Статистика на Python

Вспомним такие базовые понятия, как выборка и генеральная совокупность.

Генеральная совокупность – это множество абсолютно всех объектов, которые используются для исследования.

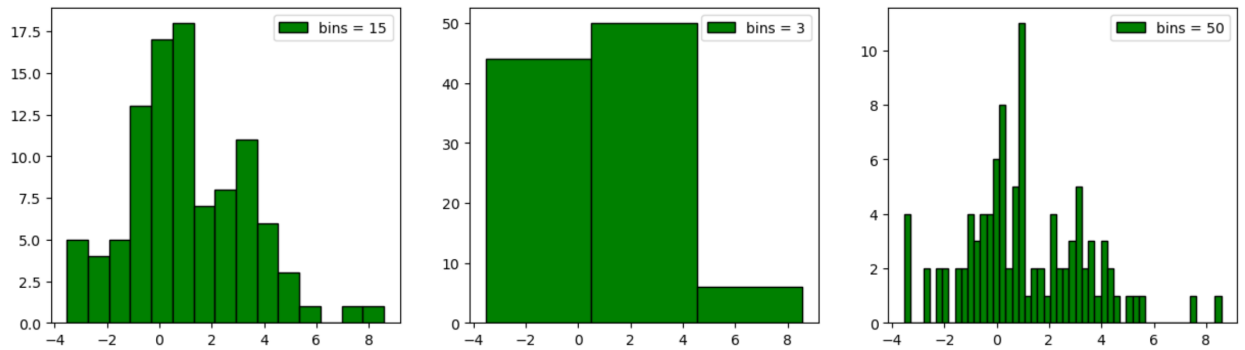
Выборка – это часть генеральной совокупности, которая выбирается для изучения. Очень важно, чтобы выборка была максимально похожа на генеральную совокупность, отражала ее свойства. Степень такой похожести называется **репрезентативностью.**

Статистика занимается тем, что исследует выборки и старается аппроксимировать полученные знания на уровне генеральной совокупности.

Например: показание температуры тела. Собрать показания температуры тела у абсолютно всех жителей России достаточно сложно. Намного проще работать с выборкой, то есть узнать показатель температуры тела у части населения России. Мы пытаемся выявить локальные закономерности исследования, которые могут быть отражены в генеральной совокупности.

Для того, чтобы исследовать форму распределения выборки, используется **гистограмма частот**. По оси абсцисс откладывается значение переменной, а по оси ординат указывается как часто значение этой переменной встречается на определенном интервале. Интервалы можно выбрать разной длины, если интервалы будут слишком большими, то гистограмма будет очень грубой и малоинформативной, если интервалы очень малы, то гистограмма будет очень разреженной. С помощью аргумента bins можно регулировать длину интервалов. Пример с использованием matplotlib:

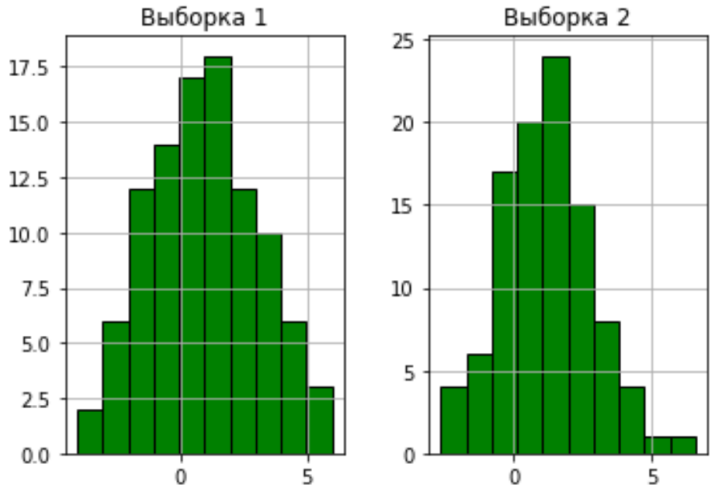
```
fig, ax = plt.subplots(1,3, figsize = (15,4))
ax[0].hist(s2, edgecolor = 'black', color = 'green',bins = 15,label = 'bins = 15')
ax[0].legend()
ax[1].hist(s2, edgecolor = 'black', color = 'green',bins = 3,label = 'bins = 3')
ax[1].legend()
ax[2].hist(s2, edgecolor = 'black', color = 'green',bins = 50,label = 'bins = 50')
ax[2].legend()
plt.show()
```



Аналогичный результат будет получен с использованием серий и датафреймов **pandas**.

```
print(dataframe)
dataframe.hist(color = 'green',edgecolor = 'black')
plt.show()
```

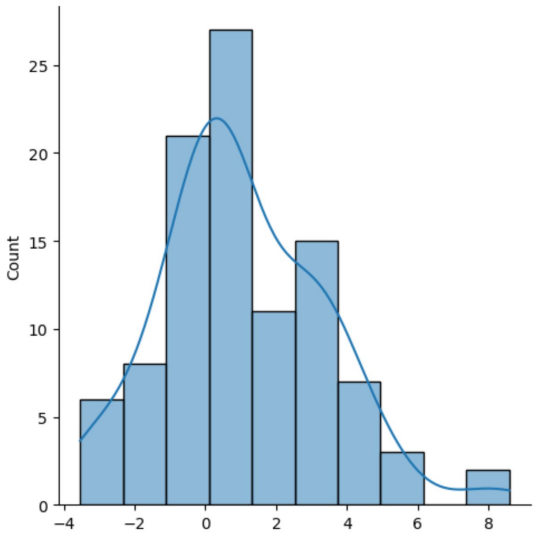
```
    Выборка 1  Выборка 2
0    1.052810   1.813691
1   -1.533349   2.382516
2   -0.635700  -0.419679
3    1.708966   1.652885
4   -1.546923   0.720153
..        ...        ...
95   3.312986   0.930325
96   0.607443  -2.639953
97  -2.633400  -0.004969
98   0.421884   1.600639
99   3.661231  -2.097273

[100 rows x 2 columns]
```



Очень часто для анализа используется библиотека визуализации **seaborn** (import seaborn as sns).

```
1 sns.displot(s2, kde = True)
```

```
<seaborn.axisgrid.FacetGrid at 0x1f6e2e68e20>
```



Непрерывную кривую, которая аппроксимирует гистограмму, можно убрать, используя kde = False. kde – это ядерное сглаживание, которое используется для гладкой оценки плотности распределения.

Одним числом данные можно описать несколькими способами:

– Меры центральной тенденции:

- **Мода.** Это значение, которое наиболее часто встречается в выборке.
- **Медиана.** Для нечетного количества элементов медиана равна центральному элементу в отсортированном массиве $(\text{sort(x)}\left[\frac{n+1}{2}\right])$.

Для четного количества элементов медиана равна среднему двух центральных элементов в отсортированном массиве $(\frac{\text{sort(x)}\left[\frac{n}{2}\right]+\text{sort(x)}\left[\frac{n+1}{2}\right]}{2})$.

- **Среднее.** Сумма значений всех элементов выборки, деленное на их количество.

Стоит отметить, что самой неустойчивой к выбросам мерой является среднее. И наоборот, самой надежной или робастной является мода.

Пример:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as sts
```

```
sp = np.array([1,2,3,2,1,2,4,5,4,100])
```

```
mean = np.mean(sp)
moda = sts.mode(sp)
med = np.median(sp)
```

```
print('Среднее = %f'%mean)
print('Мода: ',moda)
print('Медиана = %f'%med)
```

```
Среднее = 12.400000
Мода:  ModeResult(mode=array([2]), count=array([3]))
Медиана = 2.500000
```

Большая часть выборки лежит на интервале от 1 до 5, соответственно, среднее в данном примере неверно характеризует центральную тенденцию. Мода и медиана более робастны.

Библиотека **scipy** предназначена для решения различных математических задач (решение интегральный, дифференциальных

уравнений, интерполяция, оптимизация и численное решение уравнений и т.д.). Пакет **stats** содержит статистические распределения и функции.

– Меры изменчивости:

- **Размах.** Разница между максимальным и минимальным значением выборки. Очень простая мера, но она использует только два значения из всей выборки. Правильнее использовать каждое значение из выборки для расчета изменчивости данных.
- **Стандартное отклонение.** Это корень из дисперсии, которая вычисляется по формуле $\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$, где $\bar{x}$ – это среднее выборки. Это оценка для выборки. Считается, что стандартное отклонение выборки немного недооценивается, поэтому ее чуть-чуть увеличивают, делив на n-1, а не на n, как для генеральной совокупности. Для генеральной совокупности такой показатель называется **среднеквадратическим отклонением**. Этот показатель позволяет оценить, как сильно меняются данные относительно их среднего. Стандартное отклонение не устойчиво к выбросам.
- **Межквартильный размах (IQR).** Для всех выборок существуют такие отсечки, которые называются «квартили», их всего три: Q1, Q2 и Q3. Межквартильный размах – разность между Q3 (75%) и Q1 (25%), это ширина интервала, который содержит 50% данных. Это метрика полезна для описания данных, она устойчива к выбросам.

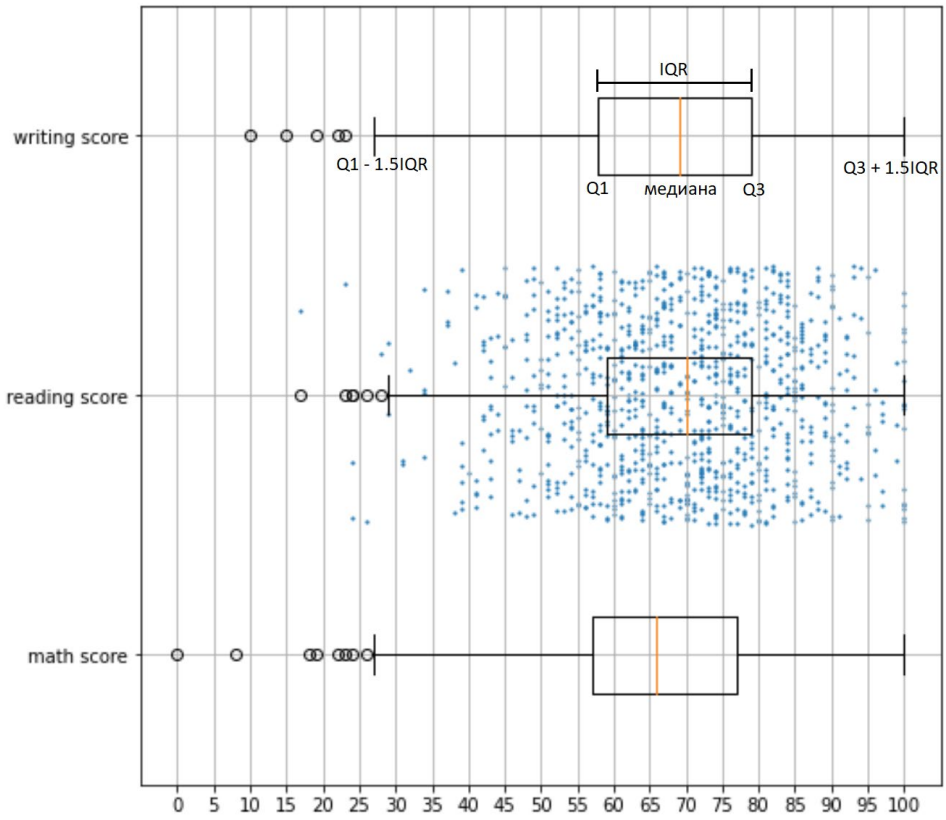
Пример:

```
std = st['writing score'].std()
raz = st['writing score'].max() - st['writing score'].min()
q1 = np.percentile(st['writing score'], 25 , interpolation = 'midpoint' )
q3 = np.percentile(st['writing score'], 75 , interpolation = 'midpoint' )
iqr1 = q3 - q1
iqr2 = sts.iqr(st['writing score'], interpolation = 'midpoint' )

print('Стандартное отклонение: ', std)
print('Размах: ', raz)
print('Межквартильный размах через numpy: ', iqr1)
print('Межквартильный размах через scipy:: ', iqr2)
```

```
Стандартное отклонение:  15.195657010869642
Размах:  90
Межквартильный размах через numpy:  21.5
Межквартильный размах через scipy::  21.5
```

Полезным графиком является «ящик с усами» или box-plot. Это диаграмма, которая используется для отображения случайной величины и несет в себе много полезной информации. Пример диаграммы для данных оценок студентов, которые содержатся в датафрейме result:

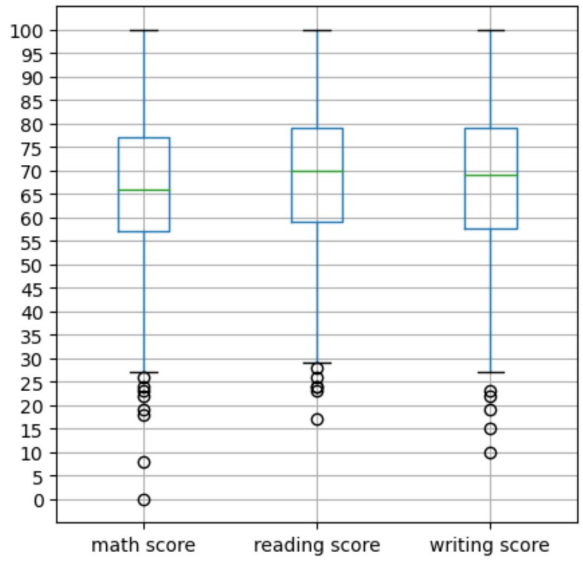
```
plt.figure(figsize=(8,8))
plt.boxplot([result['math score'],result['reading score'],result['writing score']],
            labels=['math score','reading score','writing score'],vert = False)
plt.xticks(np.arange(0,105,5))
plt.scatter(result['reading score'],rand,s=1.5)
plt.grid()
plt.show()
```



Оранжевая линия – это медиана или Q2. Границы коробки – это квартили Q1 и Q3, то есть 50 процентов выборки находится в этом диапазоне. Точки за пределами «усов» – это выбросы. Границы усов – это Q1 – 1.5IQR и Q3 + 1.5IQR в matplotlib. Еще один способ указания границ усов – это максимум и минимум выборки, тогда выбросов на такой диаграмме нет. Синие точки – это reading score, которые изображены для того, чтобы наглядно посмотреть, как box-plot соотносится с распределением данных. Если прямоугольник и усы симметричны, то данные распределяются симметрично без перекоса.

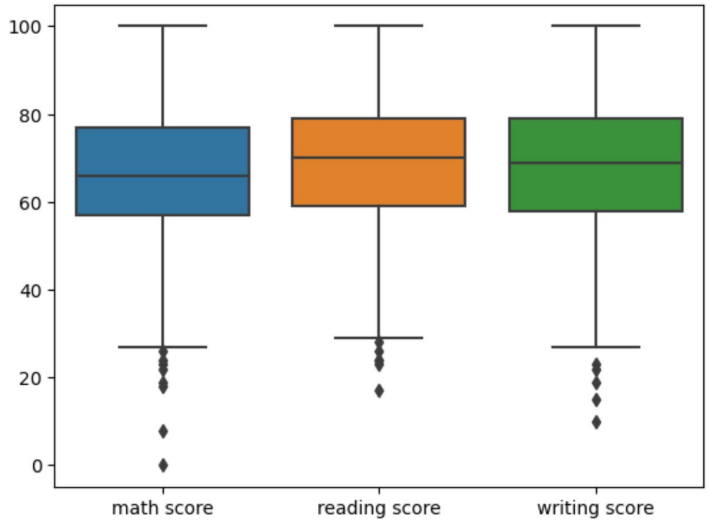
Аналогичный результат будет получен при использовании датафрейма: result.boxplot(). Для каждого столбца данных будет построен boxplot. Также можно использовать библиотеку **seaborn**.

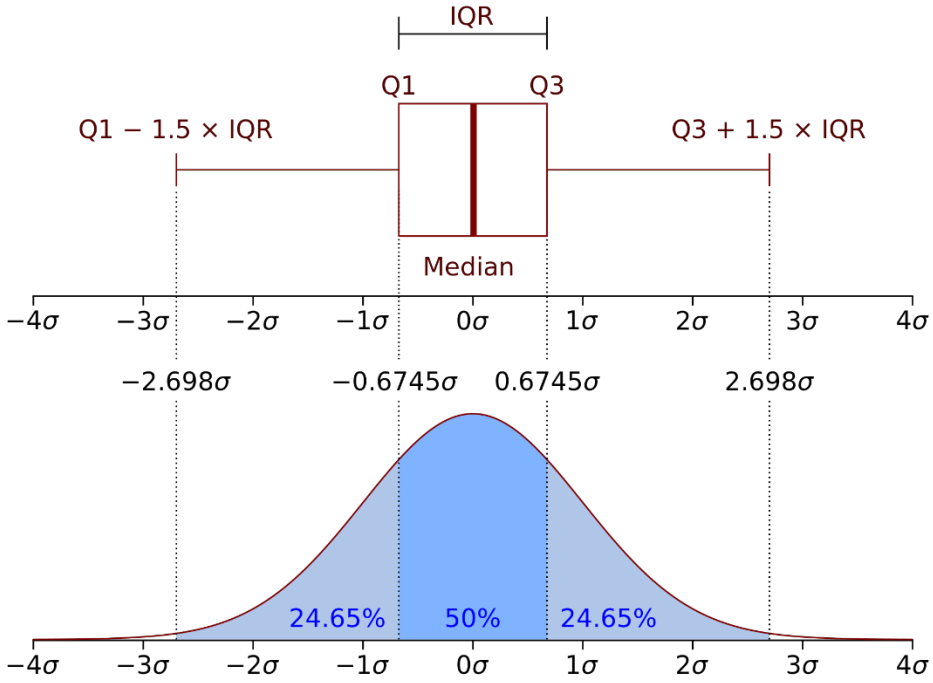
```
plt.figure(figsize=(5,5))
result.boxplot()
plt.yticks(np.arange(0,105,5))
plt.show()
```




```
sns.boxplot(data=result)
```

```
<AxesSubplot:>
```

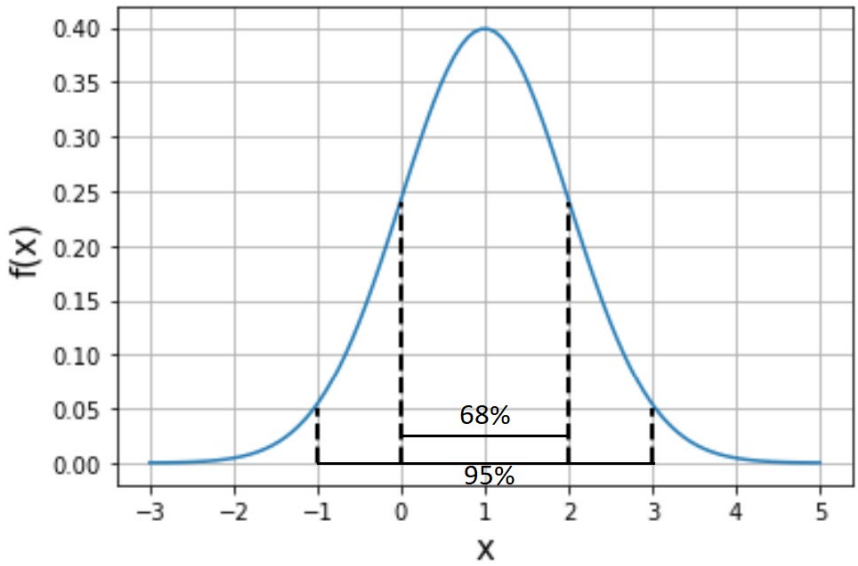




Наиболее популярным является **нормальное распределение или Гауссово распределение,** которое хорошо моделирует результат взаимодействия большого количества слабо зависимых случайных факторов. Это симметричное, унимодальное распределение, которое наиболее часто встречается в различных природных явлениях. Оно имеет два параметра – среднее и стандартное отклонение. Функция плотности вероятности нормального распределения выглядит следующим образом:

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

Функция плотности нормального распределения, где среднее и стандартное отклонение равно 1, представлено ниже.



Одна из важнейших теорем статистики – **центральная предельная теорема**. Допустим, у нас есть некоторая генеральная совокупность с распределением F. Из этой генеральной совокупности мы получаем N выборок. Если для каждой такой выборки мы посчитаем выборочное среднее, то распределение этих средних будет **нормальным**. Чем больше будет длина выборок n, тем больше распределение средних будет унимодальным, тем лучше такое распределение будет аппроксимироваться нормальным распределением со следующими параметрами:

$$\tilde{X}_n = N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

где μ_X – это среднее генеральной совокупности;

σ_X^2 – это дисперсия генеральной совокупности.

Среднее значение всех средних будет очень близко к среднему значению исходной генеральной совокупности.

Стандартное отклонение $\frac{\sigma_X}{\sqrt{n}}$ полученного распределения называется **стандартной ошибкой среднего (SE)**. Оно показывает на сколько в среднем выборочное значение отличается от среднего генеральной совокупности. Чем больше длина выборок n и чем меньше дисперсия исследуемых данных, тем

меньше будет стандартная ошибка среднего. Если количество элементов выборки n > 30 и выборка является репрезентативной, то мы можем вместо среднеквадратического отклонения генеральной совокупности σ_x использовать стандартное отклонение выборки sd_x для оценки стандартной ошибки. Благодаря этому, мы можем узнать стандартную ошибку среднего, взяв только одну выборку длины n из генеральной совокупности и найти ее стандартное отклонение sd_x.

Это работает не только с непрерывными распределениями, но и с дискретными.

Построение доверительных интервалов для среднего значения

Доверительный интервал – это интервал, в пределах которого с заданной вероятностью лежат оценки некоторых статистических характеристик. Две статистики $\hat{\theta}_1$ и $\hat{\theta}_2$ определяют границы доверительного интервала для параметра θ с коэффициентом доверия $1-\alpha$. Вероятность того, что θ лежит между этих двух статистик, больше или равна $1-\alpha$.

$$P\left(\hat{\theta}_1 \le \theta \le \hat{\theta}_2\right) \ge 1-\alpha,$$

где θ – это параметр, который оценивается с помощью интервала;

$1-\alpha$ – уровень доверия;

$\hat{\theta}_1$– нижний доверительный предел;

$\hat{\theta}_2$ – верхний доверительный предел.

Часто на практике коэффициент α принимают равным 0.05. Как правило, длина доверительного интервала возрастает при увеличении коэффициента доверия $1-\alpha$ и стремится к нулю с ростом размера выборки n.

Если повторять эксперимент по построению интервала бесконечно, то в $100(1-\alpha)\%$ случаев, этот интервал будет накрывать истинное значение θ. Это называется 95 % доверительный интервал.

Очень часто исследователей интересует среднее значение исследуемого признака во всей генеральной совокупности.

Например, у нас имеется выборка баллов по экзамену 1000 студентов. Ее среднее $\bar{x}$ равно 68.054 балла, стандартное отклонение *sd* равно 15.19. Но нам интересно узнать, чему равно среднее не в этой выборке, а во всей генеральной совокупности. Но собрать данные всех студентов мы не можем, поэтому абсолютно точное значение среднего генеральной совокупности получить невозможно. Но можем получить интервал, в который будет включено истинное среднее значение. Тут нам помогает центральная предельная теорема. Множество раз извлекаем из генеральной совокупности выборки длины n. Для каждой выборки рассчитываем среднее значение и свой

доверительный интервал, которые рассчитывается по следующей формуле: $\bar{x} \pm 1{,}96\ SE$. 95 % построенных доверительных интервалов содержат истинное значение среднего генеральной совокупности.

Рассчитаем стандартную ошибку среднего для примера с баллами студентов:

$$SE = \frac{sd_x}{\sqrt{n}} = \frac{15.19}{\sqrt{1000}} = 0.48.$$

Далее рассчитаем доверительный интервал, нижняя граница = $\bar{x} - 1{,}96\ SE$, верхняя граница $\bar{x} + 1{,}96\ SE$. Тогда доверительный интервал равен $[67.11, 68.99]$. Мы на 95 % уверены, что такой интервал содержит среднее значение генеральной совокупности. Так же мы можем рассчитать 99% доверительный интервал, где нижняя граница = $\bar{x} - 2.58\ SE$, верхняя граница $\bar{x} + 2.58\ SE$, такой интервал будет шире.

Статистическая проверка гипотез

Допустим, исследователи произвели новый препарат, который позволяет спортсменам улучшить свои результаты. Для проведения эксперимента было отобрано 50 спортсменов. Пусть их средний лучший результат $\bar{x}$ составляет 20 условных единиц. А стандартное отклонение *sd* равно 3. После приема препарата среднее значение результата составило 18 единиц. Возникает вопрос, действительно ли новый препарат позволяет улучшить результаты, или же это статистическая случайность и для всей генеральной совокупности спортсменов улучшения результатов не будет.

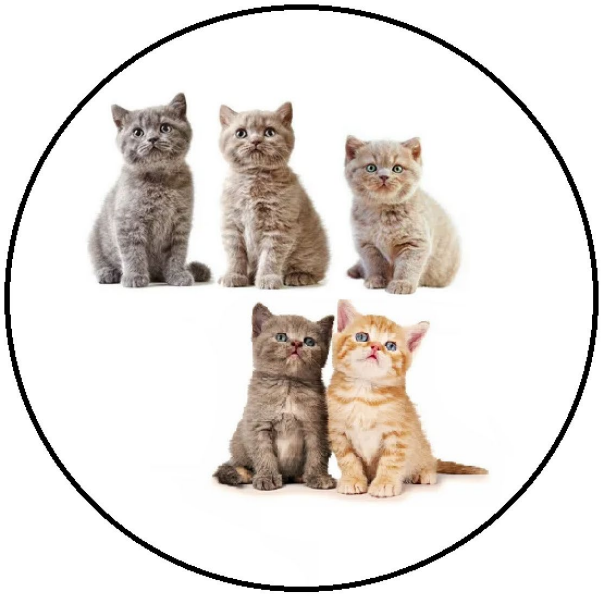
Для начала считается, что результаты никак не отличаются друг от друга, то есть прием нового препарата не дает никаких результатов – это предположение является **нулевой гипотезой (H_0, гипотеза отсутствия различий)**. С другой стороны, есть гипотеза о том, что прием нового препарата дает некие результаты, это **альтернативная гипотеза (H_1, гипотеза о значимости различий).**

Далее рассчитывается так называемый **p-уровень значимости** с помощью различных статистических критериев. Это минимальный уровень, при котором гипотеза отвергается. По сути, p-уровень значимости – это вероятность получить такие же или большие отклонения при условии нулевой гипотезы. То есть чем меньше p-уровень значимости, тем больше оснований отклонить нулевую гипотезу. Как правило, если **p-уровень значимости** менее 0.05, то нулевая гипотеза отвергается и принимается **альтернативная** гипотеза, что выборки все-таки различны. Если **p-уровень значимости** больше 0.05, то нулевая гипотеза не отвергается. То есть наши данные неплохо

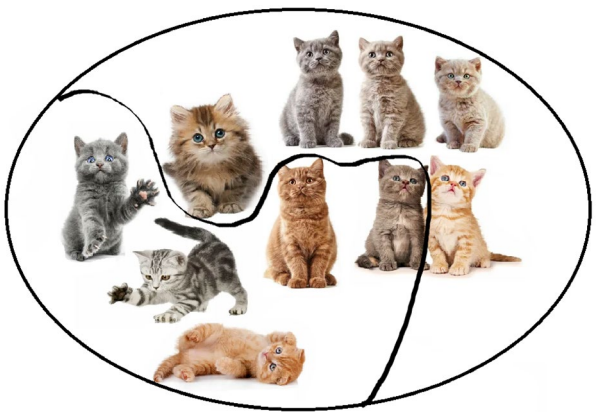
согласуются с нулевой гипотезой и недостаточно оснований для ее отброса. Низший уровень статистической значимости: $p \leq 0{,}05$; достаточный: $p \leq 0{,}01$; высший: $p \leq 0{,}001$. Бывают исследования, когда этот уровень может варьироваться.

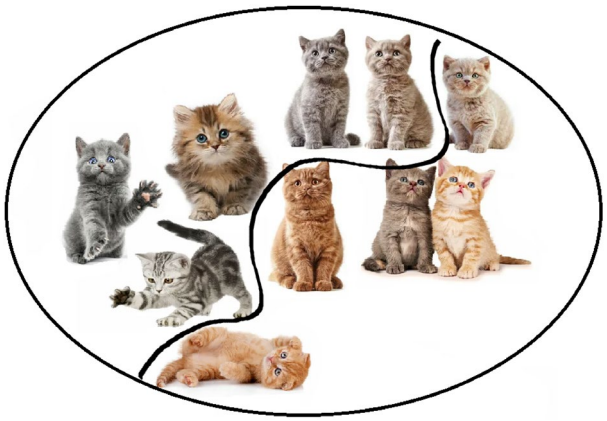
Пример. Имеется две группы котов. Одним давали новый корм, другим нет. Средний вес котов, которым не давали корм, равен 8 кг. Средний вес котов, которым давали корм, равен 5.5 кг, разница 2.5 кг. Вопрос, действительно ли при употреблении нового корма заметно снижается вес животного?

Средний вес = 8



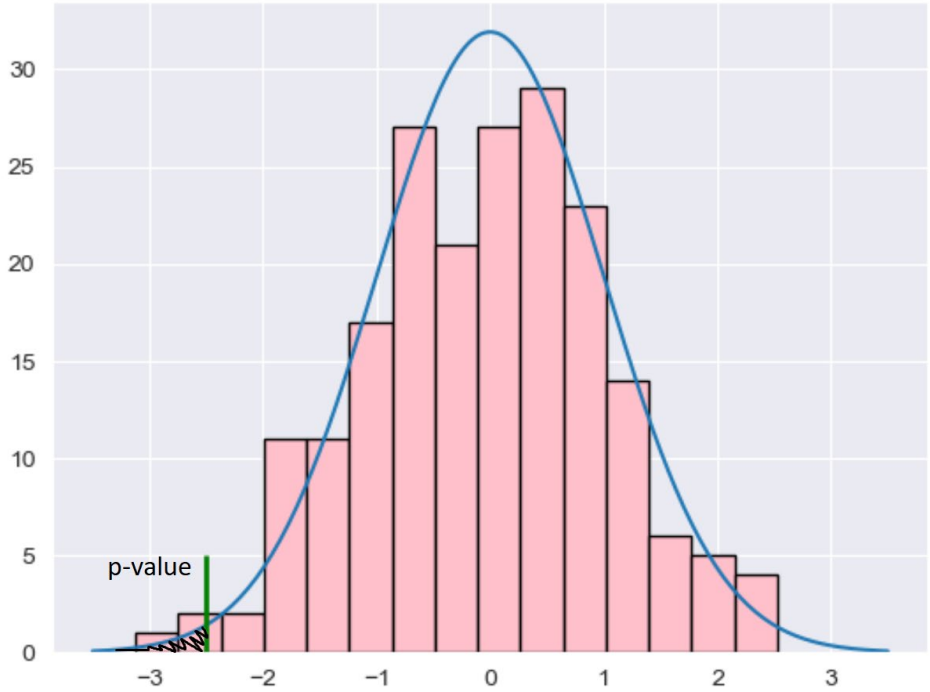
Средний вес = 5.5





Если мы поделим котов на две другие выборки, то получим разницу средних, равную 1 кг., если разделим другим образом, то получим 1.5 кг. То есть прием корма не влияет на вес, это просто разные выборки из одной группы.

Таким образом мы можем разделить исходные данные n раз и изобразить распределение разниц средних значений веса.



Имея ввиду нулевую гипотезу о том, что наша выборка изначально не имеет различий, то есть нет влияния нового корма на котов, мы смотрим, какова вероятность того, что при случайных разбиениях выборки мы получим отклонение большее или равное -2.5. Это и есть p-значение или p-уровень значимости. Если получается p-уровень значимости меньше 0.05, то мы можем отклонить нулевую гипотезу и принять альтернативную. То есть корм оказывает влияния на вес котов.

Гипотезы бывают следующих видов:

- гипотезы о законах распределения;

Статистический вывод – это утверждение, сделанное о параметрах генеральной совокупности, которое основывается на результатах исследования выборки из генеральной совокупности.

Существует два рода ошибок статистического вывода при проверке статистической гипотезы:

- **ошибка I рода (α-ошибка)** – отклоняется нулевая гипотеза, но она была верна. Вероятность ошибки первого рода называют уровнем значимости и обозначают α.
- **ошибка II рода** – нулевая гипотеза не отклоняется при том, что она не верна, а альтернативная гипотеза является верной. Вероятность ошибки второго рода обозначается буквой β.

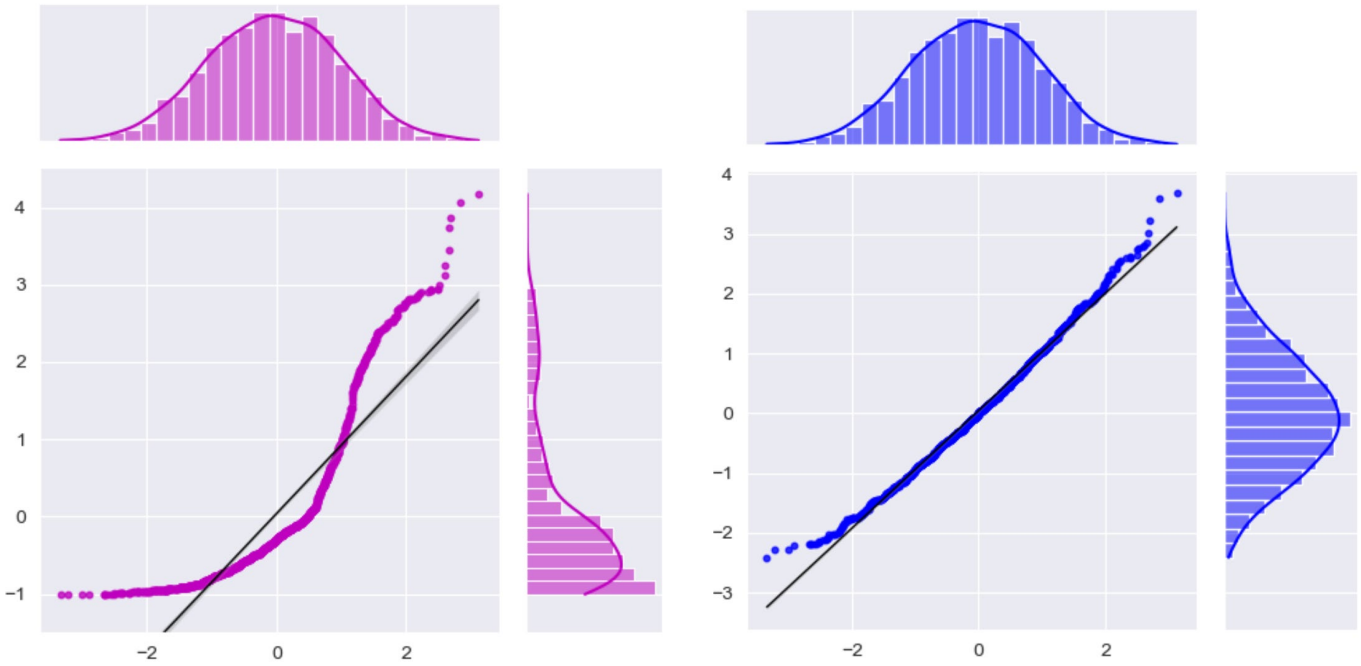
Проверка на нормальность

Достаточно часто в статистике присутствует требование к нормальному распределению при использовании различных методов. Посмотрим, как можно определить, насколько сильно распределение исследуемых данных отличается от нормального теоретического распределения.

Первый способ – это поверх гистограммы частот построить теоретическую кривую нормального распределения. Также можно воспользоваться box-plot для оценки симметричности распределения. Если медиана находится в центре прямоугольника и «усы» симметричны, то это нормальное распределение. Можно оценить моду, медиану и среднее выборки, как известно, у нормального распределения меры центральной тенденции равны.

Еще один способ – это график Q-Q plot (Quantile-Quantile plot). Представляет собой зависимость исходных значений выборки и значений идеального нормального распределения. Если наблюдается идеальная прямая, то данные следуют нормальному закону, если наблюдается отклонение выше прямой, то исходные значения выше, чем нормальные, и наоборот. Удобно использовать Q-Q plot, когда данных немного. Также Q-Q plot позволяет определить асимметрию в данных. Пример с использованием библиотеки seaborn:

```
g = sns.jointplot(x=q2, y=q12_s,
                  kind="reg", truncate=True,
                  color="b", height=5,ratio=3,
                  scatter_kws={"s": 10,}, line_kws={"lw": 1, 'color':'black'})
```



По оси абсцисс откладываются значения стандартного нормального распределения, по оси ординат – распределение исследуемой выборки. Левый график показывает, что распределение исследуемой выборки сильно отличается от нормального. На правом графике середина распределения следует нормальному закону, но его концы отклоняются от него.

Существует множество тестов для проверки распределения на нормальность, некоторые из них и наиболее часто используемые представлены ниже.

Тест Колмогорова-Смирнова (KS-тест)

Это непараметрический тест, который позволяет оценить существенность различий между распределениями двух выборок, например, оценка соответствия распределения исследуемой выборки закону нормального распределения. Критерий Колмогорова-Смирнова определяет расстояние между эмпирической функцией распределения выборки и функцией распределения эталонного распределения (это не обязательно распределение Гаусса). В случае проверки на нормальность распределения, выборки стандартизуются и сравниваются со стандартным нормальным распределением. Данный тест эффективен при размере выборки ≥ 50.

kstest. Пример:

```
11 test_sk = stats.kstest(ch, 'norm')
12 print(test_sk)
```

```
KstestResult(statistic=0.18846450965981876, pvalue=4.381949677777384e-42)
```

ch – это стандартизированная исследуемая выборка. Статистика теста Колмогорова-Смирнова – это максимальная абсолютная разница между двумя кумулятивными распределениями. Значение статистики необходимо

сравнивать с критическим значением из таблицы. Если полученное значение выше критического, то нулевая гипотеза может быть отброшена.

p-значение намного меньше 0.05, следовательно нулевая гипотеза отвергается и выборка не имеет нормального распределения.

Тест Андерсона-Дарлинга. Позволяет проверить, получена ли выборка данных из заданного распределения вероятностей.

Тест Лиллифорса. Это тест на нормальность, который основан на тесте Колмогорова–Смирнова.

Тест Шапиро-Уилка

Гипотеза о нормальности распределения отбрасывается, если значение p-уровня значимости меньше выбранного уровня α. Нулевая гипотеза – распределение выборки НЕ отличается от нормального, альтернативная гипотеза – распределение отличается от нормального. Данный тест дает отличные результаты на небольших размерах выборок (≤ 50).

Данный тест можно использовать с помощью scipy.stats.shapiro. Пример:

```
12 test = stats.shapiro(ch)
13 print(test)
```

```
ShapiroResult(statistic=0.814687967300415, pvalue=1.150477698013898e-36)
```

statistic – это статистика критерия W, которая вычисляется по следующей формуле:

$$w = \frac{1}{s^2}[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)]^2$$

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

Коэффициенты a_{n-i+1} берутся из таблицы. Значение тестовой статистики сравнивается с критическим значением для данного размера выборки и ранее определенного уровня значимости. Для критических значений существуют готовые таблицы. Если значение тестовой статистики больше критического значения, нулевая гипотеза не отклоняется. Статистику теста можно интерпретировать как коэффициент корреляции, который может принимать значения от 0 до 1. Чем ближе статистика теста к 1, тем меньше отклонений фактической дисперсии от гипотетической дисперсии при нормальном распределении.

На скриншоте выше p-значение намного меньше 0.05, это высший уровень статистической значимости, следовательно, нулевая гипотеза может быть отброшена, распределение выборки не является нормальным. Если

размер выборки достаточно велик, этот тест обнаруживает незначительные отклонения от нулевой гипотезы, то есть p-значение будет очень маленьким и нулевая гипотеза будет отброшена, при том, что она верна.

Практическая работа

1. Загрузить данные из файла “insurance.csv”.
2. С помощью метода describe() посмотреть статистику по данным. Сделать выводы.
3. Построить гистограммы для числовых показателей. Сделать выводы.
4. Найти меры центральной тенденции и меры разброса для индекса массы тела (bmi) и расходов (charges). Отобразить результаты в виде текста и на гистограммах (3 вертикальные линии). Добавить легенду на графики. Сделать выводы.
5. Построить box-plot для числовых показателей. Названия графиков должны соответствовать названиям признаков. Сделать выводы.
6. Используя признак charges или imb, проверить, выполняется ли центральная предельная теорема. Использовать различные длины выборок n. Количество выборок = 300. Вывести результат в виде гистограмм. Найти стандартное отклонение и среднее для полученных распределений. Сделать выводы.
7. Построить 95% и 99% доверительный интервал для среднего значения расходов и среднего значения индекса массы тела.
8. Проверить распределения следующих признаков на нормальность: индекс массы тела, расходы. Сформулировать нулевую и альтернативную гипотезы. Для каждого признака использовать KS-тест и q-q plot. Сделать выводы на основе полученных p-значений.
9. Оформить отчет на основе проделанной работы. Написать выводы.