Which Stats Method Should I Use?

Richard White

2017-08-12

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.

For example:

- Temperature
- Weight
- Height
- BMI
- Blood pressure

2.2 Come up with 5 of your own examples:

-
-
-
-

2.2.1 Binary variables

A variable is binary if it can only hold two values.

For example:

- 0 or 1
- True or false
- Male or female
- Sick or healthy
- Born in Norway vs Born outside of Norway

2.2.2 Categorical variable

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

For example:

- Sick or healthy
- Born in Norway vs Born outside of Norway
- Cancer stage (I, II, III, or IV)
- BMI category (underweight, normal, or overweight)

2.2.3 Censored variables

Censored variables are a subset of continuous variables. They are artificially cutoff ("censored") at some point.

For example:

- Height – if everyone over 175cm is recorded as "175+"
- Age – if everyone under 10 years old is recorded as "<=10"
- Time alive since receiving illness diagnosis if there is loss to followup (i.e. we know that the person has lived at least 4 years before we lost track of them)

2.2.4 Count variables

Count variables are a subset of continuous variables. They can only have integer values (e.g. 0, 1, 2, 3).

For example:

- Number of cars that use the parking lot in a day
- Number of influenza patients who use the hospital every day
- Number of tuberculosis patients who are screened every year

2.3 Independent vs Dependent variables

An independent variable is often called an exposure or predictor variable. In an experiment, this variable is manipulated by the researcher.

A dependent variable is often called the outcome. In research, we generally want to see if (the following all mean the same thing):

- The dependent variable is dependent on the independent variable
- The predictor variable predicts the outcome.
- The exposure affects the outcome

For ease of understanding, we will use the terms "outcome" and "exposure" for the rest of this course.

2.4 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.

2.5 Which method to use?

Deciding on the appropriate statistical method is fairly easy. You just look at the:

- Outcome type (continuous, binary, categorical, censored, count)
- Exposure (type)
- Dependencies in the data

And then essentially use a flowchart.

2.6 One sample t-test

A one sample t-test tests if the mean of a continuous variable differs from a specified value (generally zero)

$H_0 : \mu = 180$

$H_1 : \mu \neq 180$

Or rephrased:

H_0: The average height of men is equal to the 180cm

H_1: The average height of men is not equal to 180cm

2.6.0.1 Assumptions

Aim: test if the mean of a continuous variable differs from a specified value

Outcome: continuous variable, all observations independent, distributed as a Normal distribution

Exposure: Does not exist

2.6.0.2 Non-parametric equivalent

This is to be used when the Normal distribution assumption does not hold

2.7 Two sample t-tests

2.7.1 What is a two sample t-test?

A t-test tests if the mean of a continuous variable differs between 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

There are two kinds of two-sample t-tests: paired and unpaired.

2.7.2 Two-sample paired t-test

A paired t-test is a special case where we have N participants, and each participant has two observations (generally "before experiment" and "after experiment"). We want to test if the mean of outcome variable differs between "after" and "before".

For example, in a weight-loss experiment, we have N participants and we want to see if the average "after weight" is different from the average "before weight".

This is done by subtracting the outcome from one group ("before weight") from the outcome in the other group ("after weight") for each person ("difference in weight"), and then performing a one-sample t-test to see if the mean of this variable is different from zero.

$H_0 : \mu_{\text{after}-\text{before}} = 0$

$H_1 : \mu_{\text{after}-\text{before}} \neq 0$

2.7.2.1 Assumptions

Aim: test if the mean of a continuous variable measured twice for each participant differs between "before" and "after"

Special preprocessing of data: for each participant subtract the "before" observation from the "after" observation

Outcome: ("after weight" minus "before weight") continuous variable, all observations within each group independent, distributed as a Normal distribution

Exposure: $\text{group}_{\text{after}}$ vs $\text{group}_{\text{before}}$

2.7.2.2 Non-parametric equivalent

Wilcoxon signed-rank test

2.7.3 Two-sample unpaired t-test

An unpaired t-test is where we have two independent groups of N_1 and N_2 participants, and we want to test if the mean of the outcome variable differs between group_1 and group_2.

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

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

2.7.3.1 Assumptions

Aim: test if the mean of a continuous variable differs between group_1 and group_2.

Outcome: continuous variable, all observations within each group independent, distributed as a Normal distribution

Exposure: group_1 vs group_2

2.7.3.2 Non-parametric equivalent

Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test)

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

Non-parametric t-test equivalents should be used when the Normality distribution fails.

2.9 ANOVA

ANOVA

2.9.0.1 Assumptions

2.9.0.2 Non-parametric equivalent

Kruskal–Wallis test

2.10 Identifying when linear regression should be used

2.10.0.1 Assumptions

Aim: test if the mean of a continuous variable differs between group_1 and group_2.

Outcome: continuous variable

Exposure:

- Continuous
- Binary (0 or 1)
- Categorical (0, 1, 2, …)
- Count data

2.11 Identifying the similarities between t-tests, ANOVA, and linear regression

t-tests are ANOVA with only two groups

t-tests are linear regressions with a binary (0/1) exposure

ANOVA is a lienar regression with a categorical exposure

2.12 Identifying when logistic regression models should be used

- When you have a binary (0/1) outcome
- When you are doing a case-control study [case control studies can ONLY be analysed using logistic regression]

2.13 Identifying when Poisson/negative binomial models should be used

When your outcome is count data

2.14 Cox regression models should be used

When you have survival data

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

When you have a categorical outcome and a categorical exposure

Chapter 3

Lecture 2

3.1 What is data with dependencies

3.1.1 What is independent data

Broadly, having knowledge about one observation should not give you knowledge about other observations in your dataset.

For example, if we flip a coin ten times, knowing the result of the first coin toss (heads) will not give us knowledge about the subsequent 9 coin tosses.

3.1.2 What is data with dependencies

In reality, most data have dependencies, so we will focus on some of the most important kinds that will severely impact your analyses if you do not identify them.

3.1.3 Repeated measures/longitudinal data

If you have a dataset with repeated measures (e.g. some people in your cohort have more than one observation), then the repeated observations on each person cause dependencies in the data. That is, if a person has their weight measured five times, then just by knowing their first weight you can have a good guess at what their subsequent weights will be.

3.1.4 Grouped/clustered data

Repeated measures data is a type of clustered data, where each person is their own cluster. There can be many kinds of clusters, for example:

- Data sampled from multiple hospitals could have the hospital as the cluster variable
- Data sampled from multiple countries could have the country as the cluster variable
- Data sampled from multiple counties/municipalities could have the country/municipality as the cluster variable
- If it is a study of children, and multiple children from each mother are included, then the mother could be the cluster variable

3.1.5 Matched data

Inside case-control studies, for each case a control (or multiple cases) can be selected to have similar attributes. For example, for each case, a control can be selected with a similar age. These controls have been "matched" to a case, and have introduced dependencies into the data.

3.2 Analysing data with dependencies

3.2.1 Mixed effects regression

Mixed effects regression should be used for grouped/clustered data (and subsequently repeated measures/longitudinal data)

3.2.2 Identifying when conditional logistic regression models should be used

Conditional logistic regression should be used for matched data

3.3 (TBD) Understanding the different imputation methods used when lab data is below the limit of detection (LOD)

3.4 (TBD) Understanding the best practices for data files and project folders

Bibliography