Class 1: Basics: t-tests, sample sizes, and experimental design

Andrew Parnell, School of Mathematics and Statistics,
University College Dublin

Learning outcomes

- Know how to create and interpret a two-sample t-test
- Understand what a p-value means
- Be able to perform a simple sample size calculation
- Understand the basics of experimental design

General goal for the course: be able to create a statistical model for a medical test in R and check that it is robust

Course details

- More lectures in the morning (9:30 - 1pm), more practicals in the afternoon (2pm - 5pm). More details in the **timetable**.
- All course notes, code and data sets available on **Github page**
- All Slides available in pdf or RMarkdown (Rmd) format which can be opened in Rstudio

Some basic concepts:

- One way data can be grouped is either *continuous* (e.g. age, weight), or *discrete* (disease state, Gleason grade, etc)
- You can divide continuous into *interval* (temperature) or *ratio* (age, weight)
- You can divide discrete into *ordinal* (e.g. Gleason grade) or *nominal* (disease state, eye colour)

The type of statistical model we fit is almost entirely dependent on the type of data we have

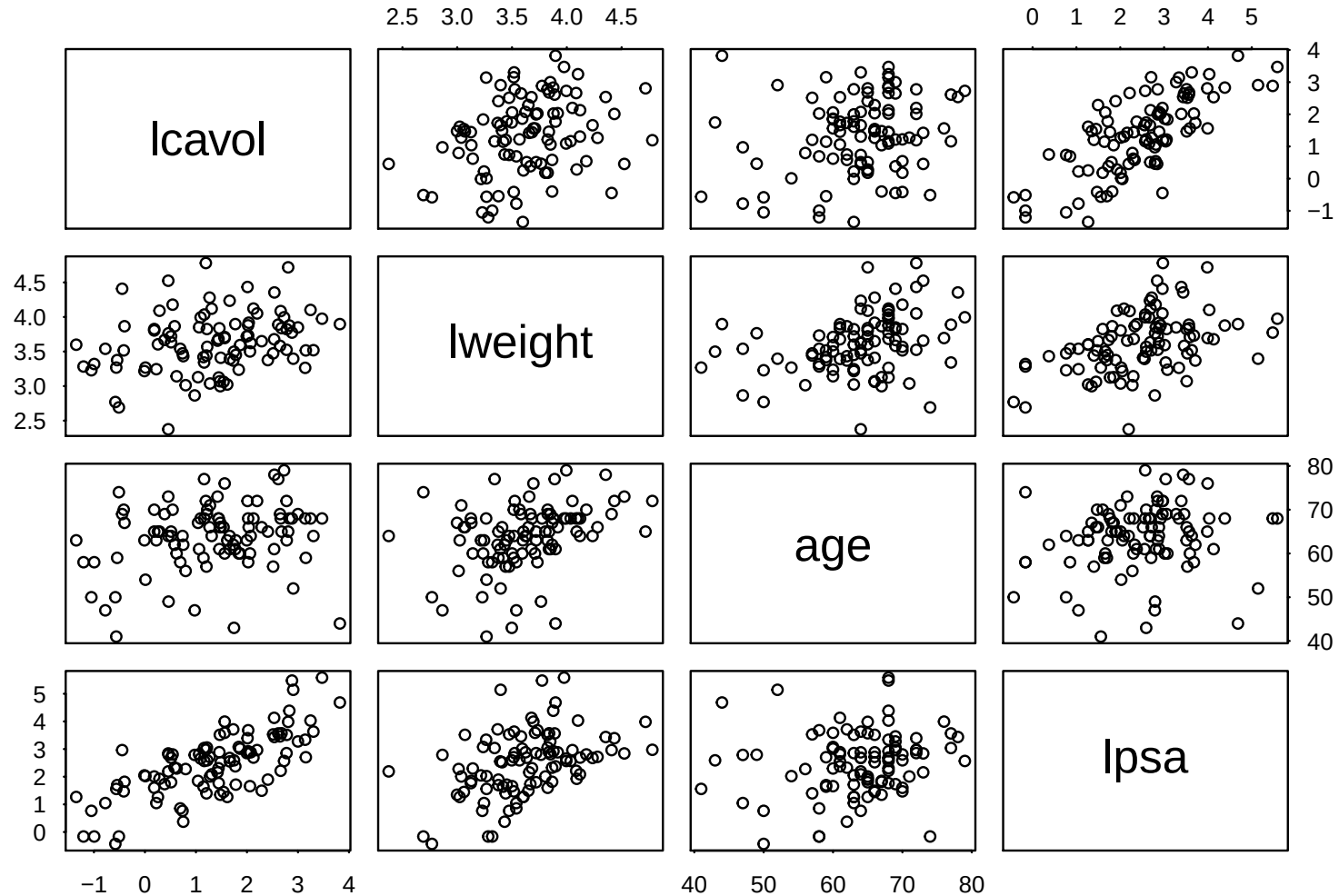
Two examples: prostate cancer (regression)

Prostate cancer data set:

- `lcavol`: log(cancer volume)
- `lweight`: log(weight)
- `age`: age
- `lbph`: log(benign prostatic hyperplasia amount)
- `svi`: seminal vesicle invasion
- `lcp`: log(capsular penetration)
- `gleason`: grade of cancer
- `pgg45`: percentage Gleason scores 4 or 5
- `lpsa`: outcome variable - log prostate specific antigen
- `train`: whether the observation should be included in the training or test set

Task: predict `lpsa` based on other variables for the training set, and check performance on the test set

Prostate example matrix scatter plot



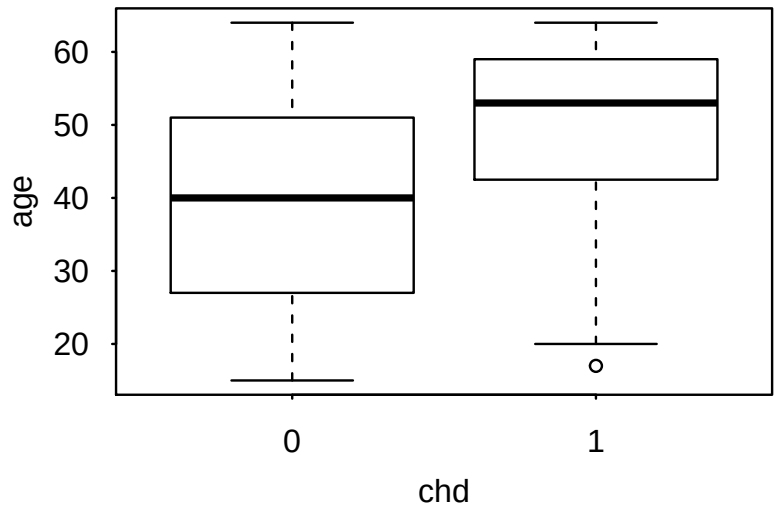
Example 2: South African Heart Rate data (classification)

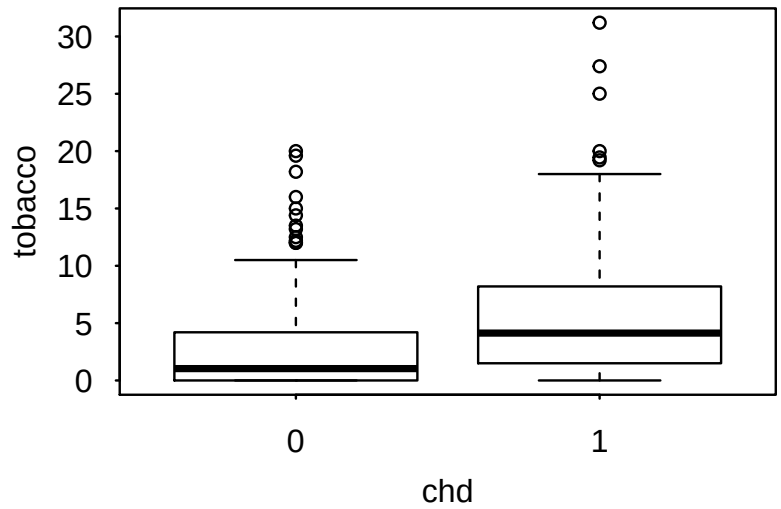
462 observations, with 10 variables:

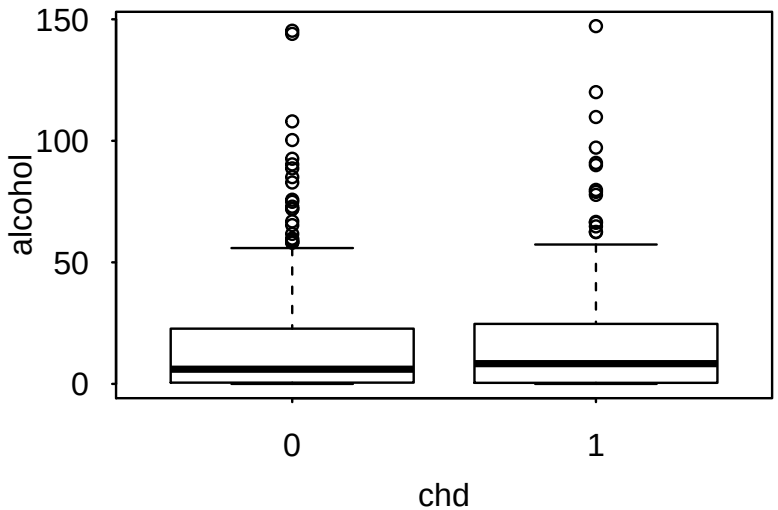
- `sbp` - systolic blood pressure
- `tobacco` - cumulative tobacco (kg)
- `ldl` - low density lipoprotein cholesterol
- `adiposity` - approx percentage body fat
- `famhist` - family history of heart disease (Present, Absent)
- `typea` - type-A behavior
- `obesity` - a measure of obesity
- `alcohol` - current alcohol consumption
- `age` - age at onset
- `chd` - output variable - coronary heart disease

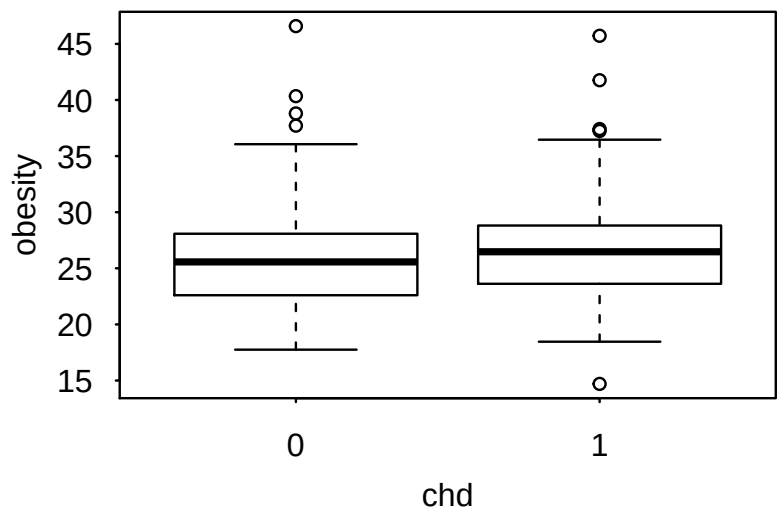
Task: predict probability of `chd` based on other variables

Heart rate data plots









Testing differences between groups; the two-sample t-test

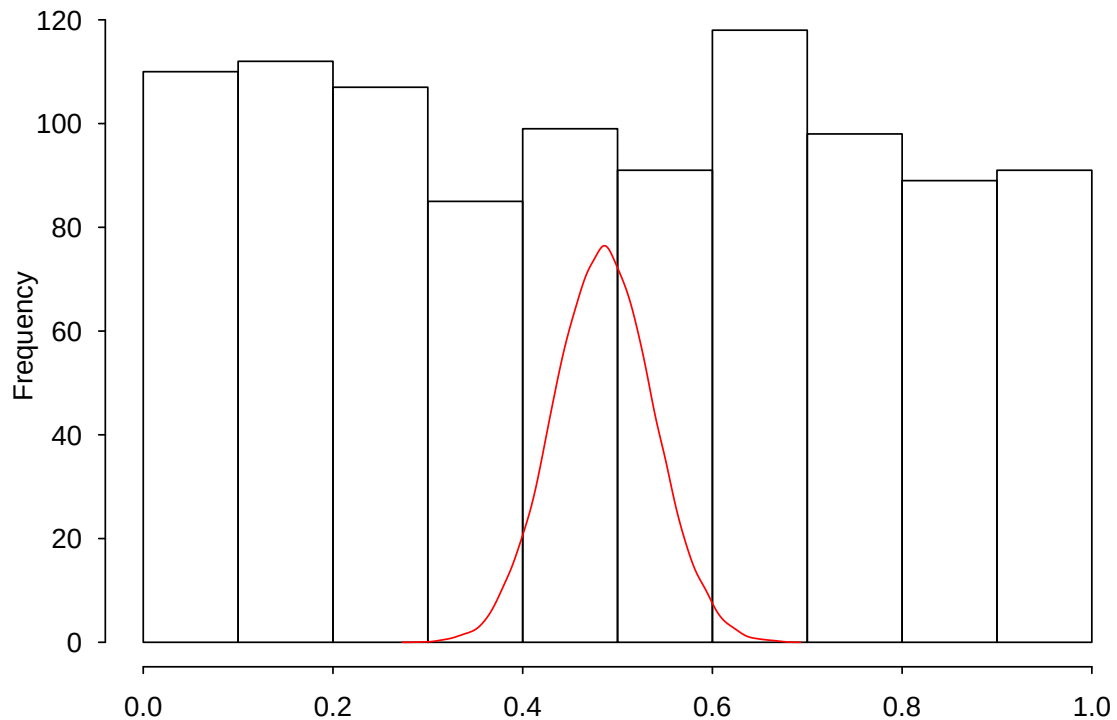
- Goal: test whether the mean of one group is equal to the mean of another group
- Obviously we only have a sample of data, not all the potential data (this is generally impossible)
- Use the mathematics of sampling distributions to determine whether the data look 'unlike' a situation where the two means are equal

Sampling distributions of data

- If we re-ran the experiment we would get different data. What might the sample means of these data sets look like?
- Amazingly, no matter what the shape of the original data, the sample mean will always follow a *normal distribution*
- The mean of this normal distribution will be the mean of the population, and the standard deviation (known as the *standard error*) will be the same as the population standard deviation divided by the square root of the sample size

Sampling distributions in pictures

```
population = runif(1000) # 1000 uniform(0,1) numbers
sample_size = 30 # A sample size
sample_mean = rep(NA, 10000) # Create 10k samples
for(i in 1:10000) {
  current_sample = sample(population, sample_size)
  sample_mean[i] = mean(current_sample)
}
```



Sampling distributions in theory and in practice

- It's nice to know that in theory if we took thousands of samples we would end up with a normally distributed sample mean
- However, we usually only take 1 sample, so we don't know what the standard deviation of this sampling distribution really is
- The usual shortcut is to use the sample standard error (i.e. the standard deviation of the sample divided by the sample size)
- This shortcut allows us to quantify our sampling variability and therefore decide whether any differences between samples occur because of sampling, or because there is a real difference between the sample means

Null and alternative hypotheses

- The usual way to run a two-sample t-test is to define a *null hypothesis* that says both population means are equal, and an *alternative hypothesis* that states that they are not
- We then create a sampling distribution of the difference between the two samples
- If the two sample means are sufficiently different after taking account of their standard errors then we usually reject the null hypothesis

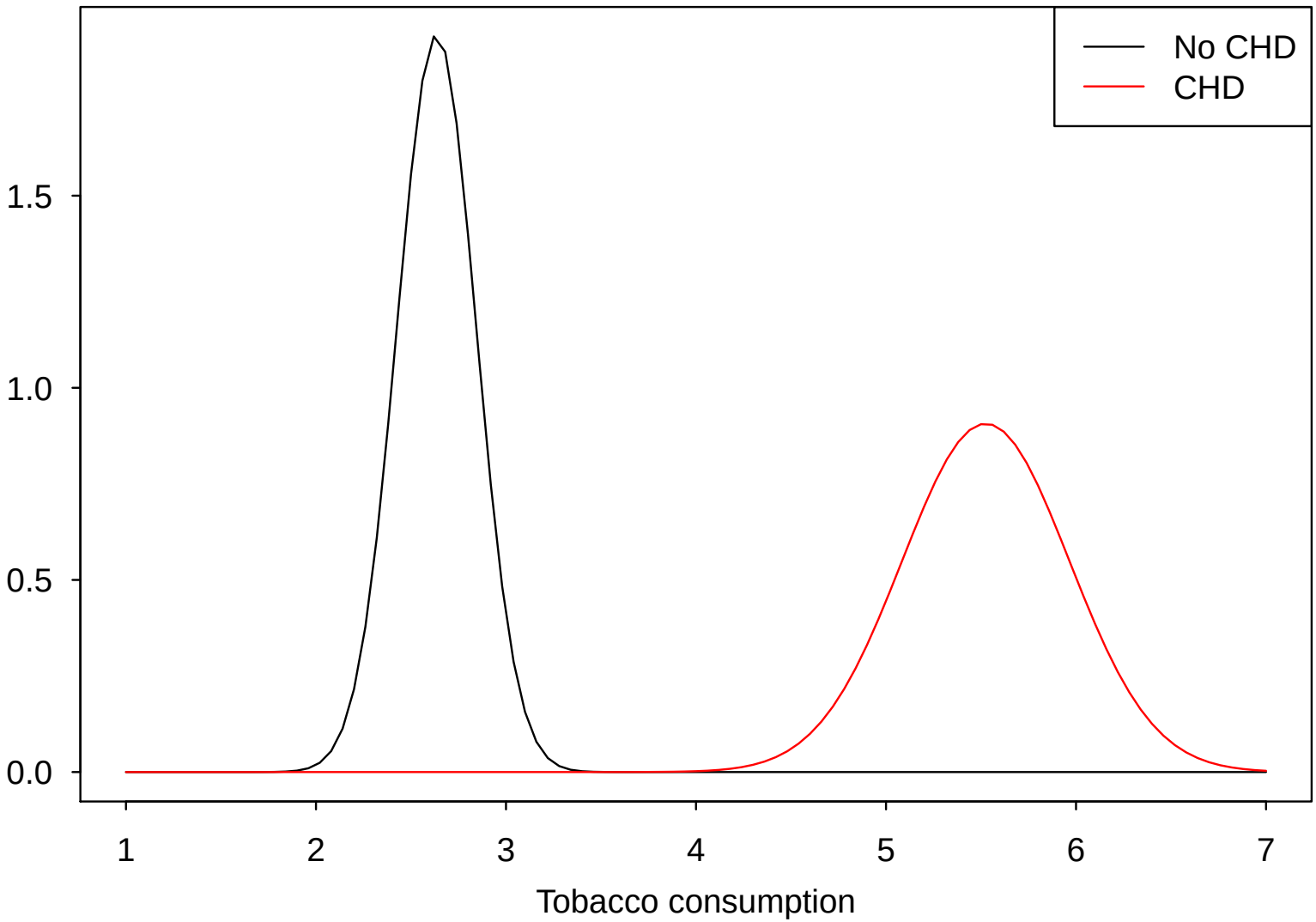
Example: the heart rate data

Suppose we wished to test whether tobacco consumption had an effect on coronary heart disease:

```
with(SA, t.test(tobacco[chd==0], tobacco[chd==1]))
```

```
data:  tobacco[chd == 0] and tobacco[chd == 1]
t = -5.9396, df = 231.8, p-value = 1.038e-08
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -3.848845 -1.931434
sample estimates:
mean of x mean of y
 2.634735  5.524875
```

Drawing pictures



Getting and understanding the p-value

- Most people look for the p-value. A small p-value (often, for no reason, smaller than 0.05) is considered to be a 'statistically significant result'
- The meaning of the word significant here is that of *signifying something*, not that it is necessarily important
- It is often far more helpful to look at the *confidence interval* which is a measure of effect size, than the p-value

Warnings about p-values

- p-values are almost universally mis-used in science (and medicine in particular)
- A small p-value just means that you have quantified an effect well, and is usually just a function of the sample size
- The null hypothesis is almost never true, so it's easy to manipulate your experiment to get small p-values
- From the **American Statistical Association statement** on p-values:

  > *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

Sample size calculations

Introduction to sample size calculations

- The t-test (and the formula behind it) is often more useful for deriving a sample size for an experiment to quantify a given effect
- A commonly used formula is:

$$N > \frac{2\sigma^2(z_{\alpha/2} + z_\beta)^2}{d^2}$$

- where:
  - N is the sample size required
  - σ is the unknown population standard deviation
  - d is the *clinically significant* difference
  - $z_{\alpha/2}$ and z_β are the cut-off values for a given *type 1* and *type 2* error

Type 1 and Type 2 error

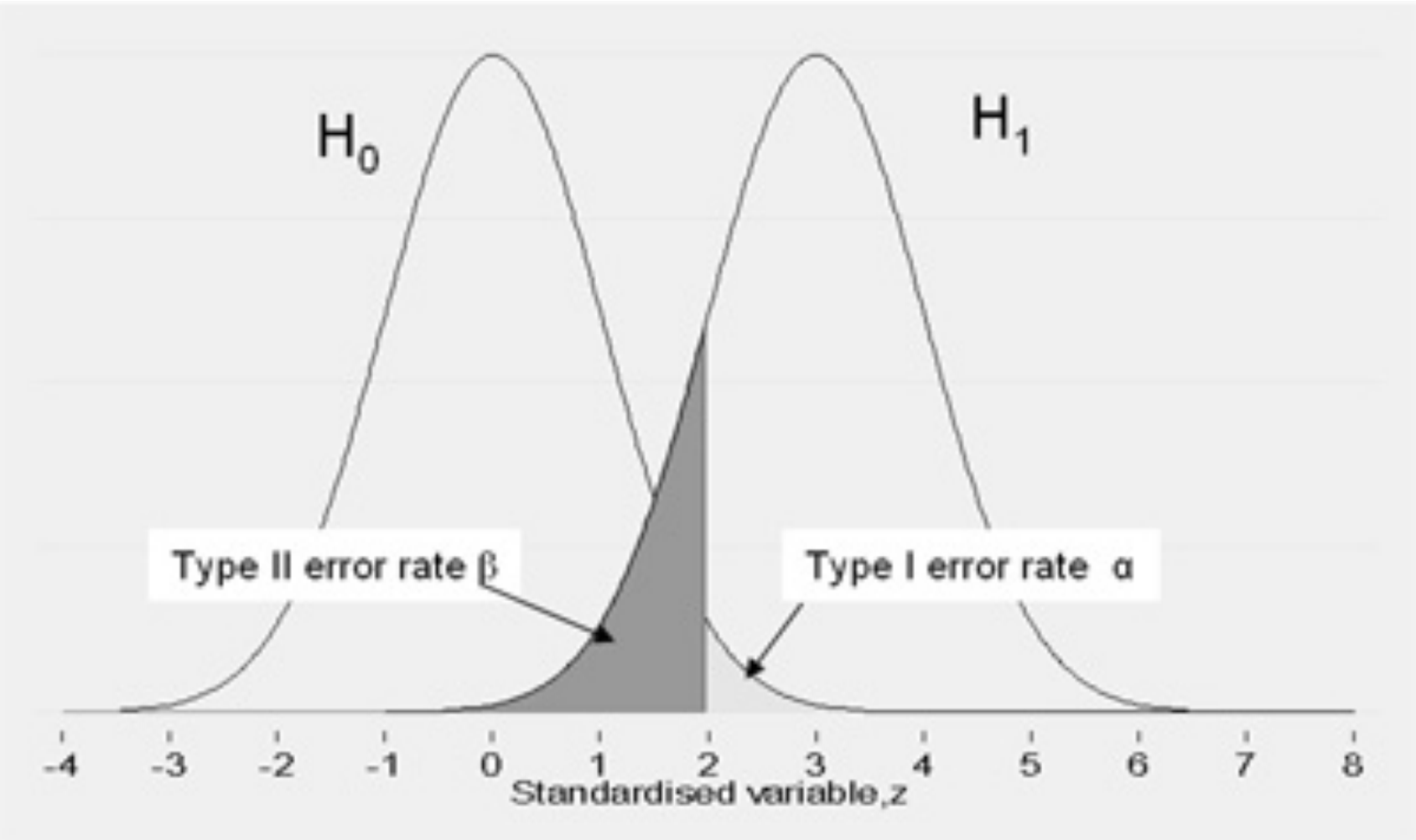


Figure 1: Type 1 and Type 2 error

Getting the values to put in to the formula

- You can usually find a good value of σ from a previous experiment
- d should be easy to choose if you are familiar with the research
- $z_{\alpha/2}$ and z_{β} are harder to choose. Many people set $\alpha = 0.05$ and $\beta = 0.2$ which gives $z_{\alpha/2} = 1.96$ and $z_{\beta} = 0.842$

Once you have all these values you can plug them into the formula

Example

- Let's suppose we wanted to conduct a new version of the test of tobacco levels on coronary heart disease. We might guess the population standard deviation to be:

```
sd(SA$tobacco)
```

```
## [1] 4.593024
```

- Suppose a difference of 2 is considered to be clinically significant, then:

```
N = (2 * sd(SA$tobacco)^2 * (1.96 + 0.842)^2 ) / ( 2^2 )
N
```

```
## [1] 82.81399
```

So we need at least 83 samples in each group

Final notes about sample size calculations

- Many people just plug in values to the above formula until they get a number they are happy with. This will often lead to a useless experiment!
- Be especially careful choosing the value of σ - the population standard deviation. Previous experiments are likely to have under-estimated it
- Be even more careful when performing comparisons between multiple groups, the type 1 and type 2 error terms (α and β) may need to be changed
- There are many more complicated and interesting versions of sample size formulae

Experimental design

Introduction

- In statistics, most experiments are not designed, and we have to pick apart the effects of different variables according to the data we are presented with.
- A problem we often face is that of *confounding* where multiple factors have changed our outcome variable and we cannot pick apart which is the cause of the change
- For example, in the CHD data most of those with CHD have adiposity scores, and consume more tobacco. If adiposity was really the key factor we have no way of separating it out from tobacco consumption
- If it were ethical to design an experiment here we could, for example, force there to be some non-smokers with high adiposity in the CHD and no-CHD groups

The golden rule of designing an experiment

Block everything you can control, randomise over everything else, and replicate as much as possible

Blocking

- A *block* is simply a variable in an experiment you have control over, e.g. temperature, sex, age, etc.
- The idea is that in each block the people in the sample are broadly similar across the treatment groups
- When there are multiple factors we might have a more complex design, such as a Latin square or similar

Randomisation

- When we can't control a variable, or we have so many variables that we can't control them all, we rely on *randomisation*, i.e. randomly allocating people to treatment groups
- *Randomisation* helps by reducing the effect of confounding
- A related concept is that of *blinding* where the subjects/experimenters do not know which group they will be put in

Replication

- It's all very well designing a beautiful experiment, but if you only end up with 5 observations at the end it will be hard to produce meaningful results
- The more replicates you have the more chance of identifying the effect size
- There are lots of different ways to replicate, including taking multiple observations on people (repeated measures) or taking them over time (longitudinal analysis)
- For more complicated experiments, simple t-tests will not work well, but regression and classification models still do!

Summary

- Two sample t-tests not ideal for most proper data sets
- Beware of mis-interpretations of p-values
- Sample size calculations are a good idea
- Always design an experiment if possible