KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 4 урок > СТАТИСТИКА

> Оглавление

> Нормальное распределение и ограниченность количества наблюдений

Предположим, что у нас есть генеральная совокупность (ГС), где среднее равно 0, а стандартное отклонение 1. Если мы многократно извлекаем выборки из ГС, то

все средние значения этих выборок распределятся нормальным образом вокруг среднего ГС, со стандартным отклонением (стандартной ошибкой среднего), которую можно рассчитать разделив стандартное отклонение ГС на корень из числа наблюдений:

$$se = \frac{\sigma}{\sqrt{n}}$$

Это удобно до тех пор, пока у нас есть большое количество наблюдений. В таком случае стандартное отклонение по выборке хорошо описывает соответствующие параметры генеральной совокупности, что позволяет преобразовать формулу в

$$se = \frac{sd}{\sqrt{n}}$$

И именно при большом количестве наблюдений все выборочные средние будут вести себя в соответствии с нормальным распределением.

Все становится интересней, когда кол-во наблюдений не так велико (особенно важно, когда оно меньше 30). В таком случае:

- стандартное отклонение по выборке – не самый хороший показатель соответствующих параметров ГС
- нарушается предположение о том что все выборочные средние будут вести себя в соответствии с нормальным законом

Почему так происходит? Предположим, мы извлекаем большие выборки из ГС (более 100 наблюдений в каждой). Тогда все выборочные средние распределятся нормальным образом вокруг среднего ГС. Это будет нормальное распределение, т.е. 95% всех выборочных средних будут лежать в диапазоне $\pm 2\sigma$.

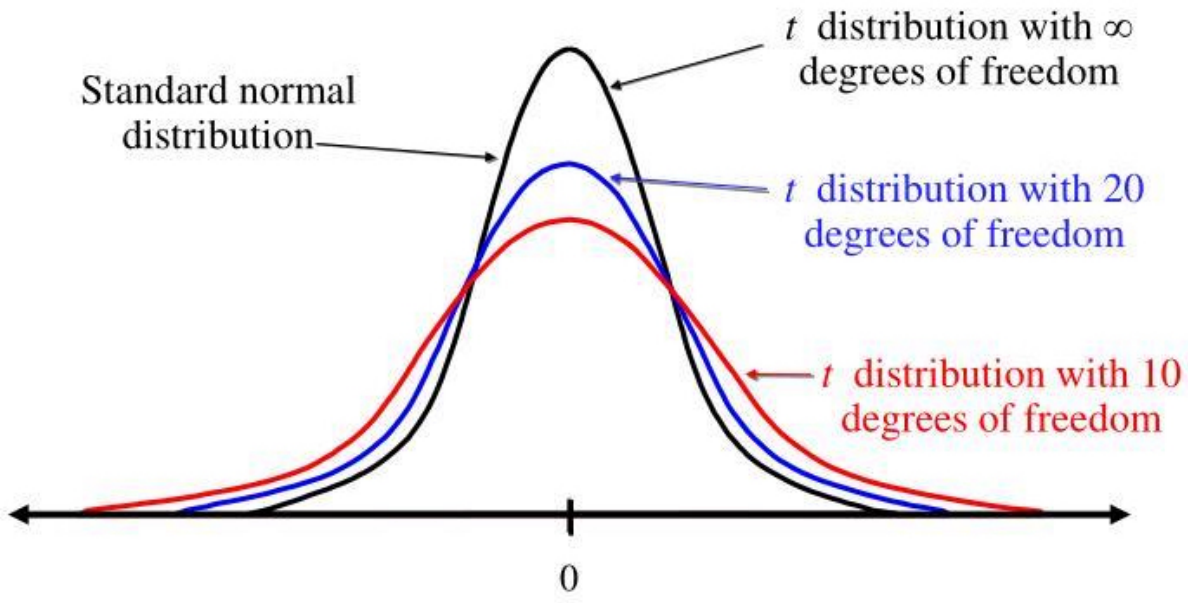
Теперь представьте, что мы очень сильно снизили объем выборок и многократно извлекаем выборки по 5 элементов в каждой. В такой ситуации мы будем гораздо чаще получать выборочные средние, которые довольно далеко отклоняются от среднего ГС. Например, в ГС среднее равно 10, а мы взяли всего 5 элементов. Неудивительно, что в полученной выборке мы можем получить среднее = 20. Или наоборот равное 3. То есть:

- Чем меньше выборка, тем больше довольно сильных отклонений мы будем получать от среднего ГС

> Распределение Стьюдента (Т-распределение)

Если число наблюдений в выборке невелико и σ(стандартное отклонение генеральной совокупности) неизвестно (почти всегда), используется распределение Стьюдента (T-distribution), чтобы описать, как будут себя вести все выборочные средние.

- Унимодально
- Симметрично
- Но: наблюдения с большей вероятностью попадают за пределы ±2σ от M



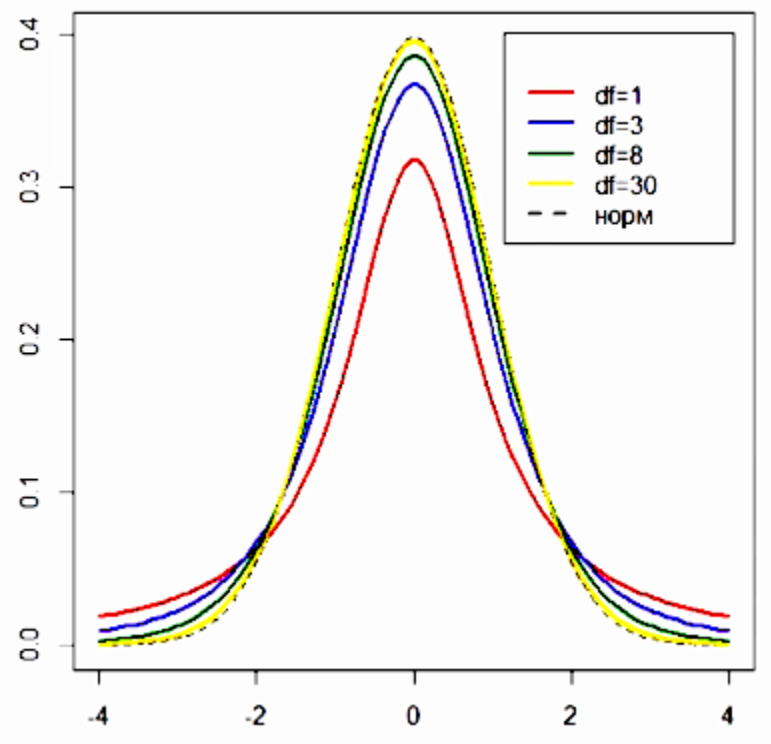
Посмотрим на отличия от нормального распределения более подробно:

Основное отличие: у распределения Стьюдента более "высокие" хвосты распределения. Это означает, что в диапазоне, превышающем 2 стандартных

отклонения вправо будет лежать больше наблюдений, чем у норм распределения, и аналогично – в диапазоне минус 2 стандартных отклонения будут чаще встречаться такие сильно выраженные отклонения от среднего ГС.

Важный параметр распределения Стьюдента: число степеней свободы. Оно зависит от количества наблюдений в нашей выборке. Так:

- Чем больше степеней свободы и чем больше наблюдений, тем всё больше распределение становится похожим на нормальное



Например, всего два наблюдения:

- $df = n - 1 = 2 - 1 = 1$
- Красное распределение имеет очень высокие хвосты в левой и правой части
- В диапазоне от 2 до 4 ст. отклонений будем иметь довольно большой процент наблюдений
- Выборочные средние будут очень далеко отклоняться от реального среднего в ГС

По мере добавления большего кол-ва наблюдений в выборку, распределение Стьюдента начинает потихоньку подбираться к нормальному, и наблюдается всё меньше сильных отклонений.

Так например, возьмем выборку из 31 наблюдения: видим, что в такой ситуации Т-распределение с 30 степенями свободы очень близко подобралось к реальному нормальному распределению. При этом мы замечаем, что хвосты распределения будут расположены чуть выше, чем у нормального распределения. В этой ситуации сильное отклонение от среднего мы будем получать чаще, чем в ситуации когда распределение абсолютно нормально.

В отличие от нормального распределения, где отклонение от среднего строго регламентировано, форма Т-распределения будет изменяться в зависимости от числа степеней свободы. Получить довольно экстремальные отклонения от среднего значения будет более или менее вероятно в зависимости от того, какая большая выборка по объему.

Предположим, что в ГС $\mu = 10$, а в выборке $\bar{X} = 10.8, sd = 2, N = 25$. Если бы мы использовали стандартную формулу, то сказали бы, что в соответствии с ЦПТ, все выборочные средние распределились бы нормально вокруг среднего ГС. Стандартную ошибку среднего (ст. отклонение этого распределения) мы могли бы рассчитать, разделив ст. отклонение выборки на корень из числа элементов:

$$se = \frac{sd_x}{\sqrt{n}} = \frac{2}{\sqrt{25}} = 0.4$$

Найдем Z-значение:

- Из нашего среднего вычтем среднее ГС
- Разделим на стандартное отклонение распределения всех выборочных средних

$$Z = \frac{10.8 - 10}{0.4} = \frac{0.8}{0.4} = 2$$

Мы получили отклонение от предполагаемого среднего на 2 стандартных отклонения вправо. Найдем вероятность получить такое или более выраженное отклонение.

Сайт: https://gallery.shinyapps.io/dist_calc/

Если мы допустили, что все выборочные средние будут распределены нормальным образом, то вероятность получить отклонения, превышающие 2*σ*, как в левую, так и в правую стороны, составит 0.0455. Это означает, что p-уровень значимости будет меньше, чем 0.05. Мы можем отклонить H_0, согласно которой наша выборка принадлежит ГС со средним 10.

Но: при небольшом объеме выборки распределение выборочных средних будет отличаться от нормального, и вероятность получить более выраженное отклонение от среднего станет выше. Рассчитаем данную вероятность, если мы предположим, что мы работаем с т-распределением, и т.к. у нас 25 элементов в выборке, то df=24 степени свободы. В этом случае получить отклонение на 2 сигмы в ту или иную сторону составит уже 0.0569. А значит, что если бы мы пользовались т-распределением, то H0 отклонить мы бы уже не смогли.

Т-критерий рассчитывается как и Z-значения:

- Из выборочного среднего вычитаем среднее ГС
- Делим на стандартное отклонение выборки, деленное на корень из N

Однако, если бы мы в этом случае получили 2, то в Т-распределении с 24 степенями свободы p-уровень значимости составил бы уже 0.056, и H_0 мы бы не смогли отклонить.

$$t = \frac{\bar{X} - \mu}{\frac{sd}{\sqrt{n}}}$$

Подробнее остановимся на теоретической части.

В лекциях было сказано, что мы используем Т-распределение в ситуации небольшого объема выборки. Необходимо более подробно пояснить, зачем это нужно.

Вернемся к предельной центральной теореме, мы уже узнали, что если некий признак в генеральной совокупности распределен нормально со средним *μ* и стандартным отклонением *σ*, и мы будем многократно извлекать выборки одинакового размера *n*, и для каждой выборки рассчитывать, как далеко выборочное среднее $\bar{X}$ отклонилось от среднего в генеральной совокупности в единицах стандартной ошибки среднего:

–

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

то эта величина z будет иметь стандартное нормальное распределение со средним равным нулю и стандартным отклонением равным единице.

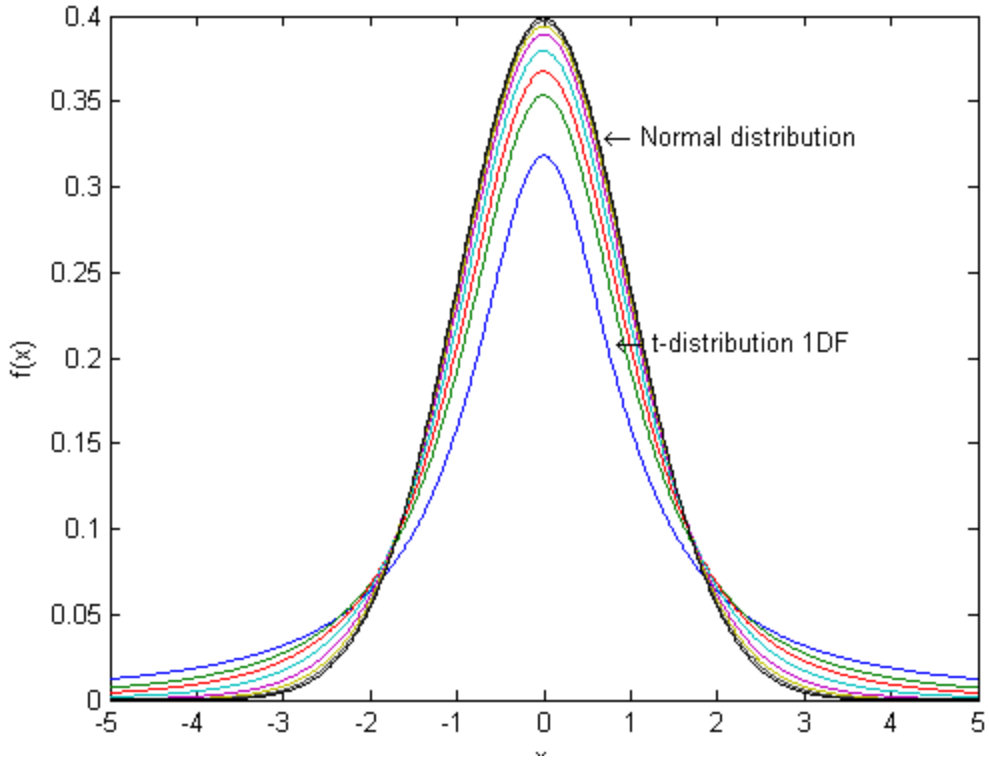
Обратите внимание, что для расчета стандартной ошибки мы используем именно стандартное отклонение в генеральной совокупности - *σ*. Ранее мы уже обсуждали, что на практике *σ* нам практически никогда не известна, и для расчета стандартной ошибки мы используем выборочное стандартное отклонение.

Так вот, строго говоря, в таком случае распределение отклонения выборочного среднего и среднего в генеральной совокупности, деленного на стандартную ошибку, теперь будет описываться именно при помощи t - распределения.

$$t = \frac{\bar{X} - \mu}{\frac{sd}{\sqrt{n}}}$$

Таким образом, в случае неизвестной *σ* мы всегда будем иметь дело с t-распределением. На этом этапе вы должны с негодованием спросить, почему же мы применяли z - критерий в первых уроках для проверки гипотез, используя выборочное стандартное отклонение?

Мы уже знаем, что при довольно большом объеме выборки (обычно в учебниках приводится правило, n > 30) t - распределение совсем близко подбирается к нормальному распределению:



Поэтому иногда, для простоты расчетов говорится, что если n > 30, то мы будем использовать свойства нормального распределения для наших целей. Строго говоря, это конечно неправильный подход, который часто критикуют. В до компьютерную эпоху этому было некоторое объяснение, чтобы не рассчитывать для каждого n больше 30 соответствующее критическое значение t - распределения, статистики как бы округляли результат и использовали нормальное распределение для этих целей. Сегодня, конечно, с этим больше никаких проблем нет, и все статистические программы, разумеется, без труда рассчитают все необходимые показатели для t - распределения с любым числом степеней свободы. Действительно при выборках очень большого объема t - распределение практически не будет отличаться от нормального, однако, хоть и очень малые но различия все равно будут.

Поэтому, правильнее будет сказать, что мы используем t - распределение не потому что у нас маленькие выборки, а потому что мы не знаем стандартное отклонение в генеральной совокупности. Поэтому в дальнейшем стоит использовать t - распределение для проверки гипотез, если нам неизвестно стандартное отклонение в генеральной совокупности, необходимое для расчета стандартной ошибки, даже если объем выборки больше 30.

> Понятие числа степеней свободы

Число степеней свободы – количество элементов, которые могут варьироваться при расчете некоторого статистического показателя. Например, если у нас есть 10 наблюдений и мы знаем среднее значение по этим 10 наблюдениям, то нам достаточно знать среднее и только 9 из них, чтобы узнать, чему равен 10 оставшийся элемент. Т.е. у него нет никакой возможности варьировать свои значения.

- В случае Т-распределения зависит от количества наблюдений

Важно понимать, сколько элементов информации мы использовали для расчета того или иного показателя. Представьте, что мы знаем среднее значение ГС, а в нашей выборке мы получили отклонение на 2σ вправо. Что же это может значить? Совершенно разные вещи в зависимости от того, как много элементов мы использовали, чтобы получить это Т значение. Если у нас Т-распределение с 30 степенями свободы, то отклонение на 2σ или более будет приводить к определенным результатам (например, отклонению H_0). Если такой результат получится в случае, когда есть 3 степени свободы, то это приведет к абсолютно другим выводам.

Таким образом, важно учитывать как много независимых элементов мы используем для того, чтобы получить это значение. В статистике в большинстве методов всегда будет указываться число степеней свободы. В большинстве случаев это будет связано именно с размером выборки (как много элементов в нашей выборки позволили нам сделать оценку того или иного статистического параметра).

Дополнительная статья на Habr: https://habr.com/ru/company/stepic/blog/311354/

> Сравнение двух средних

Критерий, который позволяет сравнивать две выборки между собой (два выборочных средних), называется парный t-тест (t-test), или просто t-критерий Стьюдента.

Как же он работает? Предположим, мы хотим сравнить два средних выборочных значения: X_1, рассчитанное по выборке со sd_1 и числом элементов n_1, и выборочное значение X_2 с sd_2 и n_2.

Гипотезы:

H_0 – в генеральной совокупности никакого различия между средними значениями нет

H_1 – средние в генеральной совокупности не равны (альтернативная гипотеза)

Для начала принимаем за правду факт того, что нулевая гипотеза верна. Если это так, то при многократном повторении нашего эксперимента (извлекали выборки и рассчитывали разность между двумя выборочными средними значениями), $X_1 - X_2$ распределилась бы симметрично. Если бы мы предположили, что на самом деле в ГС два средних равны, то и среднее значение разности этих значений равнялось бы нулю. При этом стандартное отклонение (стандартная ошибка) рассчитывалась бы по формуле:

$$se = \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$$

Как мы видим, стандартная ошибка среднего для первого среднего и стандартная ошибка для второго среднего вносят свой вклад. Даже если мы вычисляем разность между двумя средними значениями, чем больше стандартная ошибка среднего для каждого из них, тем больше возможных комбинаций может принять разность между этими средними значениями и тем больше будет вариативность такого показателя.

При достаточно большом кол-ве наблюдений мы могли бы сказать, что распределение разности между двумя средними значениями приняло бы нормальный вид, что будет правдой в соответствии с ЦПТ. Еще более точно будет сказать, что такое распределение будет соответствовать Т-распределению с числом степеней свобод, которое рассчитывается по след формуле

$$df = n_1 + n_2 - 2$$

то есть

$df_1 = n_1 - 1, df_2 = n_2 - 1$

Основываясь на этой информации, мы можем рассчитать, насколько далеко конкретно наша разность между двумя средними отклонилась от предполагаемого

показателя ГС, тем самым рассчитать вероятность получить такие же или еще более выраженные отклонения при условии что на самом деле верна H_0.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}$$

Поскольку $\mu 1 - \mu 2 = 0$ (в соотв. с H_0), то формула:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}$$

Рассчитав соответствующее t значение и зная число степеней свободы, мы можем рассчитать соответствующий p-уровень значимости, который скажет нам какая вероятность получить такое или еще более выраженное отличие между 2 средними, если на деле верна нулевая гипотеза.

Применяя Т-критерий Стьюдента нужно помнить о:

1. **Дисперсии** внутри наших групп должны быть примерно одинаковы (требование гомогенности дисперсий). Проверить можно с помощью критерия Левена и критерия Фишера
2. Особенно важный вопрос - это **требование к нормальности** данных обеих групп при применении t-теста. Во многих учебниках можно встретить довольно жесткое требование к нормальности данных по причине возможного завышения вероятности ошибки I рода.

NB! На практике t-тест может быть использован для сравнения средних и при ненормальном распределении, особенно на больших выборках и если в данных нет заметных выбросов. Однако при этом вы выходите на очень тонкий лёд - перед использованием t-теста на ненормальных данных дважды подумайте о своих жизненных решениях. Возможно, непараметрический тест или бутстрап окажутся лучше и адекватнее (о них будет позже). Как вариант, можно преобразовать переменную, например, логарифмировать, чтобы сделать распределение более симметричным. Подробнее об этом всём смотрите тут.

В python:

```
from scipy import stats

stats.ttest_ind(a, b)
```

Результат выполнения функции может быть похож на это:

```
Ttest_indResult(statistic=-15.386195820079404, pvalue=6.892546060674059e-28)
```

Первое число - это значение тестовой статистики, оно нам не так интересно. Второе число - это p-значение, и вот оно позволяет нам принять решение об отклонении или неотклонении нулевой гипотезы. Оно должно быть **меньше 0.05**. В данном случае оно намного меньше 0.05, поэтому мы отклоняем нулевую гипотезу и делаем вывод, что средние в группах значимо различаются.

Важно учитывать, что функции SciPy используют так называемую научную нотацию чисел - это способ отображения очень больших или очень маленьких чисел через экспоненту. В данном случае используется такая её разновидность, как E-нотация: в конце числа ставится буква *e* и пишется степень, в которую возведено число. Если степень отрицательная, то речь об очень маленьком числе и это указывает количество знаков после запятой.

В нашем случае у нас число в минус 28 степени, то есть 0.0000000000000000000000000689.

> Построение графиков

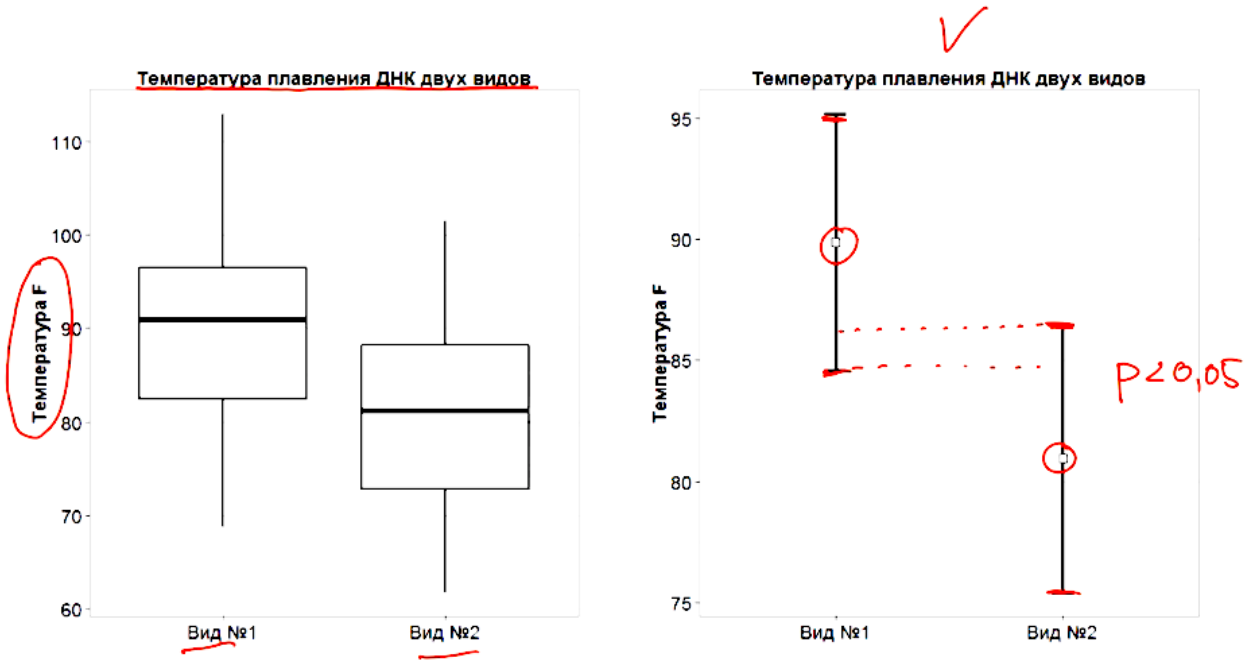
Надо:

- указать название графика
- подписать оси
- указывать меру изменчивости данных (напр. доверительные интервалы, либо сразу использовать боксплот)

Средние значения также не принято отображать столбиками (в виде барплота) → boxplot.

- видно чем занимались и кого сравнивали

- показывает значимость различий → если среднее №2 не попадает в интервал среднего №1 (и наоборот), то такие различия будут достигать уровня статистической значимости.

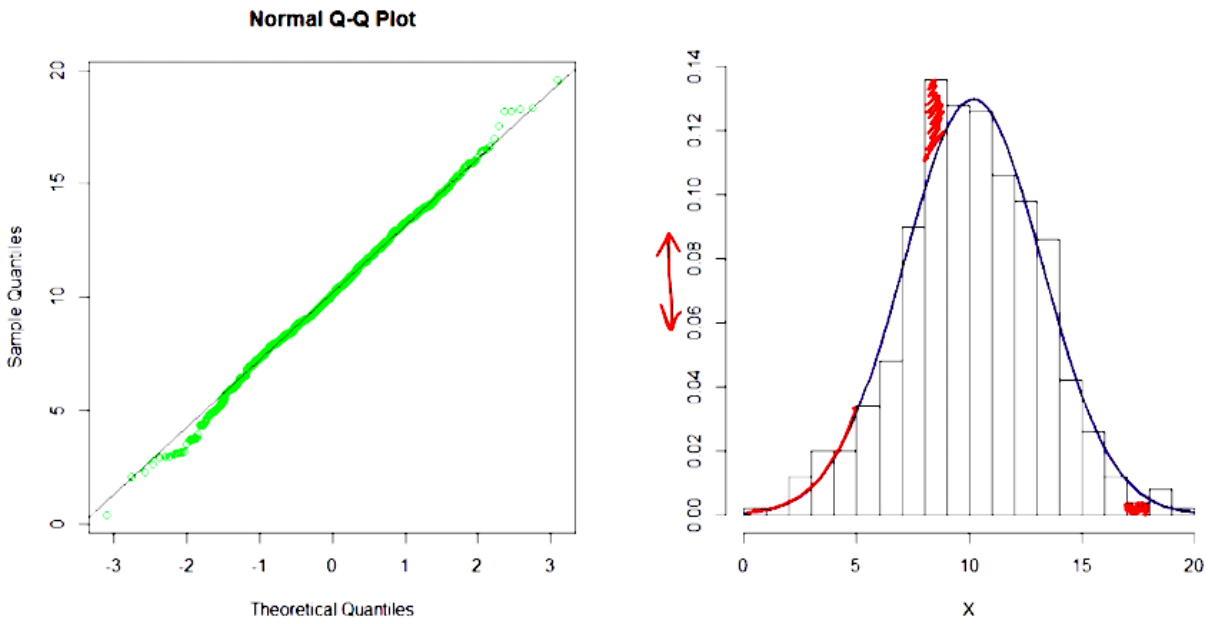


> Сравнение распределения с нормальным, QQ Plot

Как можно оценить насколько имеющееся распределение отличается от теоретического нормального?

Первый вариант – построить гистограмму частот нашего признака, поверх которой наложить кривую идеального нормального распределения (где по оси у будут значения плотности)

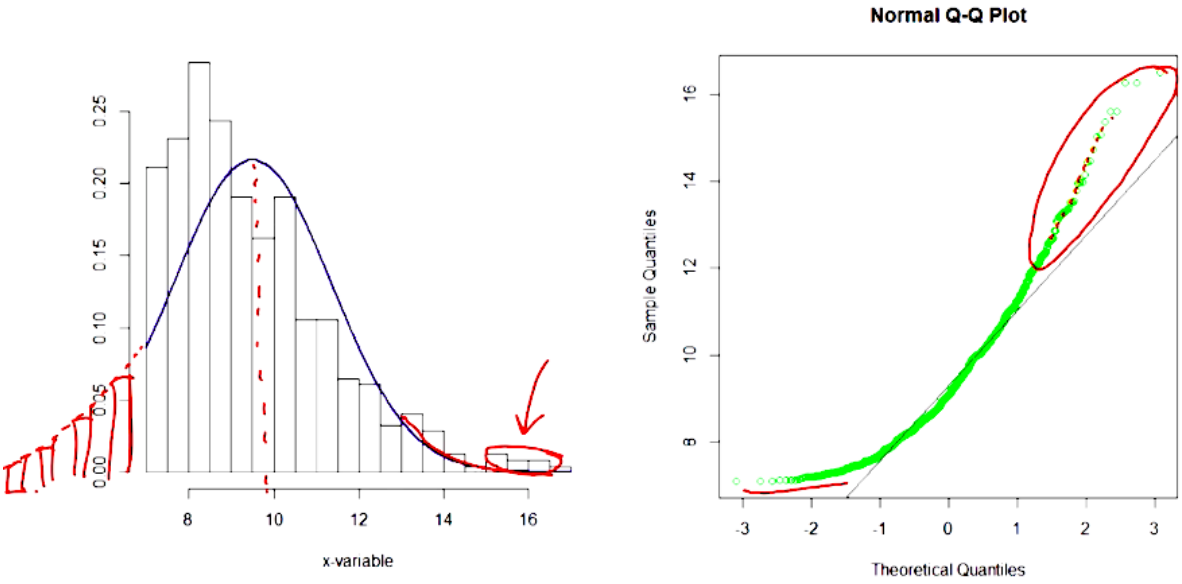
- Где-то есть отклонения чуть выше, чем должно было быть
- Где-то наоборот
- В целом признак распределен нормальным образом



Второй – quantile-quantile plot (QQ-Plot) – показывает, насколько выборочные значения хорошо соответствуют предсказанным значениям, если бы наше распределение было идеально нормальным

- Все точки лежат на прямой? → реальные значения хорошо согласуются с предсказанными значениями идеального нормального распределения
- Кто-то над прямой? → получаем слишком высокие значения, чем должны
- Под? → слишком маленькие значения, чем должны были бы быть в случае нормального распределения

Ноль – середина распределения. По правую сторону от нашего распределения все точки постепенно начинают уползать вверх – значит, максимальные значения слишком высоки для нормального распределения. При этом в левой стороне – минимальные значения также слишком высоки для нормально распределения (не хватает сильно выраженных минимальных значений).



Тесты для проверки нормальности распределения:

- Тест Колмогорова-Смирнова
- Тест Шапиро-Уилка

```
from scipy import stats
stats.shapiro(x)
```

Проверяют гипотезу о том, что выборка изъята из генеральной совокупности, где распределение признака соответствует нормальному. Здесь мы как раз хотим получить p-уровень значимости больше 0.05, поскольку если он меньше, то новость плохая: мы тестируем гипотезу о том, что распределения значимо не отличаются от нормального, поэтому отклонять *H*0 мы не хотим.

Результат выполнения функции может выглядеть так:

```
ShapiroResult(statistic=0.945556104183197, pvalue=0.0004277393454685807)
```

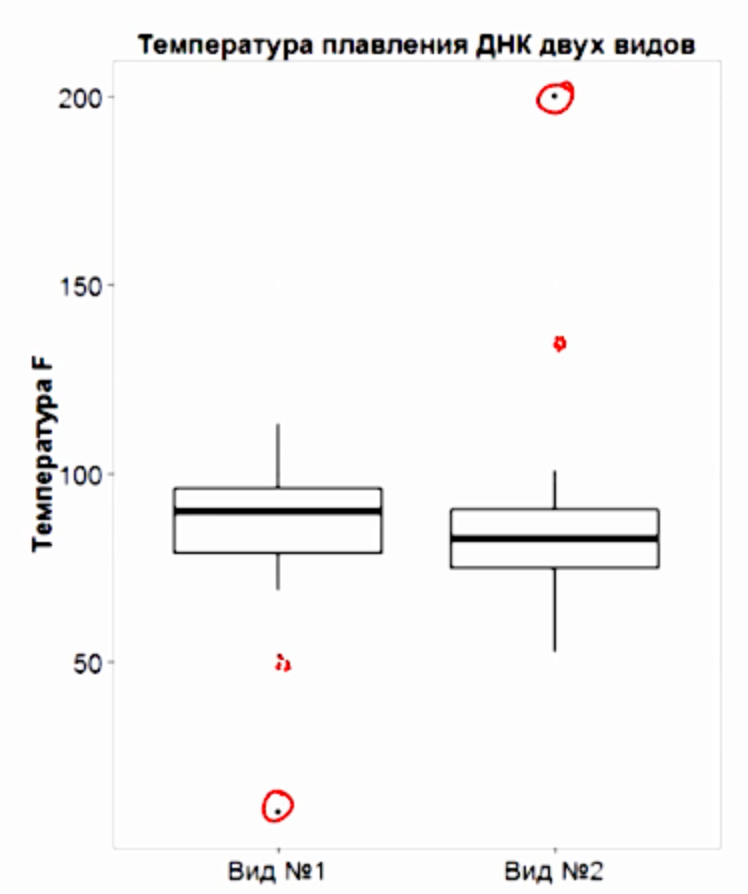
Как и в случае t-критерия, первое число - значение тестовой статистики, второе - p-значение. Так как нулевая гипотеза в случае Шапиро-Уилка - нормальность

распределения, то здесь мы можем сделать вывод, что распределение отклоняется от нормального. Функция также использует научную нотацию.

> Выбросы и U-критерий Манна-Уитни

Почему отклонения от нормального распределения могут негативно повлиять на результаты исследования?

Выбросы – экстремально высокие или экстремально низкие значения. Представьте, что мы добавим очень маленькое наблюдение в 1 выборку, и очень большое во 2. На боксплоте и на QQ plot мы бы заметили, что максимальное значение в выборке слишком максимально для нормального (слишком далеко отклонилось от среднего значения).



Всего два значения уничтожат все значимые результаты: Т-критерий Стьюдента скажет, что вероятность получить такие или еще более выраженные различия

составит 97%, нулевую гипотезу мы отклонить не сможем.

$$t = 0.03, p = 0.97$$

Если распределение признака отличается от нормального, можно использовать непараметрический аналог – U-критерий Манна-Уитни. Он переводит все данные в ранговую шкалу (ранжирует показатели температуры от 1 до последнего), после этого считает какой средний ранг оказался в первой группе и какой во второй. Этот критерий менее чувствителен к экстремальным отклонениям от нормальности и наличию выбросов. С его помощью в данном случае мы бы получили *p*=0.09, что тоже поставило бы под вопрос отклонение H_0, но явно лучше, чем *p*=0.97.