# Module 4 solutions

## Activity 4.1

In each of the following situations, what decision should be made about the null hypothesis if the researcher indicates that:

a) $P < 0.01$
b) $P > 0.05$
c) ``ns"
d) ``significant differences exist"

### Answers

a) There is strong evidence against the null hypothesis.
b) There is weak or little evidence against the null hypothesis - but the researchers should be advised to provide the actual P-value, not just $P > 0.05$.
c) Traditionally, 'ns' stands for not significant (for the set level of significance mentioned in the study, usually 0.05). You might still come across this term in some journal articles but this is not best practice for most journals these days. Researchers should always state the P-value (not just whether or not it was significant).
d) This would imply that the P-value is less than the set level of significance mentioned in the study (usually, 0.05). As such, we would conclude that there was evidence against the null hypothesis. However, the researchers should be advised to always state the P-value (not just whether or not it was significant).

## Activity 4.2

For the following hypothetical situations, formulate the null hypothesis and alternative hypothesis and write a conclusion about the study results:

a) A study was conducted to investigate whether the mean systolic blood pressure of males aged 40 to 60 years was different to the mean systolic blood pressure of females aged 40 to 60 years. The result of the study was that the mean systolic blood pressure was higher in males by 5.1 mmHg (95% CI 2.4 to 7.6; $P = 0.008$).

b) A case-control study was conducted to investigate the association between obesity and breast cancer. The researchers found an OR of 3.21 (95% CI 1.15 to 8.47; P = 0.03).
c) A cohort study investigated the relationship between eating a healthy diet and the incidence of influenza infection among adults aged 20 to 60 years. The results were RR = 0.88 (95% CI 0.65 to 1.50; P = 0.2).

### Answers

a) **$H_0$**: There is no difference in the mean systolic blood pressure between males and females aged 40-60 years.

   **$H_A$**: There is a difference in the mean systolic blood pressure between males aged 40-60 years and females aged 40 to 60 years.

   **Conclusion**: The mean SBP was 5.1 mmHg (95% CI: 2.4 to 7.6 mmHg) higher in males aged 40-60 years compared to females aged 40-60 years. There is strong evidence of a difference in the mean SBP of males and females aged 40-60 years (P=0.008).

b) **$H_0$**: There is no association between obesity and breast cancer.

   *[An alternative way of saying this is that there is no difference in the odds of exposure to obesity among cases of breast cancer and controls i.e. OR = 1].*

   **$H_A$**: There is an association between obesity and breast cancer.

   *[An alternative way of saying this is that there is a difference in the odds of exposure to obesity among cases and controls i.e. OR ≠ 1].*

   **Conclusion**: The odds ratio is estimated as 3.21 (95% CI: 1.15 to 8.47), indicating a positive association between the study factor of obesity and the outcome of breast cancer. There is evidence of a positive association between obesity and breast cancer (P=0.03).

c) **$H_0$**: There is no association between influenza infection and a healthy diet among adults aged 20-60 years.

   *[An alternative way of saying this is that there is no difference in the risk of influenza infection among adults aged 20-60 years who have a healthy diet compared to those who do not have a healthy diet. i.e. RR = 1].*

   **$H_A$**: There is an association between influenza infection and a healthy diet among adults aged 20-60 years.

   *[An alternative way of saying this is that there is a difference in the risk of influenza infection among adults aged 20-60 years who have a healthy diet compared to those who do not have a healthy diet. i.e. RR ≠ 1].*

   **Conclusion**: The relative risk is estimated as 0.88 (95% CI: 0.65 to 1.50), indicating a protective association between the study factor of healthy diet and the outcome of influenza infection among adults aged 20 to 60 years. There is no evidence of an association between a healthy diet and influenza infection among adults aged 20 to 60 years (P=0.2).

*[Note that the 95% confidence interval includes the null value of 1, which is consistent with the P-value being greater than 0.05]*

## Activity 4.3

A pilot study was conducted to compare the mean daily energy intake of women aged 25 to 30 years with the recommended intake of 7750 kJ/day. In this study, the average daily energy intake over 10 days was recorded for 12 healthy women of that age group. The data are in the the Excel file Activity_4.3.xls. Import the file into Stata or R for this activity.

a) State the research question
b) Formulate the null hypothesis
c) Formulate the alternative hypothesis
d) Analyse the data in Stata or R and report your conclusions

## Answers

a) Is the mean daily energy intake of women aged 25-30 years different to the recommended daily intake of 7750 kJ/day?
b) $\mathbf{H_0}$: the mean daily energy intake of women aged 25-30 years is the same as the recommended daily intake of 7750 kJ/day.
c) $\mathbf{H_A}$: the mean daily energy intake of women aged 25-30 years is not same as the recommended daily intake of 7750 kJ/day.
d) The mean daily energy intake of the 12 women is 6856 kJ (95% CI: 6131 to 7581 kJ). There is evidence that the mean daily energy intake of women aged 25-30 years is lower than the recommended daily intake of 7750 kJ/day ($t = -2.71$ with 11 DF, $P = 0.02$).

## Process

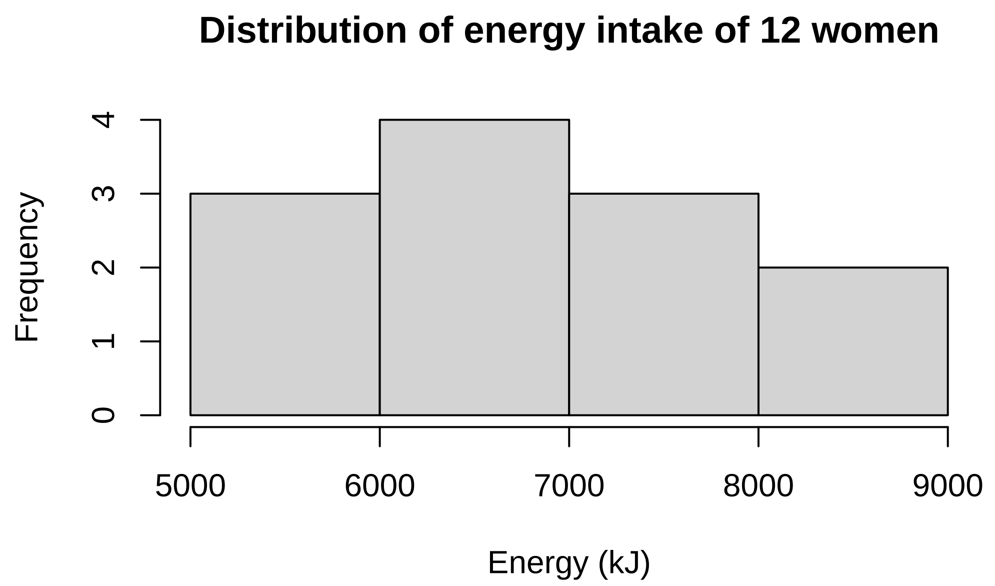
As we are comparing a continuous distribution to a hypothesised mean, we will use a one-sample t-test to conduct this analysis. As the one-sample t-test assumes our data follow a Normal distribution, we should assess this using a histogram.

```
library(readxl)

energydata <- read_excel("data/activities/Activity_S4.3.xls")

hist(energydata$Energy,
```

```
    main="Distribution of energy intake of 12 women",
    xlab="Energy (kJ)")
```




It is very difficult to assess the shape of a distribution with only 12 observations, but here we can see that the distribution looks roughly symmetric. In this case, we will assume Normality.

We can use the `t.test` function to obtain the sample mean, its 95% confidence interval and perform a one-sample t-test to compare the variable Energy to the hypothesised mean of 7750 kJ/day:

```
t.test(energydata$Energy, mu=7750)
```

```

	One Sample t-test

data:  energydata$Energy
t = -2.7141, df = 11, p-value = 0.02014
alternative hypothesis: true mean is not equal to 7750
95 percent confidence interval:
 6131.023 7580.977
sample estimates:
mean of x 
     6856 
```

### Activity 4.4

Which procedure gives the researcher the better chance of rejecting a null hypothesis?

a) comparing the data-based p-value with the level of significance at 5%
b) comparing the 95% CI with a nominated value
c) neither procedure

Both 'a' and 'b' would give the same chance to reject the null hypothesis. This is because both 'a' and 'b' are giving you the same information in a different way. In 'a' you will get the probability of observing the difference you see in your data by chance and if it is <0.05 you will reject the null hypothesis at the 5% level. Whereas in 'b' you will see whether the null value (value of no difference) lies within the range which you are 95% confident contains the true value. If the null value falls outside the 95% CI, you would have less than 5% (100-95 = 5%) probability seeing the observed difference in your data if there were no difference.

However, the best approach is to describe the P-value in terms of the strength of evidence against the null hypothesis. That is, the P-value should be interpreted based on Table 4.1 of the course notes.

### Activity 4.5

Setting the significance level at $P < 0.10$ instead of the more usual $P < 0.05$ increases the likelihood of:

a) a Type I error
b) a Type II error
c) rejecting the null hypothesis
d) Not rejecting the null hypothesis

Setting the significance level cut-off at 0.10 instead of the more usual 0.05 increases the likelihood of

a. Type I error and c. rejecting the null hypothesis

The cut-off of 0.10 increases the chance of a Type I error from 5% to 10% (the chance of making a Type I error is the same as the significance level). If the significance level is higher, then there is a higher probability of rejecting the null hypothesis if there no effect in reality.

### Activity 4.6

For a fixed sample size setting the significance level at a very extreme cutoff such as $P < 0.001$ increases the chances of:

a) obtaining a significant result
b) rejecting the null hypothesis

c) a Type I error
d) a Type II error

Setting the significance level at a very extreme cut-off (such as 0.001) increases the chances of:

d. a Type II error

For a given sample, if the significance level is set very small it will make it harder to find evidence against the null hypothesis. In other words, it will be difficult to detect an effect if an effect exists in reality. In other words, the probability of Type II error will increase: you will not be able to reject the null hypothesis when a real difference exists.