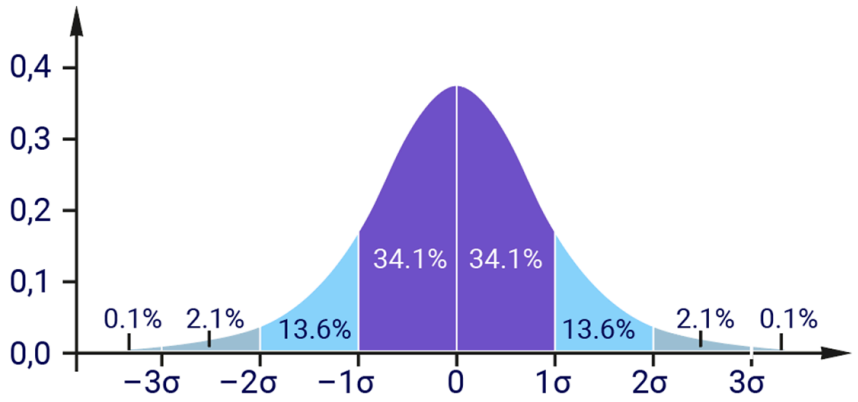
$$(X_{outliner} \in [\hat{x}_{0.25} - 1.5IQR; \hat{x}_{0.75} + 1.5IQR])$$.

- Дисперсия: $\sigma^2 = \frac{\Sigma(x_i - x_{min})^2}{n-1}$.

- Стандартное отклонение: (SD) $SD = \sqrt{\sigma^2}$.

- Среднее абсолютное отклонение (MAD): $MAD = \frac{\Sigma|x_i - x_{min}|}{n}$.

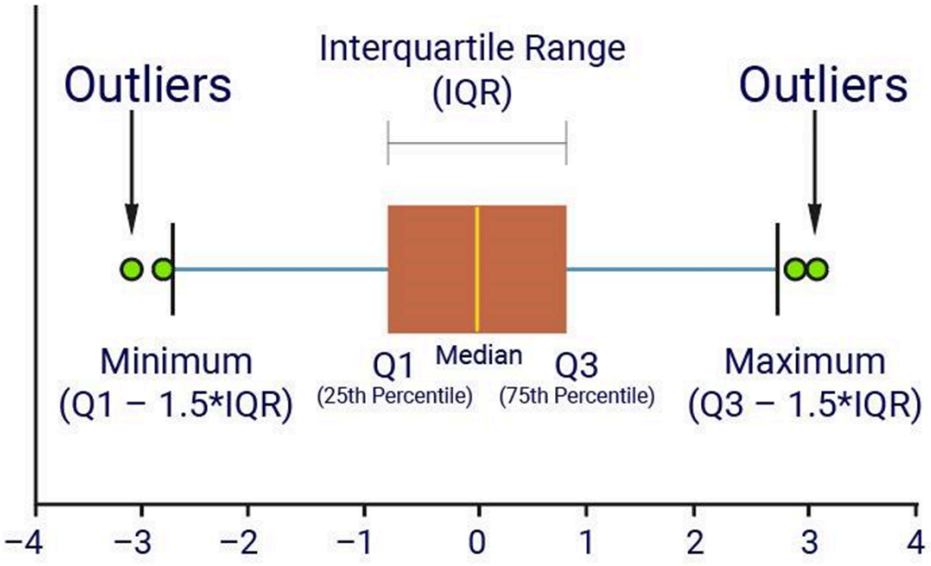
Меры вариации, в частности, стандартное отклонение, играют важную роль в статистике. Например, стандартное нормальное распределение:



Здесь с каждым стандартным отклонением удаления от среднего процент наблюдения в выборке уменьшается. Если распределение данных близко к нормальному, то практически никакие значения не будут превышать модуля суммы среднего и трех стандартных отклонений.

5. Ящик с усами

Достаточно большое количество метрик мы можем визуализировать с помощью специального графика — «ящика с усами».



Ящик может быть как горизонтальным, так и вертикальным. Мы можем на нем видеть большое количество метрик. Вдоль ящика располагаются все возможные значения переменных. Нанесено значение медианы переменной (желтая полоса).

Границы ящика определяются нижней и верхней квартилью. Высота ящика — межквартильный размах. «Усы» ящика — границы типичных наблюдений. Слева от левого уса и справа от правого — выбросы или outliers. Как правило, выбросы обозначаются точками.

6. Вычисление ключевых метрик в Pandas

Метрики можно вычислять по отдельности.

Рассмотрим наш пример с датафреймом фруктов:

```
print('Mean total: ', df['total'].mean())
print('Median total: ', df['total'].median())
print('Lower Quartile of total: ', df['total'].quantile(0.25))
print('Upper Quartile of total: ', df['total'].quantile(0.75))
print('Interquartile Range of total ', df['total'].quantile(0.75)
- \
      df['total'].quantile(0.25))
```

```
  Mean total:  23.727272727272727
  Median total:  15.0
  Lower Quartile of total:  11.0
  Upper Quartile of total:  32.0
>> Interquartile Range of total  21.0
```

На основе этого мы можем выделить критерий для определения выбросов:

```
uq = df['total'].quantile(0.75)

lq = df['total'].quantile(0.25)

IQR = df['total'].quantile(0.75) - df['total'].quantile(0.25)

print('Outliers thresholds: Lower - ', lq - 1.5*IQR, 'Upper - ',
uq +1.5*IQR)
```

```
>> Outliers thresholds: Lower -  -20.5 Upper -  63.5
```

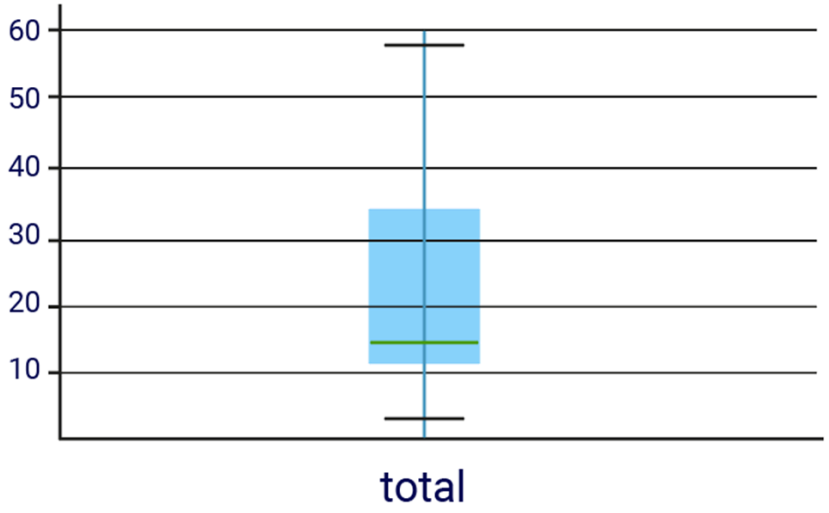
Рассчитаем некоторые метрики вариаций:

```
print('Range of total: ', df['total'].max() - df['total'].min())

print('Variance of total: ', df['total'].var())

print('SD of total: ', df['total'].std())

print('MAD of total: ', df['total'].mad())
```

```
  Range of total:  58
  Variance of total:  389.4181818181818
  SD of total:  19.73368140561162
  MAD of total:  15.652892561983471
>>
```

Ящик с усами:

```
df.boxplot(column =['total'])
```



>>

7. Асимметрия и эксцесс

При изучении распределения данных необходимо посмотреть на ряд других параметров.

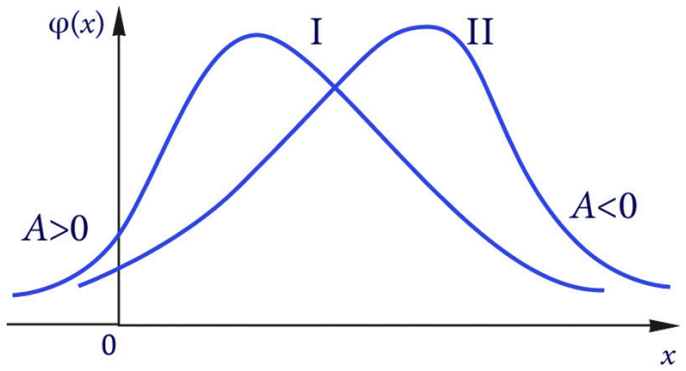
Асимметрия — отношение центрального момента третьего порядка к кубу стандартного отклонения. Фактически она показывает, насколько наше эмпирическое распределение отклоняется влево или вправо от **идеальной модели** — нормального распределения.

Асимметрия **положительна**, если «длинная часть» кривой распределения расположена справа от математического ожидания. Асимметрия **отрицательна**, если «длинная часть» кривой расположена слева от математического ожидания.

Формула для расчета коэффициента асимметрии:

$$a_s = \frac{\Sigma\left(x_i - x_{min}\right)^3}{\sigma^3}$$

Репрезентация различных распределений с точки зрения их коэффициента асимметрии A:



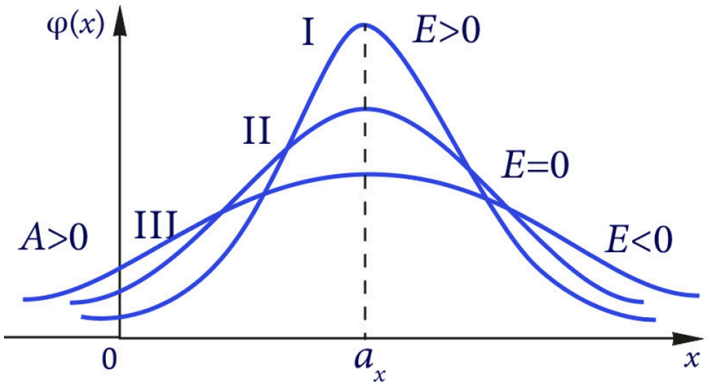
Коэффициент асимметрии в Pandas можно реализовать через метод **skew()**.

Коэффициент эксцесса показывает «остроту» распределения. Аналогично асимметрии, эксцесс показывает отклонение от идеальной модели нормального распределения, но уже не по горизонтали, а по вертикали (вверх-вниз). Кривые, более островершинные, чем нормальная, обладают **положительным** эксцессом, более плосковершинные — **отрицательным** эксцессом.

Формула расчета коэффициента эксцесса:

$$e_s = \frac{\Sigma\left(x_i - x_{mean}\right)^4}{\sigma^4} - 3$$

Пример:



Для расчета эксцесса в Pandas реализован метод **kurtosis()**.

8. Выборка и генеральная совокупность

Как отличить выборку от генеральной совокупности

Соотношение концепции выборки и генеральной совокупности являются ключевыми для многих как научных сфер и прикладных аналитических направлений исследовательской деятельности.

Генеральная совокупность (N) — это весь набор объектов, о которых мы хотим получить информацию. Набор этих объектов зависит от тех или иных исследовательских задач.

Выборка (n) — подмножество объектов из генеральной совокупности. Как правило, это те объекты, которые могут быть непосредственными участниками нашего исследования.

В большинстве случаев выборка и генеральная совокупность не совпадают по размеру.

При решении задач анализа данных мы чаще работаем с выборкой. Например, при исследовании благосостояния россиян мы не можем провести опрос 140+ миллионов человек. В то же время мы не располагаем публичной базой данных по гражданам России с такой информацией. Соответственно, нужно взять выборку и спроецировать те или иные закономерности на всю генеральную совокупность.

Примеры выборки для разных исследований:

- Респонденты маркетингового исследования привлекательности товара X.
- Совокупность посещений сайта Y во временные периоды $T_1 - T_n$, собранная для тестирования конверсии нового лендинга.

- База клиентов банка S, использованная для решения задачи кредитного скоринга.

Количество респондентов в выборке зависит от целей исследования.

Пример. Представим, что мы пытаемся предсказать целесообразность построения станции метро в микрорайоне и проводим уличный опрос около местного автовокзала.

Вопрос. Важно ли нам количество опрошенных респондентов?

Ответ. Конечно, ведь большинство методов количественного анализа данных просто не будут давать валидные оценки на слишком маленькой выборке.

Как выбрать достаточный размер выборки

Одна из общепринятых формул расчета необходимой выборки (для прикладных опросных исследований):

$$n = \frac{Z^2 pq}{\Delta^2}$$

n — объем выборки,

Z — коэффициент, зависящий от доверительного уровня,

p — доля респондентов с наличием целевого признака,

$q = 1 - p$ — доля респондентов без целевого признака,

Δ — предельная ошибка выборки.

Вернемся к задаче предсказания целесообразности построения станции метро. Представим, что нам нужно определить размер выборки для исследования с доверительным уровнем в 95% и ошибкой не более 4%. При том что мы не знаем, как будет распределяться целевой признак среди респондентов.

Доверительный уровень — это вероятность того, что реальный параметр лежит в границах полученного доверительного интервала: значение параметра ± ошибка выборки (Δ). Доверительный уровень устанавливает сам исследователь в соответствии со своими требованиями к надежности полученных результатов. Чаще всего применяются доверительные уровни равные 0,95 или 0,99.

Импортируем специальную библиотеку:

```
import scipy.stats as st
```

Scipy (Scientific Python) — библиотека Python, специализирующаяся на научных инженерно-математических и статистических исследованиях. Ее функционал хорошо подходит для статистических задач, которые мы применяем в аналитике данных.

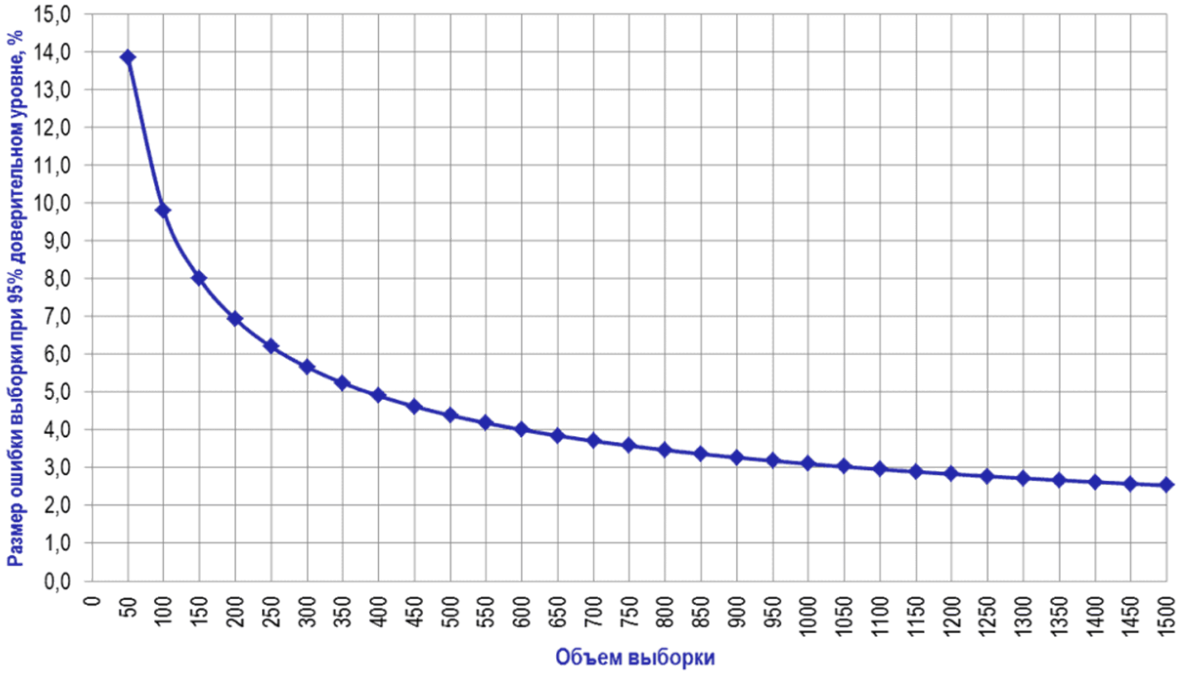
```
conf = 0.95 #доверительный уровень в долях
Z = st.norm.ppf(conf + (1-conf)/2)
p = 0.5
q = 1 - p
delta = 0.04 #ошибка в долях
print(round(((Z**2)*p*q)/delta**2))
```

>>> 600

Объем выборки vs Ошибка

Размер выборки не единственный параметр, по которому мы можем улучшить наше предсказание и сделать его более точным.

На графике показано соотношение размера ошибки выборки при 95% доверительном уровне от объема выборки. С ростом выборки в области минимальных значений падение ошибки существенно. А далее работает **закон убывающей предельной полезности**.



Как определить достаточность выборки

Часто на практике интересующие характеристики могут быть более сложными (их может быть больше).

Вопрос. Как охарактеризовать генеральную совокупность (ГС) «Все жители России» одним бинарным признаком?

Ответ. Никак. Нам необходима **репрезентативность** нашей выборки — отсутствие значимых различий в характеристиках выборки и ГС. В данном случае одного параметра будет недостаточно.

Вернемся к примеру, где мы пытаемся предсказать целесообразность построения станции метро в микрорайоне и проводим уличный опрос около местного автовокзала.

На первом этапе мы попробовали подсчитать оптимальный размер выборки. Получили, что нам нужно опросить 600 человек.

Однако этого мало для проведения исследования, так как исходный дизайн исследования не предполагает репрезентативности. Это во многом связано с методом сбора данных.

ГС людей, которые по разным причинам находятся у автовокзала, очень далека от ГС жителей микрорайона:

- разная причастность к микрорайону;
- разный образ жизни.

Идеально, если выборка формируется случайным алгоритмом. То есть вероятность попадания каждого члена ГС будет одинакова.

+ Возможно, когда мы имеем в распоряжении данные по всей ГС (например, A/B-тестирование на базе метрик нашего сайта).

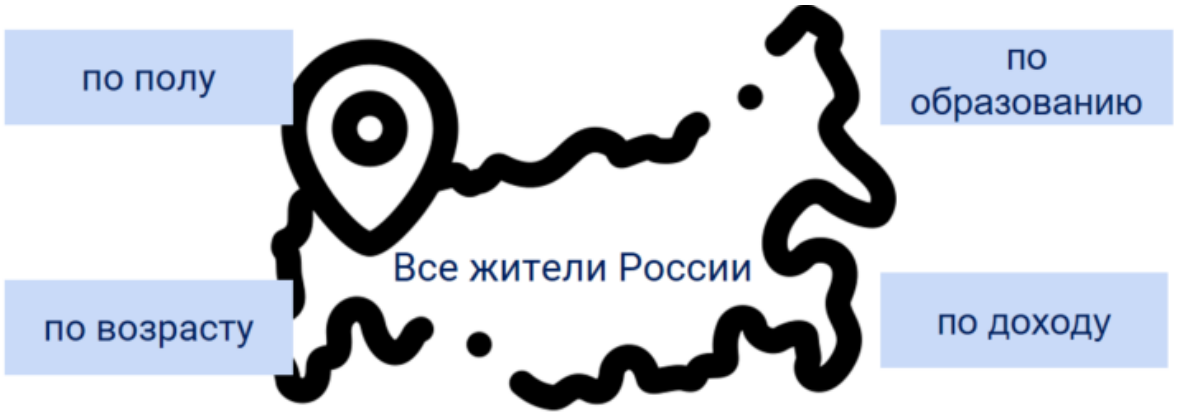
- Невозможно, когда нет технической и финансовой возможности достигнуть каждого потенциального члена ГС с равной вероятностью.

В случаях, когда нет возможности сформировать случайную выборку, можно схитрить: обратиться к отдельной науке о том, как формировать выборки, из которых можно делать более-менее репрезентативные результаты.

Стратификация — выделение подгрупп (страт) на основе важных признаков.

Пример. Благосостояние жителей России зависит от множества параметров:

- От пола. Проблема гендерного равенства до сих пор существует.
- От образования. Как правило, более высокий уровень образования скоррелирован с более высоким уровнем благосостояния.
- От возраста.
- От дохода.



В соответствии с этими признаками мы можем разработать определенные страты и попытаемся в нашей выборке реплицировать все эти страты по долям.

Вернемся к задаче предсказания целесообразности построения станции метро.

Какие стратификационные критерии выбрать:

- Наличие прописки, регистрации, фактического проживания в микрорайоне.
- Потребление транспортных услуг (личный автомобиль, общественный транспорт).
- Социально-демографические критерии (пол, возраст, образование).

Конечного списка здесь предложить нельзя. Нужно опираться на похожие исследования, здравый смысл и метод проб и ошибок.

9. Распределения

Пусть нам удалось сформировать выборку.

Для того чтобы количественно анализировать собранные данные, понимать описательные статистики и правильно их интерпретировать, необходимо знать о концепте распределения.

Случайная величина

Случайная величина — это такая величина, для которой возможное значение в результате эксперимента зависит от такого большого количества разных факторов, что предсказать ее с ходу невозможно.

Пример:

- Значение на верхней плоскости «правильного кубика» в результате его броска (6 значений).
- Рост случайного выбранного студента в аудитории (бесконечное множество всех возможных значений).

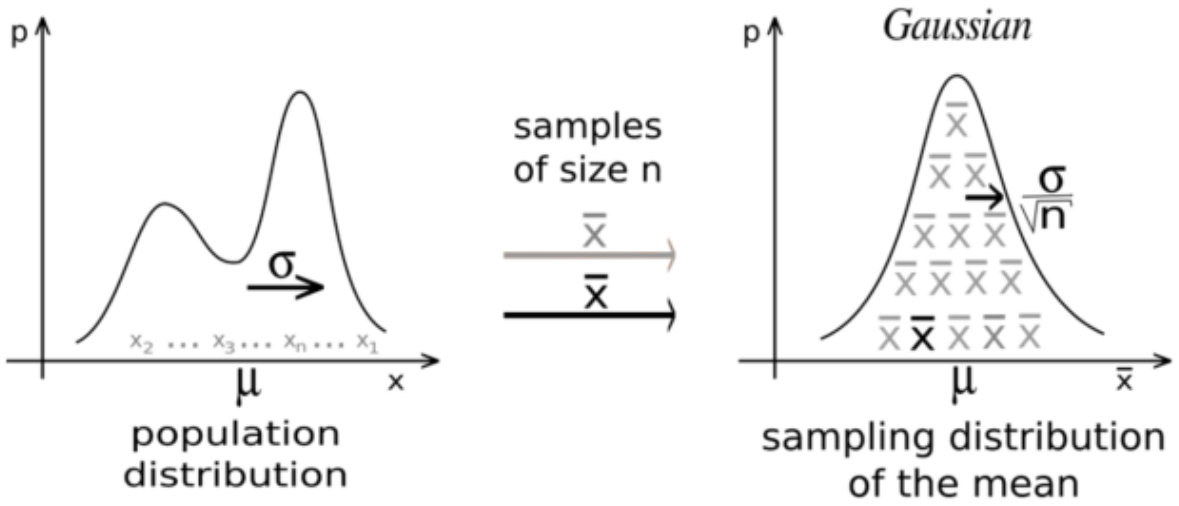
Существует два типа случайных величин:

- **Дискретная СВ** принимает отдельно взятые, изолированные значения (но их число может быть бесконечным).
- **Непрерывная СВ** может принимать абсолютно все числовые значения (на промежутке).

На практике мы часто работаем с суммой сразу всего множества случайных величин, поэтому здесь важно поговорить о **центральной предельной теореме**.

Центральная предельная теорема и нормальное распределение

Центральная предельная теорема (ЦПТ) гласит, что сумма независимых одинаково распределенных случайных величин имеет распределение, близкое к нормальному. Поэтому и распределение параметров этих величин будет нормальным.



Нормальное распределение описывает плотность распределения многих случайных непрерывных величин.

Плотность распределения — это взаимоотношение между значением величины и вероятностью того, что величина примет именно это значение (первая производная от ее функции распределения). То есть плотность распределения показывает частоту встречаемости того или иного значения для нашей случайной величины.

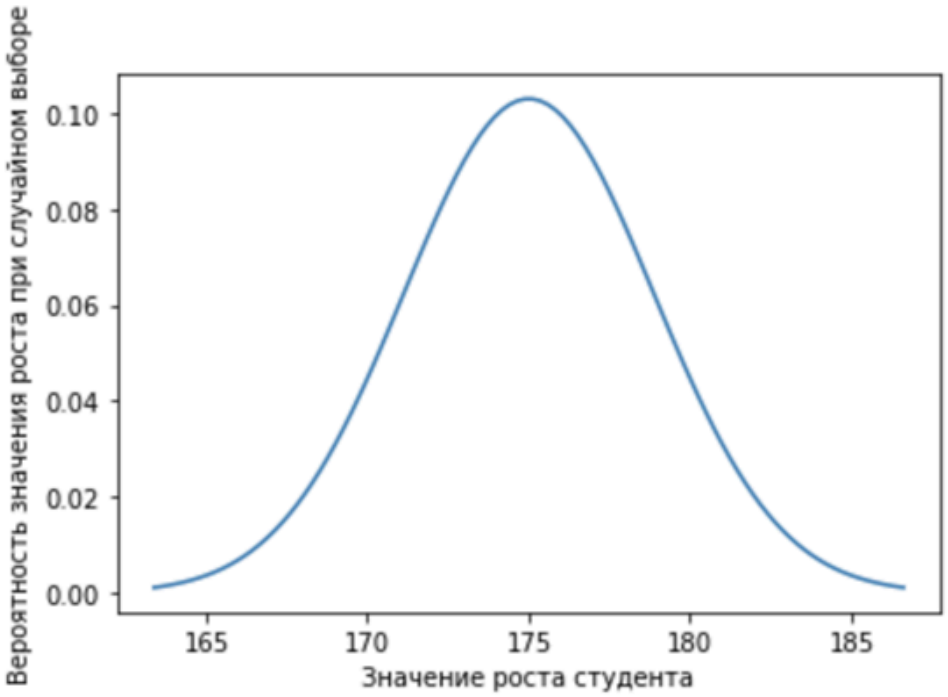
Функция плотности нормального распределения:

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{\frac{(-x-\mu_x)^2}{2\sigma_x^2}}$$

Пример. Представим, что мы хотим посмотреть на распределение роста студентов на потоке.

Итоговое распределение — сумма множества случайных одинаково распределенных случайных величин (рост конкретного человека).

По ЦПТ получим следующую плотность распределения:



У нормального распределения есть свои свойства:

1. Кривая нормального распределения выпукла (колоколообразная), с симметрией относительно среднего, с точками перегиба в значениях абсцисс суммы и разности среднего и стандартного отклонения.
2. Нормальное распределение определяется двумя параметрами: значением генерального среднего (μ) и генерального стандартного отклонения (σ).
3. Медиана и мода нормального распределения совпадают и равны μ.
4. Коэффициенты асимметрии и эксцесса нормального распределения равны нулю.

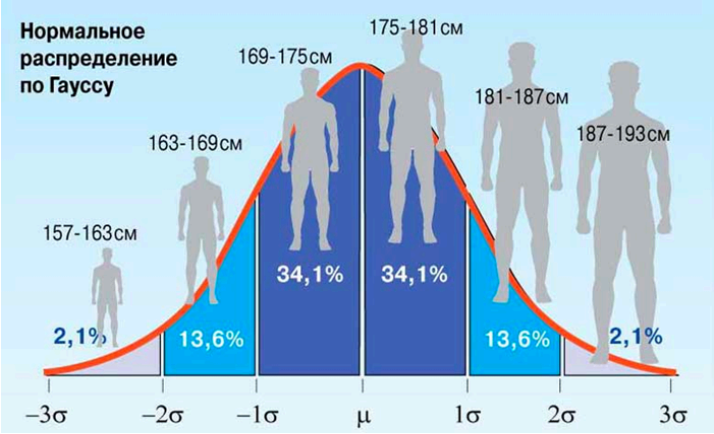
Таким образом, нормальное распределение — идеальная модель, под которую мы подстраиваем некоторые распределения эмпирических данных.

Правило «трех сигм»

Нормальное распределение подчиняется правилу «трех сигм»:

> Если непрерывная СВ распределена нормально, то практически все ее значения лежат в промежутке от разности среднего и трех стандартных отклонений до их суммы.

Пример. Распределение роста человека:



Примерно 68% значений всех наблюдений выборки заключено на промежутке $[-\sigma, \sigma]$.

10. Оценки

Пусть мы собрали некоторую выборку для нашего исследования. Смогли изучить характеристики, построить распределения. Далее нам необходимо вычленить определенные оценки и параметры.

Как можно распространить результаты исследования по выборке на всю генеральную совокупность? Использовать оценки.

Точечная оценка

Точечная оценка параметра — число, оцениваемое на основе наблюдений, предположительно близкое к оцениваемому параметру.

Цель нахождения точечной оценки — это восстановить параметр генеральной совокупности по выборке.

Генеральный параметр	Формула точечной оценки
Среднее μ	$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$
Дисперсия σ^2	$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n-1}$
Доля π	$\hat{\pi} = \frac{\sum_{i=1}^{n} x_i}{n}$

Для полного представления о качестве оценок необходимо знать три свойства, которыми они должны обладать.

1. **Состоятельность.** Оценка стремится к значению генерального параметра.
2. **Несмещенность.** Ее математическое ожидание равно оцениваемому параметру генеральной совокупности.
3. **Эффективность.** Оценка, которая имеет при заданном объеме выборки n наименьшую дисперсию среди всех возможных несмещенных точечных оценок.

Можно ли доверять точечным оценкам? Любая точечная оценка будет иметь погрешность, поскольку в большинстве задач размер выборки и размер генеральной совокупности значимо отличаются.

Поэтому лучше в явном виде смоделировать эту погрешность, чтобы показать, насколько точечная оценка потенциально точно описывает возможный параметр генеральной совокупности.

Пример. Мы провели репрезентативный опрос по поводу строительства новой станции метро в микрорайоне на выборке в 600 человек и выяснили, что 75% жителей поддерживают строительство. Но мы опросили только часть жителей. Даже при идеальном дизайне не может быть 100% уверенности в точности параметра. Поэтому оценим интервал для параметра, чтобы смоделировать погрешность.

Доверительный интервал

Доверительный интервал — это такой интервал, который покрывает неизвестный параметр с заданной надежностью. Чем меньше выборка, тем в большей степени доверительный интервал предпочтительнее точечной оценки.

Общая формула для доверительного интервала:

$$CI = (\hat{\theta} - \epsilon; \hat{\theta} + \epsilon)$$

Надежность — вероятность того, что оценка параметра принадлежит доверительному интервалу. Часто также используют термин статистическая значимость

(*1 — надежность*).

Пример. Надежность 0.99 соответствует уровню значимости 0.01 (также могут задаваться в процентах).

Доверительный интервал для среднего

При нахождении доверительного интервала для среднего для какой-то выборки нужно обратить внимание на ее размер.

Маленькая выборка (не более 30 наблюдений):

$$CI_{mean}^{n \leq 30} = (\hat{\mu} - t\frac{\hat{\sigma}}{\sqrt{n}}; \hat{\mu} + t\frac{\hat{\sigma}}{\sqrt{n}})$$

Большая выборка (более 30 наблюдений):

$$CI_{mean}^{n > 30} = (\hat{\mu} - z\frac{\hat{\sigma}}{\sqrt{n}}; \hat{\mu} + z\frac{\hat{\sigma}}{\sqrt{n}})$$

Здесь:
μ — точечная оценка среднего,
σ — точечная оценка стандартного отклонения,
n — размер выборки,
t — табличное значение, распределение Стьюдента,
z — табличное значение стандартного нормального распределения.

Пример. Дана выборка ростов студентов потока:

```
data = [187, 185, 165, 145, 152, 168, 172, 179, 180, 195, 168,
168, 170, 172, 160]
```

Узнаем, в каком промежутке будет находиться среднее генеральной совокупности с 90, 95 или 99% надежностью.

```
#90% надежность

print(st.t.interval(alpha=0.90, df=len(data)-1,
loc=np.mean(data), scale=st.sem(data)))
```

>>> (165.11615700777816, 177.01717632555517)

```
#95% надежность

print(st.t.interval(alpha=0.95, df=len(data)-1,
loc=np.mean(data), scale=st.sem(data)))
```

>>> (163.82059816147947, 178.31273517185386)

```
#99% надежность

print(st.t.interval(alpha=0.99, df=len(data)-1,
loc=np.mean(data), scale=st.sem(data)))
```

>>> (161.00953301227102, 181.1238003210623)

Для случая большой выборки будет использоваться аналогичный метод:

```
st.norm.interval(alpha=a, loc=np.mean(data), scale=st.sem(data))
```

Доверительный интервал для доли

Рассчитывается по следующей форме:

$$CI_{proportion} = (\hat{\pi} - z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}; \hat{\hat{\pi}} + z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}})$$

Пример. Мы провели репрезентативный корпоративный опрос в компании X. На выборке в 150 респондентов 93 респондента заявили о недовольстве введением фиксированного времени обеда. В каком промежутке с 95% надежностью будет находиться доля недовольных во всей компании:

```
from statsmodels.stats.proportion import proportion_confint

proportion_confint(93, 150, alpha=(1 - 0.95))
```

>>> (0.5423234184516537, 0.6976765815483463)

11. Тестирование гипотез

Для более глубокого изучения данных точечных и интервальных оценок недостаточно.

Как сравнивать параметры между собой? Тестирование гипотез — один из ключевых процессов анализа данных.

Откуда брать гипотезы? Источником гипотез может быть фантазия исследователя, а также устоявшиеся практики в области.

Важно!

- Даже самую креативную и свежую гипотезу нужно уметь переводить на язык статистики.
- Мы должны понимать условия фальсификации и верификации гипотезы.

Тестирование гипотез можно проводить по единому алгоритму.

1. Необходимо определиться с исследовательским вопросом и сформулировать гипотезы на языке статистики. Всегда должна быть сформулирована **нулевая гипотеза** и **альтернатива**, они должны быть взаимоисключающими. Например, возьмем какой-нибудь параметр μ:

    - $H_0: \mu=x, H_1: \mu\neq x.$
    - $H_0: \mu\leq x, H_1: \mu>x.$
    - $H_0: \mu\geq x, H_1: \mu<x.$

    Мы видим, что сумма вероятностей условий происхождения нулевой гипотезы и альтернативной равна 1.

2. Чтобы решить, использовать ли параметрическую или непараметрическую версию теста, мы должны проверить конкретные требования, перечисленные ниже:

    - Наблюдения в каждой выборке независимы и одинаково распределены (IID).
    - Наблюдения в каждой выборке распределены нормально.
    - Наблюдения в каждой выборке имеют одинаковую дисперсию.

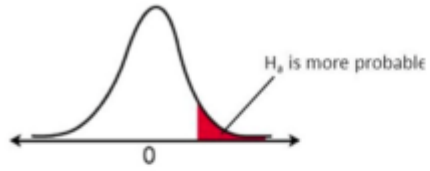
3. Выбор теста.
4. Имплементация теста с выбранными гиперпараметрами (в частности, степень надежности).
5. Интерпретация — статистический вывод.

Статистический вывод делается на основе статистики конкретного теста. Но тестов много, запомнить пограничные значения для каждой спецификации практически невозможно. Поэтому каждый тест предоставляет **p-value (p-значение)** — вероятностную оценку того, что статистика принимает то или иное значение случайно. То есть, чем меньше p-value, тем больше вероятность того, что какая-то закономерность не случайна.

P-value для теста необходимо сравнить с **пороговым значением значимости** (например, 0,05) или статистической надежности (0,95). Превышение порогового значения дает основания отвергнуть нулевую гипотезу в пользу альтернативной.

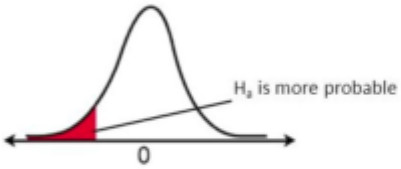
Для одного и того же уровня значимости может существовать множество алгоритмов проверки гипотез.

Гипотеза может быть правосторонней, левосторонней и двусторонней:



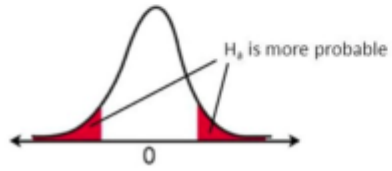
Right-tail test

$H_a: \mu > value$



Left-tail test

$H_a: \mu < value$



Two-tail test

$H_a: \mu \neq value$

Особенностью тестирования гипотез будет заключаться в том, что при двусторонней гипотезе значение статистической значимости будет делиться пополам.

Тестов огромное множество, поскольку есть очень много разных логик постановки гипотез.

Рассмотрим некоторые важные тесты, которые часто встречаются при анализе данных:

- на нормальность данных;
- на равенство дисперсий;
- на сравнение средних;
- на взаимозависимость переменных.

Гипотеза о нормальности

Большое количество тестов часто предполагает нормальность распределения наших данных, поэтому предварительное тестирование на нормальность может быть необходимо.

Тестов, проверяющих на нормальность много, самый популярный — **тест Шапиро-Уилка**.

Предположение:

- Наблюдения в каждой выборке независимы и одинаково распределены (iid).

Интерпретация:

- H_0: выборка имеет нормальное распределение.
- H_1: выборка не имеет нормального распределения.

Пример. Проверим произвольно загруженные данные на нормальное распределение:

```
from scipy.stats import shapiro

data = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360,
-1.478, -1.637, -1.869]

stat, p = shapiro(data)

print('stat=%.3f, p=%.3f' % (stat, p))

if p > 0.05:

 print('Данные скорее всего распределены нормально')

else:

 print('Данные скорее всего распределены не нормально')
```

>>> stat=0.895, p=0.193

>>> Данные скорее всего распределены нормально

Гипотеза о равенстве дисперсий

Также большое количество тестов предполагает равенство дисперсий между двумя выборками (**параметрические** тесты).

Другие не делают такого предположения (**непараметрические** тесты).

Один из вариантов проверки — **тест Левена**.

Интерпретация:

- H_0: две выборки имеют одинаковые дисперсии.
- H_1: две выборки имеют разные дисперсии.

Пример. Представим, что мы хотим проверить, являются ли дисперсии возрастов в двух отделах одной компании одинаковыми:

```
from scipy.stats import levene

group_1 = [14, 34, 16, 43, 45, 36, 42, 43, 16, 27]

group_2 = [34, 36, 44, 18, 42, 39, 16, 35, 15, 33]

alpha =0.05

stat, p =levene(group_1,group_2, center ='mean')

print('stat=%.3f, p=%.3f' % (stat, p))


if p > alpha :

  print("Дисперсии двух выборок скорее одинаковы")

else:

  print("Дисперсии двух выборок скорее различны")
```

>>> stat=0.545, p=0.470

>>> Дисперсии двух выборок скорее одинаковы

T-тест Стьюдента

Перейдем к тестам на сравнение среднего.

T-тест Стьюдента проверяет, существенно ли различаются средние значения двух независимых выборок.

Предположения:

- Наблюдения в каждой выборке независимы и одинаково распределены (iid).
- Наблюдения в каждой выборке распределены нормально.
- Наблюдения в каждой выборке имеют одинаковую дисперсию.

Интерпретация:

- H_0: средние значения выборок равны.
- H_1: средние значения выборок не равны.

Пример. Профессор университета читал онлайн-лекции вместо очных занятий из-за COVID-19. Позже он загружал в облако записанные лекции для студентов, которые асинхронно слушали курс (для тех, кто не посещал урок, но позже смотрел записи). Однако он считает, что учащиеся, которые посещают занятия в учебное время и участвуют в процессе, более успешны. Поэтому он записал средние оценки студентов в конце семестра.

Данные:

```
sync = np.array([94. , 84.9, 82.6, 69.5, 80.1, 79.6, 81.4, 77.8,
81.7, 78.8, 73.2,87.9, 87.9, 93.5, 82.3, 79.3, 78.3, 71.6, 88.6,
74.6, 74.1, 80.6])

asyncr = np.array([77.1, 71.7, 91. , 72.2, 74.8, 85.1, 67.6,
69.9, 75.3, 71.7, 65.7, 72.6, 71.5, 78.2])
```

Предварительный анализ:

```
print(check_normality(sync))

print(check_normality(asyncr))

print(check_variance_homogeneity(sync, asyncr))
```

>>> p-value: 0.6556

>>> Не отвергаем нулевую гипотезу

>>> Данные нормальны

>>> p-value: 0.0803

>>> Не отвергаем нулевую гипотезу

>>> Данные нормальны

>>> p-value: 0.8149

>>> Не отвергаем нулевую гипотезу

>>> Дисперсии одинаковы

Интерпретация. Попробуем сформулировать гипотезы для данного теста.

Нулевая гипотеза: нет различий в средних между группами (синхронные vs асинхронные слушатели).

Альтернативная гипотеза — синхронные слушатели учатся лучше:

```
ttest,p_value = stats.ttest_ind(sync,asyncr)

print("p value:%.8f" % p_value)

print("Односторонння гипотеза >> делим p value на 2 >>
односторонний p value :%.4f" %(p_value/2))

if p_value/2 <0.05:

     print("Средние равны")

else:

     print("Средние не равны")
```

>>> p-value: 0.00753598

>>> Односторонняя гипотеза >> делим p-value на 2 >> односторонний p-value: 0.0038

>>> Средние не равны

Парный t-тест Стьюдента

Парный t-тест Стьюдента проверяет, существенно ли различаются средние значения двух парных выборок.

Предположения:

- Наблюдения в каждой выборке независимы и одинаково распределены (iid).
- Наблюдения в каждой выборке распределены нормально.
- Наблюдения в каждой выборке имеют одинаковую дисперсию.
- Наблюдения двух выборок являются парными

Интерпретация:

- H_0: средние значения выборок равны.
- H_1: средние значения выборок не равны

Пример. Центр здоровья университета диагностировал у восемнадцати студентов высокий уровень холестерина в предыдущем семестре. Медицинский персонал рассказал этим пациентам об опасности высокого уровня холестерина и прописал диету. Через месяц больные пришли на контроль, и у них повторно проверили уровень холестерина. Проверьте, есть ли разница в уровнях холестерина у пациентов.

В соответствии с этой информацией проведите проверку гипотезы, чтобы проверить, есть ли снижение уровня холестерина у пациентов после диеты, используя уровень значимости 0,05.

```
test_stat, p_value_paired =
stats.ttest_rel(test_results_before_diet,test_results_after_diet)

print("p value:%.6f" % p_value_paired , "односторонний p
value:%.6f" %(p_value_paired/2))

if p_value_paired <0.05:

     print("Средний уровень холестерина значимо упал")

else:

     print("Нет разницы в уровне холестерина")
```

>>> p-value: 0.000008, односторонний p-value: 0.000004

>>> Средний уровень холестерина значимо упал

Другие тесты на проверку средних

Мы рассмотрели два теста:

- Равенство дисперсий — независимые выборки.
- Равенство дисперсии — парные выборки.

В случае если условие равенства дисперсий не удовлетворено, будем использовать другие тесты:

- Независимые выборки — тест Манна-Уитни **mannwhitneyu(data1, data2)**.
- Парные выборки — тест Уилкоксона **wilcoxon(data1, data2)**.

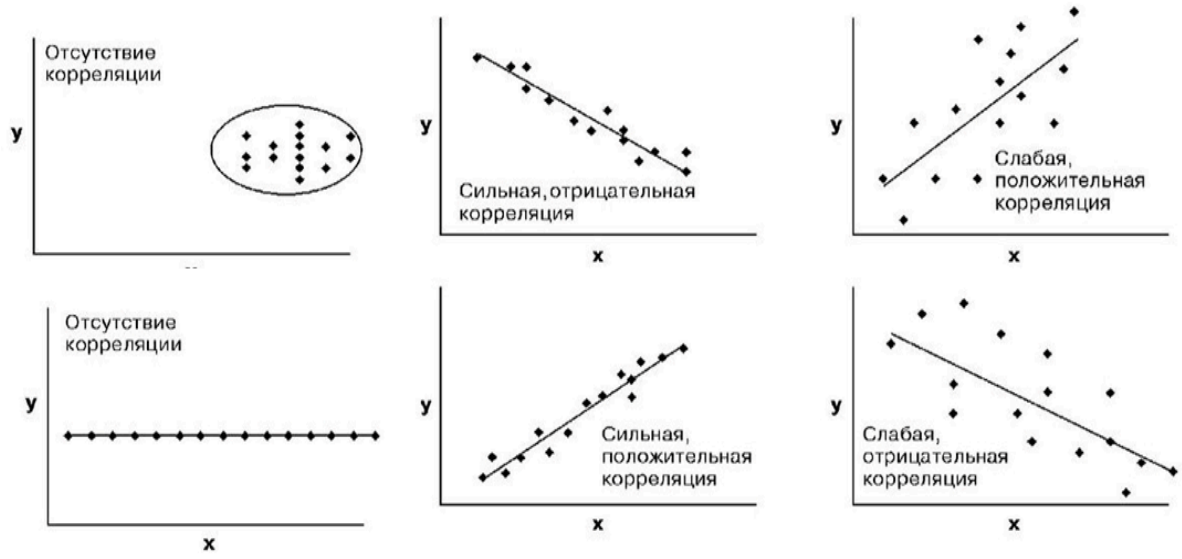
Существует ли взаимосвязь между переменными

Часто нам интересно посмотреть на взаимосвязь между двумя переменными: как связано увеличение/уменьшение одной переменной с увеличением/уменьшением значения другой переменной?

Стандартизированной мерой взаимосвязи переменных является **корреляция**.

Корреляция измеряется от —1 до 1:

- Если коэффициент положителен, то переменные имеют положительную взаимосвязь.
- Если коэффициент отрицателен — обратную взаимосвязь.



Гипотеза о корреляции

Для проверки гипотезы корреляции есть два популярных теста — на значимость корреляции Пирсона (случай равных дисперсий и нормальности данных) и Спирмана (без этих предположений):

H_0: две выборки независимы.

H_1: существует зависимость между переменными.

Пример. Представим, что в задаче SMM-аналитики нам нужно установить, как связано изменение числа подписчиков аккаунта в социальной сети за отчетный период и изменение числа лайков под материалами аккаунта.

Есть данные за последние 10 месяцев:

```
followers_change = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436,
0.360, -1.478, -1.637, -1.869]

likes_change = [0.353, 3.517, 0.125, -7.545, -0.555, -1.536,
3.350, -1.578, -3.537, -1.579]
```

Для проверки гипотезы корреляций импортируем два теста — тест Пирсона и тест Спирмана:

```
from scipy.stats import pearsonr, spearmanr
```

Рассмотрим, как эти два теста работают на реальных данных.

Тест Пирсона:

```
print('stat=%.3f, p=%.3f' % (stat, p))

if p > 0.05:

 print('Переменные независимы')

else:

 print('Переменные зависимы')
```

>>> stat=0.688, p=0.028

>>> Переменные зависимы

Тест Спирмана:

```
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
 print('Переменные независимы')
else:
 print('Переменные зависимы')
```

>>> stat=0.855, p=0.002

>>> Переменные зависимы