Sampling and hypothesis tests

Part 3: Data distributions (1/2)

By: Noureddin Sadawi, PhD

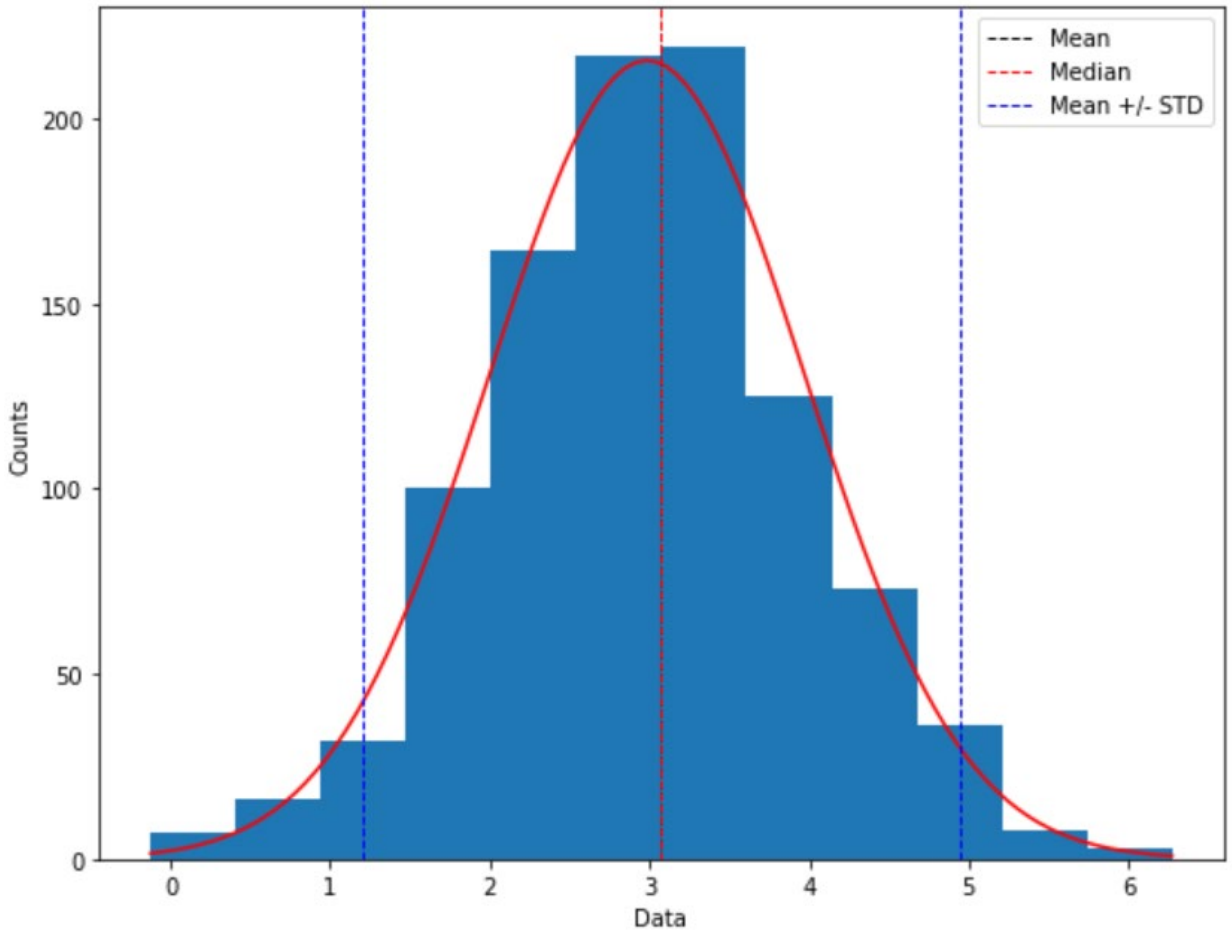
University of London

Normal distribution

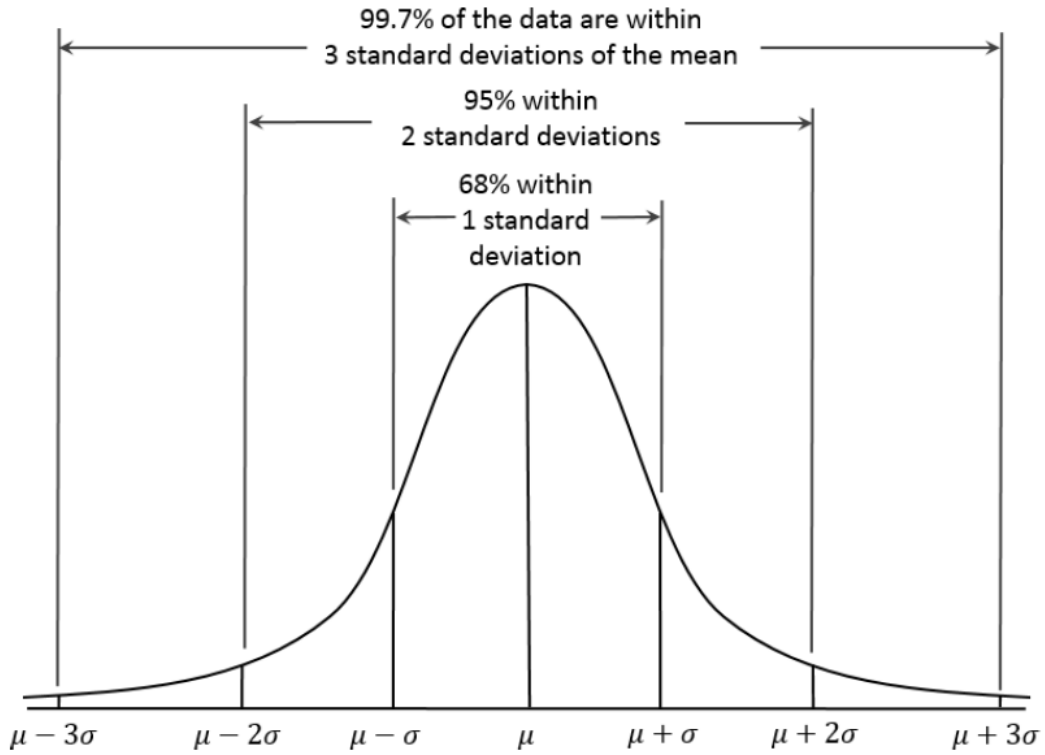
- We have seen histograms before (where values are grouped into ranges and the number of values in each range is represented as a bar).
- When the histogram has a shape of a bell, the underlying data is said to follow a normal distribution.
- Also known as the Gaussian distribution.
- This distribution can be described using two metrics: its mean and standard deviation.
- ‘In a normal distribution, 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Histogram (from Topic 1)



Normal distribution



Normal curve

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Normal distribution

- In a normal distribution:
  - Mean = Median.
  - Symmetry around the mean.
  - 50% of values are to the right of the mean (i.e. > the mean), 50% are to the left of the mean (i.e. < the mean).

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

f(x) is the probability density function

$\boldsymbol{\mu}$ is the mean

$\boldsymbol{\sigma}$ is the standard deviation.

Standard normal and standardisation

- ‘A standard normal distribution is one in which the units on the x-axis are expressed in terms of standard deviations away from the mean.
- To compare data to a standard normal distribution, you subtract the mean and then divide by the standard deviation.
- This is also called normalization or standardization.
- The transformed value is termed a z-score, and the normal distribution is sometimes called the z-distribution.’

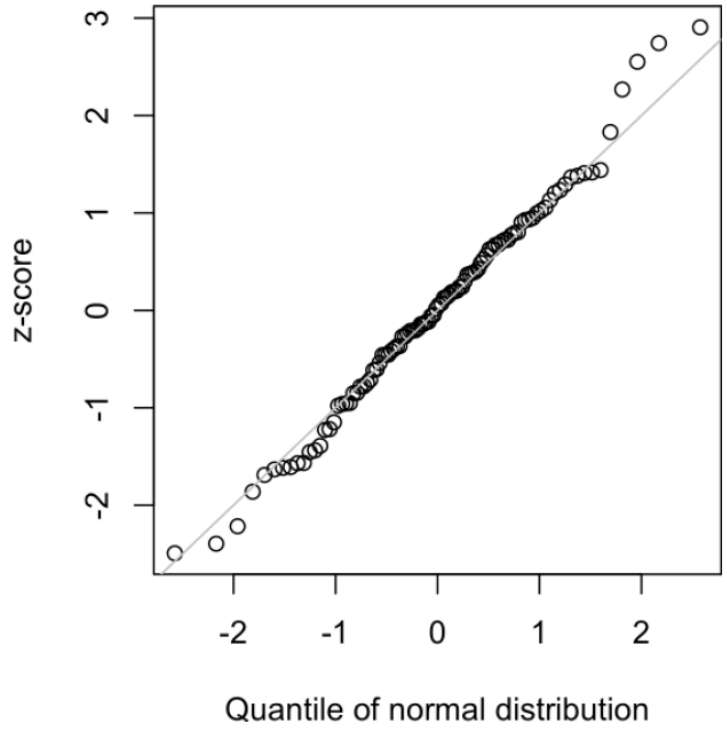
(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

QQ-plots

- ‘A QQ-Plot is used to visually determine how close a sample is to a specified distribution - in this case, the normal distribution.
- The QQ-Plot orders the z-scores from low to high and plots each value’s z-score on the y-axis; the x-axis is the corresponding quantile of a normal distribution for that value’s rank.
- Since the data is normalized, the units correspond to the number of standard deviations away from the mean.
- If the points roughly fall on the diagonal line, then the sample distribution can be considered close to normal.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Example



QQ-Plot of a sample of 100 values drawn from a standard normal distribution.

Source: Bruce and Bruce 2020

Key terms for normal distribution

‘Error

- The difference between a data point and a predicted or average value.

Standardize

- Subtract the mean and divide by the standard deviation.

z-score

- The result of standardizing an individual data point.

Standard normal

- A normal distribution with mean = 0 and standard deviation = 1.

QQ-Plot

- A plot to visualize how close a sample distribution is to a specified distribution, e.g., the normal distribution.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

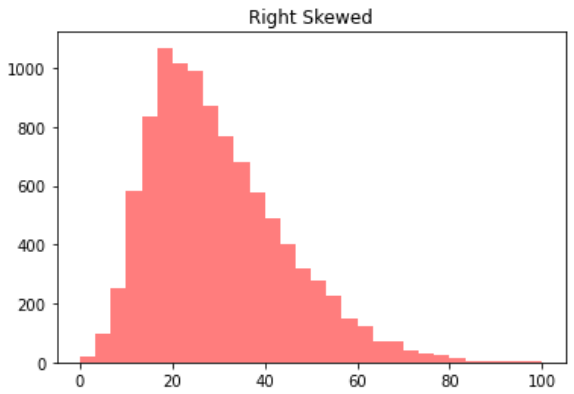
Long-tailed distribution

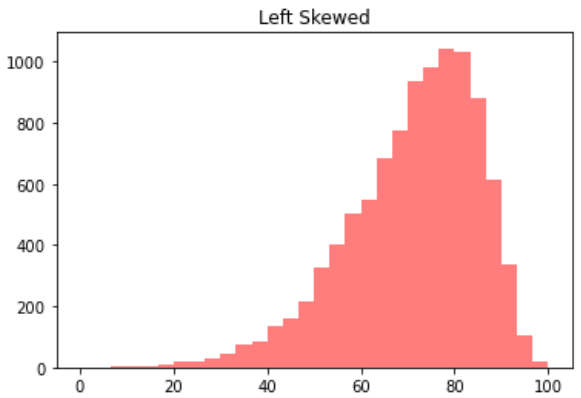
- ‘Data is generally not normally distributed.

Tail: The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.

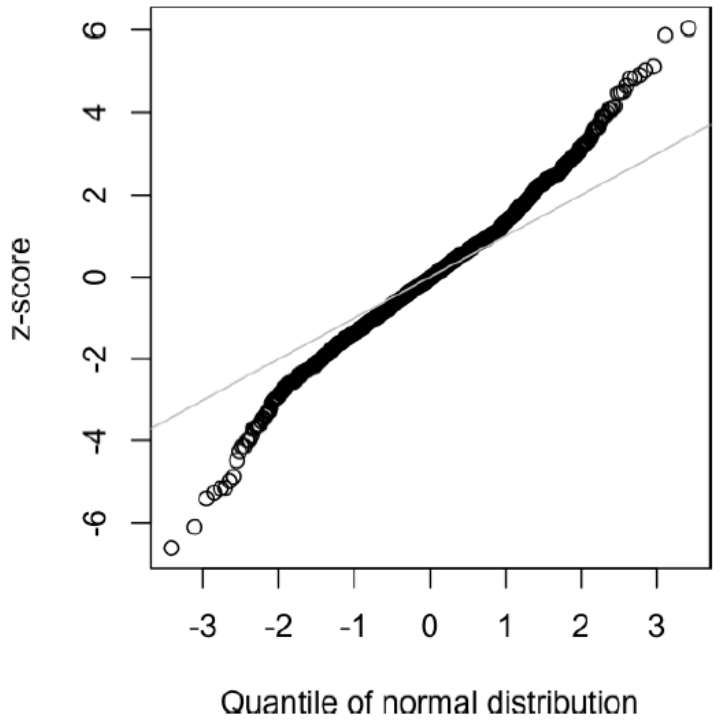
Skew: Where one tail of a distribution is longer than the other.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).





Example



QQ-Plot of a skewed distribution.

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).