Data Science 1: Non-Parametric Methods Introduction

Ann Maharaj

Introduction

- The statistical tests for quantitative variables presented in earlier modules are based on the assumption that the sample data come from normal populations and in some cases the assumption of homogeneity of variance.
- These tests are referred to as parametric tests and the associated statistics are referred to as parametric statistics.
- In certain situations:
  - The probability distribution of the variables of interest may not be normal or may be unknown.
  - The data could be qualitative and therefore unsuitable for parametric tests.
- In such cases, non-parametric methods should be used.

Introduction

- Non-parameter statistics are based on fewer assumptions about the population than parametric statistics.
- They are often referred to as distribution-free statistics because no assumption about the underlying distribution of the population is required.
- Some of the tests discussed in the module on the analysis categorical data can also be classed as non-parametric, in particular, Fisher, McNamar and Cochrane Q tests.

Introduction

Advantages of using Non-parametric Methods

- Most non-parametric procedures depend on very few assumptions.
- The computations on non-parametric statistics are usually less complicated than those for parametric statistics, particularly for small samples.
- Probability statements obtained for most non-parametric tests are exact probabilities.

Introduction

Disadvantages of using Non-parametric Methods

- Non-parametric tests can be wasteful of data if corresponding parametric tests are available for use with the data.
- Non-parametric tests are usually not as widely available and well known as parametric tests.
- For large samples, the computations for many non-parametric tests can be tedious; although the results of many tests can be generated using statistical software such as *R*.

Introduction

While there are many non-parametric methods available, we will focus on:

- A test for the median of one population which is the non-parametric equivalent of the t-test for the mean.
- A test for the equality of medians of two independent populations which is the non-parametric equivalent of the t-test for equality of two means.
- A test for the equality of medians of two related populations which is the non-parametric equivalent of the paired t-test.
- A test for the equality of medians of more than two independent populations which is the non-parametric equivalent of the completely randomised design (One-Way ANOVA).
- A test for the equality of medians of more than two related populations which is the non-parametric equivalent of the randomised block design (Two-Way ANOVA without replications).

Available materials

1. Short videos including demonstrations using R software
2. Accompanying slides
3. R scripts and relevant data sets
4. Exercises
5. Additional material
6. List of references