Вот статья, в которой автор показывает, что t-тест проваливается для не-нормально распределенных данных (<https://medium.com/statistics-experiments/%D1%87%D1%82%D0%BE-%D0%B1%D1%83%D0%B4%D0%B5%D1%82-%D0%B5%D1%81%D0%BB%D0%B8-%D0%B8%D1%81%D0%BF%D0%BE%D0%BB%D1%8C%D0%B7%D0%BE%D0%B2%D0%B0%D1%82%D1%8C-%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9-%D0%BA%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9-%D0%BD%D0%B0-%D0%BD%D0%B5%D0%BD%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE-%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%BD%D0%BE%D0%B9-%D0%B2%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B5-94be5d2afaa7>). Докажем, что это не так.

Дисклеймер: в этом описании применяются терминология, от которой тру-математики будут плакать кровавыми слезами (я не математик и пытаюсь объяснить на пальцах, имейте это в виду).

Часть первая: что мы тестируем.

Допустим, у нас есть некоторая генеральная совокупность (ГС) наблюдений (ну, скажем, данные о месячной прибыли от индивидуальных юзеров). Допустим, мы хотим проверить, что среднее нашей ГС равно некому уровню N. Для этого мы просто возьмем и вычислим среднее - у нас же есть вся ГС и посмотрим равно оно N или нет. Печаль в том, что вся ГС на практике нам недоступна, доступны только выборки.

Если мы сделаем из ГС много-много выборок, то каждая выборка даст нам отдельное значение среднего в этой выборке. Все значения средних дадут нам \_\_распределение выборочного среднего\_\_ - это отдельная случайная величина. У этой величины есть свойство - ее среднее равно среднему генеральной совокупности (на этот счет есть теорема, даже не пытайтесь заставить меня ее доказывать). Еще там есть про дисперсию, но в этот упрощенном объяснении мы сконцентрируемся на среднем.

Доказательство: <http://www.real-statistics.com/sampling-distributions/basic-concepts-sampling-distributions/>

Еще раз - если мы берем много выборок, то мы получаем много средних и эти средние составляют \_\_\_отдельную случайную величину\_\_\_. Ах как было бы здорово, если бы эта случайная величина была бы распределена нормально (мы знаем о нормальном распределении все) - мы бы тогда взяли бы ее среднее, дисперсию, построили бы доверительный интервал (примерно +- 2 сигмы от среднего) и посмотрели попадает в него N или нет. Если попадает, то гипотеза о равенстве верна, если не попадает, то гипотеза о равенстве отвергается. Это почти сущность t-теста. В реальности там все немного сложнее, применяется аппроксимация нормального распределения, но для сейчас это не важно.

Доказательство: <https://www.youtube.com/watch?v=hlM7zdf7zwU>

Доказательство про т-тест: мы не знаем дисперсию, поэтому апроксимируем нормальное распределение т-распределением <http://www.real-statistics.com/students-t-distribution/t-distribution-basic-concepts/t-distribution-advanced/>

Вывод: нам важно не столько как распределена исходная величина, сколько то, по какому закону распределено ее \_\_\_выборочное среднее\_\_\_.

Ах как было бы здорово, если бы оно было бы распределено нормально. И оно распределено - об этом нам говорит ЦПТ, но есть нюанс - ЦПТ говорит, что  \_\_\_выборочное среднее\_\_\_ будет распределено нормально, при бесконечно больших выборках. Очень практичный результат - у нас-то выборка конечна.

Доказательство: <http://www.real-statistics.com/sampling-distributions/central-limit-theorem/>

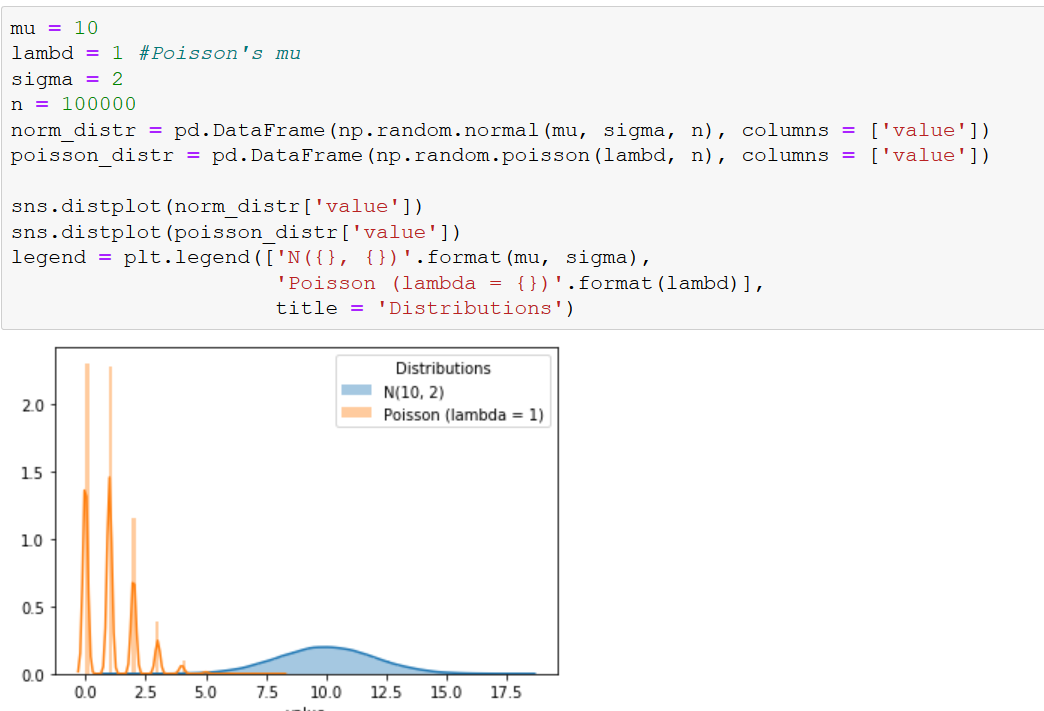
Но есть класс генеральных совокупностей, для которых нормальность распределения \_\_выборочного среднего\_\_\_ будет соблюдаться всегда, при любом размере выборки. Это генеральные совокупности, которые сами распределены нормально.

Вывод: нормальность исходного распределения важна на сама по себе, а потому, что она позволяет \_\_строго гарантировать\_\_ нормальность распределения \_\_выборочного среднего\_\_\_.

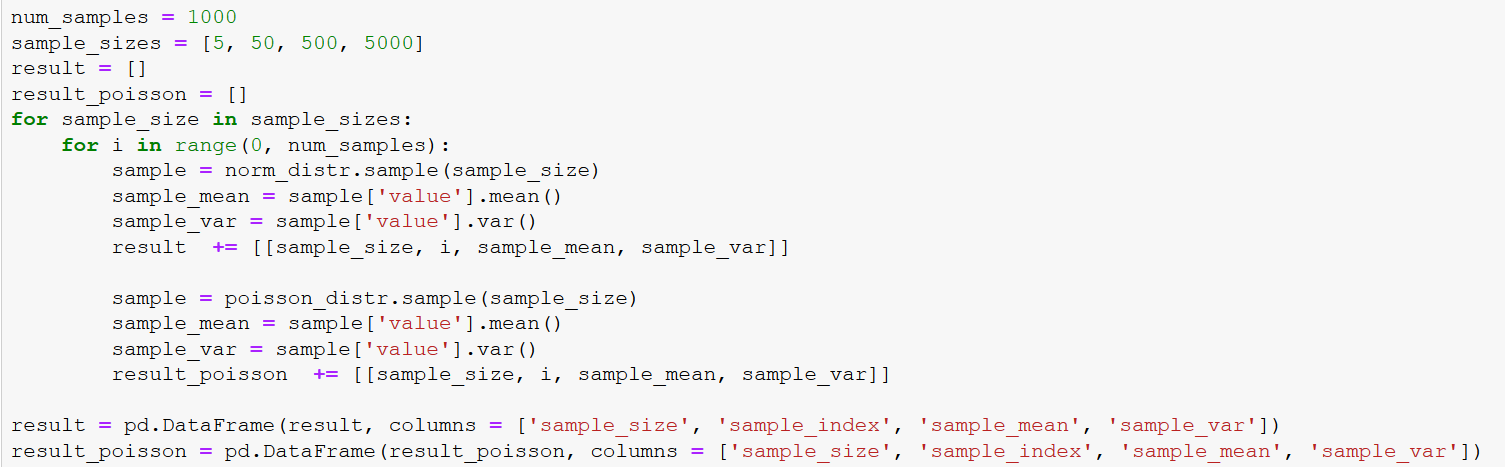
Доказательство: <http://www.real-statistics.com/students-t-distribution/t-distribution-basic-concepts/t-distribution-advanced/>

Часть вторая: вернемся к практике.

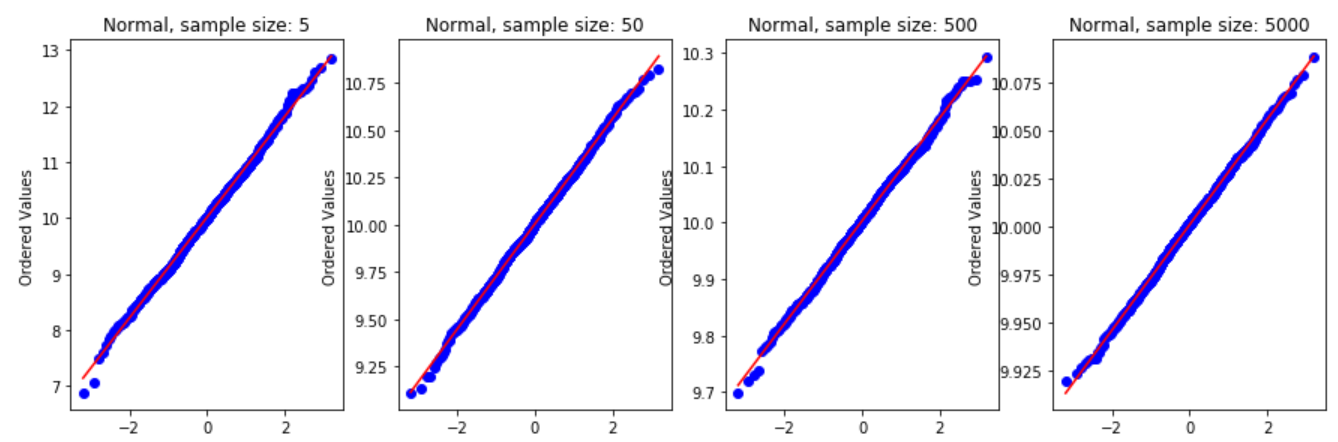
Итак, мы знаем, что для нас важна нормальность распределения \_\_выборочного среднего\_\_\_. Действительно ли нам нужны бесконечно большие выборки для достижения нормальности? Для этого построим симуляцию - создадим две случайных величины - одну нормальную, вторую явно ненормальную (пуассоновскую):



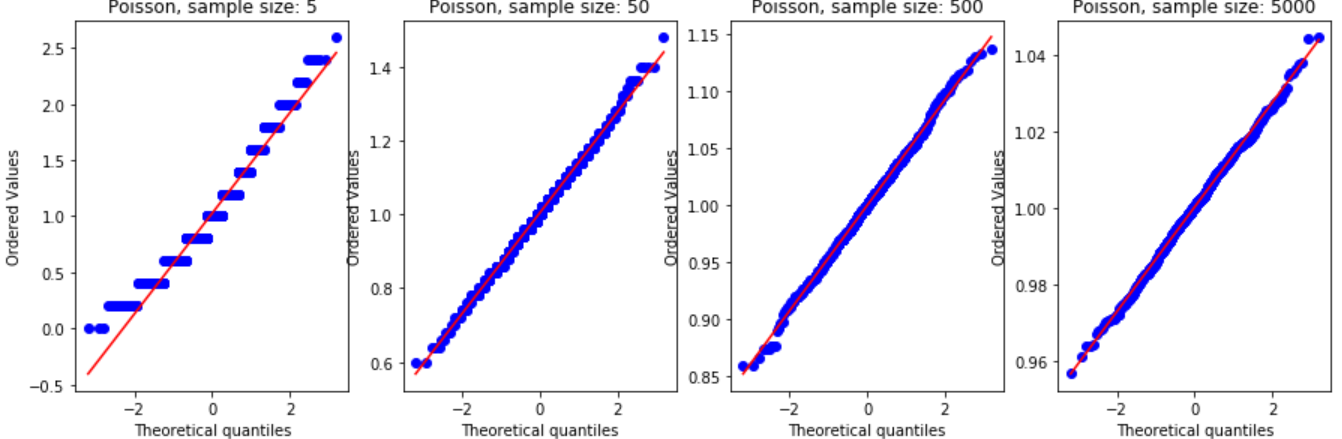
Затем, сделаем из них множество выборок разного размера:



В каждой выборке найдем среднее и используем Q-Q plot (<https://desktop.arcgis.com/ru/arcmap/10.4/extensions/geostatistical-analyst/normal-qq-plot-and-general-qq-plot.htm>) для того, чтобы визуально оценить нормальность распределения \_\_выборочного среднего\_\_\_. Сначала посмотрим как ведет себя распределение для выборок из нормальной случайной величины:



Нормально себя ведет! А теперь посмотрим, как себя ведут выборочные средние для выборок из пуассоновской случайной величины:

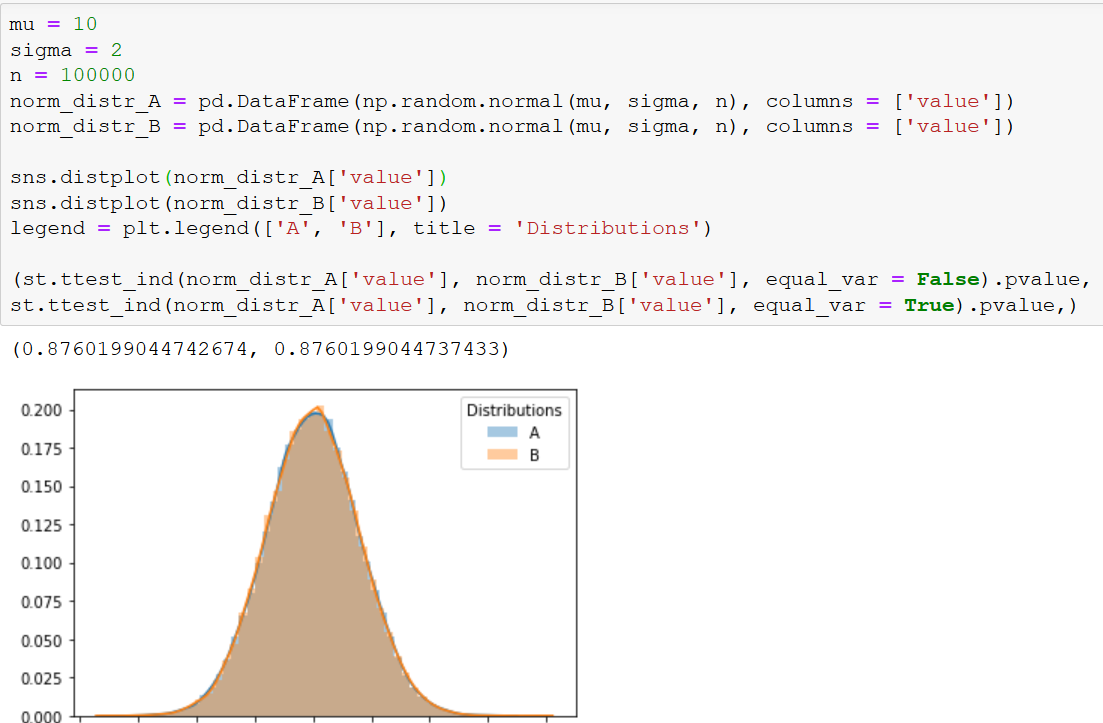


На маленьких выборках видно отклонение от нормального распределения, но при выборках размером 50 и выше наблюдений уже практически нет никаких отличий от нормального распределения.

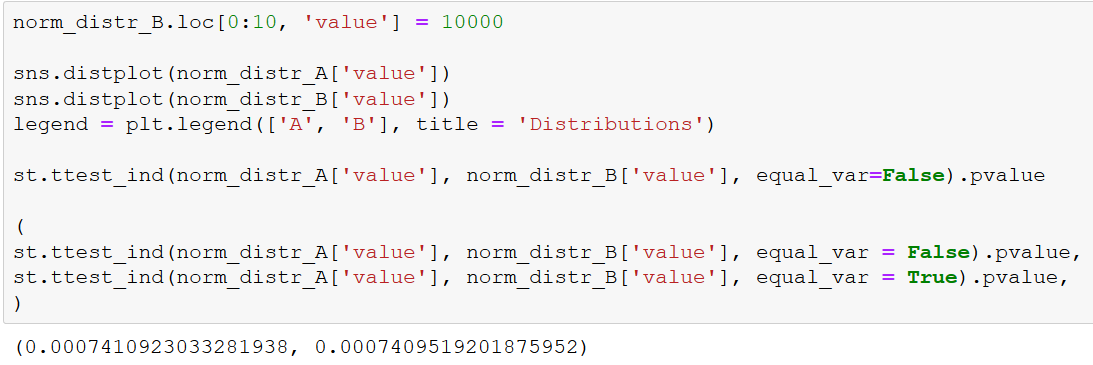
Вывод: вне зависимости от того, распределена ли исходная величина нормально или нет, на практике мы можем считать, что \_\_выборочное среднее\_\_\_ распределено нормально и мы можем применять t-тест, если размер выборки достаточно большой - скажем, больше 50. Или 100. Или 500.

Часть 3: но что же статья на Медиуме?

У меня нет исходных данных, которые использует автор, но он жалуется на выбросы и я рискну утверждать, что все его проблемы от них. Более того, я рискну утверждать, что и на нормально распределенных данных в ситуации выбросов автор получил бы точно такую же картину. Для этого проведем опыт - построим две нормально-распределенные случайные величины:



Как видите, H0 не отвергается. Теперь добавим немного выбросов:



Исходные данные вроде бы нормально распределены и на 99% идентичны, а H0 отвергается.

Вывод: удаляйте выбросы перед тестированием, анализируйте данные визуально.

Часть четвертая: у нас нормальность везде, давайте все будем тестировать t-тестом.

Не так быстро. ЦПТ говорит, что нормально распределено только выборочное среднее, а можно тестировать еще много чего - медиану, 95-ю перцентиль и т.д. Для них нам никто ничего не гарантировал. Плюс есть ситуации, когда выборки малы.