Statistics & Experimental Design with R

Barbara Kitchenham

Keele University

Keele University

Introduction

Part 1

Scope of Workshop

- Basic Statistics
  - Classical statistical methods
    - Parametric & Non-Parametric
  - Newer methods
    - Randomisation (Permutation methods)
    - Sample-based robust methods
- Experimental design
  - Experiments and Quasi-experiments

Keele University

Population and Samples

- Population
  - All participants or objects relevant to a study
    - All Java programmers
    - All software development companies
- Sample
  - A subset of subjects or objects belonging to the relevant population
- Random sample
  - Sample where member of population has same probability of being included
  - Assumption underlying many statistical methods
  - Basis of generalisations from sample to population
  - You need to be sure you know whether or not you have a random sample

Fundamental Concepts of Statistics

- Design
  - Planning & carrying out an experiment
- Description
  - Methods for summarizing data
- Inference
  - Making predictions or generalizations about the population from the sample

Keele University

Design

- Types of Study (for this tutorial)
  - Experiment
    - A test under controlled conditions to examine the validity of a hypothesis
    - Randomised experiment
      - Subjects/objects in a sample are allocated at random to one of two or more experimental treatments/interventions
    - Quasi-experiment
      - Subjects/objects cannot be allocated at random
        - » Males v. Females
        - » Project that used CMM v. those that did not
  - Observational study/Correlational Study
    - Features of a sample of subjects/objects are measured
  - You always need to be sure you know what type of study you are doing

Keele University

Description

- Descriptive statistics
  - Measures that describe or display graphically properties of the *sample*
- Measures of central tendency
  - Also called measures location
  - Aim to identify the value of a typical member of the sample
- Measures of dispersion
  - Aim to identify the spread of values in the sample
- Graphical displays
  - Aim to reveal distribution of values

Inference

- Inferential statistics
  - Often the same as descriptive statistics
  - But intended to apply to the population
- Statistical claims are based on random samples
  - Without random samples claims need to be justified
- However generalization may not cover the entire range of
  - Settings
  - Task and material complexity
  - Possible outcome measures
  - Subjects/objects of study
  - Interventions/treatments
- Random sampling does not rule out possibility of errors
  - Type 1 error α= Incorrectly rejecting the null hypothesis
  - Type 2 error β = incorrectly accepting the null hypothesis
  - Note: Power=1-β

Keele University

Statistical approaches - 1

- Classical Statistics
  - Parametric methods
    - Frequency Distributions
    - ANOVA
    - Regression & Correlation
    - Contingency Tables
- Usually based on Normal/Gaussian Distribution
  - May be unreliable if Normality assumptions don't hold
    - Starting point for developing improved methods
  - Found in all statistical packages and text books
  - Tutorial will discuss these methods

Statistical Approaches - 2

- Robust methods
  - Often based on ranks
    - Spearman's rank correlation
    - Wilcoxon Mann-Whitney test for comparing two groups
    - Kruskall-Wallis for comparing three or more groups
  - Recent studies suggest these techniques can have low power when comparing groups with different distributions
    - e.g. different variances (although they are supposed to be non-parametric)

Keele University

Statistical Approaches - 3

- Permutation/Randomisation methods
  - Used to compare different treatment groups
  - Assumes random allocation to treatment (not random sample)
  - Identifies the distribution of the null hypothesis by permuting the observations over the groups
- Very plausible method but has problem
  - For comparing two populations
    - Non-parametric but not robust if populations differ more than just wrt location

Keele University

Statistical Approaches - 4

- Bootstrapping
  - Assumes a random sample
  - Like permutation methods
    - Creates many different samples from the original data
    - But uses sampling with **replacement**
  - Non-parametric approach
    - Evidence suggests better properties than standard non-parametric tests
- Other effective non-parametric methods
  - Trimmed means
  - Kernel Density estimation
- We will cover some aspects of these methods

Statistical Approaches - 5

- Bayesian Statistics
  - Change prior probabilities that parameters take a particular value to new (posterior) probabilities
    - Based on data + prior distribution
    - Assume θ can take on *n* different values θ_i

$$Prob(\theta_i|data) = Prob(data|\theta_i)p(\theta_i) / \sum_{j=1}^{n} Prob(data|\theta_j)p(\theta_j)$$

  - Can be solved using Markov Chain Monte Carlo methods e.g. Gibbs Sampler
    - WINBUGS Software
  - Assumes
    - The prior distribution is known
    - Data are random sample from that distribution
  - Not covered in tutorial except for issues associated with logistic regression

Keele University

Design Topics

- Basic types of experimental design
  - Randomised (One factor)
  - Multiple factor (Factorials)
  - Blocking
  - Within subject v. Between Groups
  - Random v. Fixed Factors
- Quasi-experiments
  - Apply when randomisation is impossible
    - Used for assessing impact of “programs” e.g. CMM
  - Specific types of design:
    - Differences in Differences
    - Interrupted Time Series
  - Assessing Causality

Keele University

The R Statistical Language

- The examples presented in this workshop use R
- R is Open Source
- It is a very flexible language
  - Many packages are supported by leading statistical researchers
  - Many test books available
  - Easy to program your own functions
- I find it sometimes difficult to use
  - Data handling is messy
  - No consistency among different packages that perform similar functions
- But arguably the best statistical software available