Which Stats Method?

Richard White

2018-11-01

Contents

Syllabus **2**

1 Reference **5**

2 Variables **6**
2.1 Introduction . . . 6
2.2 Continuous Variables . . . 6
2.3 Binary Variables . . . 6
2.4 Categorical Variable . . . 7
2.5 Censored Variables . . . 7
2.6 Count Variables . . . 8
2.7 Independent Versus Dependent Variables . . . 8
2.8 Dataset workflow (pipeline) . . . 9

3 Simple Hypothesis Testing: Chi-Squared, T-tests, and ANOVA **9**
3.1 Hypothesis Testing . . . 9
3.2 Which Method To Use? . . . 10
3.3 Chi-Squared Test . . . 10
3.4 One Sample T-Test . . . 12
3.5 Two sample T-Tests . . . 15
3.6 Two-sample Paired T-Test . . . 15
3.7 Two-sample Unpaired T-Test . . . 20
3.8 ANOVA . . . 24

4 Simple regression (fixed effects) **27**
4.1 Regression in general . . . 27
4.2 Linear regression . . . 28
4.3 Similarities between t-tests, ANOVA, and linear regression . . . 35
4.4 Similarities between ANOVA and linear regression . . . 42
4.5 Logistic regression models . . . 44
4.6 Poisson/negative-binomial regression models . . . 47
4.7 Cox regression models . . . 49

5 Complicated regression **50**
5.1 Dependencies in your data . . . 50
5.2 Analysing data with dependencies . . . 51

5.3 (TBD) Understanding the best practices for data files and project folders . . 52

6 Good folder structure **52**
6.1 Data and results 52
6.2 Keep your source code separate from data 53

7 Examples **56**
7.1 Poisons Information Center 56
7.2 Norwegian Water Pipes 57
7.3 Early warning system (EWS) for waterborne outbreaks (part 1) 58
7.4 Early warning system (EWS) for waterborne outbreaks (part 2) 58
7.5 Incidents in the water supply system and illness 59
7.6 Compliance with boil water advisories and perception of risks 59

8 Solutions **60**
8.1 Poisons Information Center 60
8.2 Norwegian Water Pipes 61
8.3 Early warning system (EWS) for waterborne outbreaks (part 1) 62
8.4 Early warning system (EWS) for waterborne outbreaks (part 2) 62
8.5 Incidents in the water supply system and illness 63
8.6 Compliance with boil water advisories and perception of risks 64

List of Tables

List of Figures

1 img 52
2 img 53
3 img 54
4 img 55
5 img 55
6 img 56

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

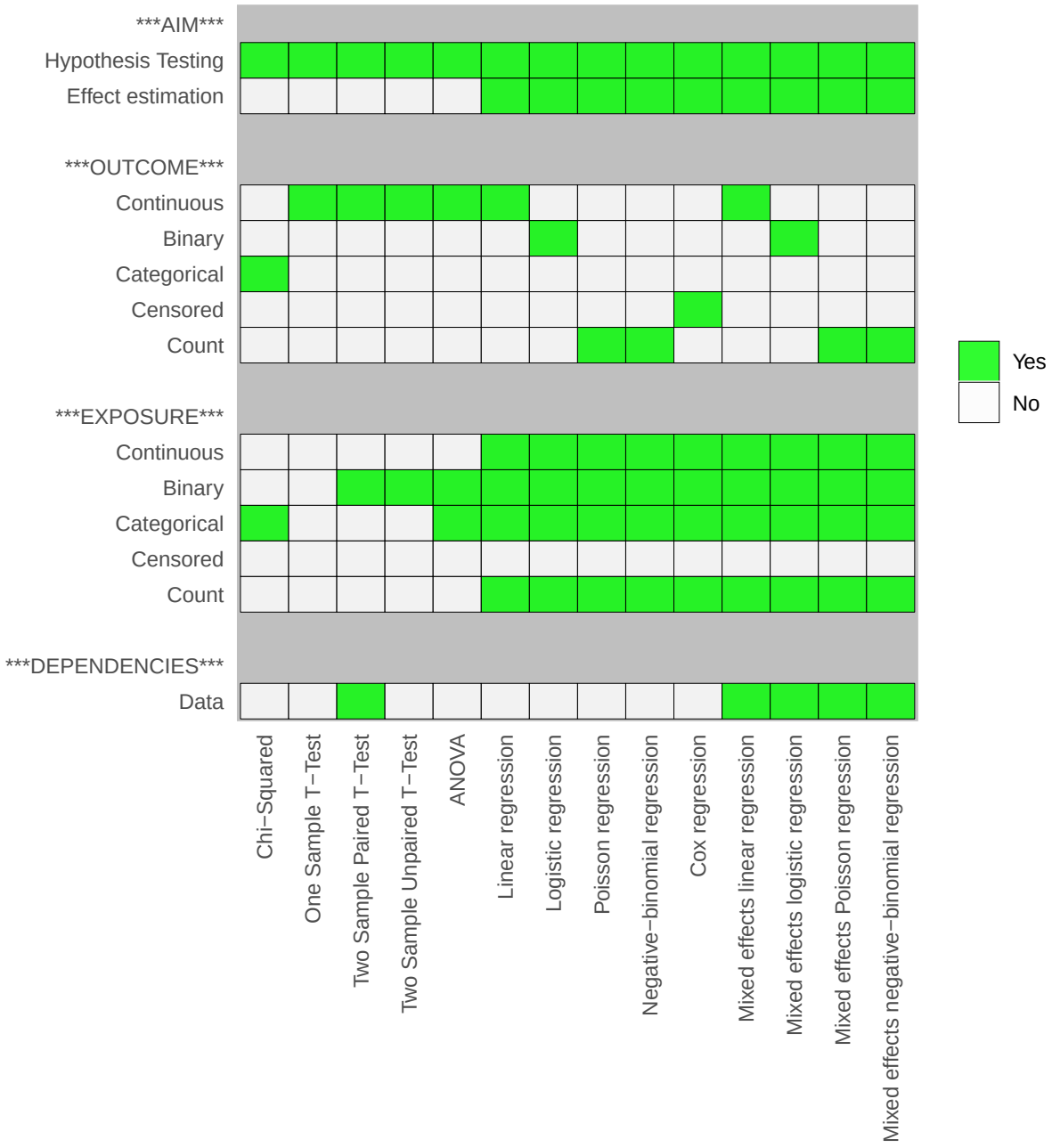
Prerequisites

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.

1 Reference


AIM
Hypothesis Testing
Effect estimation
OUTCOME
Continuous
Binary
Categorical
Censored
Count
EXPOSURE
Continuous
Binary
Categorical
Censored
Count
DEPENDENCIES
Data
Chi−Squared
One Sample T−Test
Two Sample Paired T−Test
Two Sample Unpaired T−Test
ANOVA
Linear regression
Logistic regression
Poisson regression
Negative−binomial regression
Cox regression
Mixed effects linear regression
Mixed effects logistic regression
Mixed effects Poisson regression
Mixed effects negative−binomial regression
Yes
No

2 Variables

2.1 Introduction

A variable is anything that can be measured or counted. In general, we think of our datasets as a rectangle, with a column for each variable and a row for each observation (there are some exceptions when discussing long/wide formatted data, but that is out of the scope of this course).

We care about four attributes of the variables:

- the variable's type (statistical relevance)
- the different values the variable can take (statistical relevance)
- is the variable clean (i.e. ready to use in an analysis?) (statistical relevance)
- the name of the variable (useful for us)

We can use the fourth attribute to help us remember the first three.

2.2 Continuous Variables

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

For example:

- Temperature
- Weight
- Height
- BMI
- Blood pressure
-

-

Clean continuous variables can be given the prefix `con_` to denote that they are clean. For example, `temperature` could be called `con_temperature` after it has been cleaned and is ready for analysis.

2.3 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
-

-

Clean binary variables can be given the prefix `is_` to denote that they are clean, binary, and reference the "active state". For example, an unclean variable called `sex` could be recoded as 0 for female and 1 for male, then called `is_male` to denote that it is clean (ready for analysis), binary, and "male" is the active state when `is_male=1`.

2.4 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)
-

-

Clean categorical variables can be given the prefix `cat_` to denote that they are clean. For example, `BMI category` could be called `cat_bmi` after it has been cleaned and is ready for analysis.

2.5 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)
-
-

Clean censored variables can be given the prefix `cen_` to denote that they are clean. For example, `time alive since receiving illness diagnosis` could be called `cen_time_alive` after it has been cleaned and is ready for analysis.

2.6 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
-
-

Clean count variables can be given the prefix `cou_` to denote that they are clean. For example, `number of cars` could be called `cou_num_cars` after it has been cleaned and is ready for analysis.

2.7 Independent Versus 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.

2.8 Dataset workflow (pipeline)

- We begin with a raw dataset (this is never altered)
- We clean the raw dataset and create new variables as needed
- We save a "clean" dataset (all variables have the prefixes `c_` or `is_`)
- We ONLY run analyses on the clean dataset

We do all of this in "do files" that allow us to recreate the clean dataset from the raw dataset.

Think of our analysis as making dinner, with the `do files` as our `recipe` and the `raw dataset` as our `raw ingredients`. The `recipe` tells us how to prepare the `raw ingredients` (`clean dataset`) and how to `cook` (`analyse`) the `prepared ingredients` (`clean dataset`) to produce the `food` (`results`).

All we need are `raw ingredients` and the `recipe`! The `prepared ingredients` and the `food` are downstream by-products!

This means that if our code is written correctly, we can delete our `clean datasets` and `results` without any worry, because the `raw dataset` and `do files` are sufficient.

3 Simple Hypothesis Testing: Chi-Squared, T-tests, and ANOVA

3.1 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:

$$H_0 : \text{It is summer}$$

$$H_1 : \text{It is not summer}$$

Our observed data for today is a maximum temperature of -20C. Assuming it is summer, how likely is it that today's maximum 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.

3.2 Which Method To Use?

Deciding on the appropriate statistical method is (in principle) fairly easy. You just look at the:

- Aim (hypothesis testing or estimation of effect size?)
- Outcome type (continuous, binary, categorical, censored, count)
- Exposure (type)
- Parametric assumptions
- Dependencies in the data

and we then (essentially) use a flowchart.

3.3 Chi-Squared Test

A Chi-Squared test is used to test if two categorical variables are associated with each other.

3.3.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (testing if two categorical variables are associated with each other.)

Outcome: Categorical variable

Exposure: Categorical variable

Parametric assumptions: No

Dependencies: None (all observations independent)

3.3.2 Examples

- Testing if people's country of origin (Norway/Not Norway) is associated with tuberculosis status (never had TB/has had TB)
- Testing if people's region of origin (Europe/North America/South America/Other) is associated with marital status (Single/Married/Divorced)
- Testing if county of residence (Oslo, Akershus, etc) is associated with post-surgery infection status (No infection/mild infection/deep infection)
-

-

3.4 One Sample T-Test

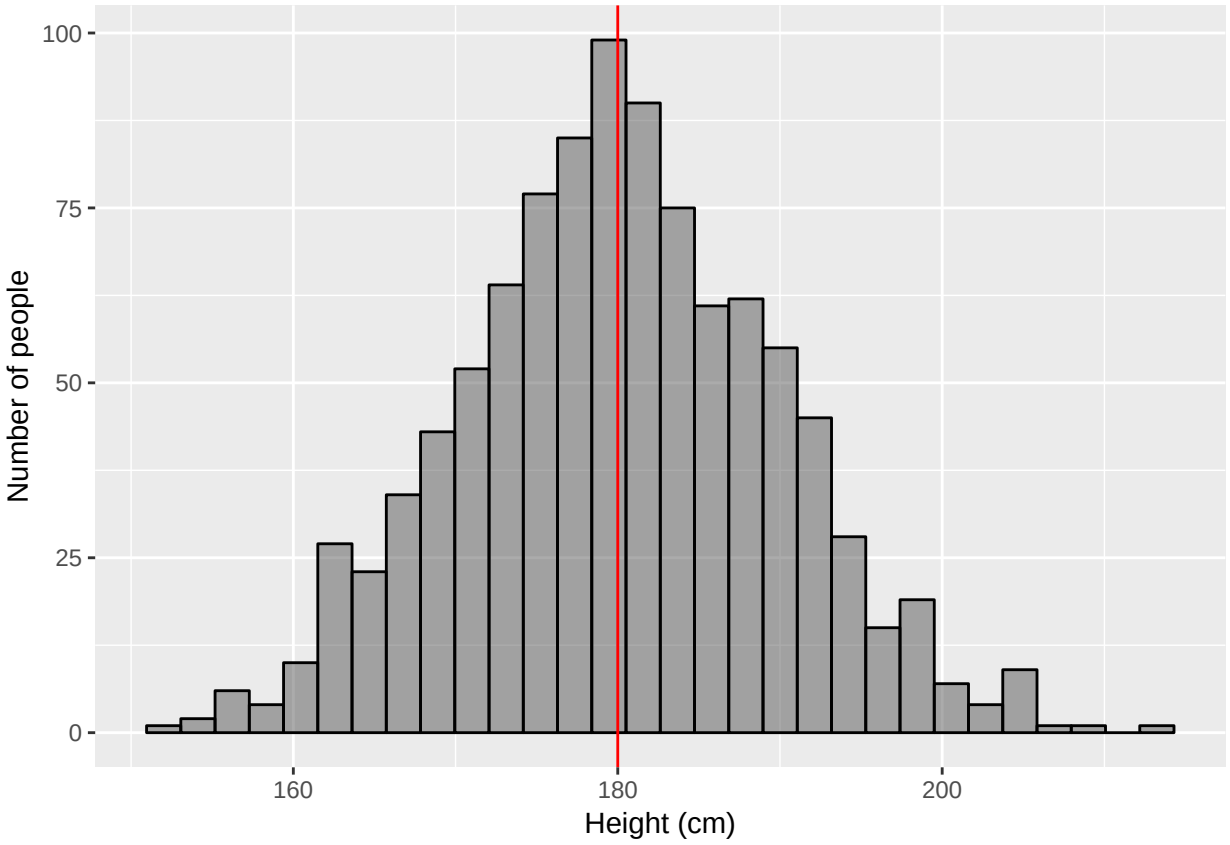
A one sample t-test tests if the mean of a continuous variable differs from a specified value (generally zero)

$$H_0 : \mu = 180$$
$$H_1 : \mu \neq 180$$

Or rephrased:

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

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



3.4.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (test if the mean of a continuous variable differs from a specified value)

Outcome: Continuous variable

Exposure: Does not exist

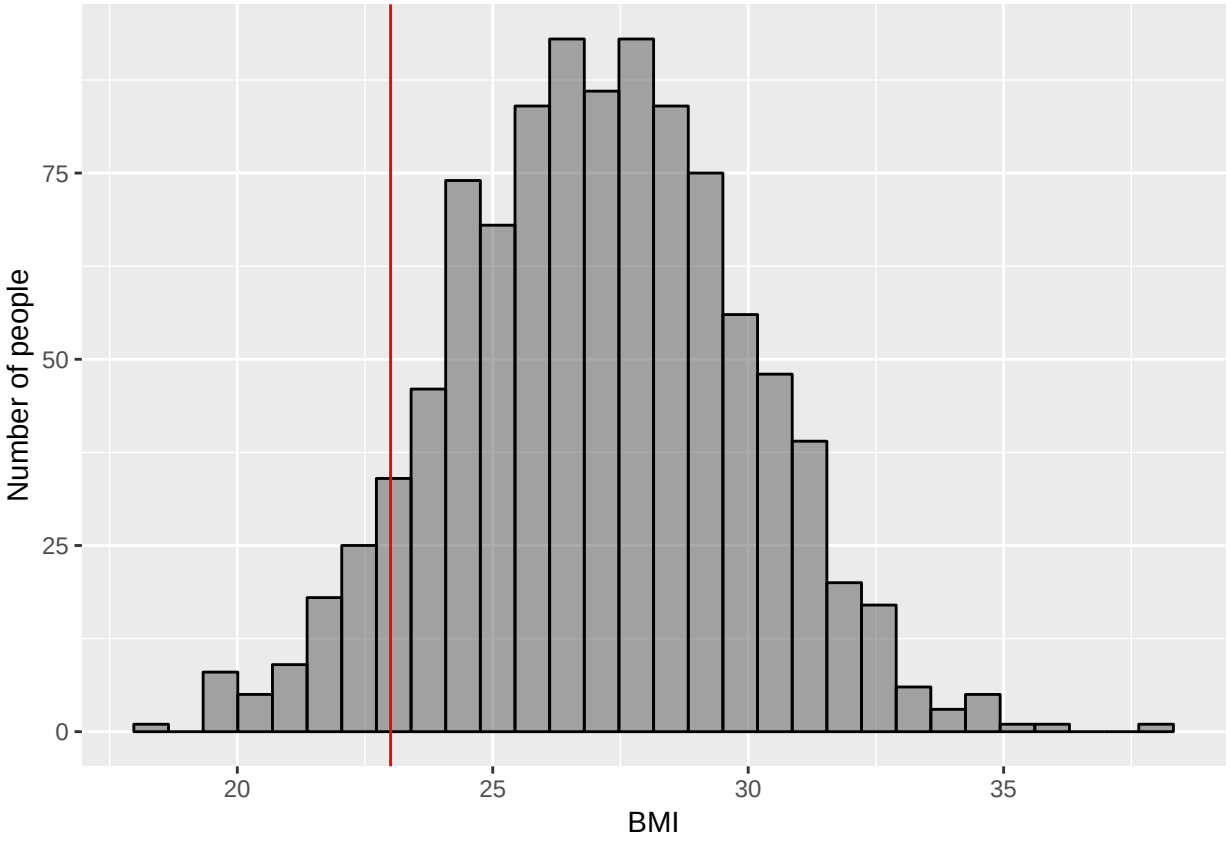
Parametric assumptions: Outcome is distributed as a Normal distribution

Dependencies: None (all observations independent)

3.4.2 Example 1

→ Testing if the average BMI of Norwegians is equal to 23

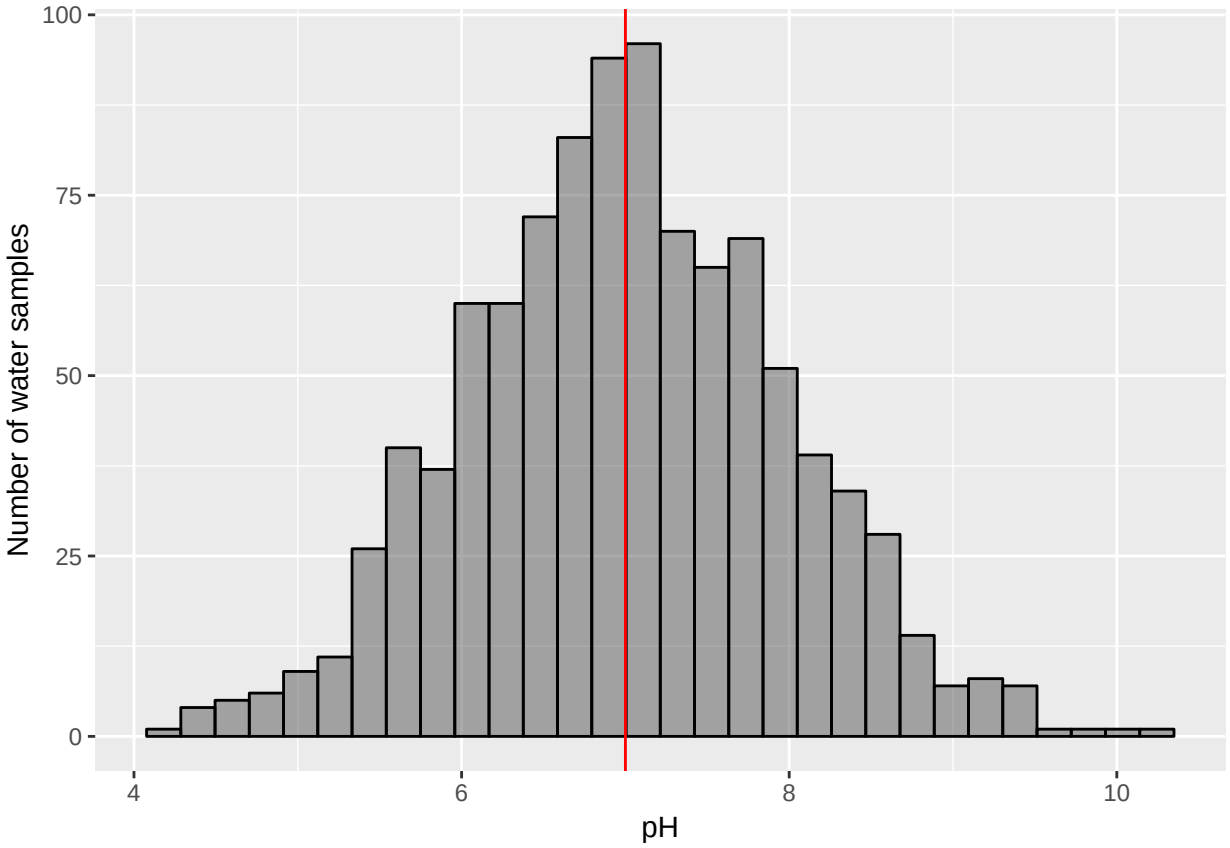
$$H_0 : \mu_{\text{bmi}} = 23$$
$$H_1 : \mu_{\text{bmi}} \neq 23$$



3.4.3 Example 2

→ Testing if the average pH of tap water is equal to 7

$$H_0 : \mu_{\text{pH}} = 7$$
$$H_1 : \mu_{\text{pH}} \neq 7$$



3.4.4 Example 3

→

H_0 :

H_1 :

3.4.5 Example 4

$\rightarrow$

H_0 :

H_1 :

3.5 Two sample T-Tests

A t-test tests if the mean of a continuous variable differs between two groups. There are two kinds of two-sample t-tests: paired and unpaired.

3.6 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.

$$\mathrm{H}_0 : \mu_{\text{after}-\text{before}} = 0$$
$$\mathrm{H}_1 : \mu_{\text{after}-\text{before}} \neq 0$$

3.6.1 Aim/Outcome/Exposure/Parametric/Dependencies

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

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

Outcome: ("after weight" minus "before weight") continuous variable

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

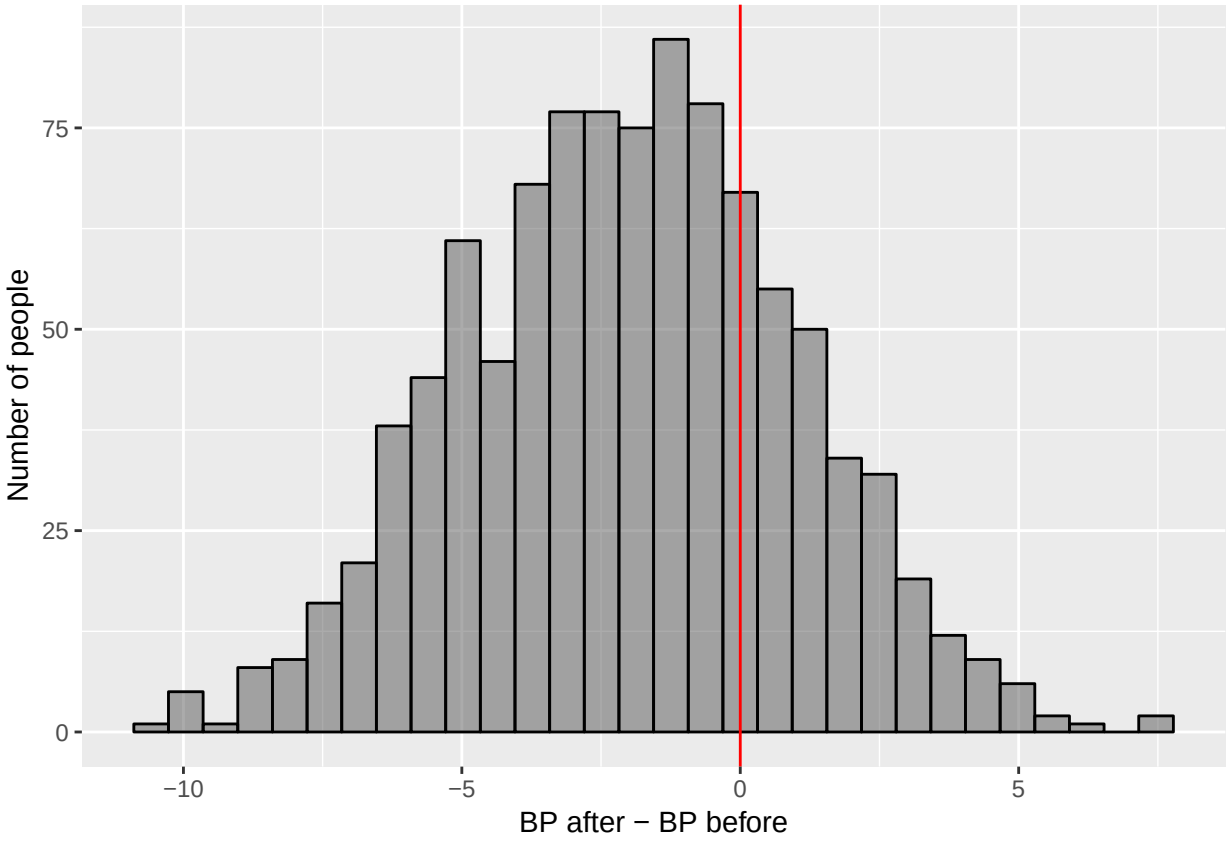
Parametric assumptions: Outcome is distributed as a Normal distribution

Dependencies: Paired data

3.6.2 Example 1

→ Testing if there is a difference in blood pressure before and after treatment (measured on the same person)

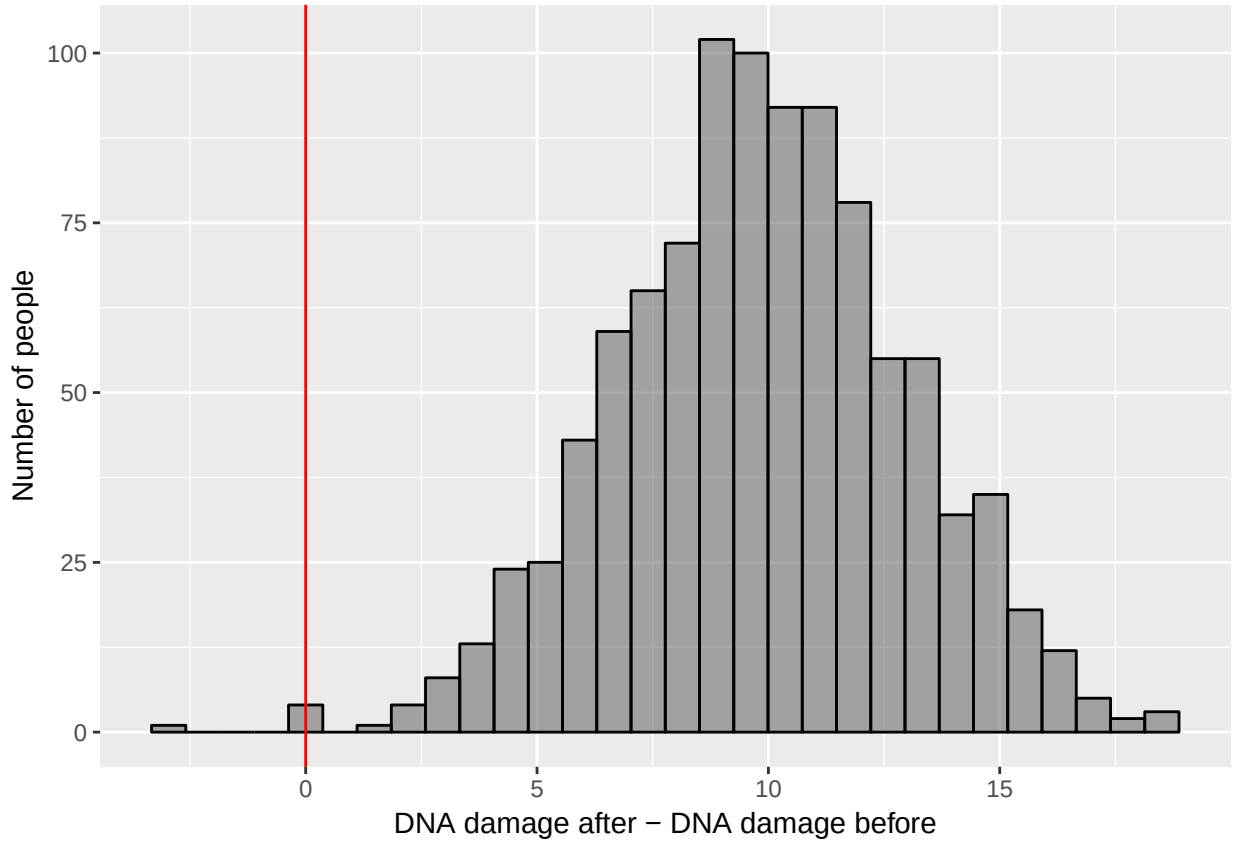
$$H_0 : \mu_{\text{BP after}-\text{BP before}} = 0$$
$$H_1 : \mu_{\text{BP after}-\text{BP before}} \neq 0$$



3.6.3 Example 2

→ Testing if there is a difference in mouse DNA damage before and after irradiating (measured on the same mouse)

$$\mathrm{H}_0 : \mu_{\text{DNA damage after}-\text{DNA damage before}} = 0$$
$$\mathrm{H}_1 : \mu_{\text{DNA damage after}-\text{DNA damage before}} \neq 0$$



3.6.4 Example 3

→

H_0 :

H_1 :

3.6.5 Example 4

$\rightarrow$

H_0 :

H_1 :

3.6.6 Non-Parametric Equivalent

Wilcoxon signed-rank test. This should be used when the Normality assumption fails.

3.7 Two-sample Unpaired T-Test

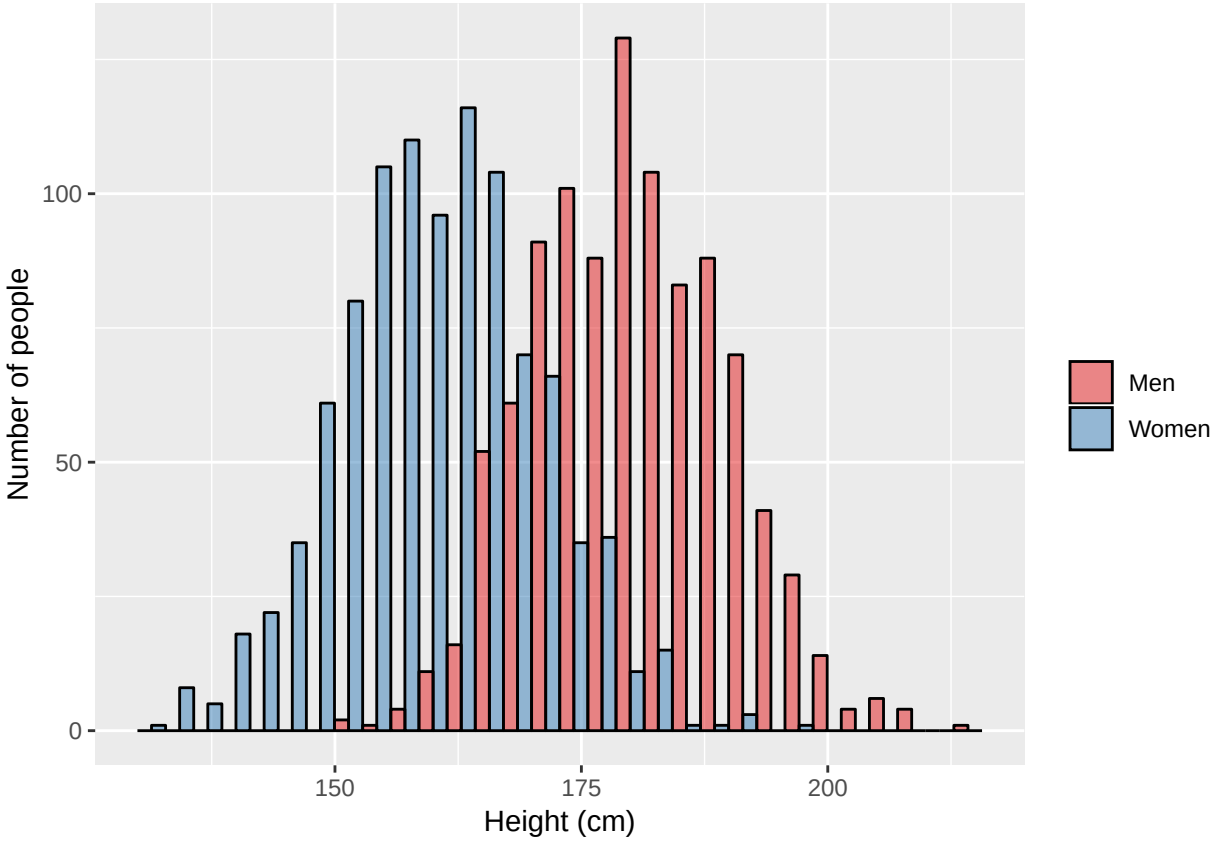
An unpaired t-test is where we have two independent groups of N_1 and N_2 participants, and we want to test if the mean of the outcome variable differs between group_1 and group_2.

$$\text{H}_0 : \mu_0 = \mu_1$$
$$\text{H}_1 : \mu_0 \neq \mu_1$$

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



3.7.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (test if the mean of a continuous variable differs between group_1 and group_2)

Outcome: continuous variable

Exposure: group_1 vs group_2

Parametric assumptions: Outcomes for each group are distributed as a Normal distribution

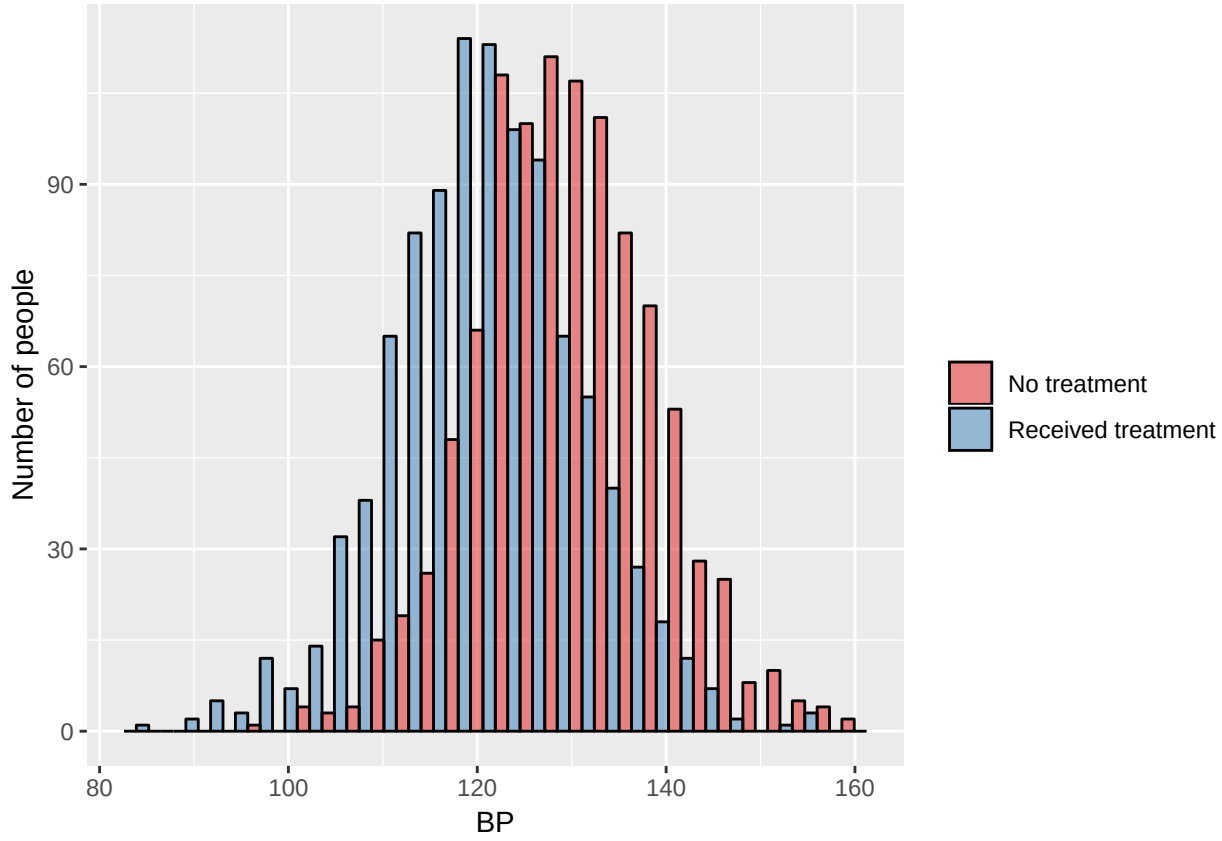
Dependencies: None (all observations independent)

3.7.2 Example 1

→ Testing if the average blood pressure in people who didn't receive treatment is different from people who did receive treatment

$$H_0 : \mu_{\text{BP treatment}} = \mu_{\text{BP no treatment}}$$

$$H_1 : \mu_{\text{BP treatment}} \neq \mu_{\text{BP no treatment}}$$

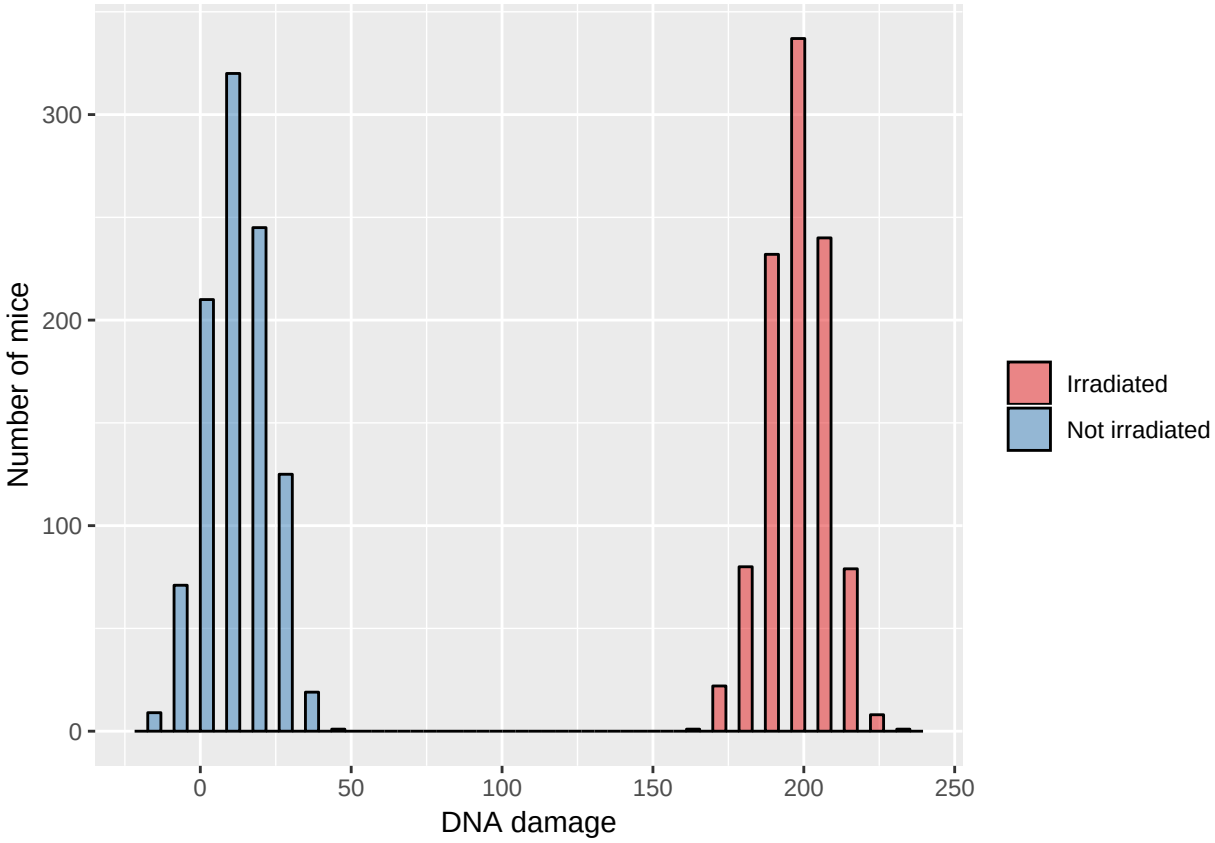


3.7.3 Example 2

→ Testing if the average DNA damage in mice that weren't irradiated is different from mice that were irradiated

$$H_0 : \mu_{\text{DNA radiation}} = \mu_{\text{DNA no radiation}}$$

$$H_1 : \mu_{\text{DNA radiation}} \neq \mu_{\text{DNA no radiation}}$$



3.7.4 Example 3

→

H_0 :

H_1 :

3.7.5 Example 4

$\rightarrow$

H_0 :

H_1 :

3.7.6 Non-parametric equivalent

Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test). This should be used when the Normality assumption fails.

3.8 ANOVA

ANOVA is an extension of the two-sample unpaired t-test. We have X independent groups of participants, and we want to test if the mean of the outcome variable differs between groups.

3.8.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing (test if the mean of a continuous variable differs between some groups)

Outcome: continuous variable

Exposure: group_1 vs group_2 (vs group_3 …)

Parametric assumptions: Outcomes for each group are distributed as a Normal distribution

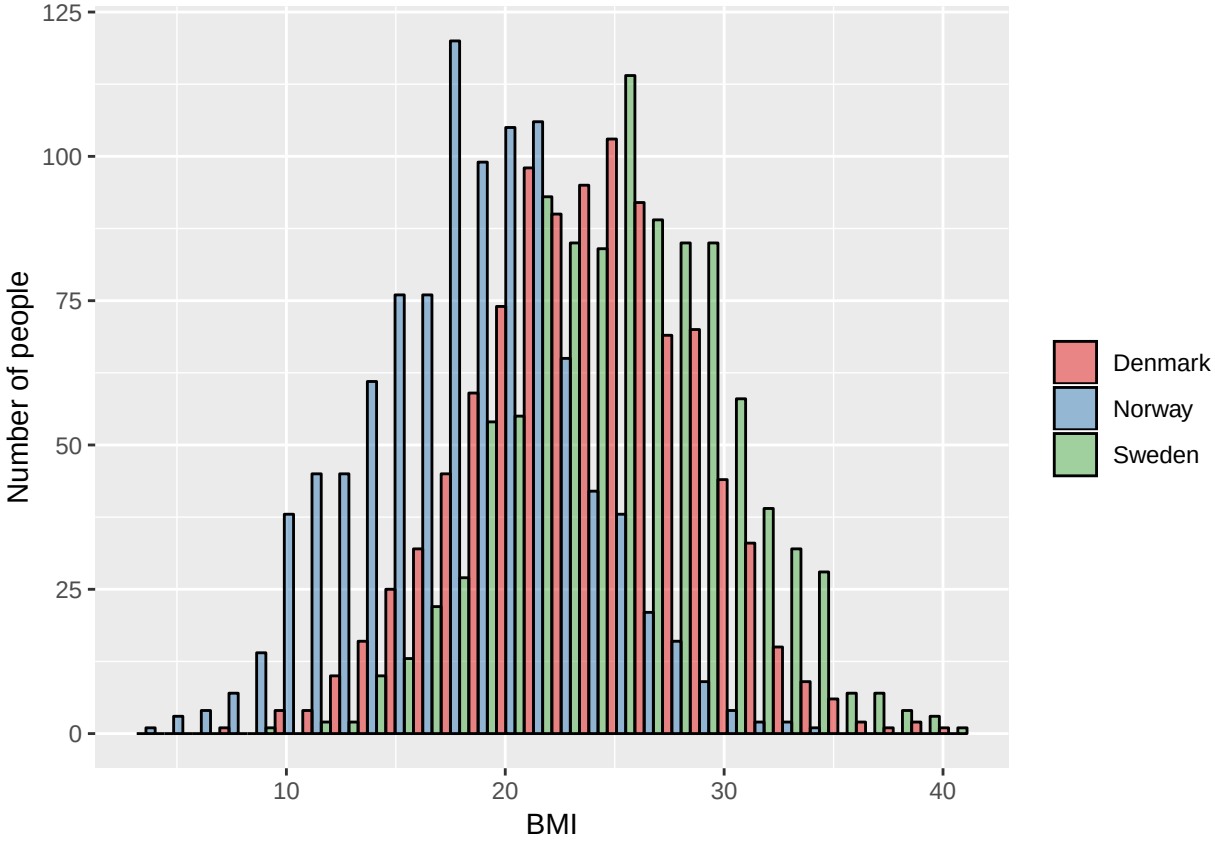
Dependencies: None (all observations independent)

3.8.2 Example 1

→ Testing if average BMI levels differ across Scandinavia

$$H_0 : \mu_{\text{Norway}} = \mu_{\text{Denmark}} = \mu_{\text{Sweden}}$$

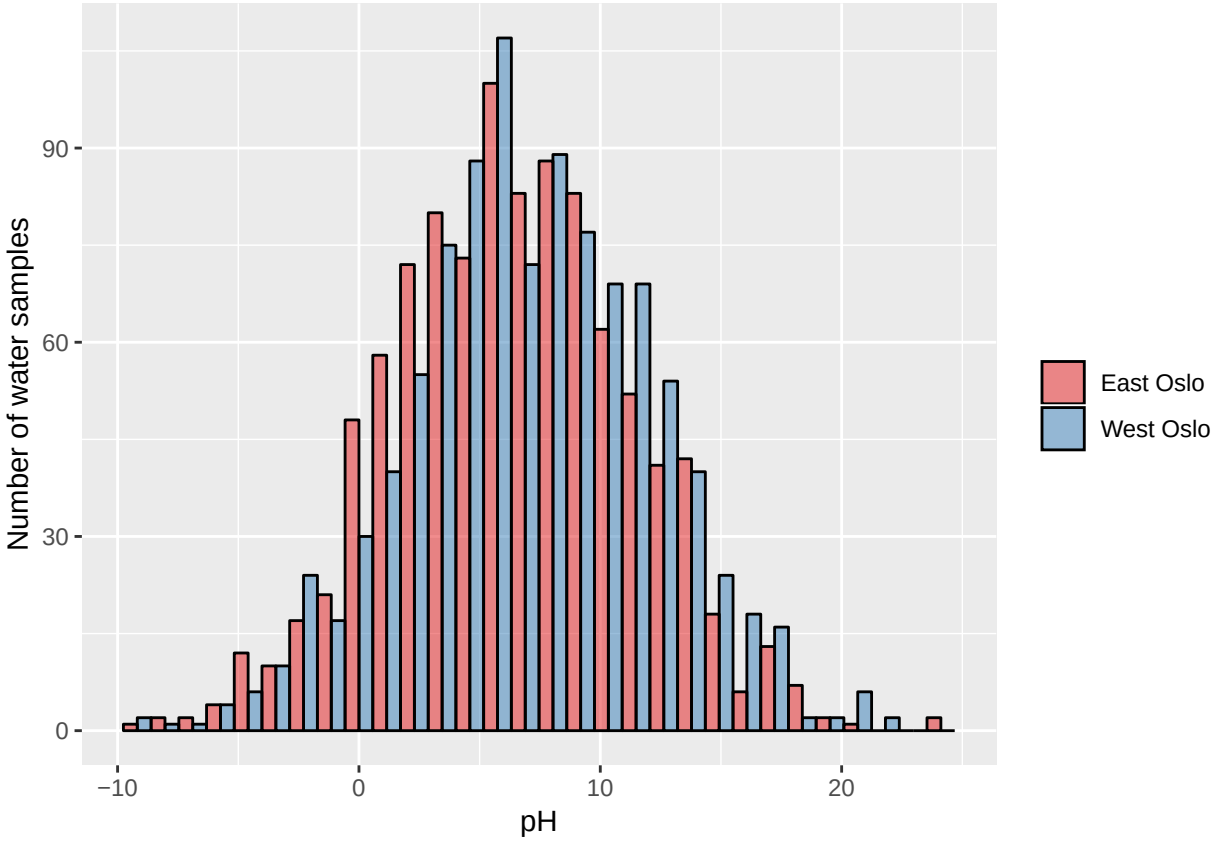
$$H_1 : \mu_{\text{Norway}} \neq \mu_{\text{Denmark}} \text{ and/or } \mu_{\text{Norway}} \neq \mu_{\text{Sweden}} \text{ and/or } \mu_{\text{Denmark}} \neq \mu_{\text{Sweden}}$$



3.8.3 Example 2

→ Testing if average water pH levels differ between East and West Oslo

$$\mathrm{H}_0 : \mu_{\text{East Oslo}} = \mu_{\text{West Oslo}}$$
$$\mathrm{H}_1 : \mu_{\text{East Oslo}} \neq \mu_{\text{West Oslo}}$$



3.8.4 Example 3

→

H_0 :

H_1 :

3.8.5 Example 4

$\rightarrow$

H_0 :

H_1 :

3.8.6 Non-parametric equivalent

Kruskal–Wallis test

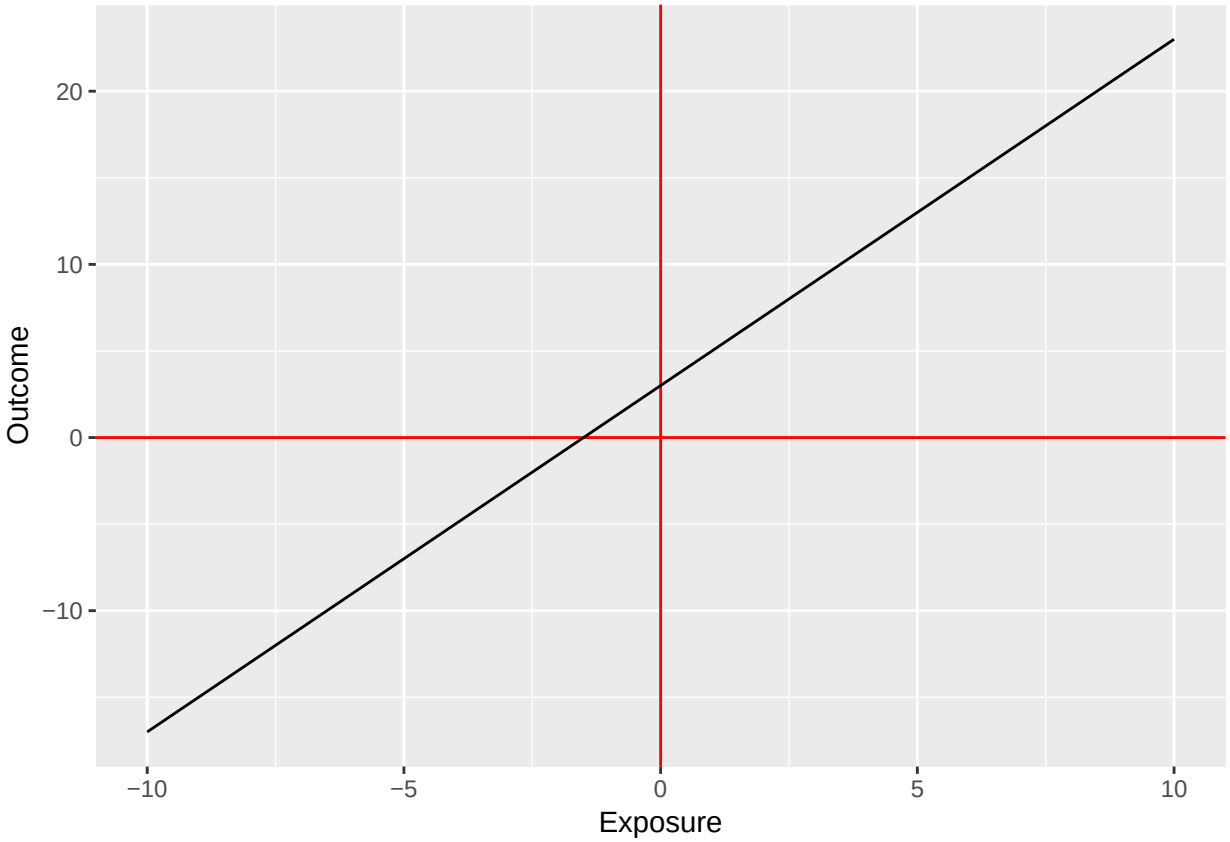
4 Simple regression (fixed effects)

4.1 Regression in general

Regression is the explicit modelling of a parametric association between an outcome and an exposure.

One such parametric association might be the following:

$$\text{outcome} = 3 + 2 \times \text{exposure}$$



Depending on the type of outcome, different types of regression will need to be used.

For all regressions, the exposure can be:

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

Regressions can both:

- Perform hypothesis testing (same as the previous tests we have learned about)
- Estimate numerically the effect size of the association between outcome and exposure (new!)

4.2 Linear regression

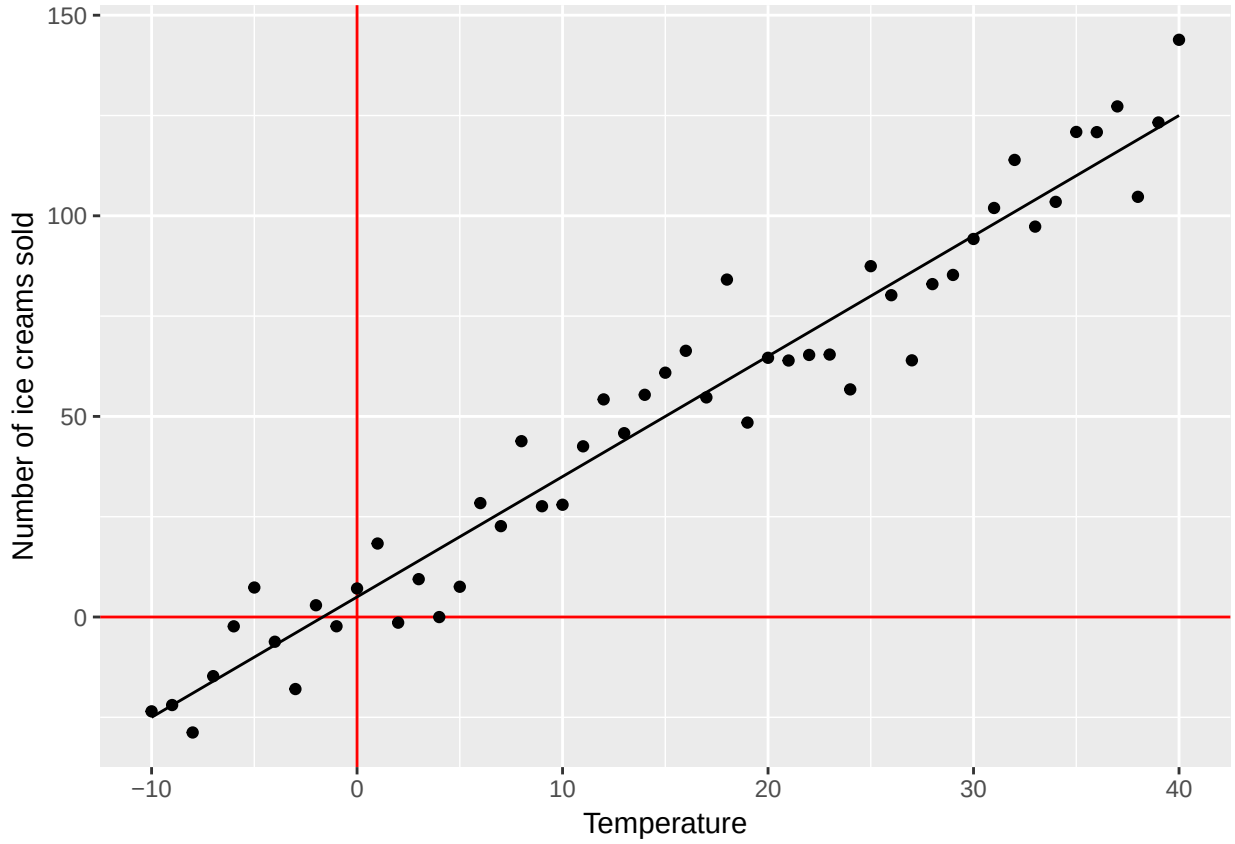
In the most basic form, we have:

$$\text{outcome} = \beta_0 + \beta_1 \times \text{exposure} + \text{error}$$

Where we aim to estimate values for β_0 and β_1.

For example, if we run an ice cream shop:

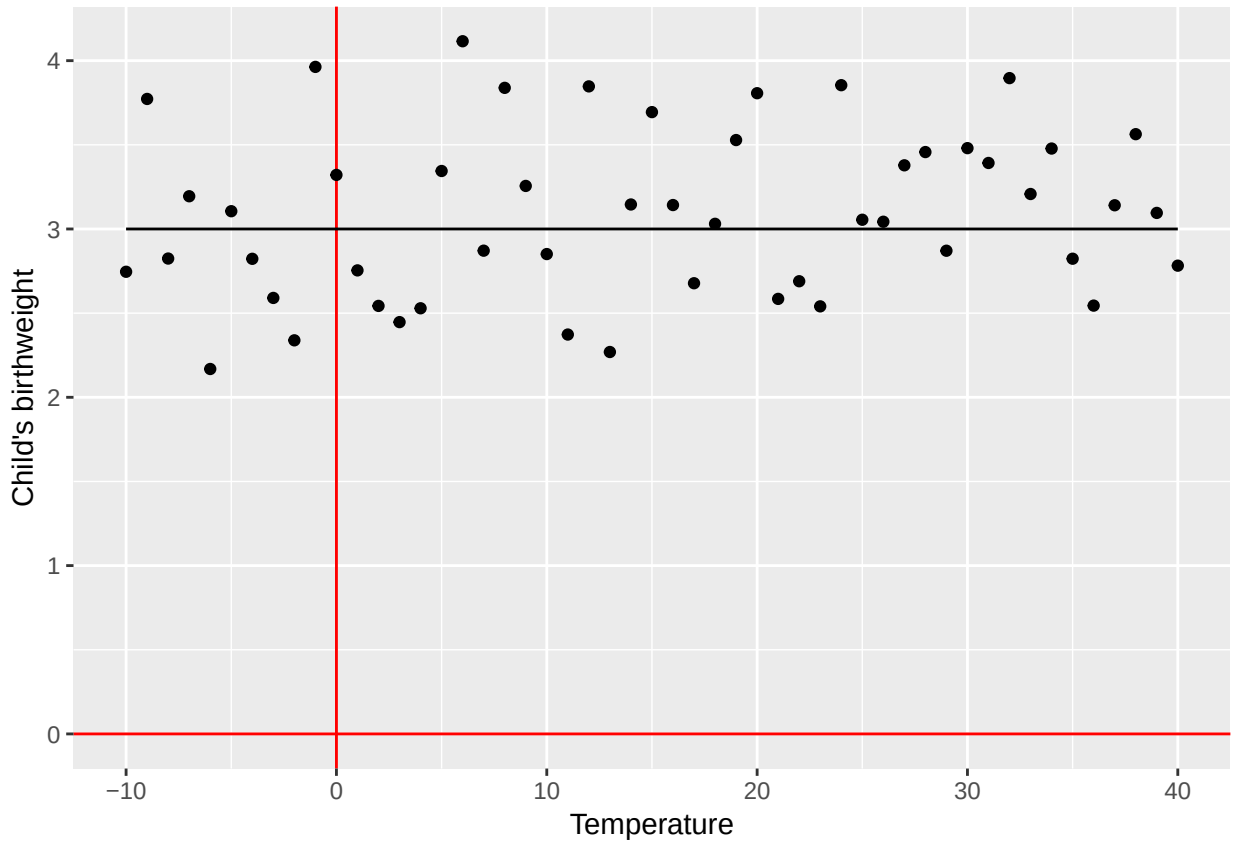
$$\text{number of ice creams sold} = 5 + 3 \times \text{temperature} + \text{error}$$



If today's temperature is 30C, we can expect our shop to sell $5 + 3 \times 30 = 95$ ice creams. Because β_1 $(= 3)$ was not zero, we have a significant association between temperature and number of ice creams sold.

Another example, if we work as a midwife:

$$\text{Child's birthweight} = 3 + 0 \times \text{temperature at day of delivery} + \text{error}$$



If today's temperature is 30C, we can expect that children born today will be (on average) $3 + 0 \times 30 = 3$ kg. If tomorrow's temperature is 10C, we can expect that children born today will be (on average) $3 + 0 \times 10 = 3$ kg. Because β_1 was zero, we do not have a significant association between temperature and birthweight.

4.2.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Continuous variable*

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions: Residuals are distributed as a Normal distribution

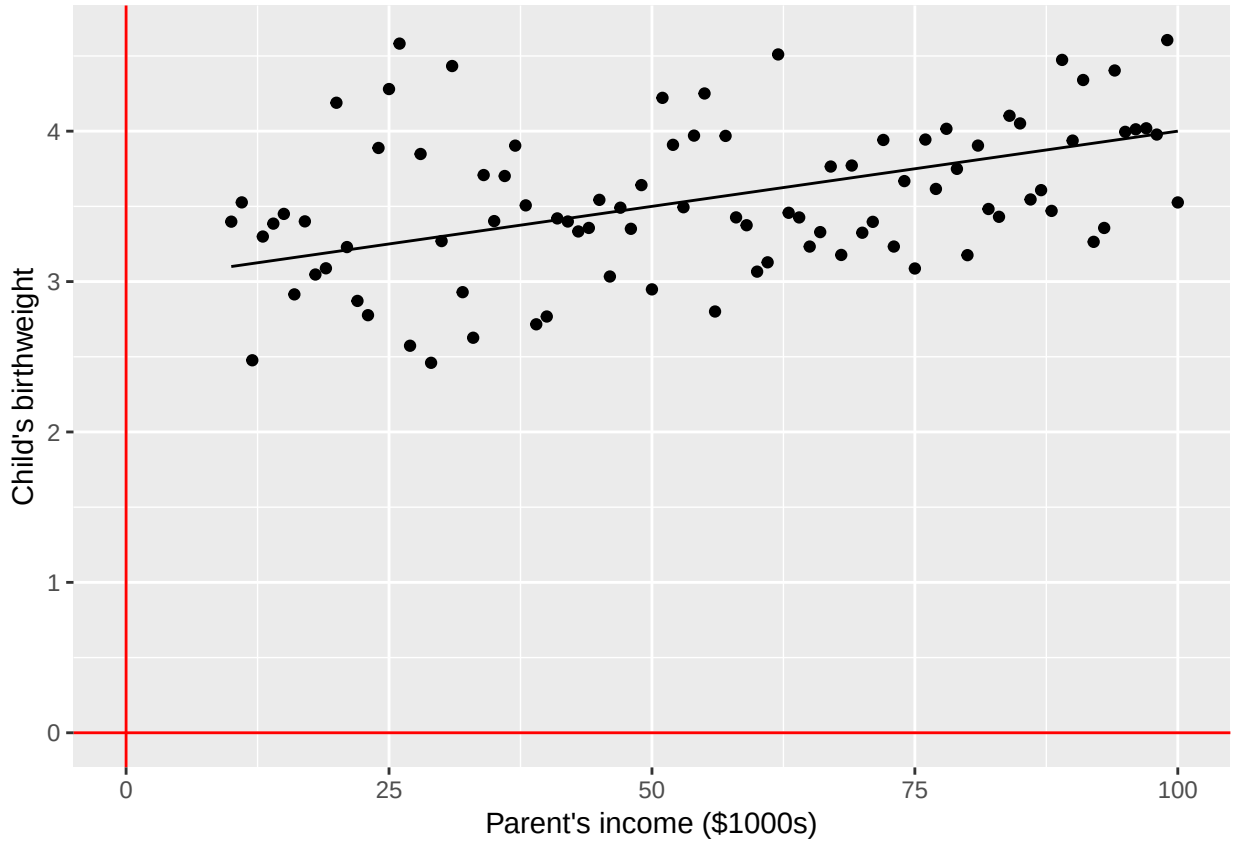
Dependencies: None (all observations independent)

4.2.2 Example 1

→ Testing if average birth weight (continuous outcome) is associated with parents' income (continuous exposure)

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{parent's income} + \text{error}$$

$$\mathrm{H}_0 : \beta_1 = 0$$
$$\mathrm{H}_1 : \beta_1 \neq 0$$

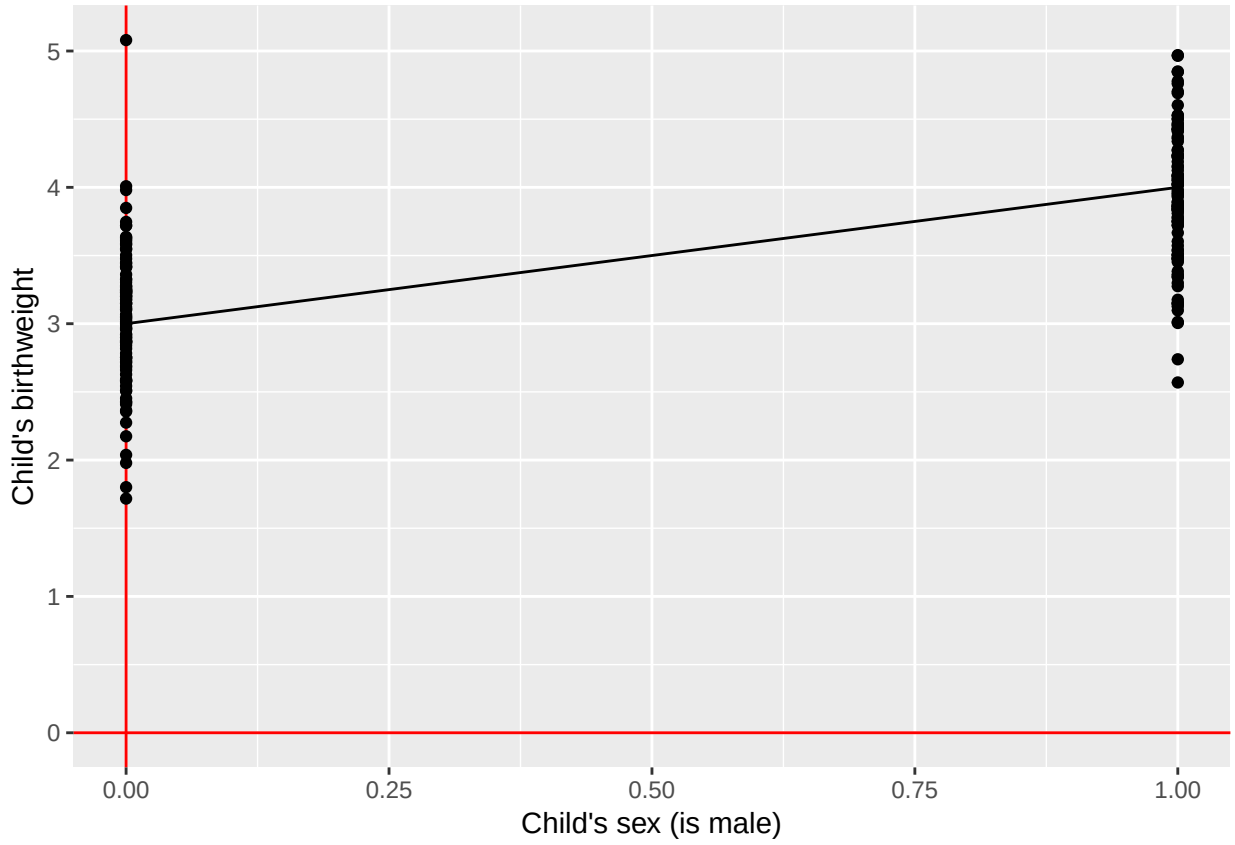


4.2.3 Example 2

→ Testing if average birth weight (continuous outcome) is associated with child's sex (binary exposure)

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is boy} + \text{error}$$

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

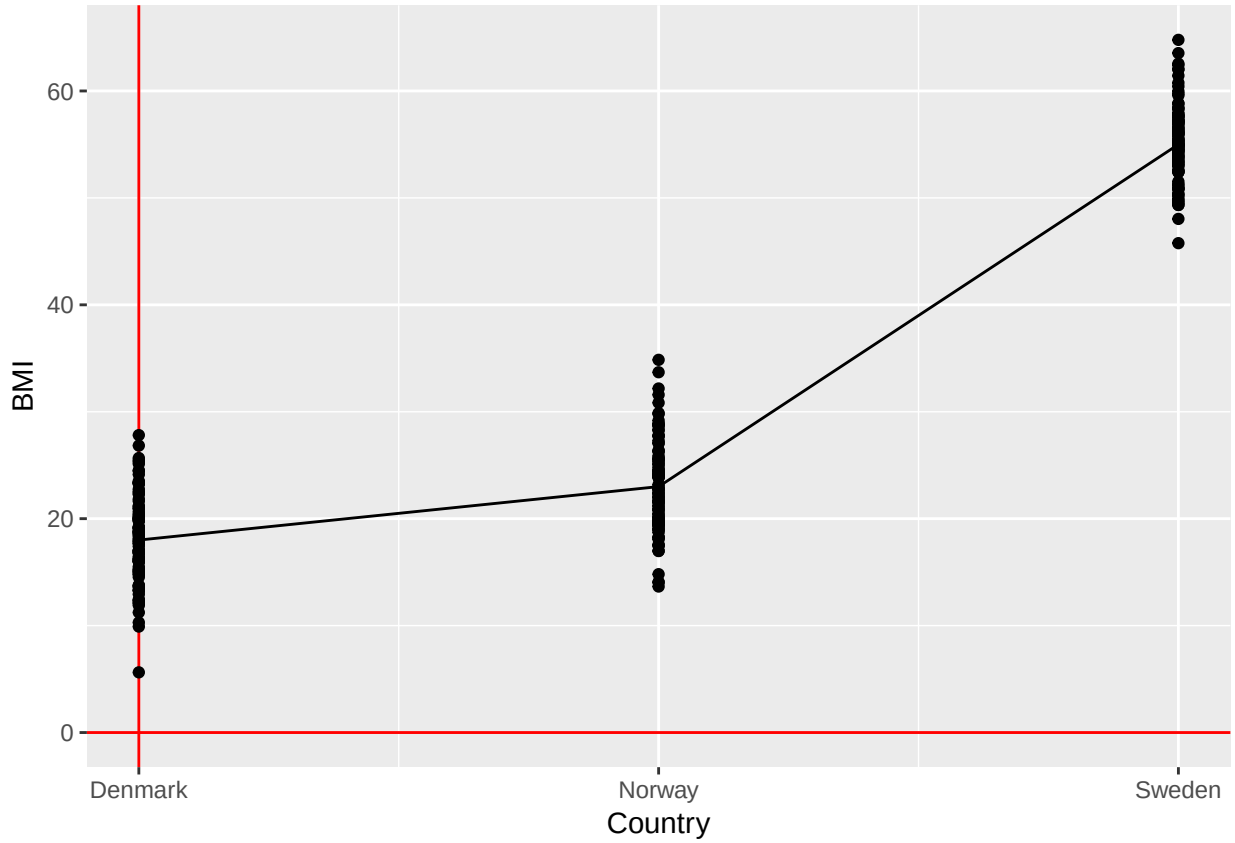


4.2.4 Example 3

→ Testing if average BMI levels (continuous outcome) differ across Scandinavia (categorical exposure)

$$\text{bmi} = \beta_0 + \beta_1 \times \text{is Norway} + \beta_2 \times \text{is Sweden} + \text{error}$$

$$\text{H}_0 : \beta_1 = \beta_2 = 0$$
$$\text{H}_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$



4.2.5 Example 4

→

H_0 :

H_1 :

4.2.6 Example 5

$\rightarrow$

H_0 :

H_1 :

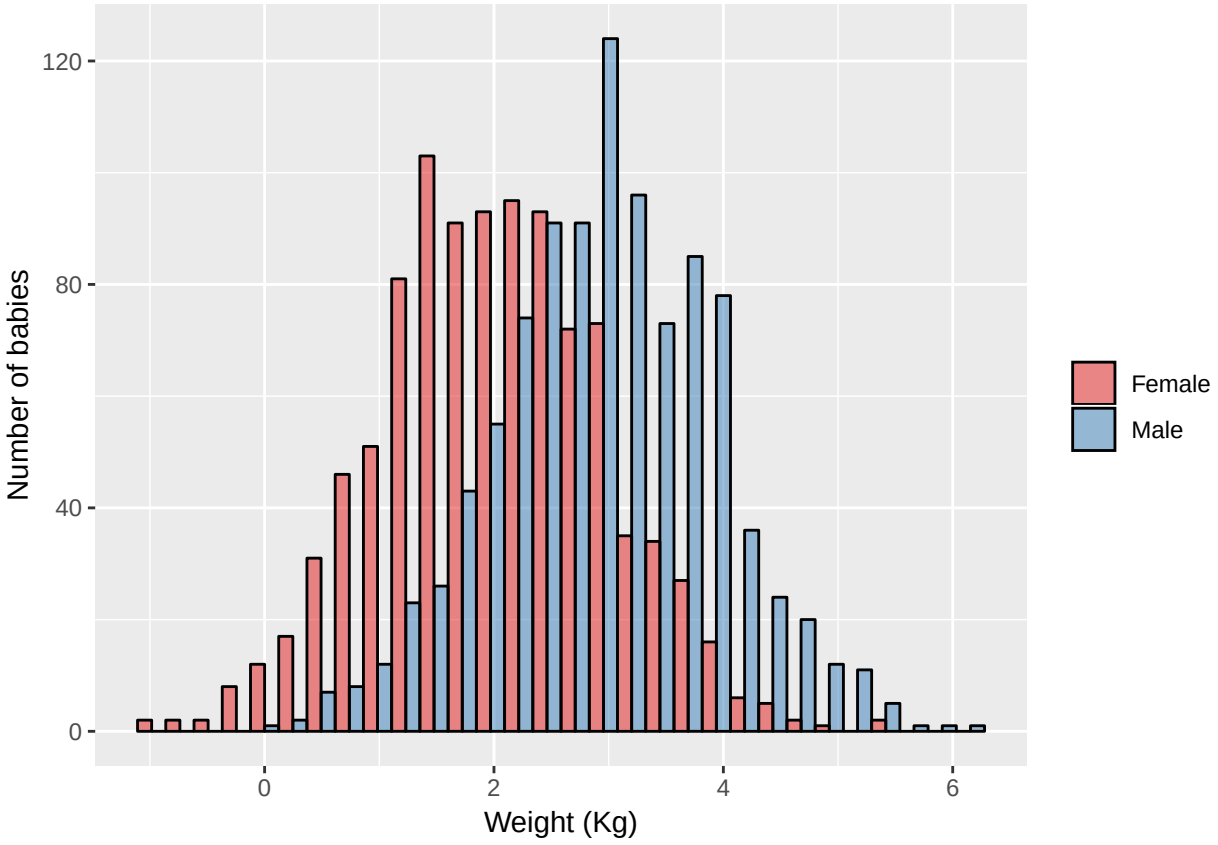
4.3 Similarities between t-tests, ANOVA, and linear regression

4.3.1 Example 1

Two-sample unpaired t-test:

→ Testing if average birth weight (continuous outcome) is different in female children versus male children

$$H_0 : \mu_{\text{boys}} = \mu_{\text{girls}}$$
$$H_1 : \mu_{\text{boys}} \neq \mu_{\text{girls}}$$

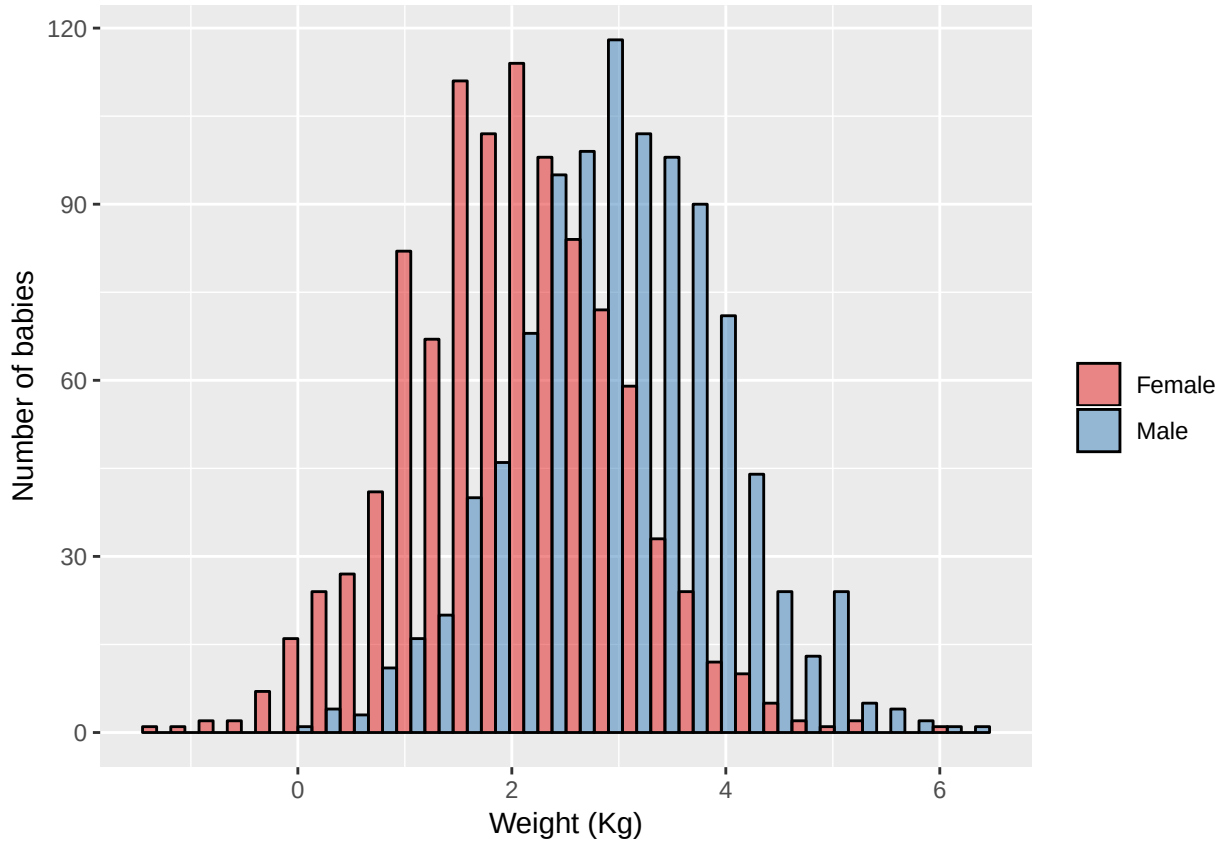


ANOVA:

→ Testing if average birth weight (continuous outcome) is different in female children versus male children

$$\mathrm{H}_0 : \mu_{\text{boys}} = \mu_{\text{girls}}$$

$$\mathrm{H}_1 : \mu_{\text{boys}} \neq \mu_{\text{girls}}$$



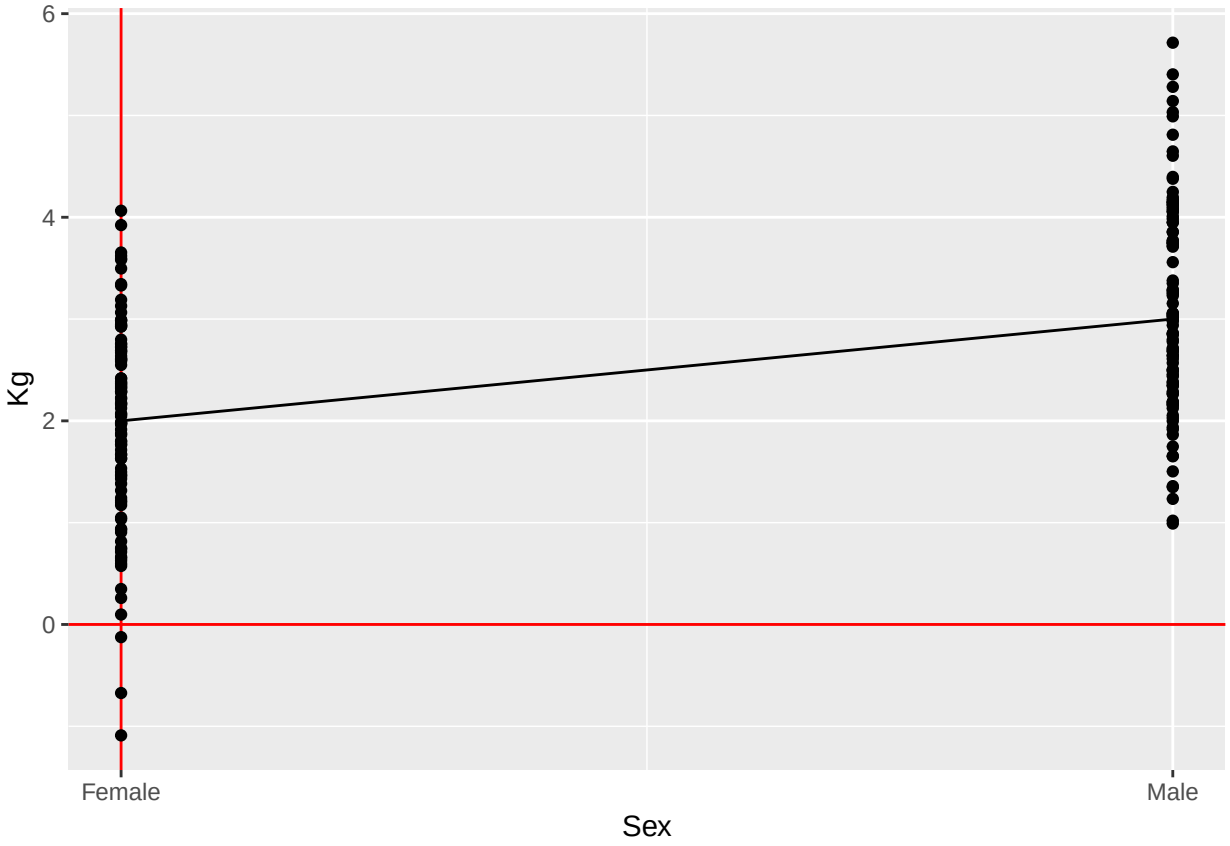
Linear regression:

→ Testing if the effect of child's sex on average birth weight (continuous outcome) is different than zero

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is boy} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$

$$\text{H}_1 : \beta_1 \neq 0$$



Conclusion:

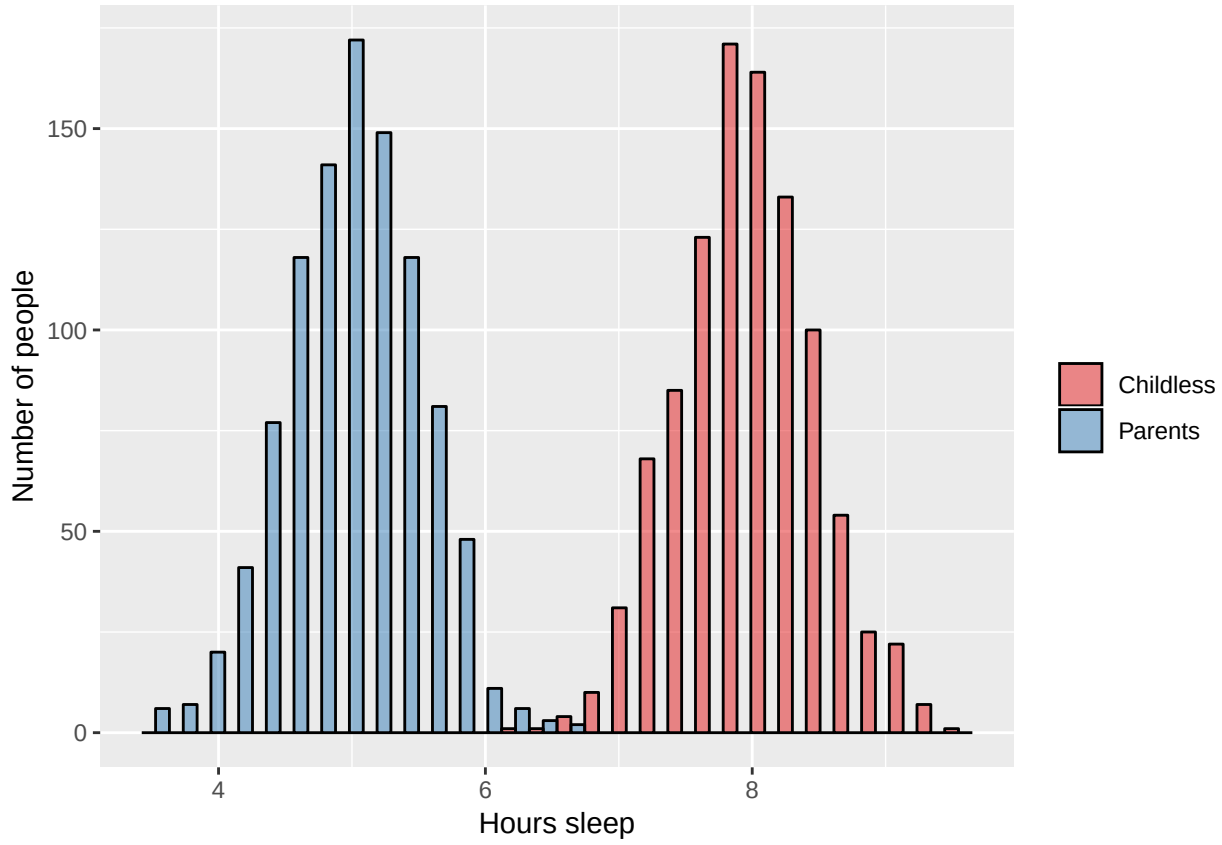
- Two-sample unpaired t-tests are ANOVAs with only two groups
- Two-sample unpaired t-tests are linear regressions with a binary (0/1) exposure
- ANOVA is a linear regression with a categorical exposure

4.3.2 Example 2

Two-sample unpaired t-test:

→ Testing if average number of hours sleep (continuous outcome) is different in adults who are parents versus those who are childless

$$H_0 : \mu_{\text{parents}} = \mu_{\text{childless}}$$
$$H_1 : \mu_{\text{parents}} \neq \mu_{\text{childless}}$$

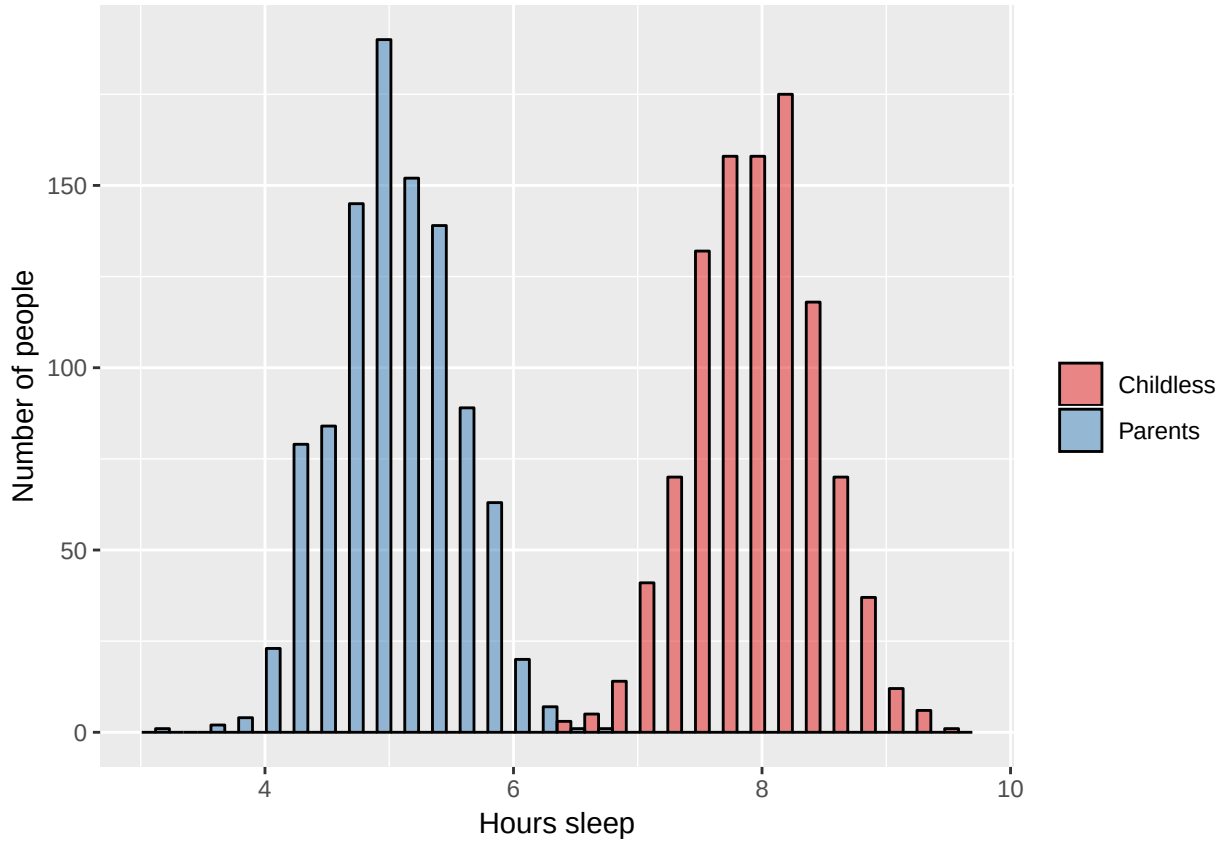


ANOVA:

→ Testing if average number of hours sleep (continuous outcome) is different in adults who are parents versus those who are childless

$$H_0 : \mu_{\text{parents}} = \mu_{\text{childless}}$$

$$H_1 : \mu_{\text{parents}} \neq \mu_{\text{childless}}$$



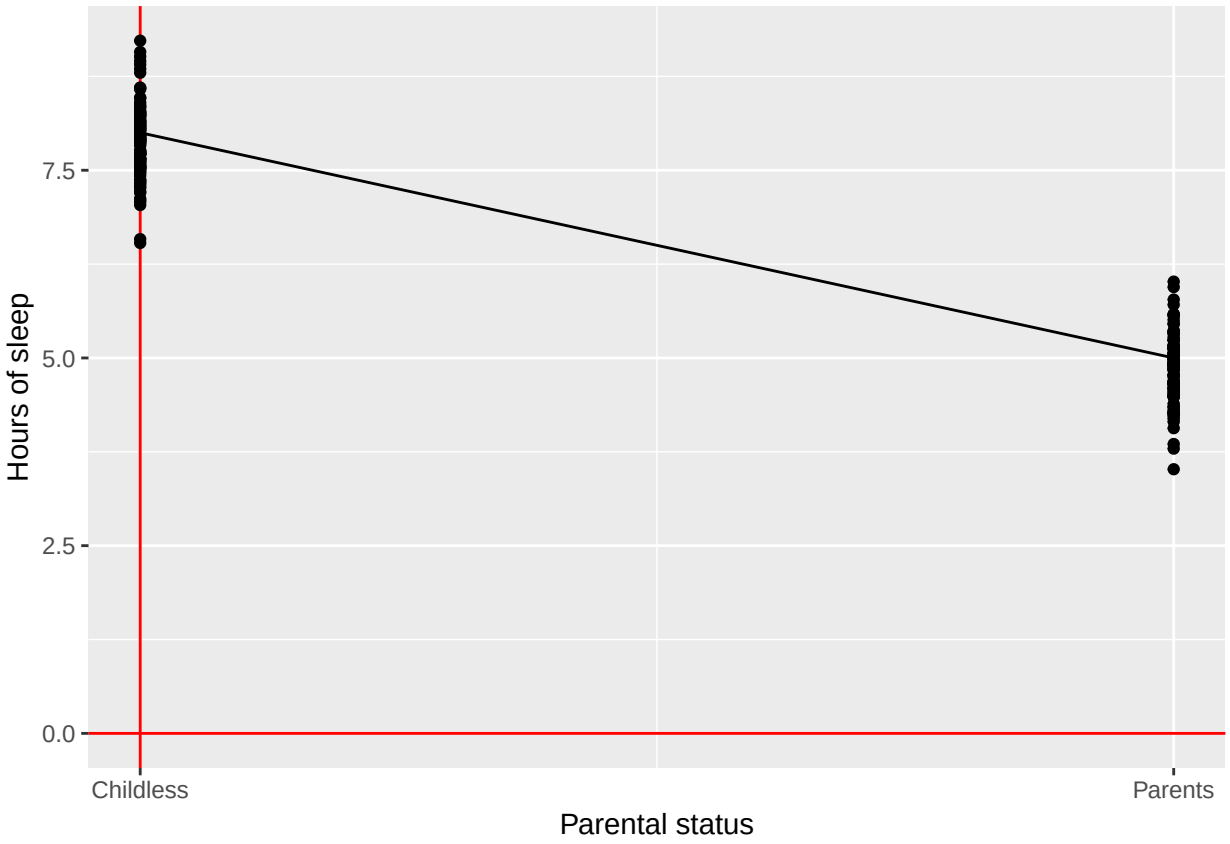
Linear regression:

→ Testing if the effect of being a parent on average number of hours sleep (continuous outcome) is different than zero

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is parent} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$

$$\text{H}_1 : \beta_1 \neq 0$$



Conclusion:

- Two-sample unpaired t-tests are ANOVAs with only two groups
- Two-sample unpaired t-tests are linear regressions with a binary (0/1) exposure
- ANOVA is a linear regression with a categorical exposure

4.3.3 Example 3

Two-sample unpaired t-test:

$\rightarrow$

H_0 :

H_1 :

ANOVA:

$\rightarrow$

H_0 :

H_1 :

Linear regression:

$\rightarrow$

H_0 :

H_1 :

4.4 Similarities between ANOVA and linear regression

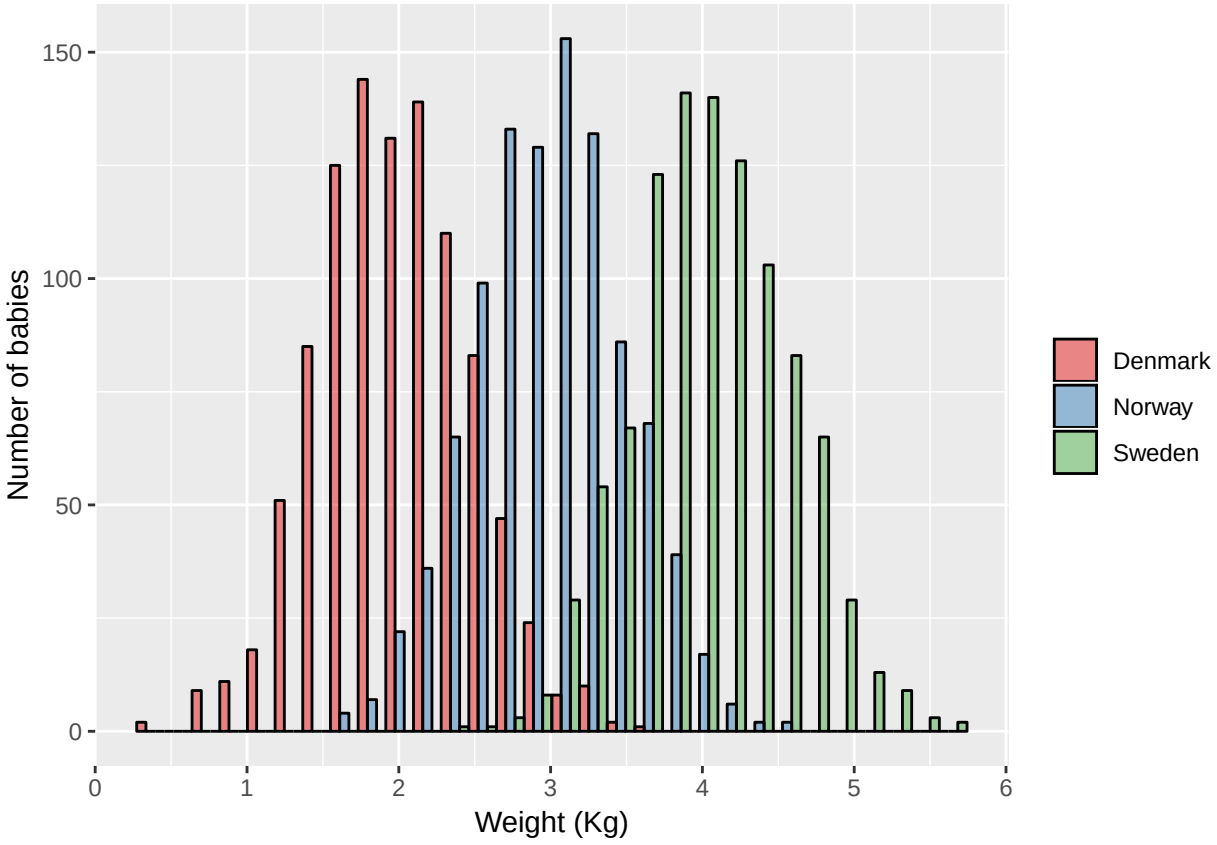
4.4.1 Example 1

ANOVA:

→ Testing if average birth weight (continuous outcome) differs between Scandinavian countries

$$H_0 : \mu_{\text{Norway}} = \mu_{\text{Denmark}} = \mu_{\text{Sweden}}$$

$$H_1 : \mu_{\text{Norway}} \neq \mu_{\text{Denmark}} \text{ and/or } \mu_{\text{Norway}} \neq \mu_{\text{Sweden}} \text{ and/or } \mu_{\text{Denmark}} \neq \mu_{\text{Sweden}}$$



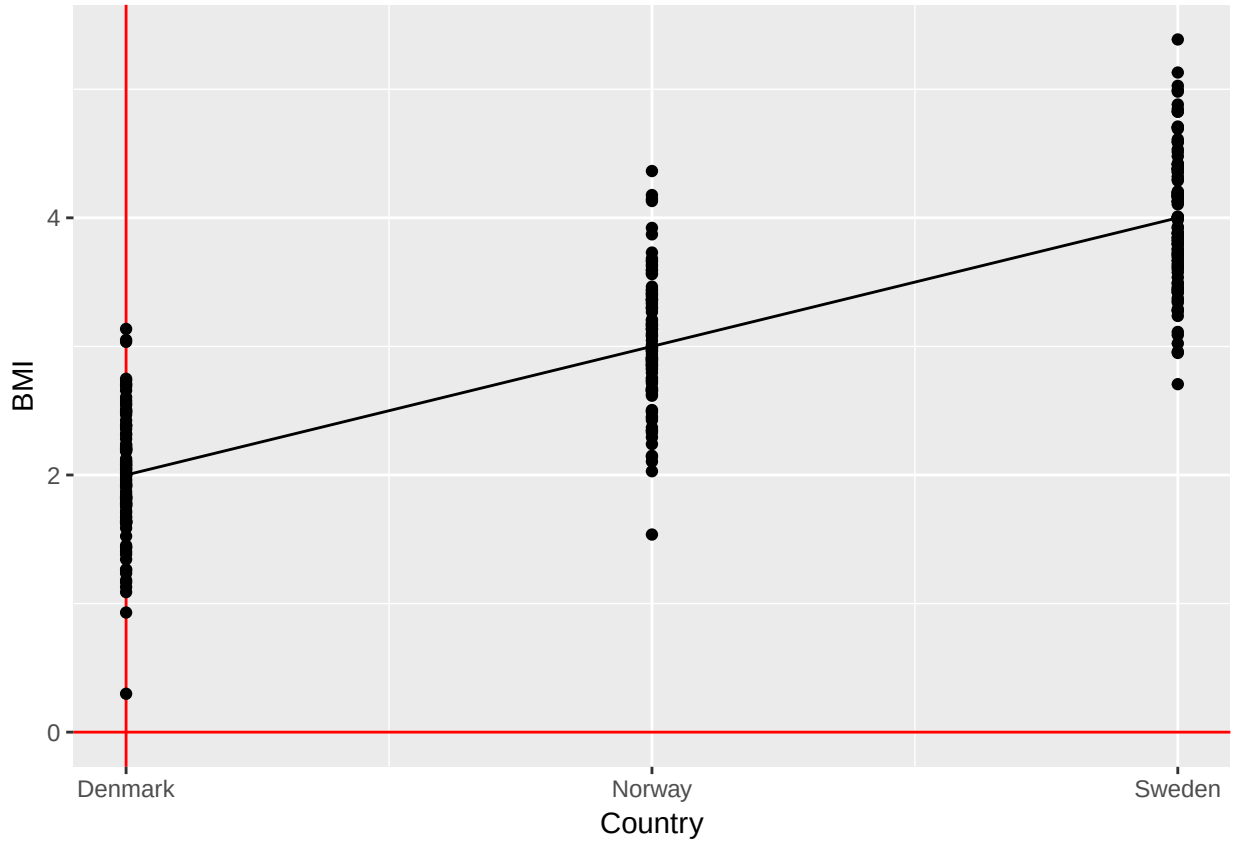
Linear regression:

→ Testing if the effect of country on average birth weight (continuous outcome) is different than zero

$$\text{birth weight} = \beta_0 + \beta_1 \times \text{is Norway} + \beta_2 \times \text{is Denmark} + \text{error}$$

$$\text{H}_0 : \beta_1 = \beta_2 = 0$$

$$\text{H}_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$



Conclusion:

- ANOVA is a linear regression with a categorical exposure

4.4.2 Example 2

ANOVA:

$\rightarrow$

H_0 :

H_1 :

Linear regression:

$\rightarrow$

H_0 :

H_1 :

4.5 Logistic regression models

Logistic regression is essentially the same as linear regression, but it is used when:

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

4.5.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Binary variable*

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions: No

Dependencies: None (all observations independent)

4.5.2 Example 1

→ Testing if percentage of women (binary outcome) differ across the bydels of Oslo (categorical exposure)

$$\log\left(\frac{\Pr(\text{Is woman})}{\Pr(\text{Is man})}\right) = \beta_0 + \beta_1 \times \text{bydel}_1 + \beta_2 \times \text{bydel}_2 + \beta_3 \times \text{bydel}_3 + \text{error}$$

$$\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$\text{H}_1 : \beta_1 \neq 0 \text{and/or} \beta_2 \neq 0 \text{and/or} \beta_3 \neq 0$$

4.5.3 Example 2

→ Testing if risk of unemployment (binary outcome) is associated with parents' income (continuous exposure)

$$\log\left(\frac{\Pr(\text{Is unemployed})}{\Pr(\text{Is employed})}\right) = \beta_0 + \beta_1 \times \text{parent's income} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$
$$\text{H}_1 : \beta_1 \neq 0$$

4.5.4 Example 3

→ Testing if risk of smoking (binary outcome) is associated with parents' smoking status (binary exposure)

$$\log\left(\frac{\Pr(\text{Is smoker})}{\Pr(\text{Is not smoker})}\right) = \beta_0 + \beta_1 \times \text{parent's are smokers} + \text{error}$$

$$\text{H}_0 : \beta_1 = 0$$
$$\text{H}_1 : \beta_1 \neq 0$$

4.5.5 Example 4

$\rightarrow$

H_0 :

H_1 :

4.5.6 Example 5

$\rightarrow$

H_0 :

H_1 :

4.6 Poisson/negative-binomial regression models

Poisson/negative-binomial regression is essentially the same as linear regression, but it is used when:

- You have a count outcome

Negative-binomial regression is a more flexible version of poisson regression. Poisson regression requires that the residual variation (after fitting the model) is equal to the expected mean. This is quite often not the case. Negative-binomial regression fits the variation and the mean separately, removing this problem. It is therefore recommended that you always use a negative-binomial regression instead of a poisson regression. The only exception is if you encounter statistical errors with the negative-binomial regression (i.e. it won't converge/run), then a poisson regression is your only option.

4.6.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Count variable*

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions for Poisson: Mean equals variable

Parametric assumptions for negative-binomial: No

Dependencies: None (all observations independent)

4.6.2 Example 1

→ Testing if average number of influenza cases (count outcome) is different between 2000-2009 and 2010-2015 (binary exposure) in Norway

$$\text{yearly number of influenza cases} = \beta_0 + \beta_1 \times \text{is 2010 to 2015} + \text{error}$$

$$\mathrm{H}_0 : \beta_1 = 0$$
$$\mathrm{H}_1 : \beta_1 \neq 0$$

4.6.3 Example 2

→

H_0 :

H_1 :

4.7 Cox regression models

Cox regression models should be used when your outcome is "time-to-event".

The most common example of this is when you are following a cohort of people over time, trying to observe an (e.g. sickness, death, response). Your outcome is "length of time until person X gets disease Y". However, a number of your participants stop responding at some point, so you only know "person X was healthy up until 200 days, when we lost contact".Thus person X's outcome has been censored at day 200.

4.7.1 Aim/Outcome/Exposure/Parametric/Dependencies

Aim: Hypothesis testing and estimating the effect size of the association between outcome and exposure

Outcome: *Censored variable* (time-to-event)

Exposure: Continuous, Binary, Categorical, Count variable

Parametric assumptions: Proportional hazards

Dependencies: None (all observations independent)

4.7.2 Example 1

→ Testing if time-to-death (outcome) is associated with having a hospital-acquired-infection after hip surgery (binary exposure)

$$\lambda(t|X_i) = \lambda_0(t) \times \exp(\beta_1 \times \text{had HAI})$$

Where $\lambda(t|X_i)$ is the hazard rate of dying at time t for subject i.

$$\mathrm{H}_0 : \beta_1 = 0$$
$$\mathrm{H}_1 : \beta_1 \neq 0$$

4.7.3 Example 2

→

H_0 :

H_1 :

5 Complicated regression

5.1 Dependencies in your data

5.1.1 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.

5.1.2 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.

5.1.3 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.

5.1.4 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.

5.1.5 Grouped/clustered data

Repeated measures data is a type of clustered data, where each person is their own cluster. Matched data is also a type of clustered data, where each group of matched controls-to-a-case is a cluster.

There can be many other 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

5.2 Analysing data with dependencies

5.2.1 Mixed effects regression

Mixed effects regression is an extension of the simple regression models (fixed effects) that we learned about in the previous chapter. You can have:

- Mixed effects linear regression
- Mixed effects logistic regression
- Mixed effects negative-binomial regression

Mixed effects models are models that have both fixed and random effects. "Effects" is the term used to describe the estimated impact a variable has on the outcome. For example, "the effect of smoking on the risk of lung cancer".

To determine if an effect is fixed or random, we introduce the concept of pooling data (i.e. sharing strength). Let us assume we have sampled 100 people per city for 9 cities, and 2 people for 1 city, and we want to estimate the average height in each city, we can do one of the following three options:

1. Within each city, take the average height of the 100 (or 2) sampled people (zero pooling)
2. Take the average height of 1000 sampled people and say each city is the same height (complete pooling)
3. Estimate how much the average height varies between the cities. If the average height doesn't vary much, give estimates close to #2. If the average height varies a lot, give estimates close to #1 (partial pooling)

Zero pooling tends to work well when you have a limited amount of effects you want to estimate, each with a good sample size (i.e. you don't need to borrow strength from other data points). Partial pooling tends to work well when you have a large number of effects you want to estimate, each with a small sample size (i.e. you need to borrow strength from other data points).

Fixed effects are estimated with zero pooling. Random effects are estimated with partial pooling.

With data that has a large number of clusters, the clustering need to be accounted for in the regression model. This is done by introducing a variable that uniquely identifies each cluster, and allows the cluster effect to be estimated (e.g. in Oslo people are 1% more likely to die

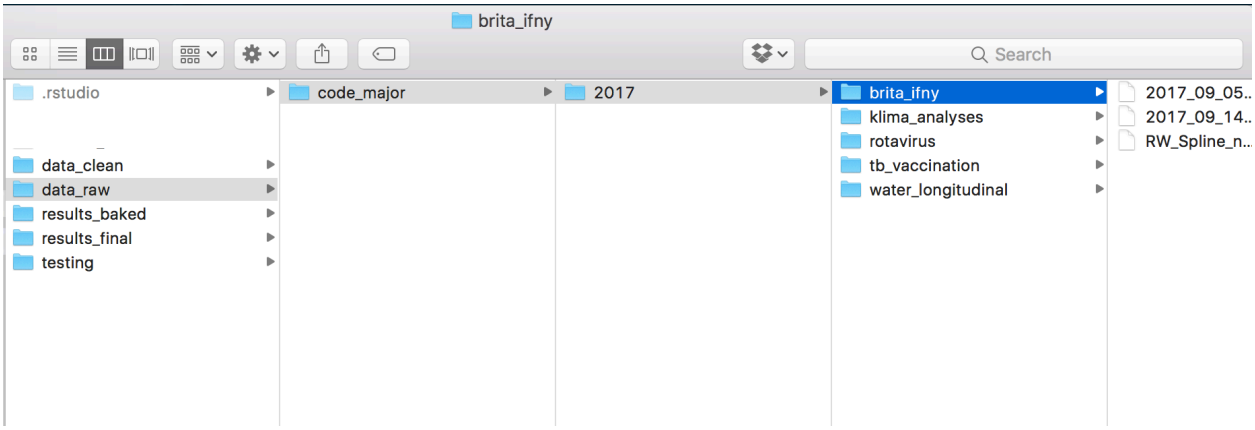


Figure 1: img

than in the rest of Norway). If there are a large number of clusters with a small amount of people in each cluster, then the cluster variable should be estimated using partial pooling (i.e. random effects).

In summary: Mixed effects regression should be used for grouped/clustered/matched data (and subsequently repeated measures/longitudinal data).

5.2.2 Conditional logistic regression

Conditional logistic regression can also be used for matched data with a binary outcome, however, it is less flexible than mixed effects regression.

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

6 Good folder structure

6.1 Data and results

- One folder for raw data
- One folder for temporary data (if needed)
- One folder for clean data

- One folder for shared results (dropbox)
- Every time you run your analyses, you should store your new results in a new day's folder, allowing you to always access your previous results
- Labelling by date (year-month-day) is a lot more intuitive than "results_1", "results_2", "results_final", "results_final_final"

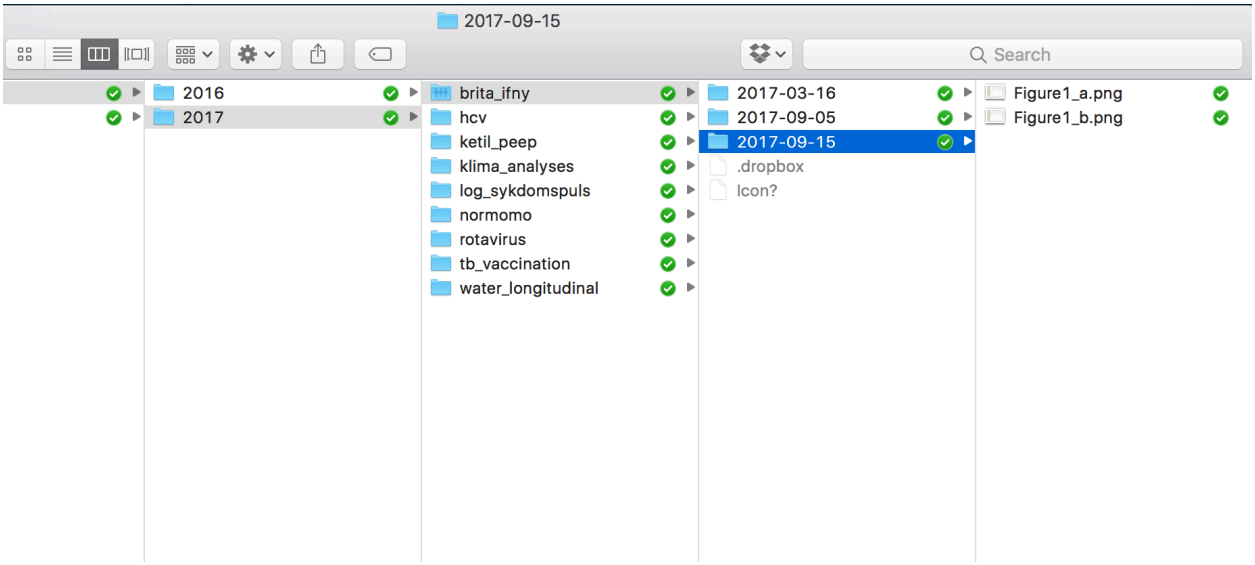


Figure 2: img

- Label in the format 2017-09-01 so that your computer can easily sort the results (the padding/leading 0s are important for sorting!)
- Do not label folders using the format 2017-SEP-1 (it doesn't sort well)
- Do not label folders using the format 1-SEP-17 (it doesn't sort well)

6.2 Keep your source code separate from data

Source code is:

- Not sensitive
- Very easy to accidentally delete and/or overwrite

This means it is perfect to be uploaded to http://github.com where you can store every version of your code, and never have to worry about losing an old copy

You can also see the differences between your versions:

It is best to keep each project in it's own folder. Within each folder, there should be a masterfile ("Run") that will perform all of the necessary tasks in the entire analysis:

- Clean data
- Run analyses

Do files/scripts should be numbered:

- 0_run_all.do
- 1_clean_lab_data.do
- 2_clean_lifestyle_data.do
- 3_merge_lab_lifestyle_data.do

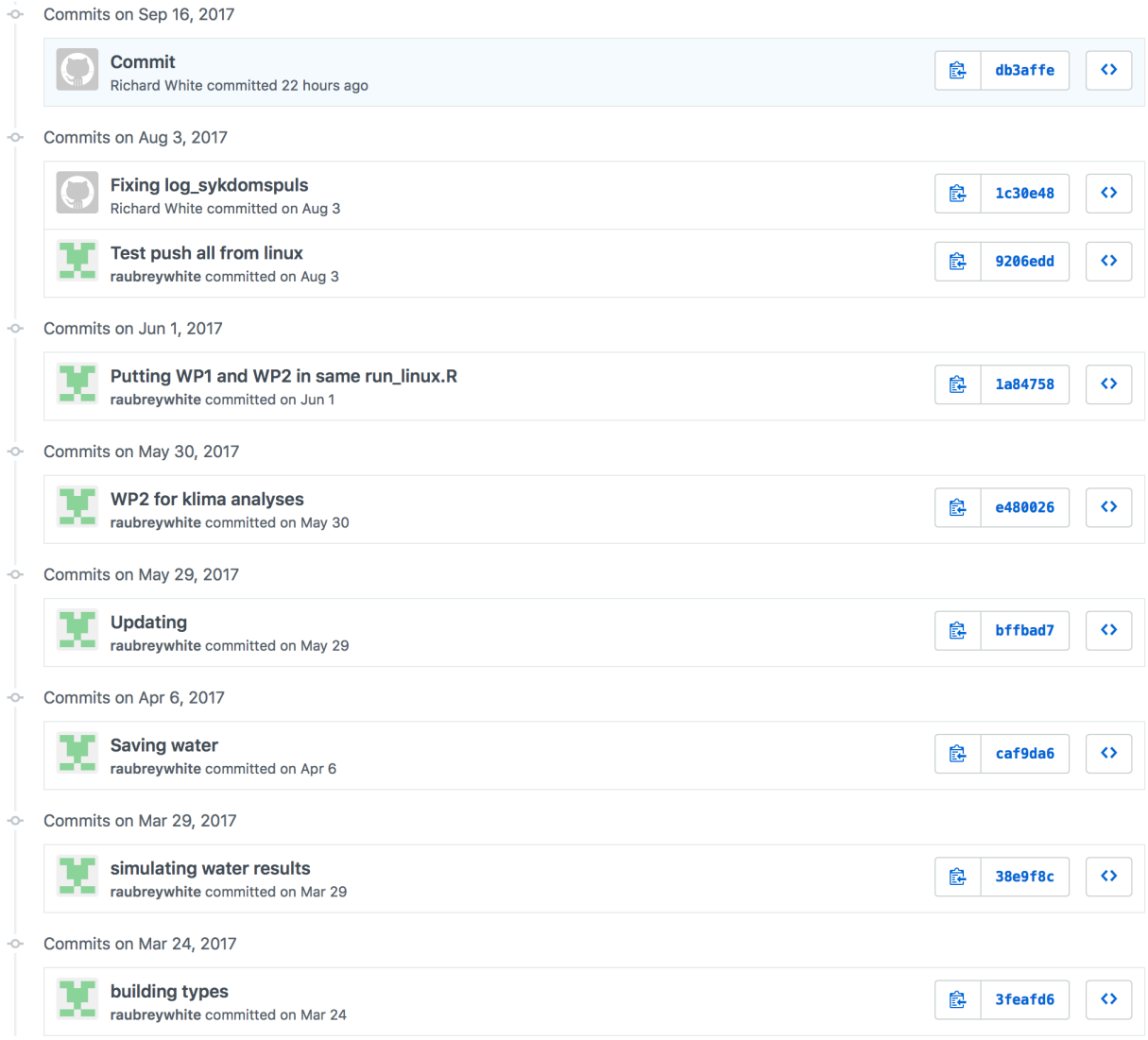

Commits on Sep 16, 2017
Commit
Richard White committed 22 hours ago
db3affe
Commits on Aug 3, 2017
Fixing log_sykdomspuls
Richard White committed on Aug 3
1c30e48
Test push all from linux
raubreywhite committed on Aug 3
9206edd
Commits on Jun 1, 2017
Putting WP1 and WP2 in same run_linux.R
raubreywhite committed on Jun 1
1a84758
Commits on May 30, 2017
WP2 for klima analyses
raubreywhite committed on May 30
e480026
Commits on May 29, 2017
Updating
raubreywhite committed on May 29
bffbad7
Commits on Apr 6, 2017
Saving water
raubreywhite committed on Apr 6
caf9da6
Commits on Mar 29, 2017
simulating water results
raubreywhite committed on Mar 29
38e9f8c
Commits on Mar 24, 2017
building types
raubreywhite committed on Mar 24
3feafd6

Figure 3: img

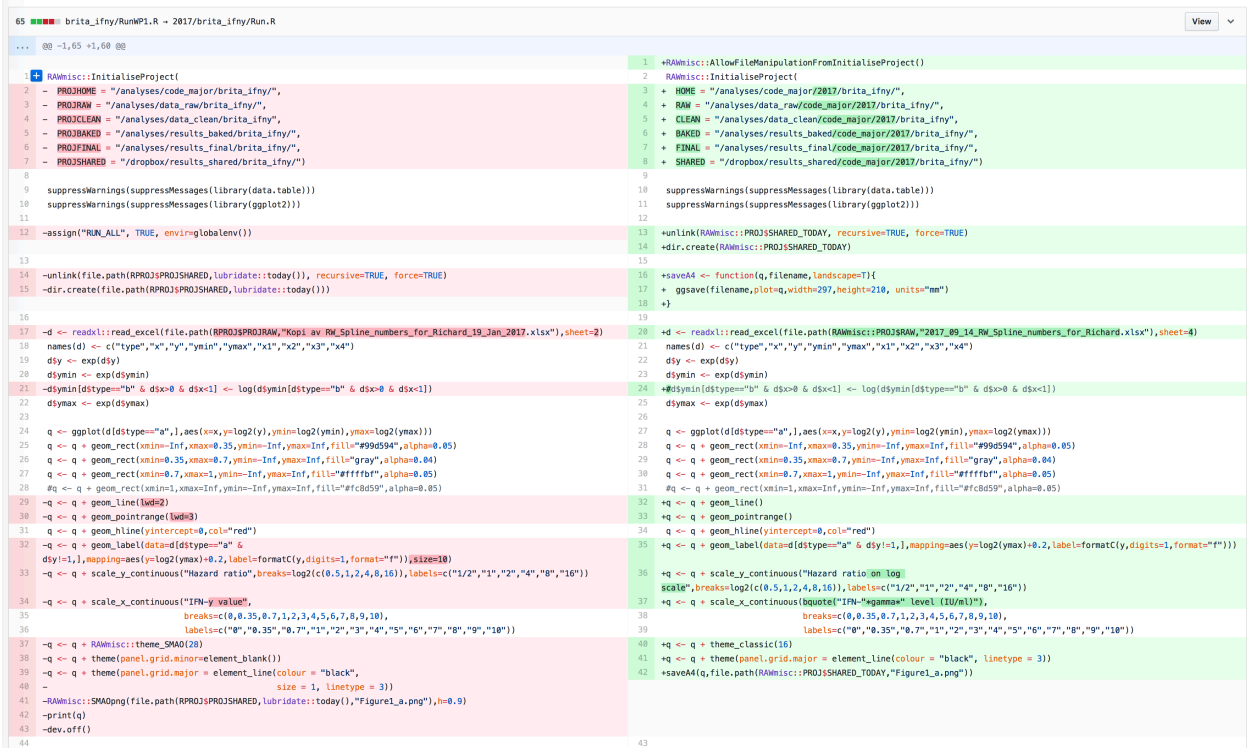


Figure 4: img

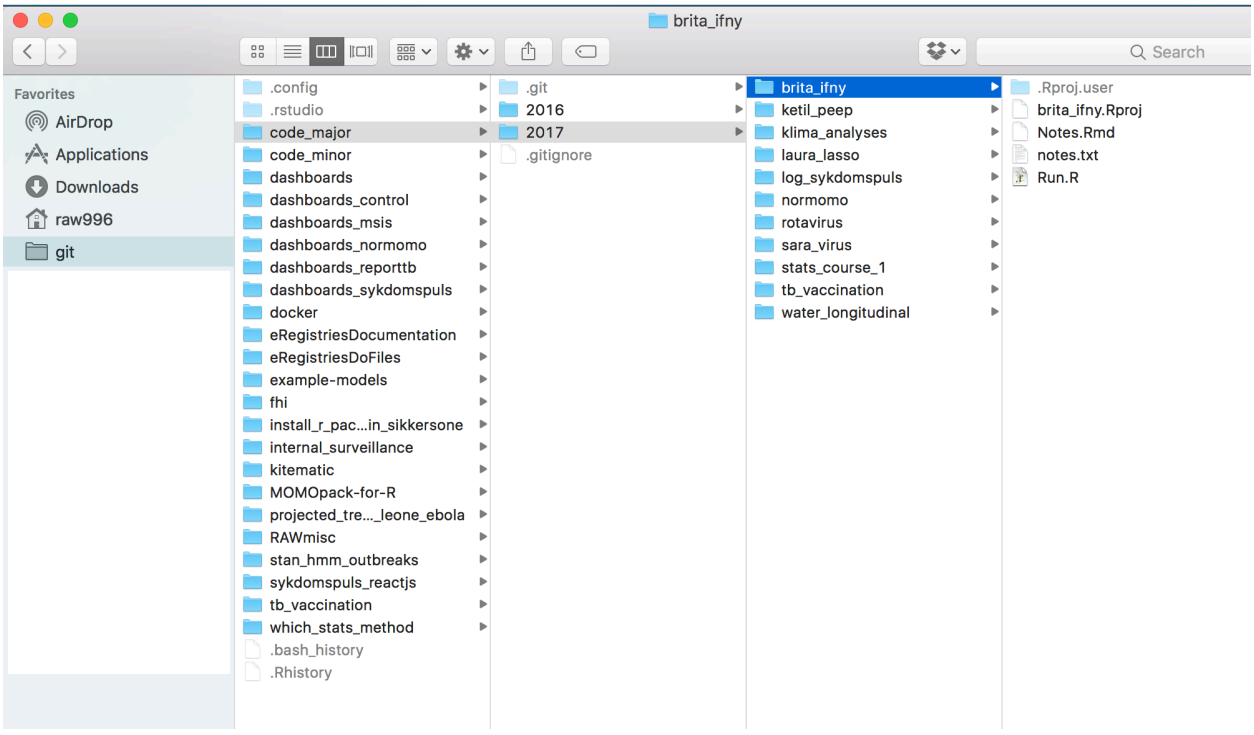


Figure 5: img

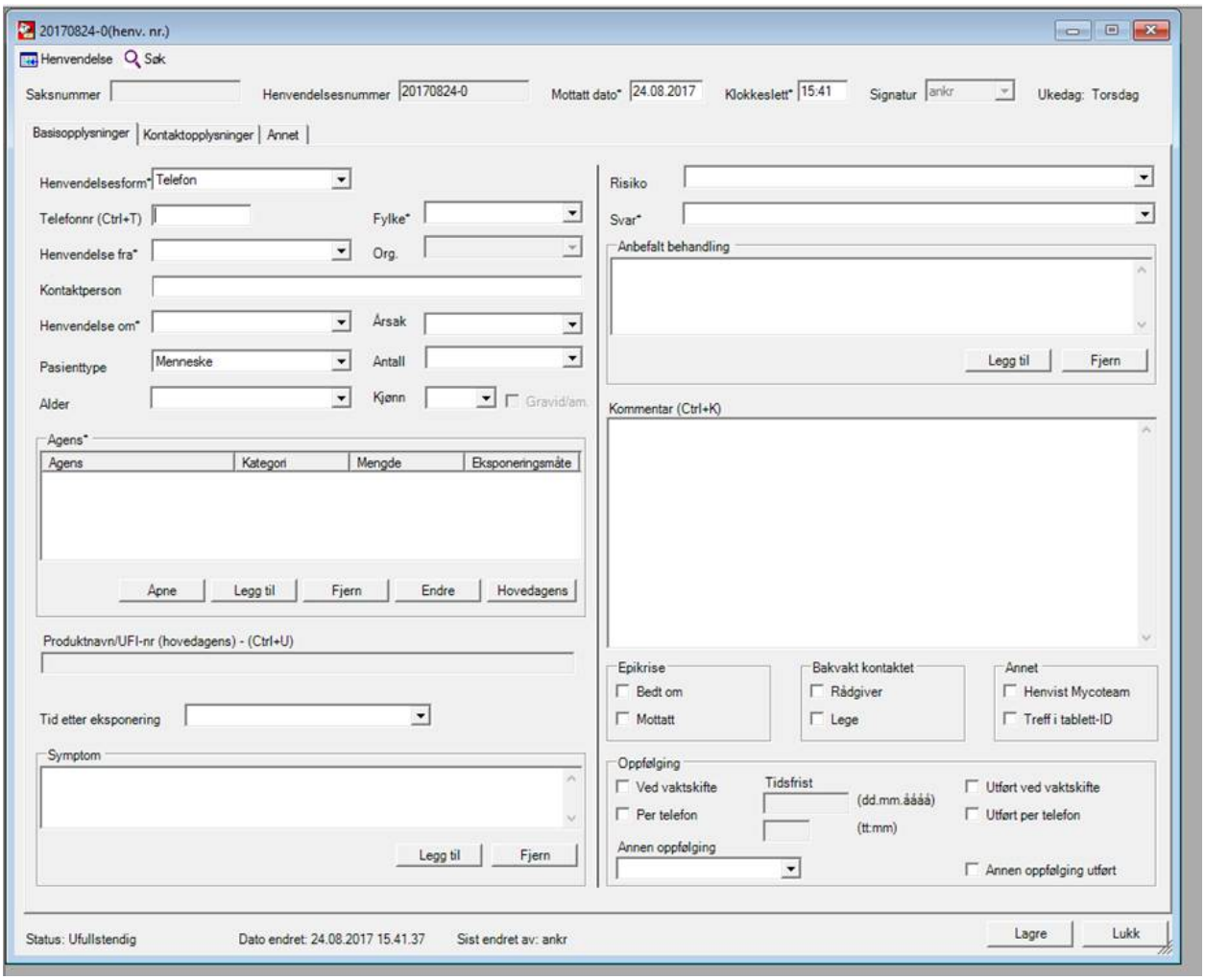


Figure 6: img

- 4_descriptive_analyses.do
- 5_regression_analyses.do

It is very very very important that analysis files only use the "clean" data (from the clean data folder), and perform ZERO changes to the data. All necessary changes and/or variable creations must be done in the data cleaning do files.

7 Examples

7.1 Poisons Information Center

Scenario:

- Approximately 40 000 calls per year

- The frequency of calls regarding different drugs/plants etc varies greatly, from a couple per year to thousands per year, so most probably one method will not cover everything.
- The registration of data is done while on the phone, and we know there are mistakes
- Only the field Kommentar has free text

Question: Is the number of calls regarding women aged 15-19 who have been exposed to paracetamol rising or falling over the years.

Aim:

Outcome:

Exposure:

Parametric assumptions:

Dependencies:

7.2 Norwegian Water Pipes

Scenario:

- 146 waterworks in 19 counties in Norway
- Each waterwork uses pipes to deliver water to households
- Information on each waterwork:
- Length of pipes made out of asbestos (in meters)
- Length of pipes made out of iron/steel (in meters)
- Length of pipes made out of PVC (in meters)
- Length of pipes made out of PE/PEH (in meters)
- Length of pipes made out of other (in meters)
- Length of pipes made out of unknown (in meters)
- Length of pipes installed before 1910 (in meters)
- Length of pipes installed in 1910-1940 (in meters)
- Length of pipes installed in 1941-1970 (in meters)
- Length of pipes installed in 1971-2000 (in meters)
- Length of pipes installed after 2000 (in meters)
- Length of pipes installed during an unknown period (in meters)
- Each year, some new pipes are laid to extend the network
- Each year, some pipes are replaced
- Interruption in water delivery is estimated in hours per calendar year
- Data is only for 2015

Question: Is there an association between “interuption in water delivery” and “type of pipe material” and “pipe installation period”

Aim:

Outcome:

Exposure:

Parametric assumptions:

Dependencies:

7.3 Early warning system (EWS) for waterborne outbreaks (part 1)

Scenario:

- NorSySS is a syndromic surveillance system for infectious diseases, run by the NIPH.
- The system is based on national ICPC coded consultation data from general practice in Norway.
- We want to study to what extent NorSySS can serve as an early warning system for local waterborne outbreaks by alerting us to increases in consultation rates for syndromes indicative of gastrointestinal diseases.

- Retrospective syndrome data (number of gastritis cases, per week, for each municipality) from NorSySS will be aligned with outbreak data (outbreak=yes/no) from the national web-based outbreak rapid alert system (Vesuv) for the period 2006-2017.

Question: Is there an association between "recorded outbreak" and "number of gastritis cases"

Aim:

Outcome:

Exposure:

Parametric assumptions:

Dependencies:

7.4 Early warning system (EWS) for waterborne outbreaks (part 2)

Scenario:

- We have weekly data on water quality from water works (e.g. pH, turbidity)
- We have weekly number of gastritis cases, per week, for each municipality
- We hope to increase the knowledge about causes of waterborne outbreaks and to develop an improved surveillance system for early detection of future outbreaks.

Question: Is there an association between "weekly number of gastritis cases" and "water quality from the water works"

Aim:

Outcome:

Exposure:

Parametric assumptions:

Dependencies:

7.5 Incidents in the water supply system and illness

Scenario:

- Data from the water works operation (pH, turbidity) will be linked to health outcome among recipients of the drinking water.
- The study will be a prospective cohort study, with data collected among a random selection of water works.
- Data from the recruited water works will be collected in the period autumn 2017 and the 12 following months.
- Approximately 350 water works will be recruited to provide monthly data on hygienic critical points related to operation and maintenance of the water supply system.
- In parallel, a cohort of approximately 9000 persons, served by water from the recruited water works, will submit monthly reports on symptoms that may indicate gastrointestinal illness.
- The data collection will start in the autumn of 2017 and continue for 12 months.

Question: Is water quality a risk factor for getting sick?

Aim:

Outcome:

Exposure:

Parametric assumptions:

Dependencies:

7.6 Compliance with boil water advisories and perception of risks

Scenario:

- In this study, the compliance and perception of risks among the public with boil water advisories (BWAs) will be examined.
- Although BWAs is a common practice among water utilities, a meta-study suggest that there is limited information and studies on the compliance of BWAs.

- This part, the compliance and perception of risks will be done by studying the perception of and adherence to BWAs among the consumers of drinking water in Bærum municipality.
- Even though the drinking water in Bærum is considered to have good quality, Bærum – like many water works – experience situations of pressure drops due to breaks and maintenance.
- Research has shown that these situations may lead to an increased risk of gastrointestinal infections, and due to this the municipality of Bærum has issued a precautionary BWAs to the affected consumers with every water outage during the last 5-6 years.
- Every year, some 12,000-22,000 consumers have received a precautionary BWA.
- The purpose of these precautionary BWAs is to prevent health consequences caused by possible contamination of water. However, we know little about the consumer's knowledge about why they receive these BWAs, as well as the way the consumers perceive and adhere to these BWAs.
- This is a cross-sectional study.
- A web-survey will be presented to a randomly selected sample of consumers who received a BWA i Bærum in 2017.
- The web-survey asks about adherence (yes/no) and demographics (e.g. age, sex, income)

Question: Estimate adherence by demographics (and identify if it differs by demographics)

Aim:

Outcome:

Exposure:

Parametric assumptions:

Dependencies:

8 Solutions

8.1 Poisons Information Center

Scenario:

- Approximately 40 000 calls per year
- The frequency of calls regarding different drugs/plants etc varies greatly, from a couple per year to thousands per year, so most probably one method will not cover everything.
- The registration of data is done while on the phone, and we know there are mistakes
- Only the field Kommentar has free text

Question: Is the number of calls regarding women aged 15-19 who have been exposed to paracetamol rising or falling over the years.

Aim: Hypothesis testing (maybe estimation of yearly effect)

Outcome: Count data (number of calls regarding women aged 15-19 who have been exposed to paracetamol)

Exposure: Continuous (year)

Parametric assumptions: None

Dependencies: None

Appropriate Method: Negative-binomial regression

Example STATA code:

```
nbreg number_of_calls year
```

8.2 Norwegian Water Pipes

Scenario:

- 146 waterworks in 19 counties in Norway
- Each waterwork uses pipes to deliver water to households
- Information on each waterwork:
- Length of pipes made out of asbestos (in meters)
- Length of pipes made out of iron/steel (in meters)
- Length of pipes made out of PVC (in meters)
- Length of pipes made out of PE/PEH (in meters)
- Length of pipes made out of other (in meters)
- Length of pipes made out of unknown (in meters)
- Length of pipes installed before 1910 (in meters)
- Length of pipes installed in 1910-1940 (in meters)
- Length of pipes installed in 1941-1970 (in meters)
- Length of pipes installed in 1971-2000 (in meters)
- Length of pipes installed after 2000 (in meters)
- Length of pipes installed during an unknown period (in meters)
- Each year, some new pipes are laid to extend the network
- Each year, some pipes are replaced
- Interruption in water delivery is estimated in hours per calendar year
- Data is only for 2015

Question: Is there an association between "interuption in water delivery" and "type of pipe material" and "pipe installation period"

Aim: Hypothesis testing

Outcome: Count (hours interruption in water delivery)

Exposure: 12 continuous variables

Parametric assumptions: None

Dependencies: No

Appropriate Method: Negative-binomial regression

Example STATA code:

```
nbreg hoursInterrupted pipesAsbestos pipesIronSteel pipesPVC ... pipes1910 ...
```

8.3 Early warning system (EWS) for waterborne outbreaks (part 1)

Scenario:

- NorSySS is a syndromic surveillance system for infectious diseases, run by the NIPH.
- The system is based on national ICPC coded consultation data from general practice in Norway.
- We want to study to what extent NorSySS can serve as an early warning system for local waterborne outbreaks by alerting us to increases in consultation rates for syndromes indicative of gastrointestinal diseases.

- Retrospective syndrome data (number of gastritis cases, per week, for each municipality) from NorSySS will be aligned with outbreak data (outbreak=yes/no) from the national web-based outbreak rapid alert system (Vesuv) for the period 2006-2017.

Question: Is there an association between "recorded outbreak" and "number of gastritis cases"

Aim: Hypothesis testing

Outcome: Binary (recorded outbreak)

Exposure: Count (number of gastritis cases)

Parametric assumptions: None

Dependencies: Yes (longitudinal data by municipality)

Appropriate Method: Mixed effects logistic regression

8.4 Early warning system (EWS) for waterborne outbreaks (part 2)

Scenario:

- We have weekly data on water quality from water works (e.g. pH, turbidity)
- We have weekly number of gastritis cases, per week, for each municipality

- We hope to increase the knowledge about causes of waterborne outbreaks and to develop an improved surveillance system for early detection of future outbreaks.

Question: Is there an association between "weekly number of gastritis cases" and "water quality from the water works"

Aim: Hypothesis testing

Outcome: Count (number of gastritis cases)

Exposure: Continuous (pH, turbidity)

Parametric assumptions: None

Dependencies: Yes (longitudinal data by municipality)

Appropriate Method: Mixed effects negative-binomial regression

8.5 Incidents in the water supply system and illness

Scenario:

- Data from the water works operation (pH, turbidity) will be linked to health outcome among recipients of the drinking water.
- The study will be a prospective cohort study, with data collected among a random selection of water works.
- Data from the recruited water works will be collected in the period autumn 2017 and the 12 following months.
- Approximately 350 water works will be recruited to provide monthly data on hygienic critical points related to operation and maintenance of the water supply system.
- In parallel, a cohort of approximately 9000 persons, served by water from the recruited water works, will submit monthly reports on symptoms that may indicate gastrointestinal illness.
- The data collection will start in the autumn of 2017 and continue for 12 months.

Question: Is water quality a risk factor for getting sick?

Aim: Hypothesis testing

Outcome: Binary (sick yes/no)

Exposure: Continuous (pH, turbidity)

Parametric assumptions: None

Dependencies: Yes (longitudinal data by person, clustered by waterwork and/or municipality)

Appropriate Method: Mixed effects logistic regression

8.6 Compliance with boil water advisories and perception of risks

Scenario:

- In this study, the compliance and perception of risks among the public with boil water advisories (BWAs) will be examined.
- Although BWAs is a common practice among water utilities, a meta-study suggest that there is limited information and studies on the compliance of BWAs.
- This part, the compliance and perception of risks will be done by studying the perception of and adherence to BWAs among the consumers of drinking water in Bærum municipality.
- Even though the drinking water in Bærum is considered to have good quality, Bærum – like many water works – experience situations of pressure drops due to breaks and maintenance.
- Research has shown that these situations may lead to an increased risk of gastrointestinal infections, and due to this the municipality of Bærum has issued a precautionary BWAs to the affected consumers with every water outage during the last 5-6 years.
- Every year, some 12,000-22,000 consumers have received a precautionary BWA.
- The purpose of these precautionary BWAs is to prevent health consequences caused by possible contamination of water. However, we know little about the consumer's knowledge about why they receive these BWAs, as well as the way the consumers perceive and adhere to these BWAs.
- This is a cross-sectional study.
- A web-survey will be presented to a randomly selected sample of consumers who received a BWA in Bærum in 2017.
- The web-survey asks about adherence (yes/no) and demographics (e.g. age, sex, income)

Question: Estimate adherence by demographics (and identify if it differs by demographics)

Aim: Estimation and hypothesis testing

Outcome: Binary (adhberence yes/no)

Exposure: Binary (sex), Categorical (age, income)

Parametric assumptions: None

Dependencies: No ("randomly selected sample of consumers")

Appropriate Method: Logistic regression