Forum 5
Inference: Testing Mean(s)
Data Skills for Scientists

2025-09-24

Inference

The 3 topics in the second module, ‘Inference’ are:

- Estimating Mean(s);
- Testing Mean(s); and
- Testing Median(s) and Categories.

In this forum we will cover the second topic, Testing Mean(s).

Key topics for Testing Mean(s)

- Hypothesis testing framework,
- Two-sample t-tests,
- Matched pairs t-tests,
- One-sample t-tests.

Hypothesis testing framework

Hypothesis testing framework

Four steps of tests of significance:

- State the null and alternative **hypotheses**.
- Calculate the value of the **test statistic**.
- Find the **p-value** for the observed data.
- State a **conclusion** (typically against a **significance level**).

Null hypothesis

Often, the null hypothesis is a statement of “no effect” or the status quo.

It is usually of the form:

$$H_0 : \mu = \mu_0.$$

The value μ_0 is called the **null value**.

Alternative hypothesis

The alternative is one-sided if it either of the form

$$H_a : \mu > \mu_0$$

or

$$H_a : \mu < \mu_0.$$

It is two-sided if it is of the form

$$H_a : \mu \neq \mu_0.$$

We will only consider two-sided hypothesis testing in this course.

Test statistic

A **test statistic** is usually of the form:

$$\text{test statistic} = \frac{\text{estimate} - \text{hypothesised value}}{\text{standard error of estimate}}.$$

They give a measure of how “different” the data is from what we would expect if the null hypothesis H_0 were true.

Large values of the test statistic show that the data are not consistent with H_0.

P-value

The probability, computed assuming H_0 is true, that the statistic would take a value as or more extreme than the one actually observed is called the P-value of the test. The smaller the P-value, the stronger the evidence against H_0.

Guidelines

P-value	conclusion
$0.05 <$ P-value ≤ 0.1	weak evidence against H_0
$0.01 <$ P-value ≤ 0.05	strong evidence against H_0
P-value ≤ 0.01	very strong evidence against H_0

State a conclusion

Statistical significance

The final step in performing a significance test is to draw a conclusion about the competing claims you were testing. We will make one of two decisions based on the strength of the evidence against the null hypothesis (and in favor of the alternative hypothesis)

- there is sufficient evidence to reject H_0
- there is insufficient evidence to reject H_0.

To do this, we first decide on a **significance level** at the start of our analysis.

Significance level

The significance level is the probability of rejecting the null hypothesis when it is in fact true.

The standard significance level is 0.05.

You may also see this written as a 5% significance level.

Other common significance levels include 0.01 and 0.10

Making a decision

- If the P-value is less than the significance level, then there is sufficient evidence to reject the null hypothesis.
- If the P-value is greater than or equal to the significance level, then there is insufficient evidence to reject the null hypothesis.

Warning

Note: A fail-to-reject H_0 decision in a significance test does not mean that H_0 is true. For that reason, you should never use language implying that you believe H_0 is true.

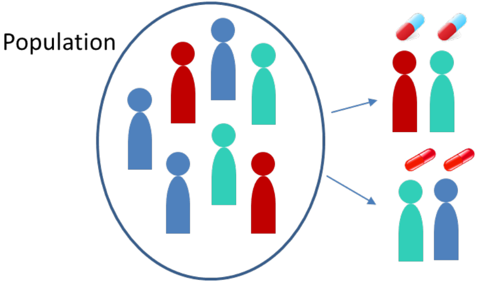
Two-sample t-test

Two-sample t-test

A two-sample t-test determines if there is a statistically significant difference between the population means of two **independent** groups.

This may be used to:

- Explore the characteristics of two distinct populations.
- Explore responses to two different treatments.



Two-sample t-test conditions

Conditions:

- We have two random samples from distinct populations,
- The samples are independent:
  - Within each group, i.e., random sampling,
  - Between each group, i.e., random allocation.
- Both populations are normally distributed, or $n > 30$ in each sample.

Example: Rats and ozone

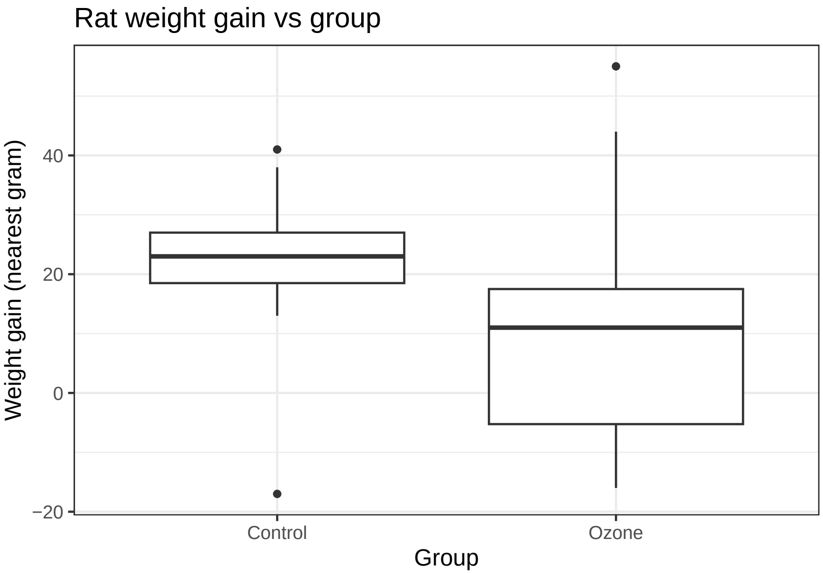
To measure the effects of ozone on weight, one group of 70-day-old rats were kept in an environment containing ozone for 7 days. A second group of rats of the same age (the control group) were kept in an ozone-free environment for the same time. The weight gains (to the nearest *gram*) were recorded.

Load and view the data

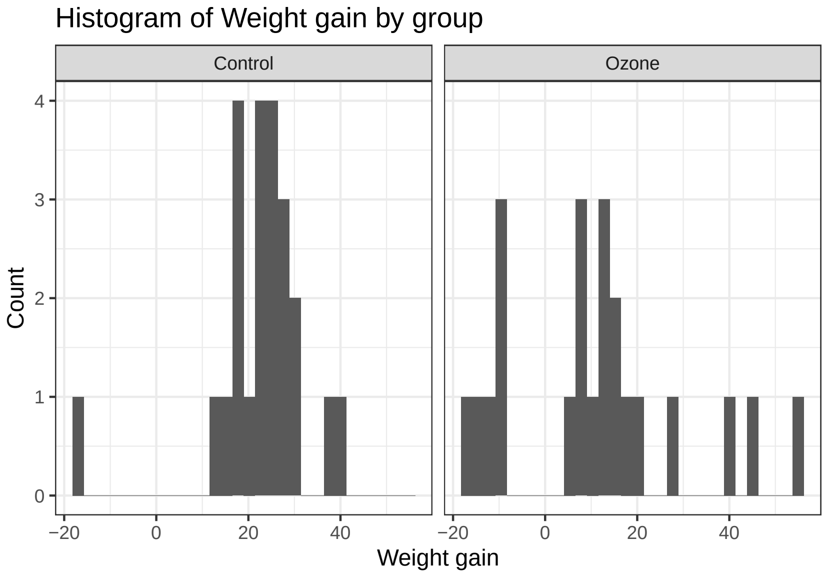
```r
rats_df <- read_excel(here::here("data", "rats.xlsx"))
head(rats_df)
```

```
# A tibble: 6 x 2
  group   weight
  <chr>    <dbl>
1 Control     41
2 Control     26
3 Control     13
4 Control    -17
5 Control     15
6 Control     22
```

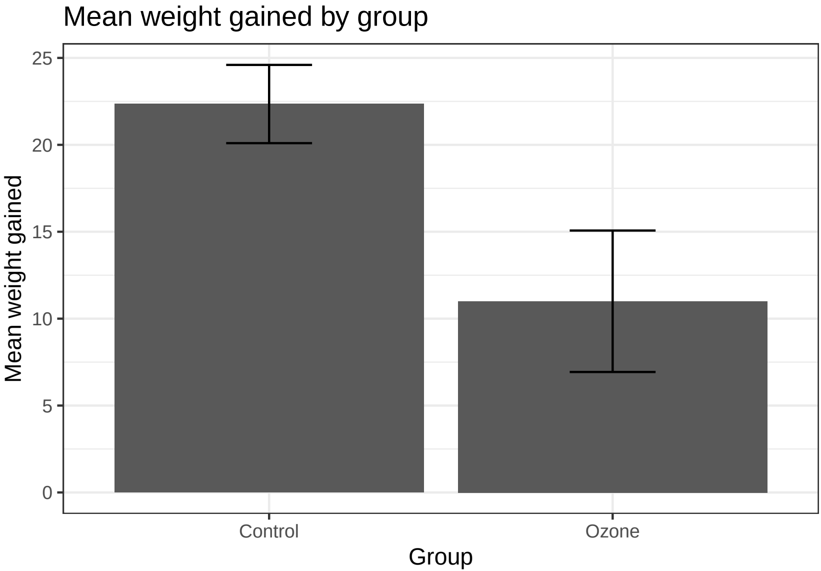
Summarise and visualise the data



Summarise and visualise the data

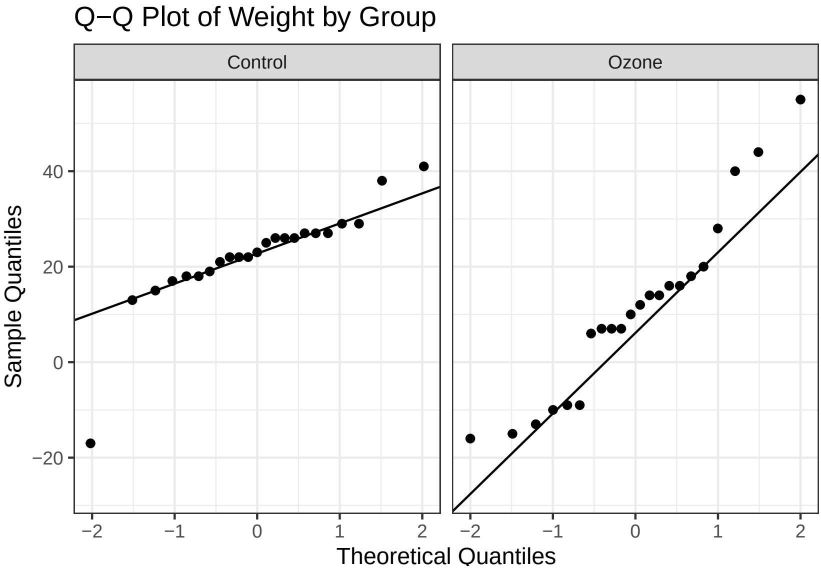


Summarise and visualise the data



Error bars represent one standard error of the mean.

Checking assumptions - Normality



Check the assumptions - Independence

- Within each group (i.e., random sampling).
- Between each group (i.e., random allocation).

State the null and alternative hypotheses

In general,

- the null hypothesis, H_0, corresponds to there is no effect or no difference,
- the alternative hypothesis, H_1, corresponds to there is an effect or a difference.

What is the null and alternative hypotheses for the rats and ozone example?

State the null and alternative hypotheses

If we have two independent samples and are interested in comparing the population means,

- $H_0 : \mu_1 - \mu_2 = 0$,
- $H_1 : \mu_1 - \mu_2 \neq 0$,

where μ_1 is the population mean of population 1 and μ_2 is the population mean of the population 2.

Perform a two-sample t-test

By default t.test in R assumes a Welch two-sample t-test where:

- population variances are different,
- the number of samples for each population is different.

Perform a two-sample t-test

```r
t.test(weight~group, data=rats_df)
```

```

	Welch Two Sample t-test

data:  weight by group
t = 2.4403, df = 32.877, p-value = 0.02023
alternative hypothesis: true difference in means between gr
95 percent confidence interval:
  1.885639 20.810013
sample estimates:
mean in group Control   mean in group Ozone
             22.34783              11.00000
```

Theory underlying t.test output - Test statistic

For the Welch two-sample t-test the test statistic is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

where,

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of populations 1 and 2, respectively,
- s_1 and s_2 are the sample standard deviations of populations 1 and 2, respectively,
- n_1 and n_2 are the number of subjects in the samples for populations 1 and 2, respectively.

Theory underlying t.test output- Test statistic

This test statistic

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

does not have a t-distribution.

However, we can obtain a good approximation with a t-distribution by carefully choosing the degrees of freedom.

In R, the number degrees of freedom of the t-distribution is calculated using the Welch-Satterthwaite equation.

If you were to perform calculations by hand you would typically choose the number degrees of freedom of the t-distribution to be equal to the smaller value of $n_1 - 1$ and $n_2 - 1$.

Theory underlying t.test output - p-value

The p-value is computed using a t-distribution with the number of degrees freedom chosen according to the previous slide.

Theory underlying t.test output - Confidence interval

The formula for a $C\%$ confidence interval for $\mu_1 - \mu_2$ is given by,

$$(\bar{x_1} - \bar{x_2}) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where

- t^* is the appropriate critical value to give a $C\%$ confidence level using the appropriate t-distribution,
- $t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is the margin of error.

Note - Populations with equal variances

As a brief aside, when the population variances are assumed to be equal one can use the Student t-test.

In R this is given by,

```r
t.test(weight~group, data=rats_df,var.equal=TRUE)
```

The Student t-test has:

- a different formula for the test statistic,
- a different formula for the confidence interval.

In this course, we will focus on the Welch two-sample t-test that assumes unequal variances.

Interpreting the results and concluding

For the rats ozone example,

- Using the p-value:

There is sufficient evidence to reject the null hypothesis since the p-value of 0.02023 is less than the significance level of 0.05.

Therefore, at a significance level of 0.05 we conclude that there is a difference between the population mean weight gained when exposed to ozone compared to the weight gained when not exposed to ozone.

Interpreting the results and concluding

- From the confidence interval:

Based on a two-sided two-sample t-test, we are 95% confident that the true difference in weight gained between the two groups lies between 1.89 and 20.8.

Since this interval does not include zero, at 5% significance level we conclude that there is a difference between the population mean weight gained when exposed to ozone compared to the weight gained when not exposed to ozone.

Summary

Hypothesis testing for two means from independent normal distributions

- **Hypotheses:**

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0, \\ H_a &: \mu_1 - \mu_2 \neq 0. \end{aligned}$$

- **Test statistic:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **P-value:** t-distribution with degrees of freedom determined by R, or by hand using the smaller of $n_1 - 1$ and $n_2 - 1$.
- **Confidence interval:**

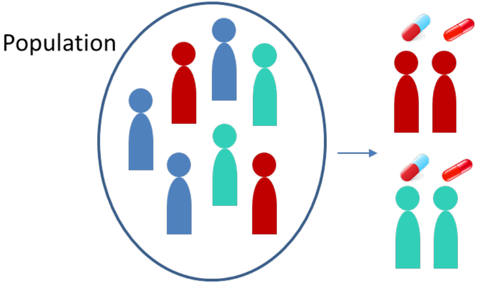
$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Matched pairs t-test

Matched pairs t-test

So far in two-sample t-tests, we have assumed the two groups are independent. We now consider matched pair designs:

- Subjects are matched.
- Subjects serve as their own control.
- Before and after studies.



Matched pairs t-test conditions

Conditions:

- We have two samples that are dependent.
- Each subject is independent.
- Each subject has two measurements.
- The population of resulting differences is normal, or the sample size $n > 30$.

Example: Diet drug

During a weight loss study, nine men, chosen at random from the target population, were each given either the active drug mCPP for two weeks and then a placebo for another two weeks, or else a placebo for the first two weeks and then mCPP for the second two weeks. As part of the study the men were asked to rate how hungry they were at the end of each two-week period.

Perform the appropriate hypothesis test at the 5% significance level to determine whether there is evidence that hunger ratings differ when given the active drug mCPP or on the placebo.

State the null and alternative hypotheses

- The null hypothesis, H_0, corresponds to there is no effect or no difference.
- The alternative hypothesis, H_1, corresponds to there is an effect or a difference.

What is the null and alternative hypotheses for the diet drug example?

Because the nine men are their own controls, the two samples of hunger ratings are dependent and our hypothesis needs to reflect this.

State the null and alternative hypotheses

In the matched pairs design, each subject has two measurements, that we can call X and Y. We then look at the differences,

$$D = X - Y.$$

So for each subject, we have one difference. Then we can test if the mean of D is different from 0, i.e.,

- $H_0 : \mu_D = 0$
- $H_1 : \mu_D \neq 0$

where μ_D is the population mean of the differences.

We can do this with a one-sample t-test.

Example: Diet drug - Null and alternative hypotheses

- $H_0 : \mu_D = 0$
- $H_1 : \mu_D \neq 0$

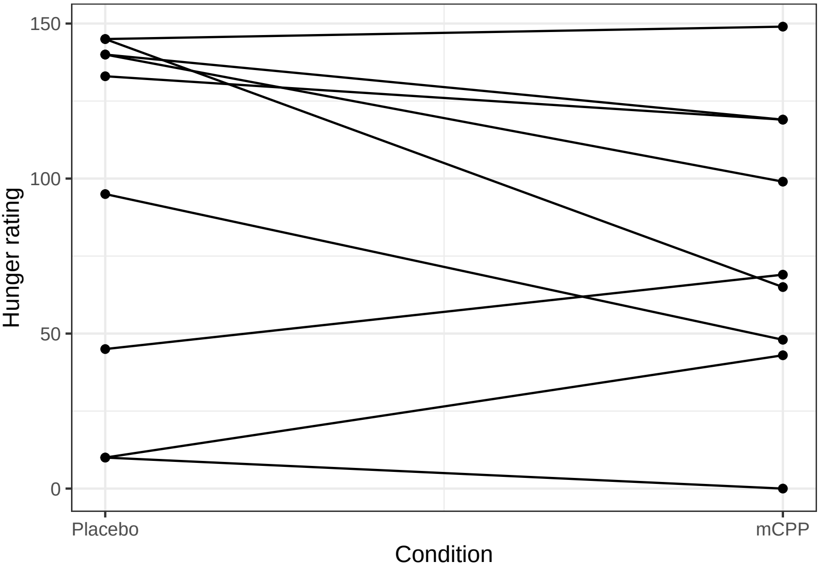
where μ_D is the population mean of the paired differences when given the active drug mCPP and when given the placebo treatment.

Load and view the data

```r
Diet_df <- read_excel(here::here("data", "Diet.xlsx"))
head(Diet_df)
```

```
# A tibble: 6 x 2
   mCPP Placebo
  <dbl>   <dbl>
1    69      45
2   119     140
3     0      10
4    48      95
5    65     145
6   119     133
```

Summarise and visualise the data



Summarise and visualise the differences data

Create a new variable for the difference,

```
Diet_df$Difference <- Diet_df$mCPP - Diet_df$Placebo
Diet_df$Difference
```

```
[1]  24 -21 -10 -47 -80 -14   4 -41  33
```

Summarise and visualise the differences data

The summary statistics of the differences are:

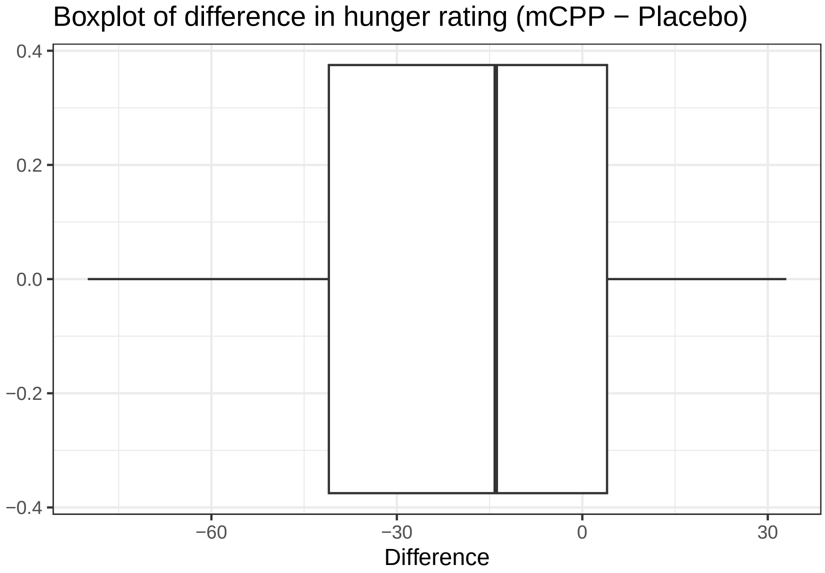
```r
summary(Diet_df$Difference)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -80.00  -41.00  -14.00  -16.89    4.00   33.00 
```

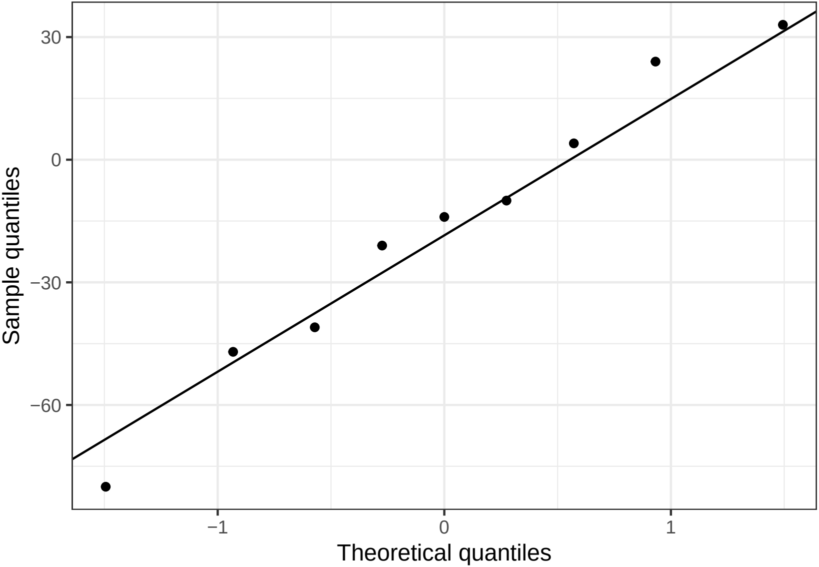
```r
sd(Diet_df$Difference)
```

```
[1] 35.60353
```

Summarise and visualise the differences data



Checking assumptions



What about the other assumptions?

Perform a one-sample t-test on the difference

```
t.test(Diet_df$Difference)
```

```

	One Sample t-test

data:  Diet_df$Difference
t = -1.4231, df = 8, p-value = 0.1925
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -44.25618  10.47840
sample estimates:
mean of x
-16.88889
```

Theory underlying t.test output- Test statistic

If we have matched pairs data and we are interested in testing if there was a difference for one treatment compared to the other, then we use

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}.$$

where

- $\bar{d}$ is the mean of the sample of differences,
- s_d is the standard deviation of the sample of differences,
- n is the number of subjects.

Theory underlying t.test output- p-value

The p-value is computed using a t-distribution with $n-1$ degrees of freedoms, where n is the number of subjects.

Theory underlying t.test output - Confidence interval

The formula for a $C\%$ confidence interval is given by,

$$\bar{d} \pm t^* \times \frac{s_d}{\sqrt{n}}$$

where t^* is the appropriate critical value to give a $C\%$ confidence level.

These confidence intervals are discussed in Forum 04.

Interpreting the results and concluding

- Using the p-value:

There is insufficient evidence to reject the null hypothesis since the p-value of 0.1925 is greater than the significance level of 0.05.

Therefore, at a significance level of 0.05 we cannot conclude that there is a difference between the hunger rating when given mCPP and the hunger rating when given the placebo.

Note: This does not mean that we accept the null hypothesis.

Interpreting the results and concluding

- From the confidence interval:

Based on a two-sided one-sample t-test, we are 95% confident that the true difference in hunger ratings (mCPP minus placebo) lies between -44.26 and 10.48.

Since this interval includes zero, at 5% significance level we cannot conclude that there is a difference between the population mean of hunger rating when given mCPP and the hunger rating when given the placebo.

Summary

Hypothesis testing for mean for matched pairs data

- **Hypotheses:**

$$H_0 : \mu_D = 0,$$
$$H_1 : \mu_D \neq 0.$$

- **Test statistic:**

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}$$

- **P-value:** Calculate using a t-distribution with $n - 1$ degrees of freedom.

- **Confidence interval**

$$\bar{d} \pm t^* \times \frac{s_d}{\sqrt{n}}.$$

One-sample t-test

One-sample t-test

A one-sample t-test determines if there is a statistically significant difference between a population mean μ and a specified value.

One-sample t-test

- In this course, **we consider the typical scenario where the population standard deviation, σ, is not known**, and use a t-test.

- If the population standard deviation, σ, is known then one can apply a z-test.

One-sample t-test conditions

Conditions:

- The sample is a simple random sample.
- Population is normally distributed, or the sample size $n > 30$.
- The samples are independent.

Example: Calcium

Estimated daily intake of calcium was recorded for 40 women between the ages of 18 and 24 years who participated in a study of women's bone health. The recommended daily intake is 1000mg.

Perform the appropriate hypothesis test using this sample to determine if the population mean μ differs significantly at the 5% significance level from the recommended value.

State the null and alternative hypotheses

- The null hypothesis, H_0, corresponds to there is no effect or no difference.
- The alternative hypothesis, H_1, corresponds to there is an effect or a difference.

What is the null and alternative hypotheses for the calcium example?

Example: Calcium - Null and alternative hypotheses

- H_0: $\mu = 1000$,
- H_1: $\mu \neq 1000$,

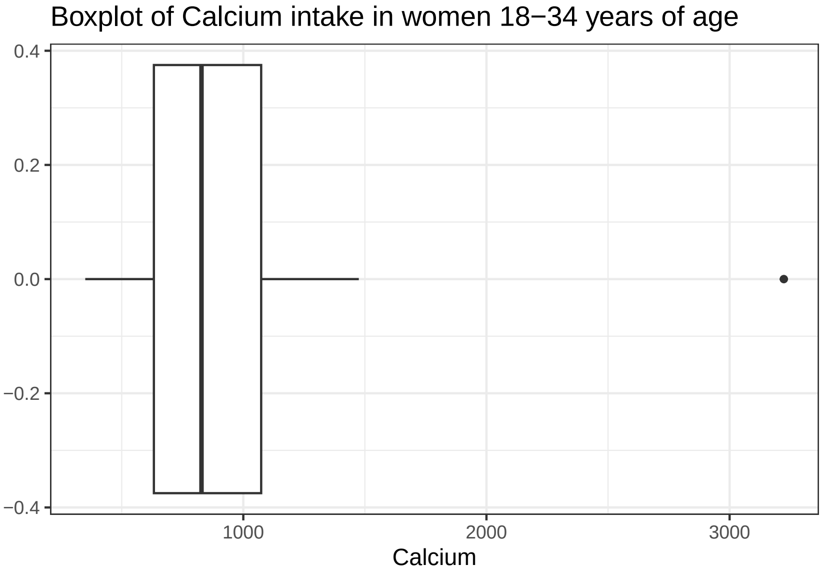
where μ is the population mean of daily intake of calcium (measured in mg) for a population of women and the recommended daily intake is 1000mg.

Load and view the data

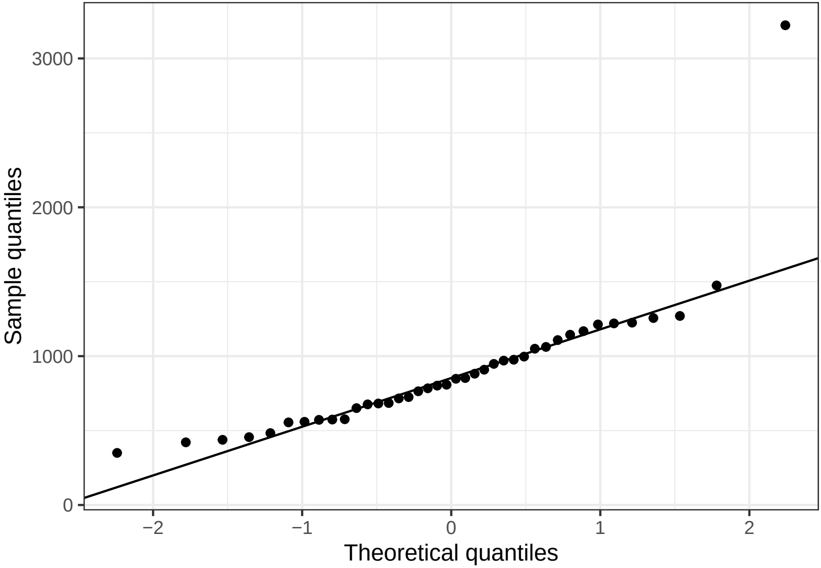
```r
Calcium_df <- read_excel(here::here("data", "calcium.xlsx")
head(Calcium_df)
```

```
# A tibble: 6 x 1
  Calcium
    <dbl>
1     725
2     764
3     853
4     559
5    1225
6     456
```

Summarise and visualise the data



Checking assumptions - Normality



Checking assumptions

Since the sample size $n = 40$ is greater than 30, this suggests that a one-sample t-test may be performed, but **the result may be affected by the presence of the outlier**.

What about the other assumptions?

Perform a one-sample t-test

```r
t.test(Calcium_df$Calcium,mu=1000)
```

```
    One Sample t-test

data:  Calcium_df$Calcium
t = -1.3267, df = 39, p-value = 0.1923
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
  752.2064 1051.4936
sample estimates:
mean of x
   901.85
```

Interpreting the results and concluding

- Using the p-value:

There is insufficient evidence to accept the alternative hypothesis since the p-value of 0.19 is greater than the significance level of 0.05.

Therefore, at a significance level of 0.05 we cannot conclude that there is a difference between the population mean of daily intake from the women and the recommended daily intake of 1000mg.

Note: This does not mean that we can accept the null hypothesis.

Interpreting the results and concluding

- From the confidence interval:

Based on the two-sided one-sample t-test, we are 95% confident that the population mean of daily intake from the women lies between 752.2 1051.4mg.

Since this interval includes the recommended daily intake of 1000mg, at 5% significance level we cannot conclude that there is a difference between the population mean of daily intake from the women and the recommended daily intake of 1000mg.

What should we do about the outlier?

The outlier invalidates the conditions for the one-sample t-test.

What should we do about the outlier?

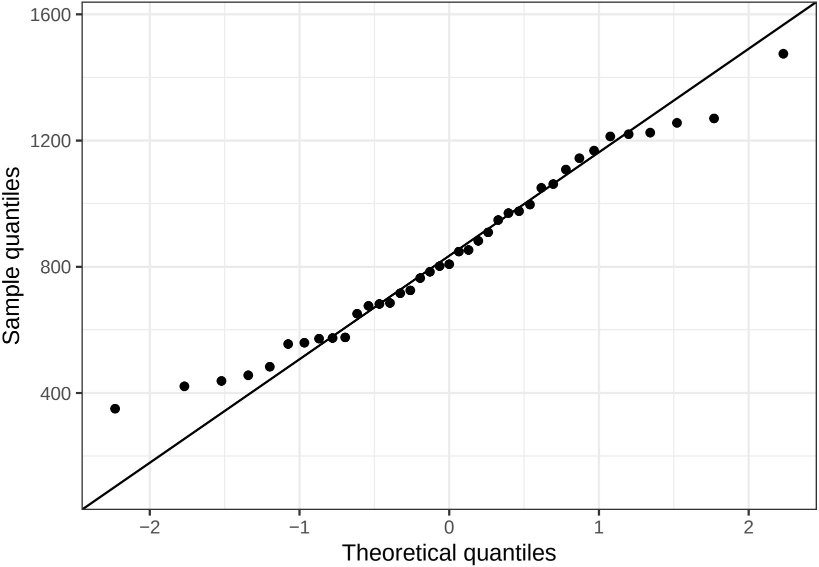
- If the outlier is a real measurement then we should explore a non-parametric hypothesis test (see Forum 6).
- If the outlier is confirmed to be an error you should aim to correct it. If correction is not possible, it could be removed but this would have to be documented and justified. The impact of the outlier should be assessed.

Re-performing the analysis with the outlier removed

Introduce a new variable which contains all measurements except the outlier. In this case the outlier is the largest value which we can remove by the following,

```
Calcium_outlierremoved_df <- Calcium_df %>%
                                arrange(Calcium) %>%
                                slice_head(n = 39)
```

Checking assumptions - Normality



What do we now notice? What about the other assumptions?

Perform a one-sample t-test

```r
t.test(Calcium_outlierremoved_df$Calcium,mu=1000)
```

```

    One Sample t-test

data:  Calcium_outlierremoved_df$Calcium
t = -3.497, df = 38, p-value = 0.001215
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
 751.0602 933.6065
sample estimates:
mean of x
 842.3333
```

Interpreting the results and concluding

- What is conclusion now?
- How much influence does the outlier have?

Summary - One-sample t-test

Hypothesis testing for mean of a normal distribution with unknown population standard deviation

- **Hypotheses:**

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

- **Test statistic:**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

- **p-value:** Calculate using a t-distribution with $n-1$ degrees of freedom (“p-value’ ’ in R output)
- **Confidence interval:**

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}.$$

Further reading/tools

- Textbook:
  - Chapter 7.1 Inference for the mean of a population
  - Chapter 7.2 Comparing two means