University of BRISTOL

Statistical hypothesis testing for unpaired data

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we cover today?

- We will begin by introducing the idea of hypothesis testing for two unpaired samples.
- We will introduce Student's t test for comparing population means in unpaired samples.
- We will discuss Welch's t test which is more robust to differences in population variances.

Drawing inferences from data

Hypothesis testing is a general methodology for drawing conclusions from data.

Consider the following example scenarios:

1. A farmer wants to know if applying different types of soil treatment will modify their crop yield.
2. A biologist wants to know if there is a difference in some morphological feature between two species.
3. A pharmaceutical company wants to know if a treatment for a medical condition is effective.
4. A retailer wants to know if there is a difference in the click-through-rate for two advertising strategies.

In each case hypothesis testing helps us to understand and control the role of statistical variation.

Hypothesis testing combined with appropriate experimental design can justify our inferences from data.

Example: Comparing Adelie & Chinstrap penguins

Suppose a biologist is investigating the morphological features of different types of penguins.

They want to know if there is a difference between average flipper lengths of Adelie & Chinstrap penguins.

The biologist can't possibly measure the flipper lengths of all the penguins.

The biologist randomly collects two samples:

A random sample of Adelie penguins.

A random sample of Chinstrap penguins.

The flipper lengths of all the penguins are measured.

Research Hypothesis: There is a difference between the average flipper lengths of Adelie & Chinstrap penguins.

Hypothesis testing

Suppose we have a clear research hypothesis and some high-quality data from a well-deigned experiment.

The key stages of statistical hypothesis testing are as follows:

1. Form our statistical hypothesis including a null hypothesis and an alternative hypothesis.
2. Apply model checking to validate any modelling assumptions.
3. Choose our desired significance level.
4. Select an appropriate statistical test.
5. Compute the numerical value of the test statistic from data.
6. Compute a p-value based upon the test statistic.
7. Draw conclusions based upon the relationship between the p-value and the significance level.

Null and alternative hypothesis

The **statistical hypothesis** frames the research question in terms of the parameters of a statistical model.

There are two hypotheses:

H_0 : The **null hypothesis** is our default position typically declaring an absence of an interesting phenomena.

H_1 : The **alternative hypothesis** is the of something interesting difference we'd like to demonstrate.

Example

We model the flipper lengths for the two groups of penguins as i.i.d. draws from two Gaussian distributions:

Group 0 (Adelie): $X_1^0, \ldots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$ (i.i.d.)

Group 1 (Chinstrap): $X_1^1, \ldots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$ (i.i.d.)

Null and alternative hypothesis

H_0 : The **null hypothesis** is our default position typically declaring an absence of an interesting phenomena.

H_1 : The **alternative hypothesis** is the of something interesting difference we'd like to demonstrate.

Example 1

We model the flipper lengths for the two groups of penguins as i.i.d. draws from two Gaussian distributions:

Group 0 : $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$ (i.i.d.)

Group 1: $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$ (i.i.d.)

Null hypothesis $\mathrm{H}_0 : \quad \mu_0 = \mu_1$

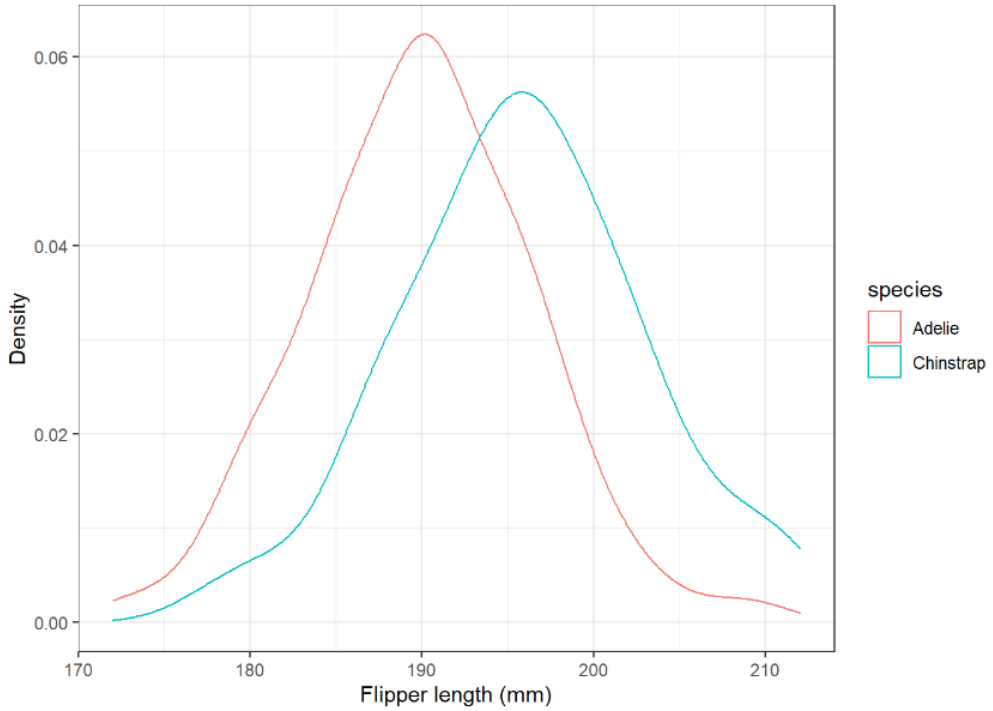
Alternative hypothesis $\mathrm{H}_1 : \quad \mu_0 \neq \mu_1$

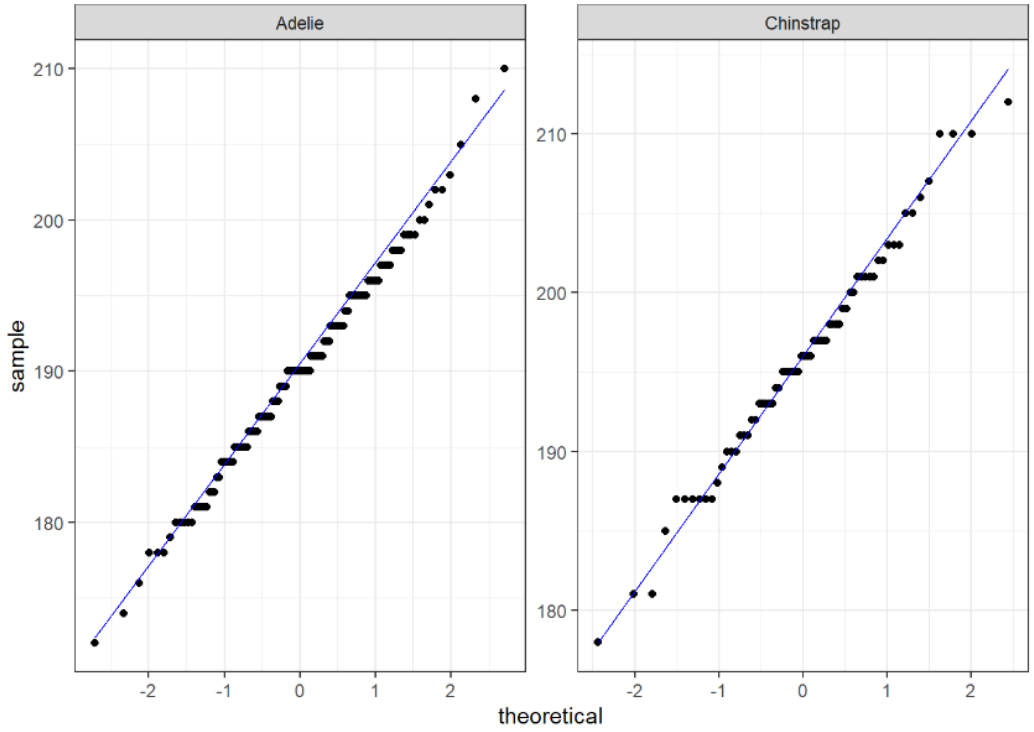
Model checking

We must check that the statistical model underlying our hypotheses is reasonable.

Example 1

Here we assumed a Gaussian model: $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$ & $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$



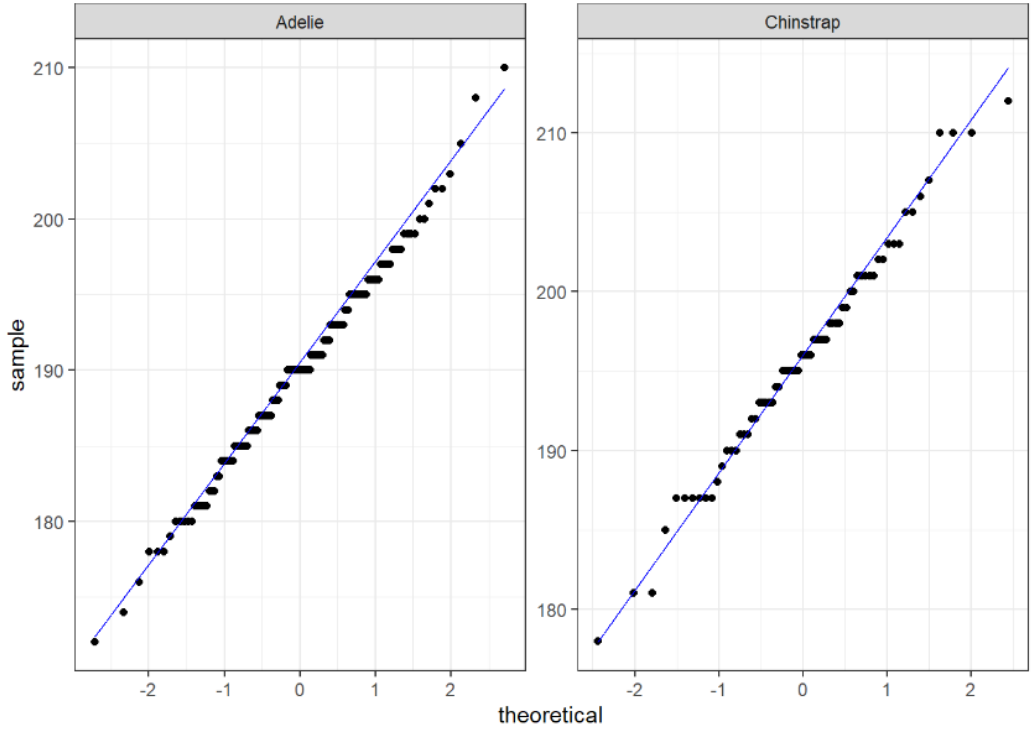


Model checking

We must check that the statistical model underlying our hypotheses is reasonable.

Example 1

Here we assumed a Gaussian model: $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$ & $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$

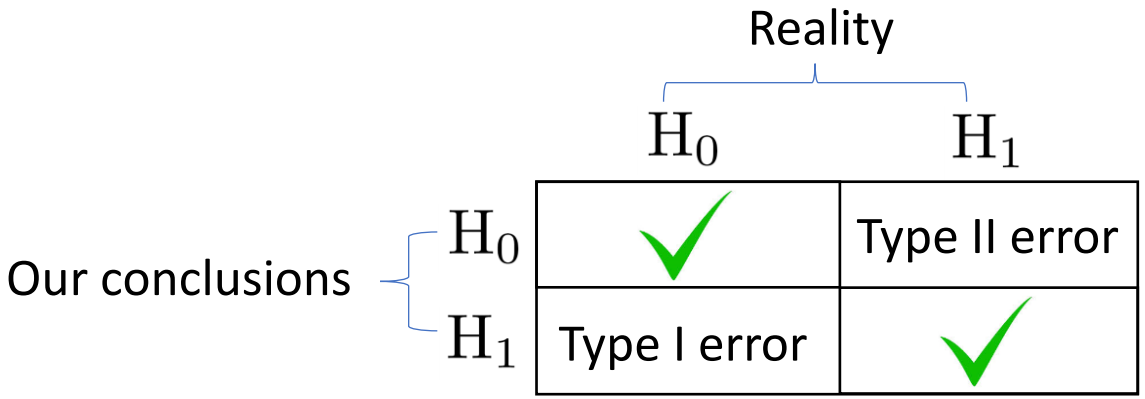


The distributions for both samples are approximately Gaussian.

We have $n_0 = 151$ & $n_1 = 68$

Hence, by the central limit theorem, the sample means will be approximately Gaussian.

Select a significance level



The **test size** of a test α_{test} is the probability of Type I error under the null hypothesis:

$$\alpha_{\text{test}} = \mathbb{P}\left(\text{Type I error} \mid \text{H}_0 \text{ is true}\right).$$

The **power** of test is $1 - \beta_{\text{test}}$ where β_{test} is the probability of Type II error under the alternative:

$$\beta_{\text{test}} = \mathbb{P}\left(\text{Type II error} \mid \text{H}_1 \text{ is true}\right).$$

The **significance level** of a test α is an upper bound on the test size $\alpha_{\text{test}} \leq \alpha$.

Let's choose a significance level of $\alpha = 0.05$.

Test statistics

A **test statistic** is a function of the data which:

a) Emphasizes differences between the null and the alternative

b) Has a known distribution under the null hypothesis $\mathrm{H}_0 : \mu_0 = \mu_1$.

Example

We have two i.i.d. samples $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$ and $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$.

For now we will make the simplifying assumption that $\sigma_0^2 = \sigma_1^2$.

Our null hypothesis is $\mathrm{H}_0 : \mu_0 = \mu_1$ and our alternative hypothesis is $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

The classical test for this setting is the **unpaired Student's t-test** for testing differences in means.

Student's t test statistic

We have i.i.d. samples $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$ and $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$

For now we will make the simplifying assumption that $\sigma_0^2 = \sigma_1^2$.

Our null hypothesis is $\mathrm{H}_0 : \mu_0 = \mu_1$ and our alternative hypothesis is $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

The classical test for this setting is the **unpaired Student's t-test** for testing differences in means:

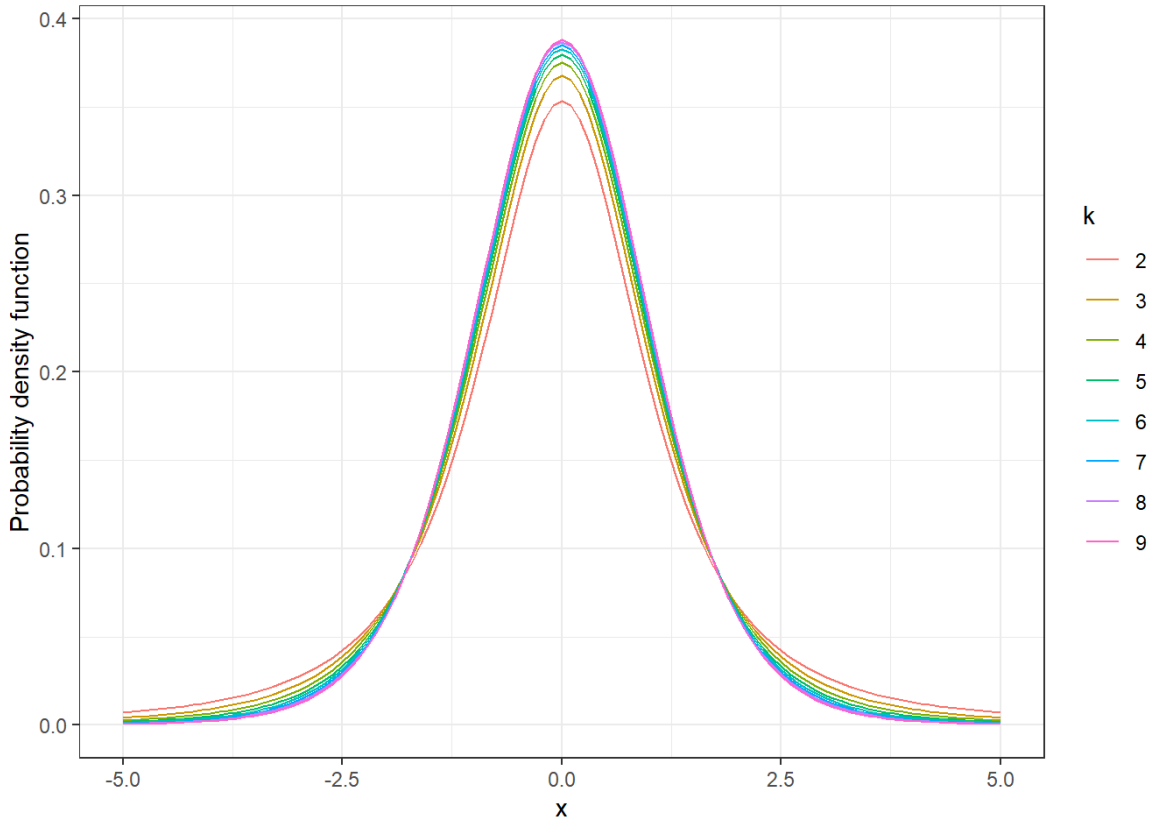
$$\hat{T} = \frac{\overline{X}_0 - \overline{X}_1}{S_{0,1}\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \qquad \text{where} \qquad \overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_i^j$$

$$S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(X_i^j - \overline{X}_j\right)^2 \qquad \text{and} \qquad S_{0,1}^2 = \frac{(n_0 - 1)S_0^2 + (n_1 - 1)S_1^2}{n_0 + n_1 - 2}$$

We expect $|T|$ to be relatively large if $\mu_0 \neq \mu_1$ and relatively small if $\mu_0 = \mu_1$

Student's t distribution

We assume that $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$, $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $\sigma_0^2 = \sigma_1^2$



Under the null hypothesis $\mathrm{H}_0 : \mu_0 = \mu_1$

Let $\overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^j$

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(X_i^j - \overline{X}_j \right)^2$$

$$S_{0,1}^2 = \frac{(n_0 - 1)S_0^2 + (n_1 - 1)S_1^2}{n_0 + n_1 - 1}$$

Then $$\hat{T} = \frac{\overline{X}_0 - \overline{X}_1}{S_{0,1}\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$$

is t-distributed with $n_0 + n_1 - 2$ degrees of freedom.

Computing the test statistc

The classical test for this setting is the **unpaired Student's t-test** for testing differences in means:

$$\hat{T} = \frac{\overline{X}_0 - \overline{X}_1}{S_{0,1}\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \qquad \text{where} \qquad S_{0,1}^2 = \frac{(n_0 - 1)S_0^2 + (n_1 - 1)S_1^2}{n_0 + n_1 - 2}$$

Example

```
peng_AC%>%head(10)
```

```
## # A tibble: 10 x 2
##    species   flipper_length_mm
##    <fct>                 <int>
##  1 Chinstrap               197
##  2 Chinstrap               197
##  3 Chinstrap               200
##  4 Adelie                  198
##  5 Chinstrap               196
##  6 Chinstrap               190
##  7 Adelie                  186
##  8 Adelie                  189
##  9 Adelie                  196
## 10 Chinstrap               202
```

Computing the test statistc

The classical test for this setting is the **unpaired Student's t-test** for testing differences in means:

$$\hat{T} = \frac{\overline{X}_0 - \overline{X}_1}{S_{0,1}\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} \qquad \text{where} \qquad S_{0,1}^2 = \frac{(n_0 - 1)S_0^2 + (n_1 - 1)S_1^2}{n_0 + n_1 - 2}$$

Example

```
mean_0<-peng_AC%>%filter(species=="Adelie")%>%pull(flipper_length_mm)%>%mean()  # Compute mean of first group
mean_1<-peng_AC%>%filter(species=="Chinstrap")%>%pull(flipper_length_mm)%>%mean() # Compute mean of second group

sd_0<-peng_AC%>%filter(species=="Adelie")%>%pull(flipper_length_mm)%>%sd()  # Compute sd of first group
sd_1<-peng_AC%>%filter(species=="Chinstrap")%>%pull(flipper_length_mm)%>%sd() # Compute sd of second group

n_0<-peng_AC%>%filter(species=="Adelie")%>%nrow()  # Compute number in first group
n_1<-peng_AC%>%filter(species=="Chinstrap")%>%nrow() # Compute number in second group

sd_combined<-sqrt(((n_0-1)*sd_0^2+(n_1-1)*sd_1^2)/(n_0+n_1-2)) # Combined sample standad deviatiation

t_statistic <- (mean_0-mean_1)/(sd_combined*sqrt(1/n_0+1/n_1)) # Conpute test statistic
```

```
t_statistic
```

```
## [1] -5.974041
```

Student's t-test

We then compute the **p-value** based on the observed test statistic. The p-value is the probability under the null hypothesis that the test statistic takes a value as extreme or more extreme than the observed value.

Example

We assume that $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$, $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $\sigma_0^2 = \sigma_1^2$.

Our test statistic is the Student's t-test statistic: $\hat{T} = \dfrac{\overline{X}_0 - \overline{X}_1}{S_{0,1}\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$.

Our null hypothesis is $\mathrm{H}_0 : \mu_0 = \mu_1$ and our alternative hypothesis is $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

Key point: Under the null hypothesis the test statistic is t-distributed with $n_0 + n_1 - 2$ degrees of freedom.

We can use $F_{(n_0+n_1-2)}(t) = \int_{-\infty}^{t} f_{(n_0+n_1-2)}(x)dx$ to compute the p-value.

Student's t-test

Example

We assume that $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$, $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $\sigma_0^2 = \sigma_1^2$.

Our test statistic is the Student's t-test statistic: $\hat{T} = \dfrac{\overline{X}_0 - \overline{X}_1}{S_{0,1}\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$.

Our null hypothesis is $\mathrm{H}_0 : \mu_0 = \mu_1$ and our alternative hypothesis is $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

Suppose that the numerical value of the test statistic based on the data is τ . To compute the p-value,

$$\begin{aligned} p &= \mathbb{P}(|T| \geq |\tau||H_0) = 2 \cdot \mathbb{P}(T \geq |\tau||H_0) \\ &= 2 \cdot \{1 - \mathbb{P}(T < |\tau||H_0)\} = 2 \cdot \left(1 - F_{(n_0+n_1-2)}(|\tau|)\right). \end{aligned}$$

Student's t-test with R

We assume that $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$, $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $\sigma_0^2 = \sigma_1^2$.

Our null hypothesis is $\mathrm{H}_0 : \mu_0 = \mu_1$ and our alternative hypothesis is $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

```
mean_0<-peng_AC%>%filter(species=="Adelie")%>%pull(flipper_length_mm)%>%mean()  # Compute mean of first group
mean_1<-peng_AC%>%filter(species=="Chinstrap")%>%pull(flipper_length_mm)%>%mean() # Compute mean of second group

sd_0<-peng_AC%>%filter(species=="Adelie")%>%pull(flipper_length_mm)%>%sd()  # Compute sd of first group
sd_1<-peng_AC%>%filter(species=="Chinstrap")%>%pull(flipper_length_mm)%>%sd() # Compute sd of second group

n_0<-peng_AC%>%filter(species=="Adelie")%>%nrow()  # Compute number in first group
n_1<-peng_AC%>%filter(species=="Chinstrap")%>%nrow() # Compute number in second group

sd_combined<-sqrt(((n_0-1)*sd_0^2+(n_1-1)*sd_1^2)/(n_0+n_1-2)) # Combined sample standad deviatiation

t_statistic <- (mean_0-mean_1)/(sd_combined*sqrt(1/n_0+1/n_1)) # Conpute test statistic

p_value<-2*(1-pt(abs(t_statistic),df=n_0+n_1-2)) # compute the p value
```

```
p_value
```

```
## [1] 9.378738e-09
```

Student's t-test with R

We assume that $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$, $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $\sigma_0^2 = \sigma_1^2$.

Our null hypothesis is $\mathrm{H}_0 : \mu_0 = \mu_1$ and our alternative hypothesis is $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

```
peng_AC%>%head(15)
```

```
## # A tibble: 15 x 2
##    species   flipper_length_mm
##    <fct>                 <int>
##  1 Chinstrap               197
##  2 Chinstrap               197
##  3 Chinstrap               200
##  4 Adelie                  198
##  5 Chinstrap               196
##  6 Chinstrap               190
##  7 Adelie                  186
##  8 Adelie                  189
##  9 Adelie                  196
## 10 Chinstrap               202
## 11 Chinstrap               206
## 12 Chinstrap               196
## 13 Adelie                  202
## 14 Adelie                  205
## 15 Chinstrap               201
```

```
t.test(flipper_length_mm~species,data=peng_AC,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  flipper_length_mm by species
## t = -5.974, df = 217, p-value = 9.379e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.806481 -3.933293
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                189.9536                195.8235
```

We record a p-value of 9.379e-09

Making inferences based on p-values

If the p value is strictly less than the significance level we reject the null hypothesis ➡ H_1

If the p value is greater than the significance level we fail to reject the null hypothesis ➡ H_0

Example

We recorded a p-value of 9.379e-09.

This is below the significance level of $\alpha = 0.05$ so we can reject the null and conclude $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

Based on the sign of our t-statistic we can conclude that $\mu_0 < \mu_1$.

The average flipper length for the Adelie penguins is below the average flipper length of Chinstraps.

However, we assumed that $\sigma_0^2 = \sigma_1^2$. Was this justified?

Welch's t-test

Student's t-test assumes that the population variances $\sigma_0^2,\ \sigma_1^2$ for the two distributions are equal.

Suppose that $X_1^0, \cdots, X_{n_0}^0 \sim \mathcal{N}(\mu_0, \sigma_0)$ and $X_1^1, \cdots, X_{n_1}^1 \sim \mathcal{N}(\mu_1, \sigma_1)$.

Welch's t-test also tests for a difference of means with $\mathrm{H}_0 : \mu_0 = \mu_1$ and $\mathrm{H}_1 : \mu_0 \neq \mu_1$.

$$T_{\mathrm{W}} := \frac{\overline{X}_0 - \overline{X}_1}{\sqrt{\frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}}} \text{ where } \overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_i^j \text{ and } S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(X_i^j - \overline{X}_j\right)^2.$$

Under the null hypothesis $\mathrm{H}_0 : \mu_0 = \mu_1$, Welch's t-statistic T_{W} is *approximately* t-distributed with

$$k = \frac{\left(\frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}\right)^2}{\frac{(S_0^2/n_0)^2}{n_0 - 1} + \frac{(S_1^2/n_1)^2}{n_1 - 1}}$$

degrees of freedom. Hence, we can compute *approximate* p-values!

Welch’s t-test in R

Welch’s t-test is robust to differences in population variances between the two distributions.

Hence Welch’s t-test is the preferred option and the default for testing differences in means in R.

Example

```
t.test(flipper_length_mm~species,data=peng_AC)
```

```
##
##  Welch Two Sample t-test
##
## data:  flipper_length_mm by species
## t = -5.7804, df = 119.68, p-value = 6.049e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.880530 -3.859244
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                189.9536                 195.8235
```

We record a p-value of 6.049e-08.

This is below the significance level of 0.05.

Hence, we reject the null-hypothesis.

Since the t-statistic is negative we conclude that $\mu_0 < \mu_1$.

Reporting experimental results and effect size

When reporting our results we should include:

1) The significance level – This must be decided before carrying out the analysis!

2) The hypothesis test and its motivation.

3) The numerical value of the test-statistic.

4) The p-value (computed based on the value of the test-statistic).

5) The effect size (this is interesting if we rejected the null and established an effect).

Small differences between two populations can yield large test statistics and small p-values, with enough data.

A treatment may cause a statistically significant change, but is this change important?

The **effect size** is a measure for quantifying the magnitude of the observed phenomena.

Cohen's d for the unpaired t-test

Suppose we carry out an unpaired t-test on a pair of samples $X_1^0, \cdots, X_{n_0}^0$ and $X_1^1, \cdots, X_{n_1}^1$.

Following our t-test we reject the null and conclude a difference of population means $\mu_0 \neq \mu_1$.

We can quantify the effect size via Cohen's d for unpaired data:

$$d = \frac{\overline{X}_0 - \overline{X}_1}{S_{0,1}} \qquad \text{where} \qquad S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(X_i^j - \overline{X}_j \right)^2$$

$$\overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^j \qquad \text{and} \qquad (S_{0,1})^2 = \frac{(n_0 - 1)S_0^2 + (n_1 - 1)S_1^2}{n_0 + n_1 - 2}.$$

```
effect_size <-(mean_0-mean_1)/sd_combined # compute the effect size
effect_size
```

```
## [1] -0.8724636
```

Comparing paired and unpaired t-tests

If our data is unpaired we must use an unpaired statistical hypothesis test.

If our data is paired we should use a paired statistical hypothesis test

Given paired data we could just ignore the pairing and apply the standard unpaired t test. This is a bad idea!

The unpaired t-test:

$$T_{\text{unpaired}} := \frac{\overline{X}_0 - \overline{X}_1}{\sqrt{(S_0^2 + S_1^2)/n}} \qquad S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^j - \overline{X}_j\right)^2.$$

The paired t-test:

$$T_{\text{paired}} := \frac{\overline{Y}}{S_Y/\sqrt{n}} \qquad (S_Y)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right) \qquad Y_i := X_i^1 - X_i^0$$

The variance of the differences $(S_Y)^2$ is often much smaller than the variance of the whole samples.

Hence, the paired t-statistic is often larger – these means greater power for the same significance level.

Common misunderstandings (I)

People often make the following mistake:

~~"The p-value is the probability that the null hypothesis is true."~~

This is incorrect!

The null hypothesis is a statement about population parameters – not random variables.

Within the classical interpretation of probability such hypotheses either hold or don't.

What's random is the data and the test statistics computed based on the data.

The **p-value** is the probability under the null hypothesis that the test statistic takes a value as extreme or more extreme than the observed value.

Common misunderstandings (II)

Suppose that we carry out a statistical hypothesis test.

We start by fixing a significance level.

We then compute our test statistic and derive our p-value.

If our p-value is above the significance level we cannot reject the null hypothesis.

This means we don’t have enough evidence to reject the null hypothesis in favor of the alternative.

It does not mean that we have good evidence for the null hypothesis!

For example, we will often have a p-value above the significance level when we have little data,

even when the alternative hypothesis is true.

What have we covered?

- We have introduced the concept of hypothesis testing for unpaired two sample data.
- We discussed the classical Student's t test for comparing population means.
- We introduced Welch's t test which is robust to differences in population variances.
- We also emphasized that if a pairing does exist it is preferrable to use a paired test.
- We concluded by noting another common misunderstanding.

University of BRISTOL

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.