Which Stats Method Should I Use?

Richard White

2017-08-11

Contents

Chapter 1

Syllabus

Instructor: Richard White [richard.white@fhi.no]

Time: 09:00 - 11:45, 18th September 2017

Location: Main auditorium, L8, Lindern Campus, Folkehelseinstittutet, Oslo

Language: English

Format and Procedures

09:00 - 10:00: Lecture 1

10:00 - 10:10: Break

10:10 - 11:10: Lecture 2

10:10 - 10:15: Break

11:15 - 11:45: Examples from FHI

Description

This course will provide a basic overview of general statistical methodology that can be useful in the areas of infectious diseases, environmental medicine, and labwork. By the end of this course, students will be able to identify appropriate statistical methods for a variety of circumstances.

This course will **not** teach students how to implement these statistical methods, as there is not sufficient time. The aim of this course is to enable the student to identify which methods are required for their study, allowing the student to identify their needs for subsequent methods courses, self-learning, or external help.

You should register for this course if you are one of the following:

- Have experience with applying statistical methods, but are sometimes confused or uncertain as to whether or not you have selected the correct method.
- Do not have experience with applying statistical methods, and would like to get an overview over which methods are applicable for your projects so that you can then undertake further studies in these areas.

Lecture 1

1. Identifying continuous, categorical, count, and censored variables
2. Identifying exposure and outcome variables
3. Identifying when t-tests (paired and unpaired) should be used
4. Identifying when non-parametric t-test equivalents should be used
5. Identifying when ANOVA should be used
6. Identifying when linear regression should be used
7. Identifying the similarities between t-tests, ANOVA, and regression
8. Identifying when logistic regression models should be used

9. Identifying when Poisson/negative binomial and cox regression models should be used
10. Identifying when chi-squared/fisher's exact test should be used

Lecture 2

1. Identifying when data does not have any dependencies (i.e. all observations are independent of each other) versus when data has complicated dependencies (i.e. longitudinal data, matched data, multiple cohorts)
2. Identifying when mixed effects regression models should be used
3. Identifying when conditional logistic regression models should be used
4. (TBD) Understanding the different imputation methods used when lab data is below the limit of detection (LOD)
5. (TBD) Understanding the best practices for data files and project folders

Prerequisites

To participate in this course it is recommended that you have some experience with either research or data.

Additional information

For the last 30 minutes of the course we will be going through examples of analyses performed at FHI and identifying which statistical methods are appropriate. If you would like your analysis to be featured/included in this section, please send an email to richard.white@fhi.no briefly describing your problem.

Chapter 2

Lecture 1

2.1 Variable types

2.1.1 Continuous variables

A variable is continuous there is a meaningful "distance" between values.

2.1.2 Categorical variable

A variable is categorical if there is no meaningful "distance" between values.

2.1.3 Censored variables

Censored variables are a subset of continuous variables

2.1.4 Count variables

Count variables are a subset of continuous variables.

2.2 Hypothesis testing

In science, we are interested in testing hypotheses. Statistics allows us to formally test our hypotheses.

In statistical testing we have a **null** hypothesis (H_0) and an **alternative** hypothesis (H_1). We assume the null hypothesis is true and try to find the probability of what we have observed (or something more extreme). If our observations are very unlikely (assuming the null hypothesis is true) then we reject the null hypothesis in favor of the alternative hypothesis.

For example:

H_0: It is summer

H_1: It is not summer

Our observed data for today is an average temperature of -20C today. Assuming it is summer, how likely is it that today's average temperature will be -20C? Not very likely! We therefore reject H_0 ("it is summer") in favor of H_1 ("it is not summer"). That is, we conclude that it is not summer today.

2.3 t-tests

2.3.1 What is a t-test?

A t-test compares the means of two groups.

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Or rephrased:

H_0: The average height of men is equal to the average height of women

H_1: The average height of men is not equal to the average height of women

2.3.2 Paired t-test

2.3.3 Unpaired t-test

2.4 Identifying when non-parametric t-test equivalents should be used

2.5 Identifying when ANOVA should be used

2.6 Identifying when linear regression should be used

2.7 Identifying the similarities between t-tests, ANOVA, and regression

2.8 Identifying when logistic regression models should be used

2.9 Identifying when Poisson/negative binomial and cox regression models should be used

2.10 Identifying when chi-squared/fisher's exact test should be used

Chapter 3

Folder structure

Here is a review of existing methods.

Bibliography