PSI:
Testing Assumptions

Technological University Dublin City Campus

School of Computer Science

Testing Assumptions

- Many statistical methods rely on the fulfilment of assumptions for their applicability.
- Here we look at the following common assumptions
  - absence of outliers
  - normality of a variable's distribution
  - homoskedasticity, i.e. the property whereby a variable has the same variance across its entire range

Testing for outliers

A good rule of thumb for identifying outliers:

- if 80 or fewer instances in sample, a z-score outside of ± 2.58 is an outlier
- if 80 instances in sample, a z-score outside of ± 3.27 is an outlier

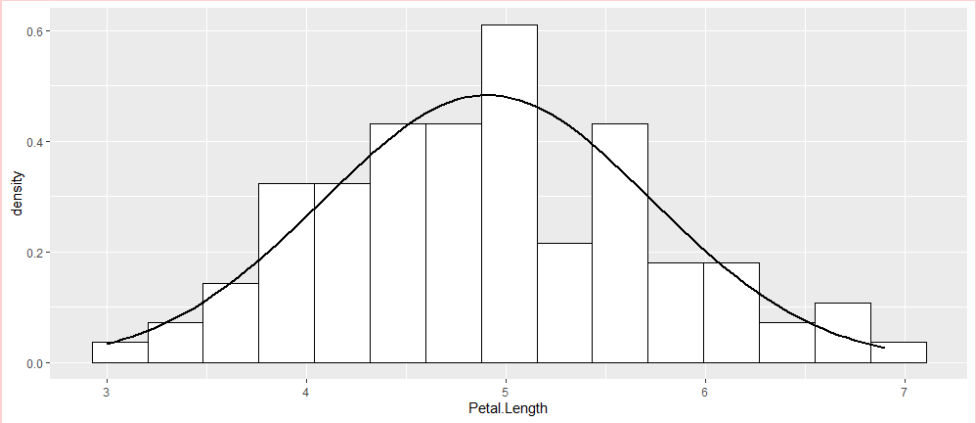
HOW TO FIX?

- remove
- replace with next highest value plus 1
- replace with value equal to 3 or 2 standard deviations
- transform the data (see fixes for non-normality)

Testing a variable for normality of distribution

Testing for normality of distribution is not a trivial task. It must be approached in several ways.

- **Graphing the distribution** is essential. Ideally, a corresponding PDF will be superimposed over the sample histogram, which gives a clearer picture as to the shape of the sample distribution in comparison with the normal distribution.



- Another type of graph that can help with assessing the closeness of a sample histogram to normality is a **Q-Q plot**. Each point on the plot is a quantile with (click for examples):
  - the x-coordinate representing its normal distribution value
  - the y-coordinate representing its value in the sample distribution that is being compared

- The **skewness** and **kurtosis** values should be tested. If these values are big, the distribution cannot be approximated as normal. A rule of thumb is to consider a distribution as close to normal if its standardised scores for both of those statistics lie within the ± 1.96 range.

- Another indication of the distribution's departure from normality would be an **excessive spread** of values. A check that can be performed is for the percentage of values found within two standard deviations from the mean. If it is a lot greater then 5%, this can be considered as a divergence from normality.

- Finally, there are goodness-of-fit tests such as **Shapiro-Wilk** and **Kolmogorov-Smirnov** test. These can be performed but keeping in mind that their high power when used with large samples (e.g. over 200 instances) means that normality is rejected quite often even though the data is still acceptable as normal for various statistical purposes.

HOW TO FIX?

- transform the data:
  - if the data is **positively skewed**, use one of:
    - log transformation
    - reciprocal transformation
    - square root transformation
  - if the data is **negatively skewed**:
    1. reverse - transform the values to change the skew
    2. then apply log, reciprocal or square root tansformation
- use a non-parametric method that does not assume normality

Testing a variable for homoskedasticity

The homogeneity of variance can be tested by a combination of visual inspection and statistical test:

- scatterplot - the dispersion should be random rather than increasing or decreasing
- Levene's test
- Hartley's test (maximal variance ratio)

HOW TO FIX?

- use non-parametric methods