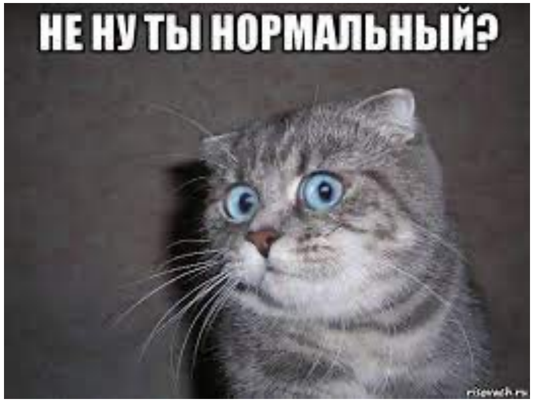
Проверка нормальности

Проверка нормальности

- Важный и большой класс задач представляет проверка принадлежности выборки семейству нормальных распределений, потому что многие мощные критерии требуют на вход именно его.
- Математически гипотеза формулируется так:

$$\mathrm{H}_0\colon F(x) \in \{N(\mu, \sigma^2)\}$$
$$\mathrm{H}_a\colon F(x) \notin \{N(\mu, \sigma^2)\}$$



Проверка нормальности

Может показаться, что задачу проверки нормальности можно решить при помощи критериев согласия, но есть два важных отличия:

1. Критерии проверки нормальности не требуют наличия оценки на параметры распределения.
2. Критерии проверки нормальности имеют бóльшую мощность для данной задачи.

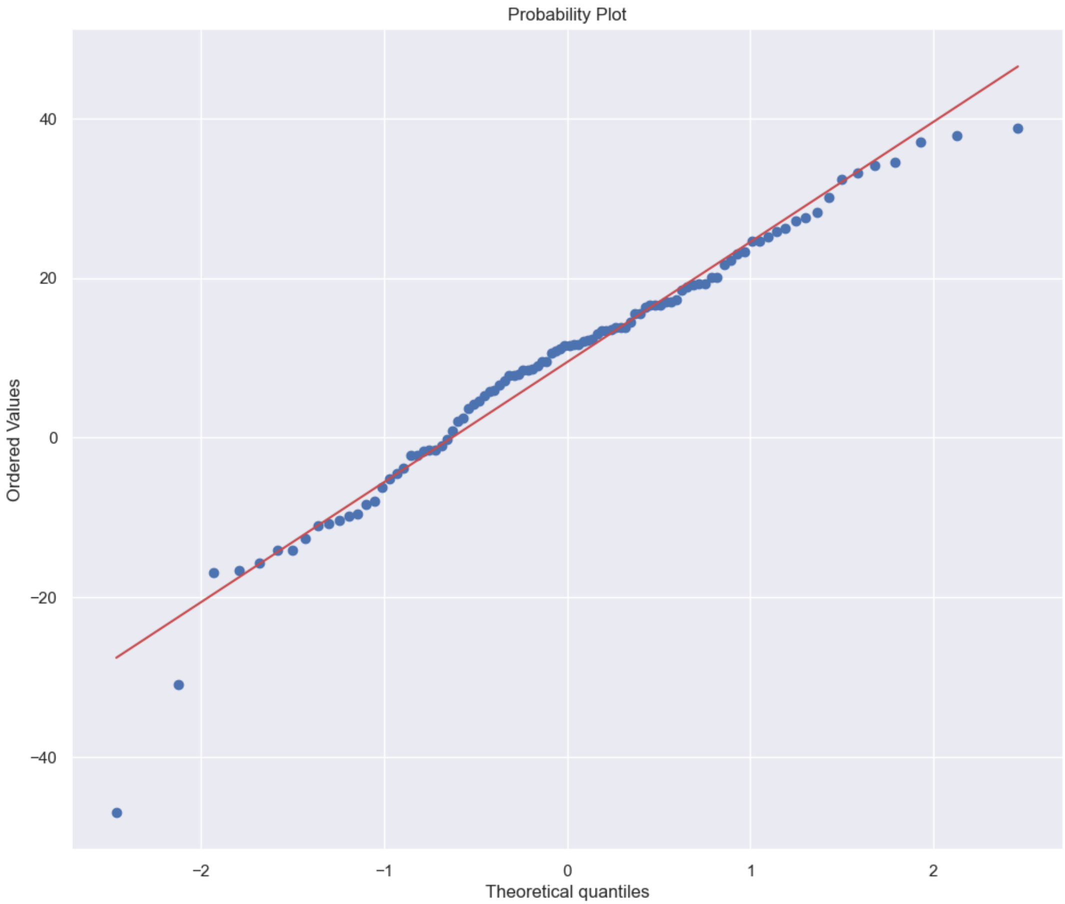
Способ 1: QQ-plot

QQ-plot («*график квантиль-квантиль*», https://en.wikipedia.org/wiki/Q–Q_plot) – визуальный способ проверки нормальности.

Способ построения: на график наносятся точки $\left(F_0\left(\frac{k-0.5}{n}\right), X_{(k)}\right)$, где F_0 - функция распределения стандартного нормального распределения. Если получилась прямая, то можно полагать, что выборка принадлежит нормальному распределению.

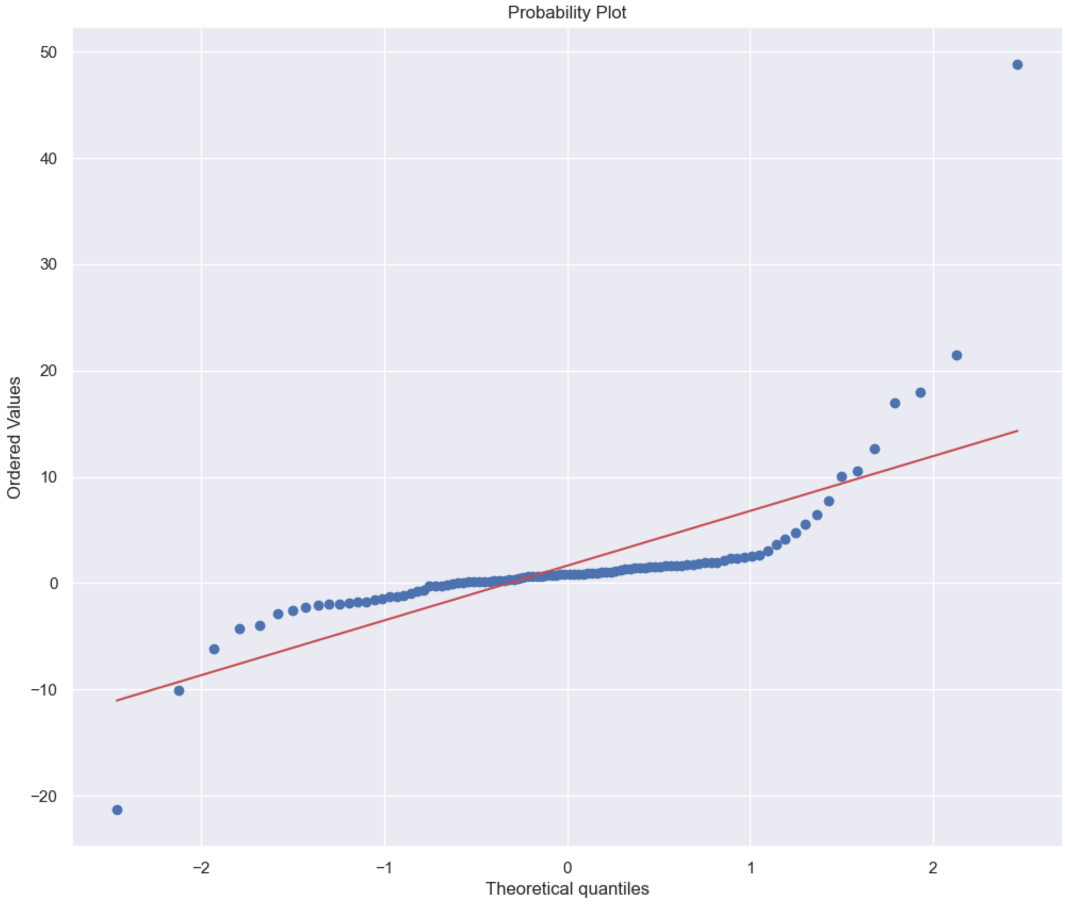
Важно: построение QQ-plot не является критерием!

QQ-plot для нормальной выборки



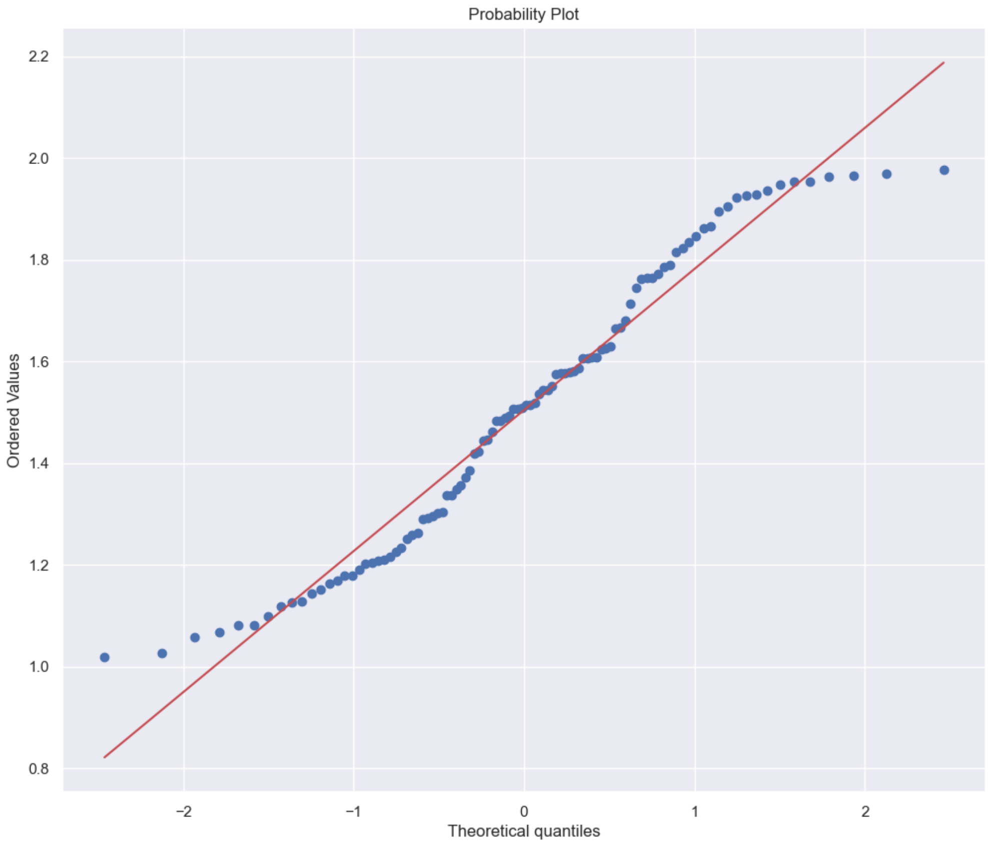
QQ-plot для выборки из распределения с тяжёлыми хвостами

Распределения Коши, Стьюдента, Парето, Лапласа…



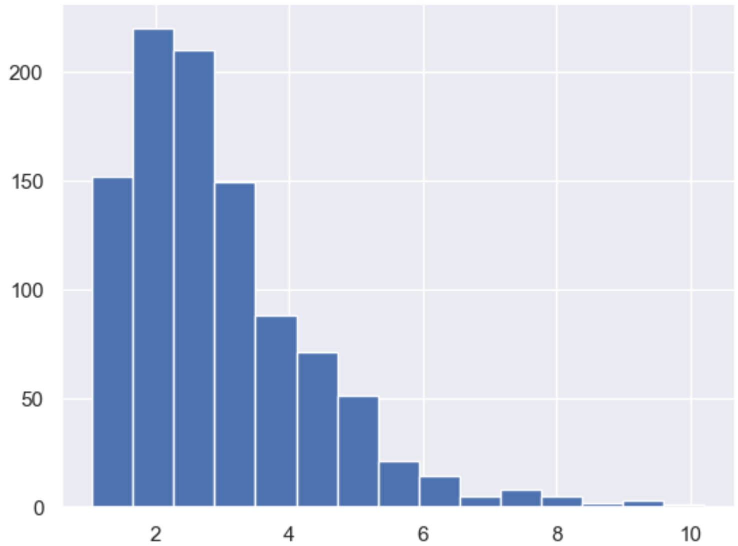
QQ-plot для выборки из распределения с лёгкими хвостами

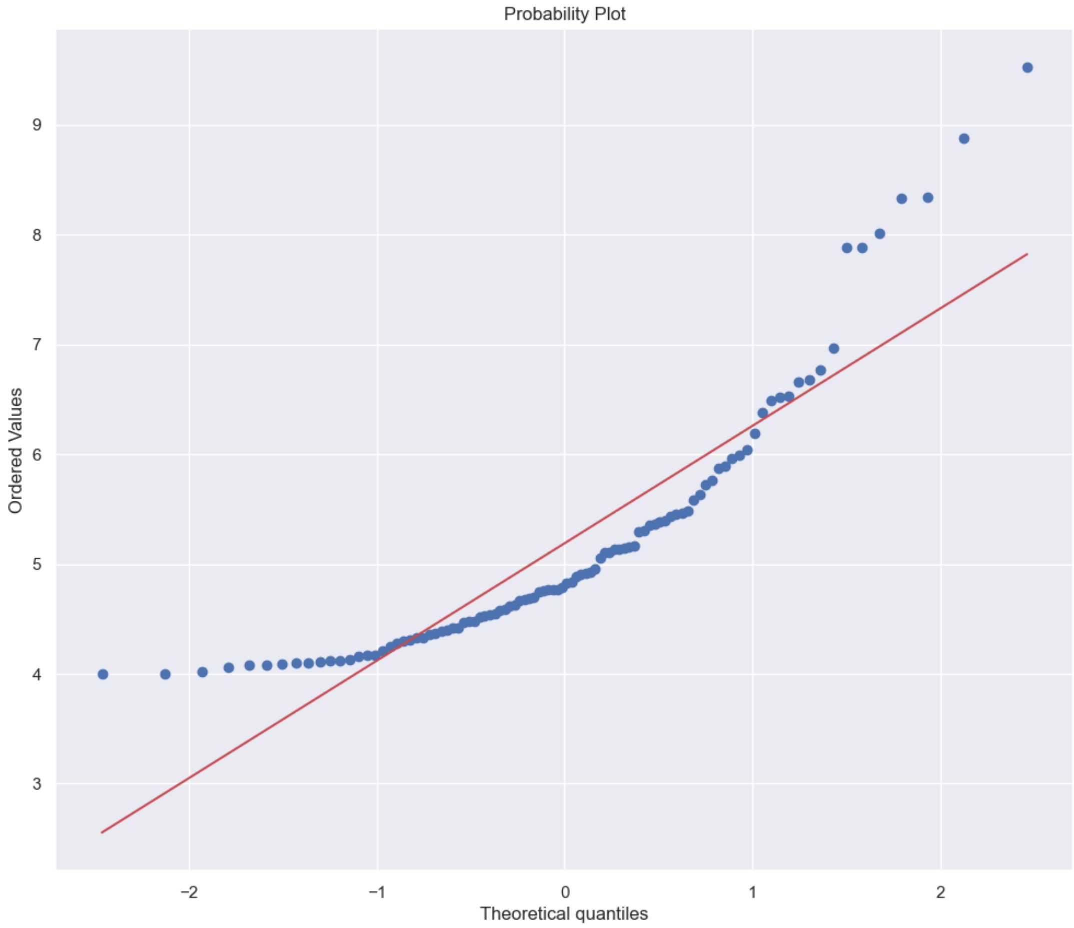
Например, распределения с ограниченным носителем (равномерное, бета-распределение,…)



QQ-plot для выборки из распределения, скошенного вправо

Экспоненциальное, Пуассона, гамма-распределение…





Что подразумевается под «скосом вправо»

Способ 2: критериями

Для проверки нормальности существует множество критериев. Рассмотрим некоторые из них:

- Критерий Лиллиефорса
- Критерий Харке-Бера
- Критерий Шапиро-Уилка

Критерий Лиллиефорса

Первая простая идея, которая приходит в голову: давайте возьмём критерий согласия (например, Колмогорова-Смирнова) и подставим в него выборочные оценки. Проблема в том, что для каждого распределения из нулевой гипотезы и для каждого вида оценки мы будем получать разное распределение статистики критерия.

Но: если взять оценки параметров μ, σ^2 в виде выборочных среднего и дисперсии и затем воспользоваться критерием Колмогорова-Смирнова, то мы получим **тест Лиллиефорса** (по имени учёного, который нашёл распределение получившейся статистики).

Критерий Харке-Бера

Вторая идея: давайте посмотрим на статистики, которые отвечают за форму распределения. Мы знаем 2 таких: коэффициент асимметрии и эксцесс (см. лекцию 1). Линейная комбинация их квадратов и представляет собой статистику **критерия Харке-Бера:**

$$\mathrm{JB} = \mathrm{n}\left(\frac{\mathrm{S}^2}{6} + \frac{(\mathrm{K}-3)^2}{24}\right)$$

На похожей идее основан тест Д’Агостино (или K^2-тест)

Критерий Шапиро-Уилка

- Один из самых мощных и часто используемых критериев проверки нормальности. Сложно выделить какую-либо простую интуицую, стоящую в реализации этого критерия, поэтому вид статистики не приводится (можно посмотреть, например, здесь https://en.wikipedia.org/wiki/Shapiro–Wilk_test)

- Важная черта: при очень больших выборках склонен детектировать любое малейшее отклонение от нормальности, поэтому часто ошибается