Преамбула

Пару дней назад @Петр Захаров задал вопрос о применимости t-теста для тестирования гипотез в условиях, когда данные распределены не нормально. Петр не удовлетворился моим дежурным объяснением по поводу центральной предельной теоремы (молодец) и нашел пример, в котором 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) и обрек меня попытки объяснить все это как можно проще, но подробно :) Т.к. писать теоретическую часть курса не входит в мои прямые обязанности, это заняло некоторое время.

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.

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.

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.

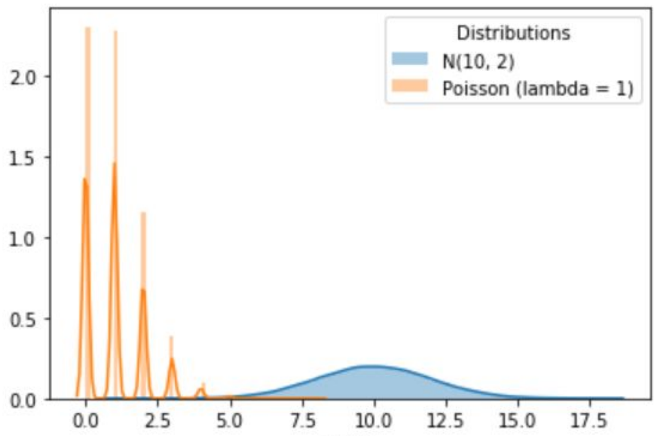
Но есть класс генеральных совокупностей, для которых нормальность распределения __выборочного среднего___ будет соблюдаться всегда, при любом размере выборки. Это генеральные совокупности, которые сами распределены нормально.

Вывод: нормальность исходного распределения важна на сама по себе, а потому, что она позволяет __строго гарантировать__ нормальность распределения __выборочного среднего___.

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):

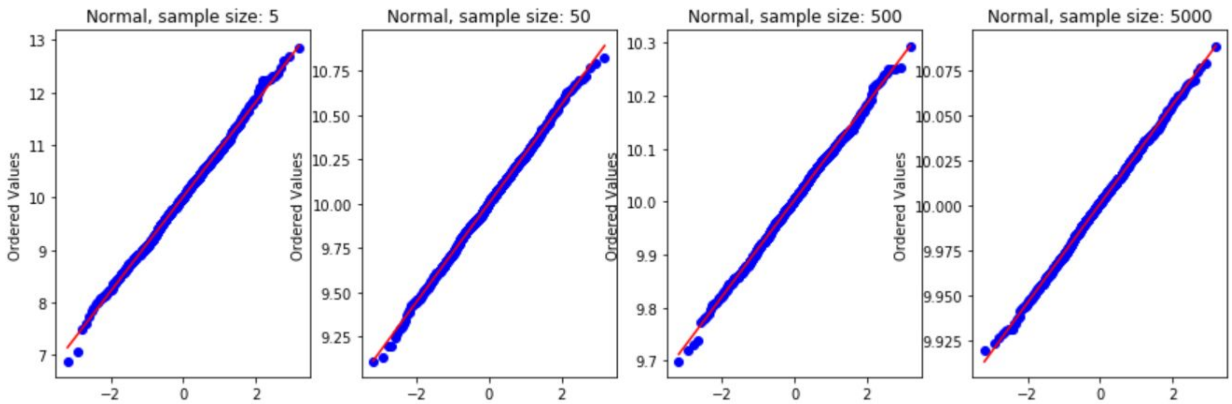
```python
mu = 10
lambd = 1 #Poisson's mu
sigma = 2
n = 100000
norm_distr = pd.DataFrame(np.random.normal(mu, sigma, n), columns = ['value'])
poisson_distr = pd.DataFrame(np.random.poisson(lambd, n), columns = ['value'])

sns.distplot(norm_distr['value'])
sns.distplot(poisson_distr['value'])
legend = plt.legend(['N({}, {})'.format(mu, sigma),
                     'Poisson (lambda = {})'.format(lambd)],
                     title = 'Distributions')
```



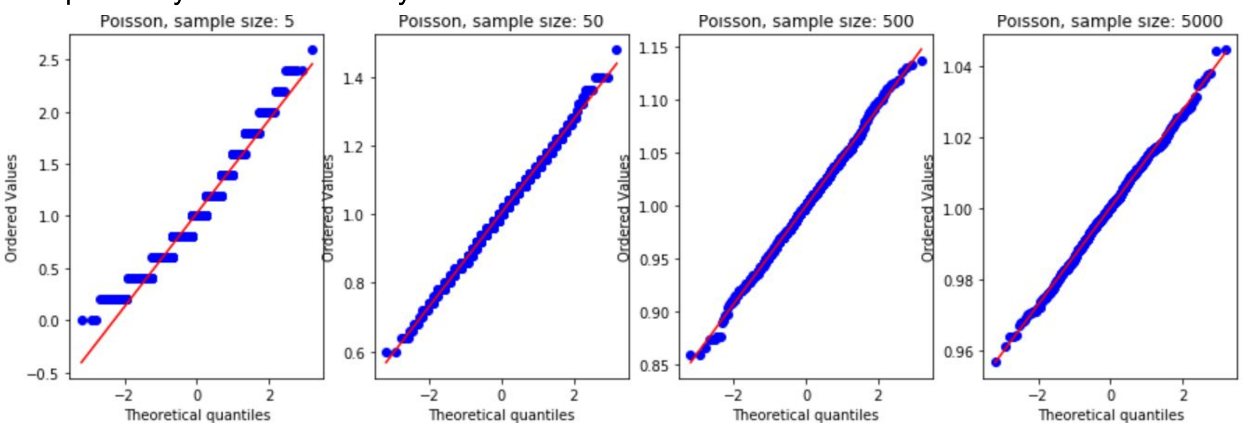
Затем, сделаем из них множество выборок разного размера:

```python
num_samples = 1000
sample_sizes = [5, 50, 500, 5000]
result = []
result_poisson = []
for sample_size in sample_sizes:
    for i in range(0, num_samples):
        sample = norm_distr.sample(sample_size)
        sample_mean = sample['value'].mean()
        sample_var = sample['value'].var()
        result  += [[sample_size, i, sample_mean, sample_var]]

        sample = poisson_distr.sample(sample_size)
        sample_mean = sample['value'].mean()
        sample_var = sample['value'].var()
        result_poisson  += [[sample_size, i, sample_mean, sample_var]]

result = pd.DataFrame(result, columns = ['sample_size', 'sample_index', 'sample_mean', 'sample_var'])
result_poisson = pd.DataFrame(result_poisson, columns = ['sample_size', 'sample_index', 'sample_mean', 'sample_var'])
```

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.

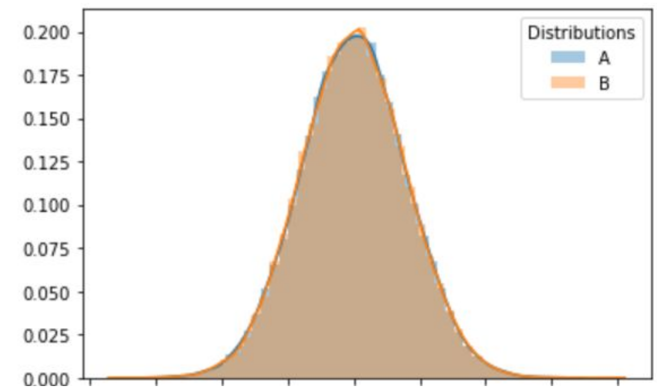
Вывод: вне зависимости от того, распределена ли исходная величина нормально или нет, на практике мы можем считать, что __выборочное среднее___ распределено нормально и мы можем применять t-тест, если размер выборки достаточно большой - скажем, больше 50. Или 100. Или 500.

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:

```
mu = 10
sigma = 2
n = 100000
norm_distr_A = pd.DataFrame(np.random.normal(mu, sigma, n), columns = ['value'])
norm_distr_B = pd.DataFrame(np.random.normal(mu, sigma, n), columns = ['value'])

sns.distplot(norm_distr_A['value'])
sns.distplot(norm_distr_B['value'])
legend = plt.legend(['A', 'B'], title = 'Distributions')

(st.ttest_ind(norm_distr_A['value'], norm_distr_B['value'], equal_var = False).pvalue,
st.ttest_ind(norm_distr_A['value'], norm_distr_B['value'], equal_var = True).pvalue,)
```

```
(0.8760199044742674, 0.8760199044737433)
```



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

```
norm_distr_B.loc[0:10, 'value'] = 10000

sns.distplot(norm_distr_A['value'])
sns.distplot(norm_distr_B['value'])
legend = plt.legend(['A', 'B'], title = 'Distributions')

st.ttest_ind(norm_distr_A['value'], norm_distr_B['value'], equal_var=False).pvalue

(
st.ttest_ind(norm_distr_A['value'], norm_distr_B['value'], equal_var = False).pvalue,
st.ttest_ind(norm_distr_A['value'], norm_distr_B['value'], equal_var = True).pvalue,
)
```

```
(0.0007410923033281938, 0.0007409519201875952)
```

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.