PHCM9795 Foundations of Biostatistics

Learning activity solutions: R version

16 July, 2022

Contents

Introduction

These notes provide R-based solutions to the learning activities in Foundations of Biostatistics. These notes are currently under development, with sections being added and revised as the course progresses.
This is the first year that R has been offered as an option. I am keen to receive feedback about the notes and your experience learning R. Please get in touch if anything is unclear, or you have any questions or suggestions.

Changelog

2022-07-16 [Added]

- Module 8: Initial release

2022-07-11 [Added]

- Module 7: Initial release

2022-07-04 [Added]

- Module 6: Initial release

2022-06-21 [Added]

- Module 5: Initial release

2022-06-12 [Added]

- Module 3: Initial release
- Module 4: Initial release

2022-06-06 [Added]

- Module 2: Initial release

[Changed]

- Various typos

2022-05-30
[Added]

- Module 1: Initial release

Module 1: Solutions to Learning Activities

Activity 1.1

25 participants were enrolled in a 3-week weight loss programme. The following data present the weight loss (in grams) of the participants:

255	198	283	312	283
57	85	312	142	113
227	283	255	340	142
113	312	227	85	170
255	198	113	227	255

a) Enter these data into R.

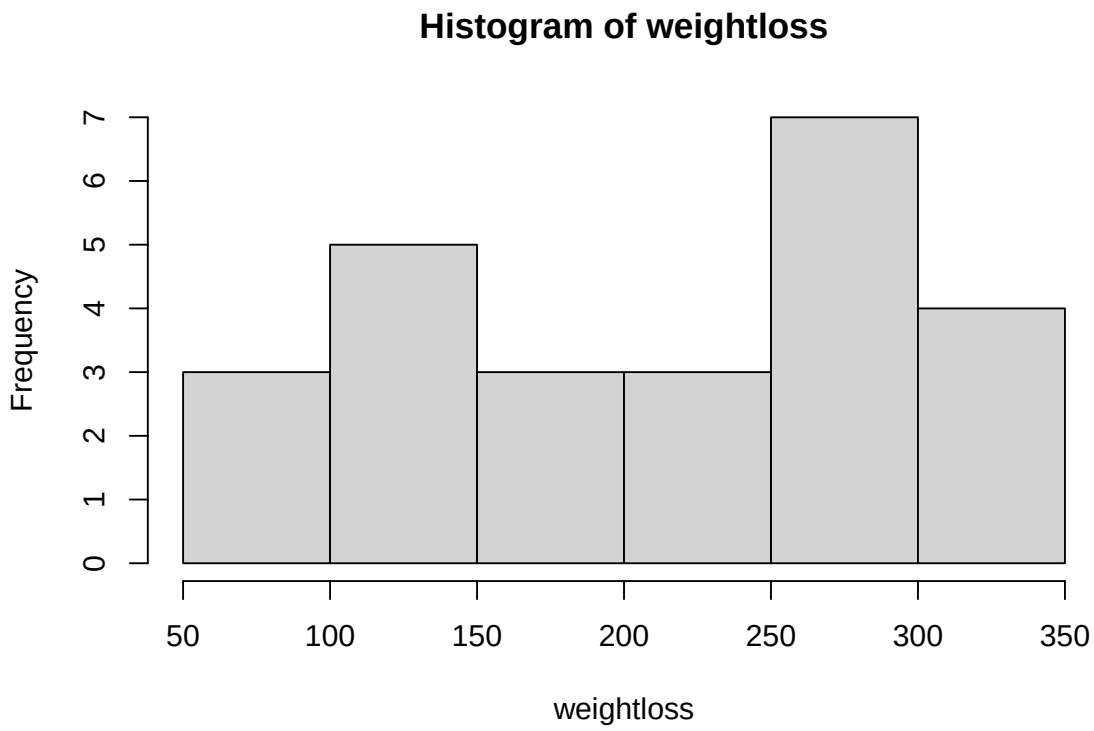
```
weightloss <- c(255, 198, 283, 312, 283, 57,  85, 312, 142, 113,
                227, 283, 255, 340, 142, 113, 312, 227,  85, 170,
                255, 198, 113, 227, 255)
```

b) What type of data are these?

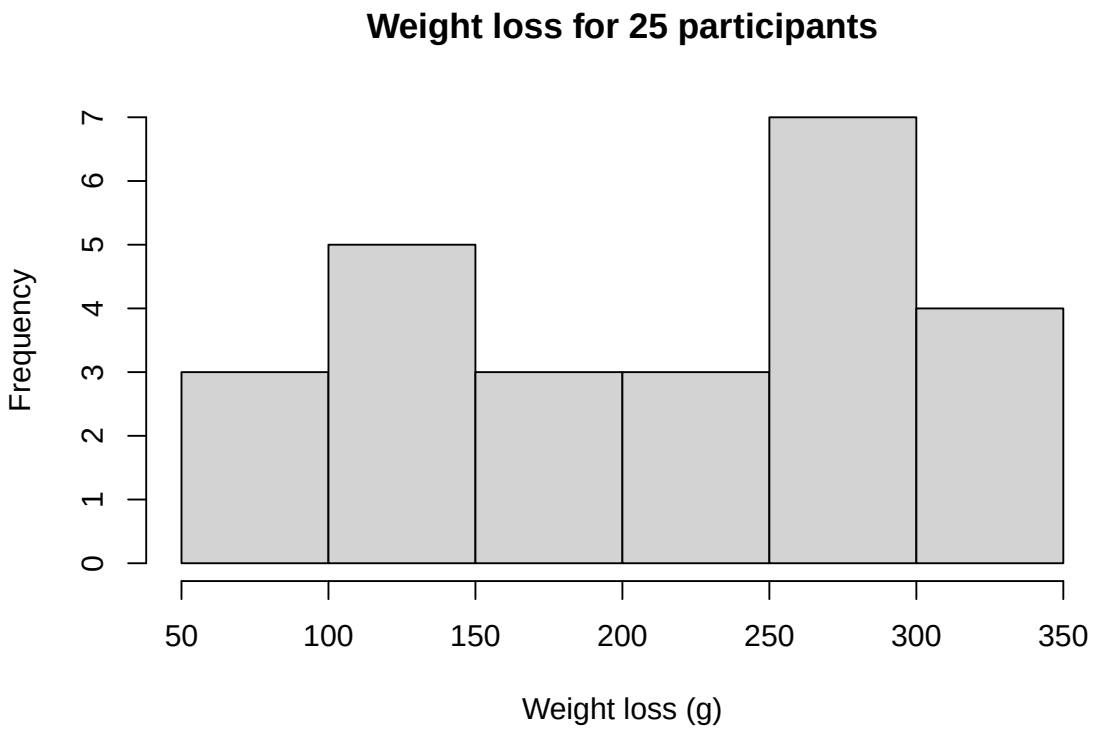
These are continuous numeric data.

c) Construct an appropriate graph to display the relative frequency of participants' weight loss. Your graph should start at 50 grams, with weight loss grouped into 50 gram bins. Provide appropriate labels for the axes and give the graph an appropriate title.

```
# Check the default histogram:
hist(weightloss)
```



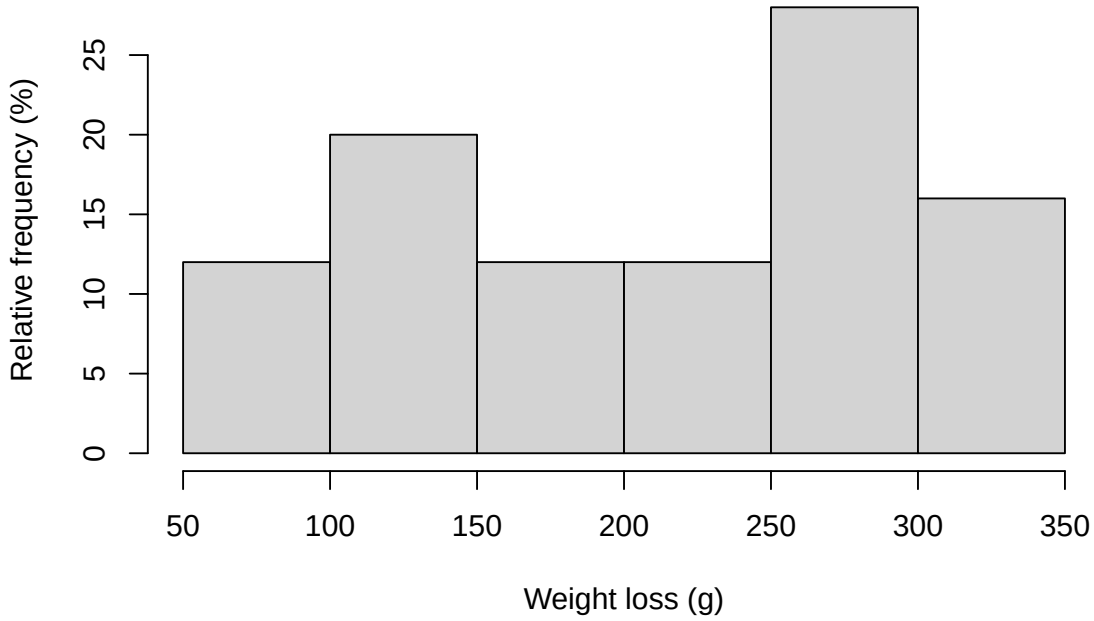

```
# The default values look ok, so let's add labels and titles
hist(weightloss, xlab="Weight loss (g)", main="Weight loss for 25 participants")
```



Note that the question requests **relative frequencies**, so we can use the code in Section 1.12 to amend this graph:

```
h <- hist(weightloss, plot=FALSE)
h$density <- h$counts/sum(h$counts)*100
plot(h, freq=FALSE,
     xlab="Weight loss (g)",
     ylab="Relative frequency (%)",
     main="Fig 1.1: Weight loss for 25 participants")
```

Fig 1.1: Weight loss for 25 participants



Activity 1.2

Researchers at a maternity hospital in the 1970s conducted a study of low birth weight babies. Low birth weight is classified as a weight of 2,500g or less at birth. Data were collected on age and smoking status of mothers and the birth weight of their babies. The file `Activity_S1.2.rds` contains data on the participants in the study. The file is located on Moodle in the Learning Activities section.
Use R to create a 2 by 2 table to show the proportions of low birth weight babies born to mothers who smoked during pregnancy and those that did not smoke during pregnancy.

```
library(jmv)

babies <-readRDS("data/activities/Activity_S1.2.rds")

# Examine the first six rows of data
head(babies)
```

```
##   AGE    AgeGrp  BWT                 LOW SMOKE
## 1  14 <20 years 2466    Low birth weight   Yes
## 2  14 <20 years 2495    Low birth weight    No
## 3  14 <20 years 3941 Normal birth weight    No
## 4  15 <20 years 2353    Low birth weight    No
## 5  15 <20 years 2381    Low birth weight    No
## 6  15 <20 years 2778 Normal birth weight    No
```

```
# Create a two-way table showing row percents
contTables(data=babies, rows=SMOKE, cols=LOW, pcRow=TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    SMOKE                    Low birth weight    Normal birth weight    Total
##
##    Yes      Observed                      30                     44           74
##             % within row            40.54054               59.45946    100.00000
##
##    No       Observed                      29                     86          115
##             % within row            25.21739               74.78261    100.00000
##
##    Total    Observed                      59                    130          189
##             % within row            31.21693               68.78307    100.00000
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²     4.923705     1    0.0264906
##    N           189
##
```

Answer the following questions:

a) What was the total number of mothers who smoked during pregnancy?

   There were 74 mothers who smoked during pregancy.

b) What proportion of mothers who smoked gave birth to low birth weight babies? What proportion of non-smoking mothers gave birth to low birth weight babies?

   41% of mothers who smoked and 25% of non-smoking mothers gave birth to low birth weight babies.

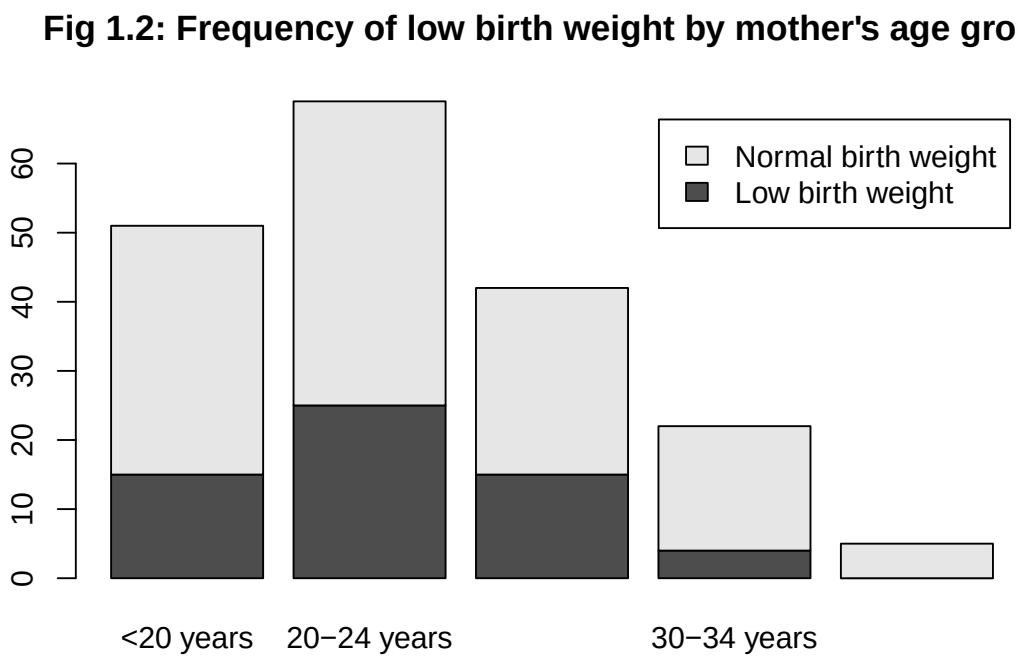
c) Use R to construct a stacked bar chart of the data to examine if there a difference in the proportion of babies born with a low birth weight in relation to mother's age? Provide appropriate labels for the axes and give the graph an appropriate title.

   We follow the instructions for creating a stacked bar chart in Module 1. First we create a table of low birth weight by mothers' age-group, and create a stacked bar chart (to check that we're on the right track):

```
counts <- table(babies$LOW, babies$AgeGrp)
counts
```

```
##
##                      <20 years 20-24 years 25-29 years 30-34 years
##   Low birth weight           15          25          15           4
##   Normal birth weight        36          44          27          18
##
##                      35 or more years
##   Low birth weight                   0
##   Normal birth weight                5
```

```
barplot(counts,
        main="Fig 1.2: Frequency of low birth weight by mother's age group",
        legend = rownames(counts), beside=FALSE)
```



We then calculate the *relative frequency* of low-birth weight by mothers' age group

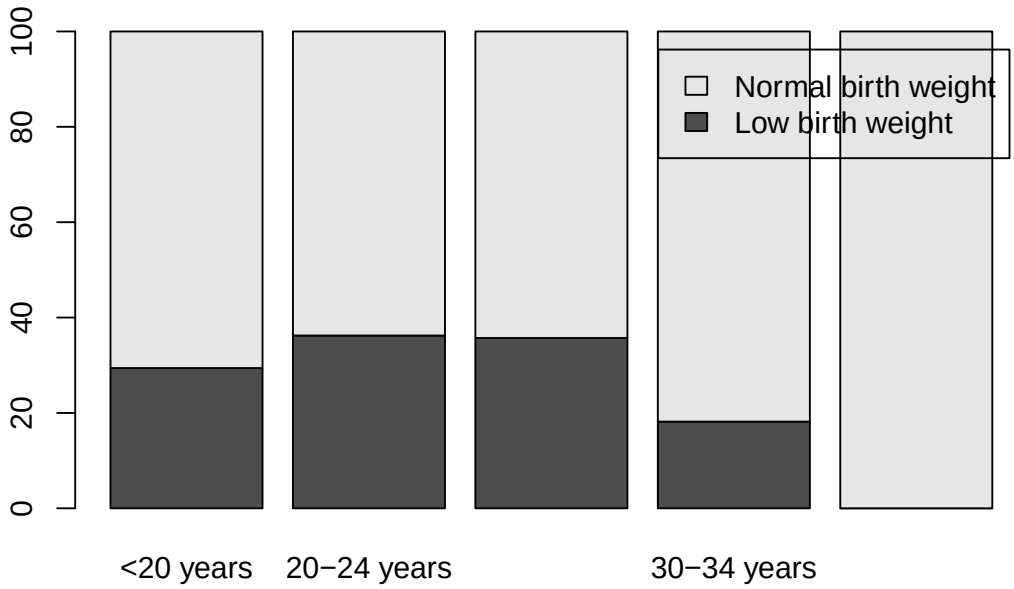
```
percent <- prop.table(counts, margin=2)*100
percent
```

```
##
##                       <20 years 20-24 years 25-29 years 30-34 years
##   Low birth weight     29.41176    36.23188    35.71429    18.18182
##   Normal birth weight  70.58824    63.76812    64.28571    81.81818
##
##                       35 or more years
##   Low birth weight             0.00000
##   Normal birth weight        100.00000
```

and use the `barplot()` command, as per the notes:

```
barplot(percent,
        main="Fig 1.3: Relative frequency of low birth weight by mother's age group",
        legend = rownames(percent), beside=FALSE)
```

Fig 1.3: Relative frequency of low birth weight by mother's age grou



d) Using your answers to the question a) and b), write a brief conclusion about the relationship of low birth weight and mother's age and smoking status.

In the study, the greatest number of babies were born to mothers in the 20-24 years age group, with the number of babies born declining with increasing maternal age for mothers older than 20-24 years (Figure 1.2). A larger proportion of mothers in the <20 years, 20-24 years and 25-29 years age groups gave birth to low birth weight babies compared to mothers aged 30-34 years. No low birth weight babies were born to mothers aged 35 or more (Figure 1.3).

A larger proportion of mothers who smoked during pregnancy gave birth to low birth weight babies compared to mothers who did not smoke during pregnancy.

NB: You will revisit two-way tables in Module 7 where you will conduct statistical tests to determine if the proportions are statistically different to each other.

Note: Coding graphs, particularly clustered and stacked bar graphs can be difficult! The site https://r-graph-gallery.com/ gives excellent instructions on constructing different types of graphs in R.

Activity 1.3

Using R, estimate the mean, median, mode, standard deviation, range and interquartile range for the data Activity_S1.3.rds, available on Moodle.

```
act1_3 <- readRDS("data/activities/Activity_S1.3.rds")

descriptives(act1_3, mode=TRUE, iqr=TRUE, pc=TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                          Lead_concn
##
##    N                            15
##    Missing                       0
##    Mean                   1.500000
##    Median                 1.500000
##    Mode                   1.900000
##    Standard deviation     0.8434623
##    IQR                    1.0000000
##    Minimum                0.1000000
##    Maximum                3.200000
##    25th percentile        0.9500000
##    50th percentile        1.500000
##    75th percentile        1.950000
##
```

We can use the `descriptives()` function to obtain summary statistics. Examining the help entry for `descriptives()` shows we can request the mode using `mode=TRUE`, the interquartile range using `iqr=TRUE` and the percentiles (by default, the quartiles) using `pc=TRUE`. The mean is estimated as 1.50, the median is 1.5 and the mode is 1.9. The standard deviation is estimated as 0.843, the range is from 0.1 to 3.2, and the inter-quartile range is from 1.0 to 2.0 (both rounded to 1 decimal place).

Note: no units were provided for the data used in this question. Summary statistics must be presented with their units where the units are available.

Activity 1.4

Data of diastolic blood pressure (BP) of a sample of study participants are provided in the dataset Activity_S1.4.rds. Compute the mean, median, range and SD of diastolic BP.

```
act1_4 <- readRDS("data/activities/Activity_S1.4.rds")

descriptives(act1_4)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                          diabp
##
##    N                       100
##    Missing                   0
##    Mean                 82.23000
##    Median               83.00000
##    Standard deviation   13.01522
##    Minimum              56.00000
##    Maximum              118.0000
##
```

The mean is 82.2 mmHg and the median is 83.0 mmHg. The range is 56.0 to 118.0 mmHg (62.0 mmHg) and the standard deviation is 13.02 mmHg.

Note that the original data have one decimal place, so we can report the median with one decimal place. Although we are justified in presenting the mean to two decimal places (1 extra than the original data), and the standard deviation with three decimal places (1 more than the mean), there is little to be gained in this level of precision when presenting summary statistics for blood pressure.

Activity 1.5

In a study of 100 participants data were missing for 5 people. The missing data points were coded as '99'. The mean of the data was estimated as 45.0 with a standard deviation of 5.6; the smallest and greatest values are 16 and 65 respectively.
If the researcher analysed the data as if the 99s were real data, would it make the following statistics larger, smaller, or stay the same?

a) Mean

   The mean will be larger.

b) Standard Deviation

   The standard deviation will be larger.

c) Range

   The range will be larger. The smallest value is still 16, but the largest is 99, and so the range is 99 − 16 = 83.

Activity 1.6

Which of the following statements are true? The more dispersed, or spread out, a set of observations are:

a) The smaller the mean value

   This is not true because the mean is not influenced by the spread of the values (if the distribution is symmetrical around the mean value)

b) The larger the standard deviation

   This is true. The larger the spread, the larger the deviations from the mean. Hence the standard deviation will be larger.

c) The smaller the variance

   This is not true. The variance will be larger if the deviations from the mean are larger.

Activity 1.7

If the variance for a set of scores is equal to 9, what is the standard deviation?

SD = $\sqrt{variance} = \sqrt{9} = 3.$

Module 1: Full script

```
# Author: Timothy Dobbins
# Date: May, 2022
# Purpose: Learning activities for Module 1

library(jmv)

### Activity 1.1

weightloss <- c(255, 198, 283, 312, 283, 57,  85, 312, 142, 113,
                227, 283, 255, 340, 142, 113, 312, 227,  85, 170,
                255, 198, 113, 227, 255)
# Check the default histogram:
hist(weightloss)

# The default values look ok, so let's add labels and titles
hist(weightloss, xlab="Weight loss (g)", main="Weight loss for 25 participants")

# Construct a relative frequency histogram
h <- hist(weightloss, plot=FALSE)
h$density <- h$counts/sum(h$counts)*100
plot(h, freq=FALSE,
     xlab="Weight loss (g)",
     ylab="Relative frequency (%)",
     main="Fig 1.1: Weight loss for 25 participants")


### Activity 1.2
babies <-readRDS("data/activities/Activity_S1.2.rds")

# Examine the first six rows of data
head(babies)

# Create a two-way table showing row percents
contTables(data=babies, rows=SMOKE, cols=LOW, pcRow=TRUE)

# Construct bar charts
counts <- table(babies$LOW, babies$AgeGrp)
counts

barplot(counts,
        main="Fig 1.2: Frequency of low birt weight by mother's age group",
        legend = rownames(counts), beside=FALSE)
```

```
percent <- prop.table(counts, margin=2)*100
percent

barplot(percent,
        main="Fig 1.3: Relative frequency of low birth weight by mother's age group",
        legend = rownames(percent), beside=FALSE)


### Activity 1.3

act1_3 <- readRDS("data/activities/Activity_S1.3.rds")

descriptives(act1_3, mode=TRUE, iqr=TRUE, pc=TRUE)


### Activity 1.4

act1_4 <- readRDS("data/activities/Activity_S1.4.rds")

descriptives(act1_4)
```

Module 2: Solutions to Learning Activities

Activity 2.1

In a Randomised Controlled Trial, the preference of a new drug was tested against an established drug by giving both drugs to each of 90 people. Assume that the two drugs are equally preferred, that is, the probability that a patient prefers either of the drugs is equal (50%). Use one of the binomial functions in R to compute the probability that 60 or more patients would prefer the new drug. In completing this question, determine:

a) The number of trials (n)

   Here, each participant represents a 'trial', so n is 90.

b) The number of successes we are interested in (k)

   We are interested in determining the probability that 60 or more participants prefer the new drug, so k is 60.

c) The probability of success for each trial (p)

   We are told to assume that the two drugs are equally preferred, so p is 0.5.

d) The form of the R function: dbinom or pbinom

   We need to calculate the probability that 60 or more participants prefer the new drug. The two R functions can be interpreted as follows: - the `dbinom` function gives the probability of observing 60 successes; - the `pbinom` function gives the probability of observing 60 or fewer successes; - the `pbinom` function with `lower.tail=FALSE` gives the probability of observing *more than* 60 successes.

   We therefore want to use `pbinom` function with `lower.tail=FALSE` here.

e) The final probability.

   To calculate the probability of obtaining 60 or more successes, we need to calculate the probabibility of observing *more than* 59 successes. So the function we use is:

```
pbinom(q=59, size=90, prob=0.5, lower.tail = FALSE)
```

```
## [1] 0.001030133
```

   Therefore, the probability that 60 or more patients would prefer the new drug is 0.001 or 0.1%.

Activity 2.2

A case of Schistosomiasis is identified by the detection of schistosome ova in a faecal sample. In patients with a low level of infection, a field technique of faecal examination has a probability of 0.35 of detecting ova in any one faecal sample. If five samples are routinely examined for each patient, use R to compute the probability that a patient with a low level of infection:

a) Will not be identified?

In all of these questions, `size` is 5 and `prob` is 0.35. Here we need to calculate the probability of P(X=0), and we can use the `dbinom` function:

```
dbinom(x=0, size=5, prob=0.35)
```

```
## [1] 0.1160291
```

The probability P(X=0) = 0.116 or 11.6%.

b) Will be identified in two of the samples?

The probability P(X=2)= 0. 336 or 33.6%:

```
dbinom(x=2, size=5, prob=0.35)
```

```
## [1] 0.3364156
```

c) Will be identified in all the samples?

The probability P(X=5) = .005 or 0.5%:

```
dbinom(x=5, size=5, prob=0.35)
```

```
## [1] 0.005252187
```

d) Will be identified in at most 3 of the samples?

"At most 3 samples" is the same as 3 or fewer samples, so we can use the pbinom function. The probability P(X≤3) = .946 or 94.6%:

```
pbinom(q=3, size=5, prob=0.35)
```

```
## [1] 0.9459775
```

Activity 2.3

If weights of men are Normally distributed with a population mean μ = 87, and a population standard deviation, σ = 8 kg:

a) What is the probability that a man will weigh 95 kg or more? Draw a Normal curve of the area represented by this probability in the population (i.e. with μ = 87 kg and σ = 8 kg).

The curve representing the desired probability is drawn below, with the region above 95kg shaded to represent the probability of interest. Note that this curve was generated by a computer: a hand-drawn figure is completely acceptable. A hand-drawn figure will probably look much less tidy, but the main thing to notice is that the shaded area looks like it would represent less than 50% of the total curve. Therefore, our final probability should be less than 0.5.

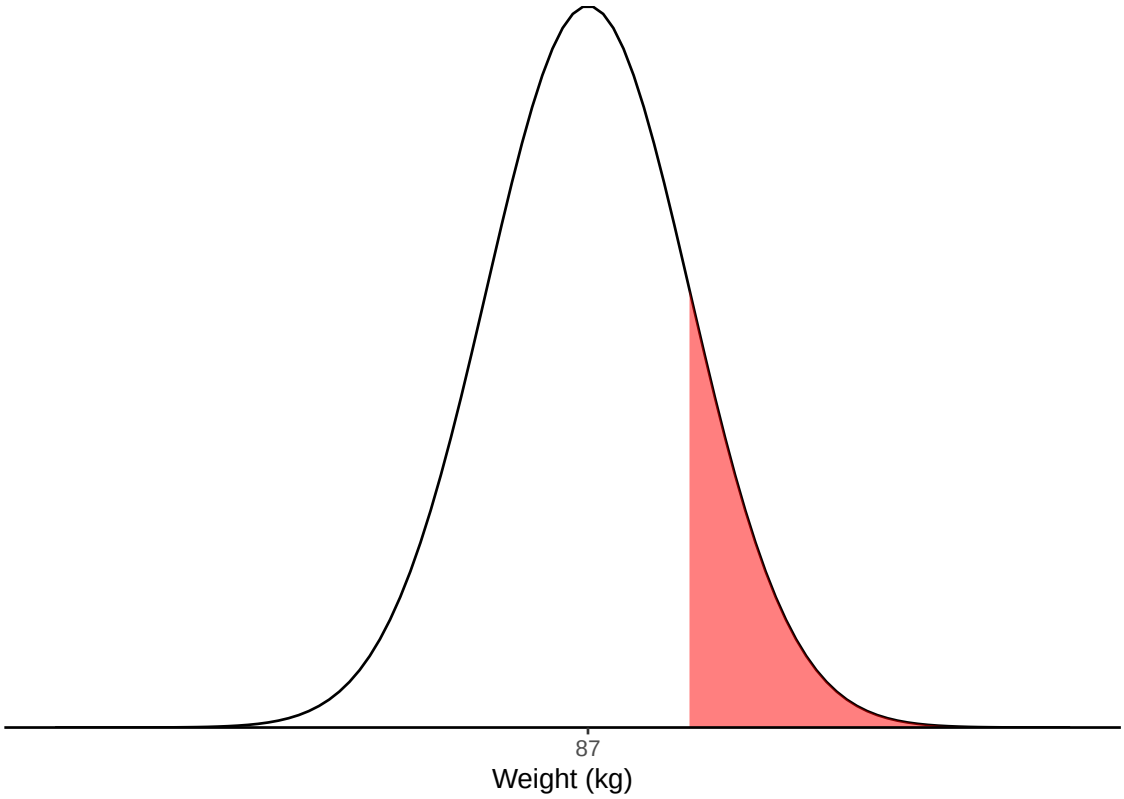


Figure 0.1: Probability that a man will weigh 95kg or more

The probability is calculated as:

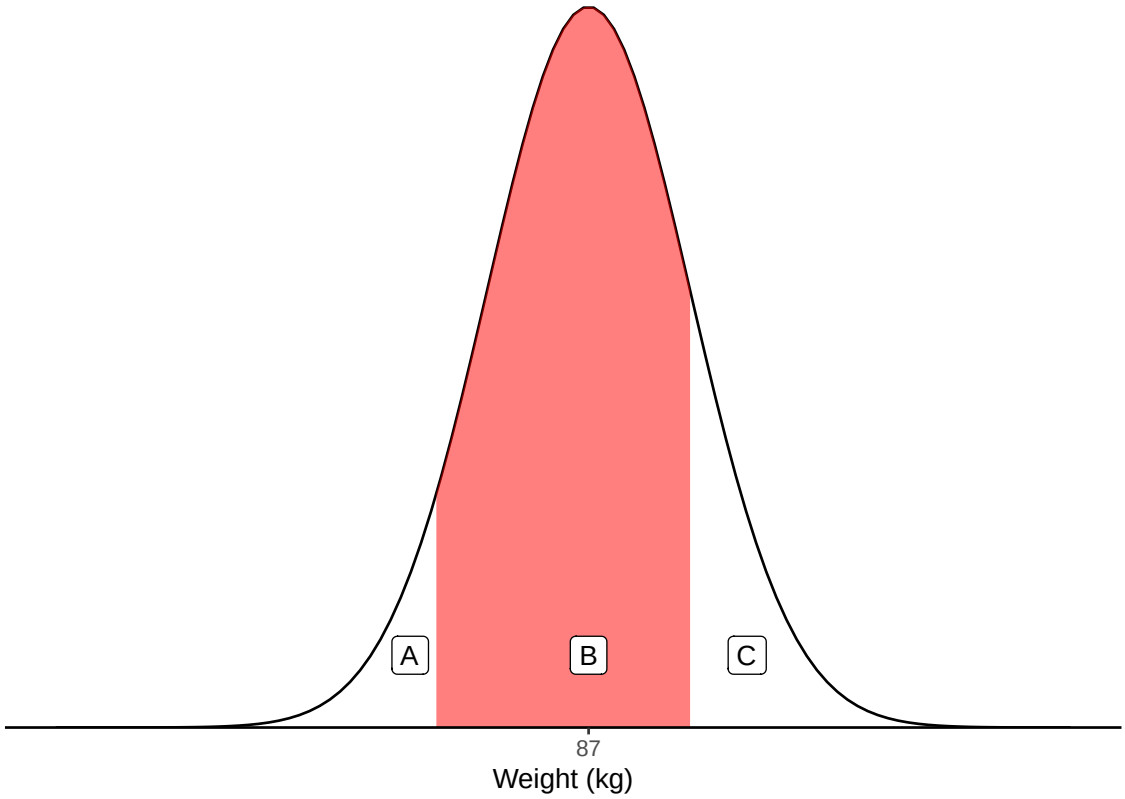
```
# Probability:
pnorm(95, mean=87, sd=8, lower.tail=FALSE)
```

```
## [1] 0.1586553
```

Therefore, the probability that a man from this population weighs 95 kg or more is 0.16 or 16%.

b) What is the probability that a man will weigh more than 75 kg but less than 95 kg? Draw the area represented by this probability on a standardised Normal curve.

The curve to represent this probability is shown below. To obtain the probability represented by the shaded region, we again use the fact that the total area under a Normal curve must add to 1. Let's break the curve into three parts, which we will call A, B and C.



We use that fact that A+B+C=1 to derive that B = 1– A – C. We have already calculated C in Part (a) of this question. To calculate A:

```
pnorm(75, mean=87, sd=8, lower.tail=TRUE)
```

```
## [1] 0.0668072
```

P(Weight < 75) = 0.0668.

The region B is calculated as: 1 - 0.1587 - 0.0668 = 0.7745.

So the probability that a man will weigh more than 75 kg but less than 95 kg is 0.77, or 77%.

Activity 2.4

Using the health survey data described in the R notes of this module, create a new variable, BMI, which is equal to a person's weight (in kg) divided by their height (in metres) squared (i.e. BMI $= \frac{\text{weight (kg)}}{[\text{height (m)}]^2}$. Categorise BMI using the WHO categories provided in the R notes. Create a two-way table to display the distribution of BMI categories by sex (sex: 1 = respondent identifies as male; 2 = respondent identifies as female). Does there appear to be a difference in categorised BMI between males and females?

```
library(readxl)
library(jmv)

survey <- read_excel("data/examples/health-survey.xlsx")
summary(survey)
```

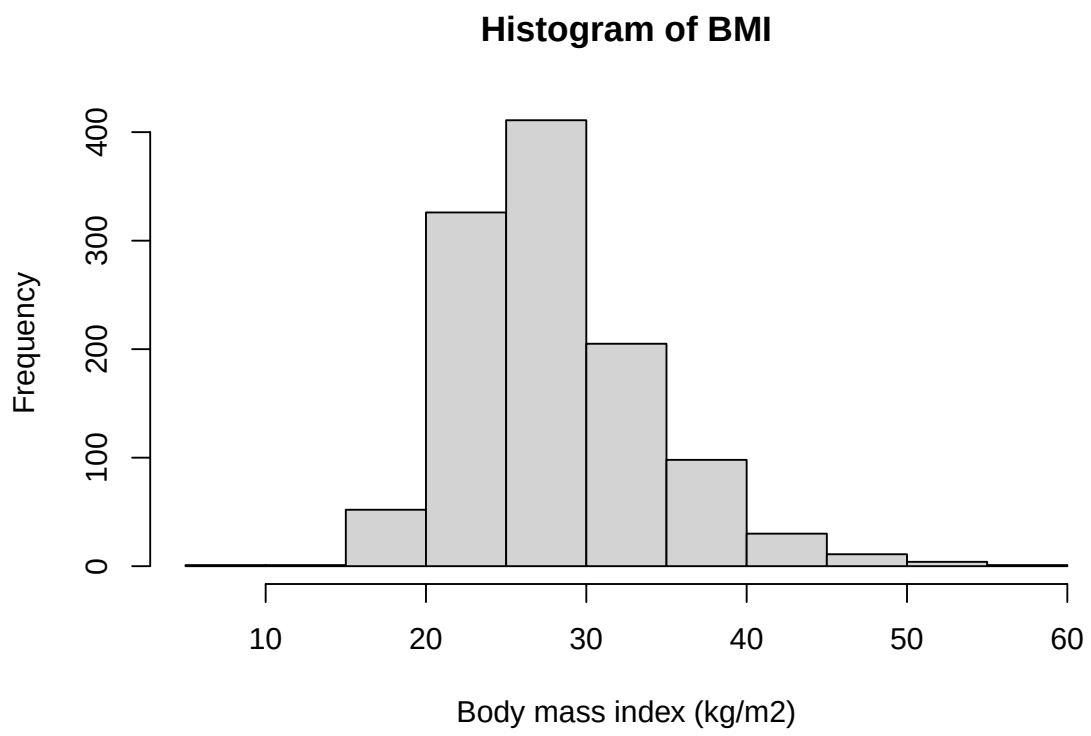
```
##       sex           height          weight
##  Min.   :1.00   Min.   :1.220   Min.   : 22.70
##  1st Qu.:1.00   1st Qu.:1.630   1st Qu.: 68.00
##  Median :2.00   Median :1.700   Median : 79.40
##  Mean   :1.55   Mean   :1.698   Mean   : 81.19
##  3rd Qu.:2.00   3rd Qu.:1.780   3rd Qu.: 90.70
##  Max.   :2.00   Max.   :2.010   Max.   :213.20
```

After reading in the data, we define sex as a factor, and create BMI:

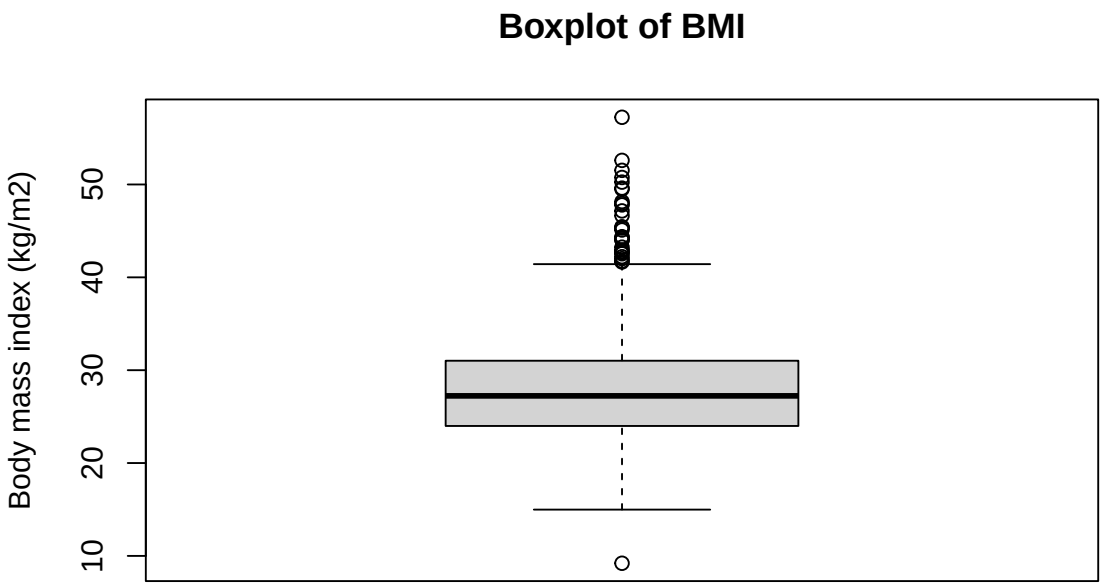
```
survey$sex <- factor(survey$sex, level=c(1,2), labels=c("Male", "Female"))

survey$bmi = survey$weight / (survey$height^2)
```

After creating BMI, we should examine its distribution using a histogram and/or a boxplot:

```
hist(survey$bmi, main="Histogram of BMI", xlab="Body mass index (kg/m2)")
```




```
boxplot(survey$bmi, main="Boxplot of BMI", ylab="Body mass index (kg/m2)")
```



The boxplot in particular shows that there are some extreme values of BMI. We can examine these records by viewing records with BMI less than, say 15, or greater than 45:

```
subset(survey, bmi<15)
```

```
## # A tibble: 2 x 4
##   sex    height weight   bmi
##   <fct>   <dbl>  <dbl> <dbl>
## 1 Female   1.57   22.7  9.21
## 2 Female   1.65   40.8 15.0
```

```
subset(survey, bmi>45)
```

```
## # A tibble: 16 x 4
##    sex    height weight   bmi
##    <fct>   <dbl>  <dbl> <dbl>
##  1 Female   1.52  105    45.4
##  2 Male     1.85  174.   50.8
##  3 Female   1.22   74.8  50.3
##  4 Male     1.93  213.   57.2
##  5 Female   1.63  127    47.8
##  6 Female   1.55  115.   48.0
##  7 Female   1.65  131.   48.2
##  8 Female   1.55  109.   45.3
##  9 Male     1.78  143.   45.1
## 10 Female   1.65  127    46.6
## 11 Female   1.63  132.   49.5
```

```
## 12 Female   1.7   152    52.6
## 13 Female   1.6   127    49.6
## 14 Female   1.5   106.   47.2
## 15 Female   1.73  154.   51.5
## 16 Female   1.6   116.   45.4
```

The smallest BMI of 9.2 kg/m2 is very low, with a weight of 22.7 kg. We should check the recorded height and weight values against the original data (paper records, survey responses) if they were available. However, as a weight of 22.7kg is not impossible, this record will not be deleted. An alternative approach would be to analyse the data including the very low BMI and again excluding the very low BMI as a sensitivity analysis. The largest BMI values are based on participants with large weights, and none of these seem biologically implausible. Therefore, no changes will be made to participants with small or large values of BMI.

We can use the `cut()` function to create the BMI categories. The WHO cutpoints are inclusive of the lower-bound, so we use `right=FALSE`. After creating the categories, it is good practice to check the resulting categories using `summary()`:

```
survey$bmi_cat <- cut(survey$bmi, c(0, 18.5, 25, 30, 35, 40, 100), right=FALSE)
summary(survey$bmi_cat)
```

```
##  [0,18.5) [18.5,25)   [25,30)   [30,35)   [35,40)  [40,100)
##        18       362       411       201       101        47
```

Finally, we can create a two-way table using the `contTables()` function within the `jmv` package. We can define the rows by BMI category, and the columns by sex:

```
contTables(data=survey,
           rows = bmi_cat,
           cols = sex)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    bmi_cat      Male    Female    Total
##
##    [0,18.5)        6        12       18
##    [18.5,25)     134       228      362
##    [25,30)       216       195      411
##    [30,35)        95       106      201
##    [35,40)        46        55      101
##    [40,100)       16        31       47
##    Total         513       627     1140
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²    22.49802     5    0.0004209
##    N         1140
##
```

To assess whether there is a difference in BMI between males and females, we should look at the within-sex relative frequencies. In other words, column percents (for this table), by specifying `pcCol = TRUE`:

```
contTables(data=survey,
           rows = bmi_cat,
           cols = sex,
           pcCol = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    bmi_cat                          Male         Female       Total
##
##    [0,18.5)     Observed                 6           12          18
##                 % within column    1.16959     1.91388     1.57895
##
##    [18.5,25)    Observed               134          228         362
##                 % within column   26.12086    36.36364    31.75439
##
##    [25,30)      Observed               216          195         411
##                 % within column   42.10526    31.10048    36.05263
##
##    [30,35)      Observed                95          106         201
##                 % within column   18.51852    16.90590    17.63158
##
##    [35,40)      Observed                46           55         101
##                 % within column    8.96686     8.77193     8.85965
##
##    [40,100)     Observed                16           31          47
##                 % within column    3.11891     4.94418     4.12281
##
##    Total        Observed               513          627        1140
##                 % within column  100.00000   100.00000   100.00000
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²    22.49802     5    0.0004209
##    N         1140
##
```

From this health survey, it appears that men are more likely to have BMIs indicating Pre-Obesity (men 42% vs women 31%) and Obesity Class I (men 19% vs women 17%), compared to women who are more likely to have BMIs indicating Normal weight (women 36% vs men 26%).

Activity 2.5

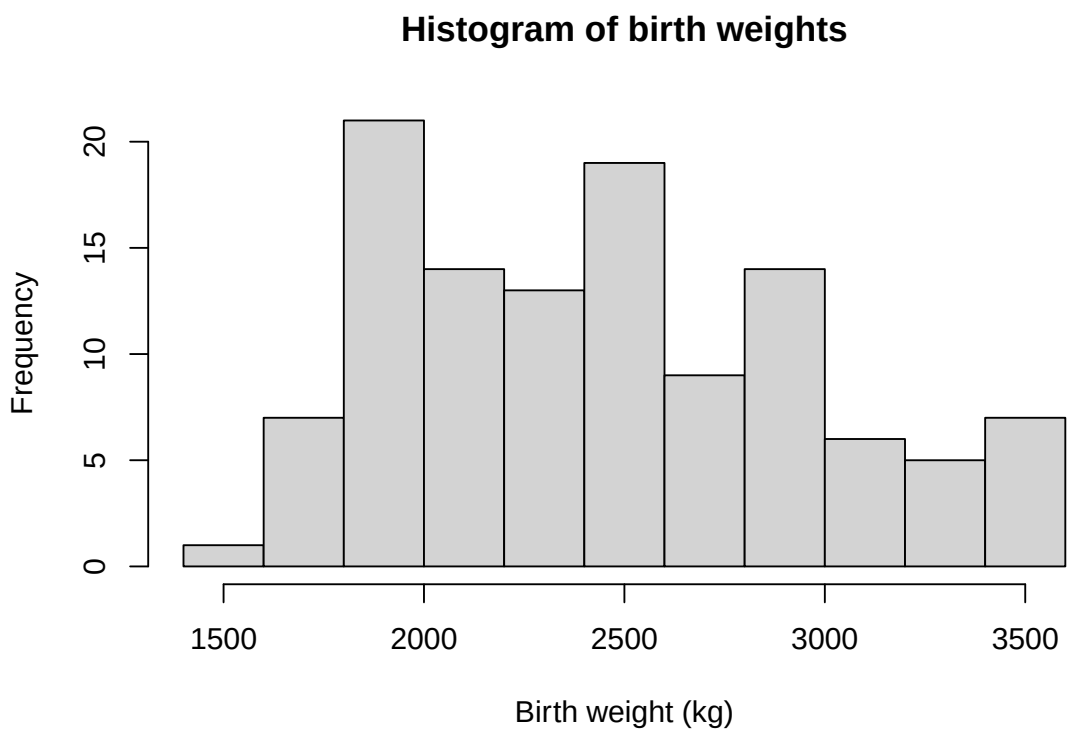
The data in the file `Activity_S2.5.rds` (available on Moodle) has information about birth weight and length of stay collected from 117 babies admitted consecutively to a hospital for surgery. For each variable:

a. Create a histogram to inspect the distribution of the variable;

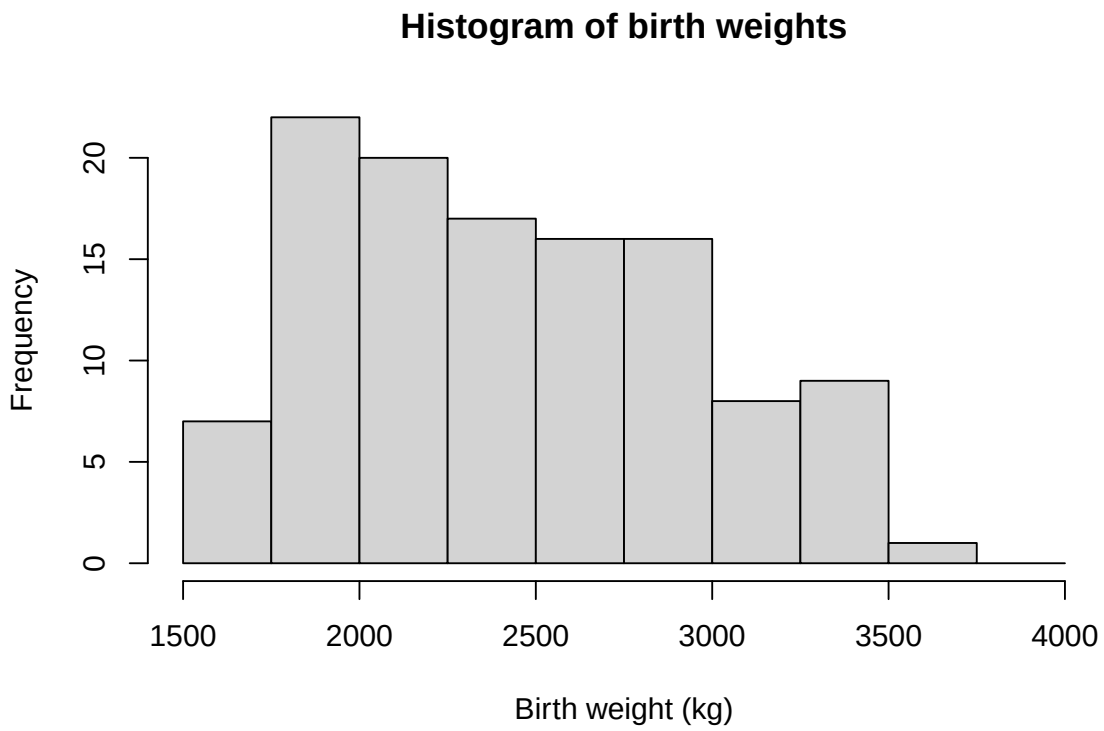
```
babies <- readRDS("data/activities/Activity_S2.5-LengthOfStay.rds")
summary(babies)
```

```
##        ID          Sex        BirthWt        GestAge        LengthStay
##  Min.   : 25   female:55   Min.   :1500   Min.   :31.00   Min.   :  0.00
##  1st Qu.: 54   male  :62   1st Qu.:2012   1st Qu.:35.75   1st Qu.: 21.00
##  Median : 83               Median :2438   Median :36.00   Median : 30.00
##  Mean   : 83               Mean   :2451   Mean   :36.56   Mean   : 41.08
##  3rd Qu.:112               3rd Qu.:2830   3rd Qu.:38.00   3rd Qu.: 43.00
##  Max.   :141               Max.   :3545   Max.   :41.00   Max.   :244.00
##                            NA's   :1      NA's   :5
```

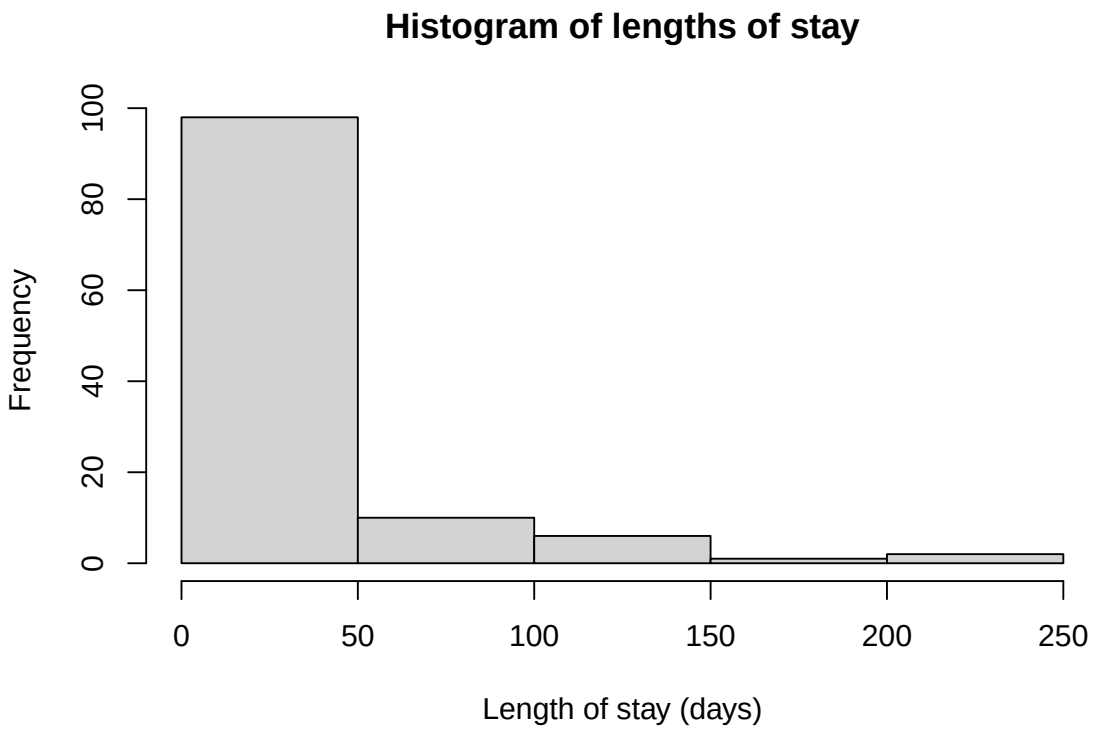
```
hist(babies$BirthWt, main="Histogram of birth weights",
     xlab="Birth weight (kg)")
```



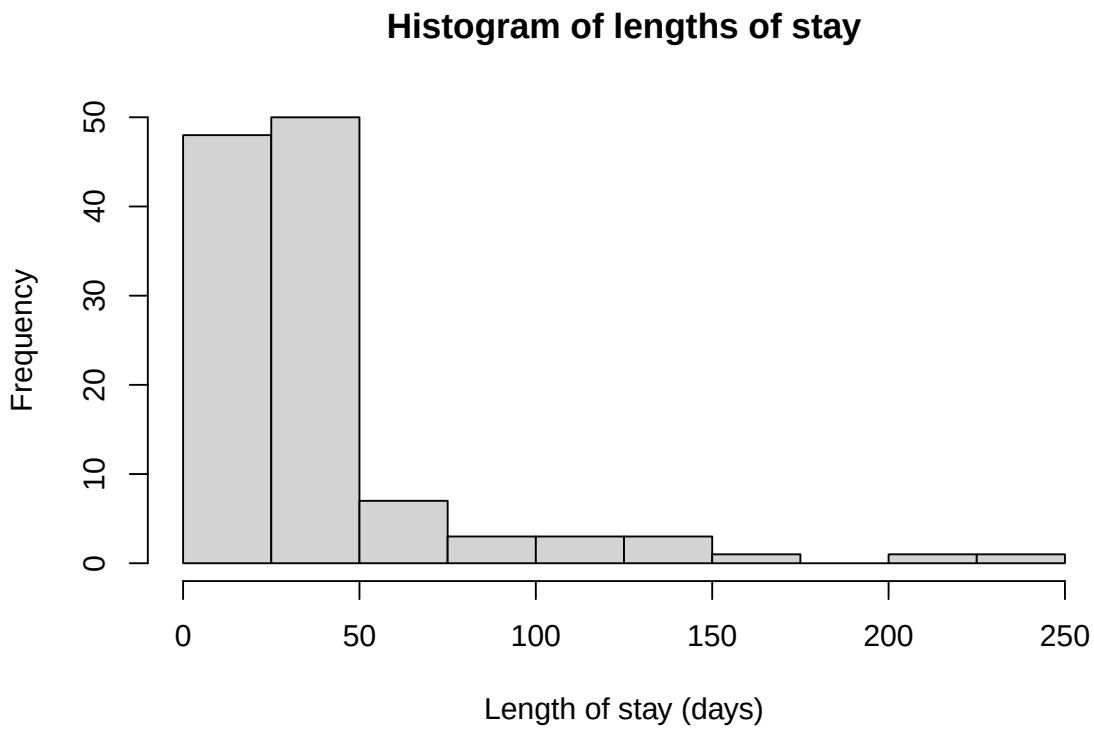

```
# We can specify our own cutpoints using the breaks command, with the seq() function:
hist(babies$BirthWt, main="Histogram of birth weights",
     xlab="Birth weight (kg)",
     breaks=seq(from=1500, to=4000, by=250))
```




```
hist(babies$LengthStay, main="Histogram of lengths of stay",
     xlab="Length of stay (days)")
```




```
hist(babies$LengthStay, main="Histogram of lengths of stay",
     xlab="Length of stay (days)",
     breaks=seq(from=0, to=250, by=25))
```



The histogram for birthweight shows a roughly symmetric distribution. The histogram for length of stay shows a highly skewed distribution (skewed to the right).

b. Complete the following summary statistics for each variable:

- mean and median;
- standard deviation and interquartile range;
- skewness and kurtosis.

```
descriptives(data = babies,
             vars = c(BirthWt, LengthStay),
             pc = TRUE,
             skew = TRUE,
             kurt = TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                            BirthWt       LengthStay
##
##    N                           116            117
##    Missing                       1              0
##    Mean                   2451.207       41.07692
##    Median                 2437.500       30.00000
##    Standard deviation      504.8221       36.92984
##    Minimum                1500.000       0.000000
```

```
##    Maximum                 3545.000      244.0000
##    Skewness                0.3548827      3.090351
##    Std. error skewness     0.2245612     0.2236233
##    Kurtosis               -0.7448547      11.56803
##    Std. error kurtosis     0.4455276     0.4436951
##    25th percentile          2012.000      21.00000
##    50th percentile          2437.500      30.00000
##    75th percentile          2830.000      43.00000
##
```

Make a decision about whether each variable is symmetric or not, and which measure of central tendency and variability should be reported.

> As birthweight follows a roughly symmetric distribution, we should present the mean and standard deviation as the appropriate measures of central tendency and spread. Notice that the mean and median are similar, which is to be expected for a symmetric distribution.
>
> Length of stay is highly skewed. In this case, the median and interquartile range are the appropriate measures to present. Notice that the mean is higher than the median, which is typical for distributions that are skewed to the right.

Activity 2.6

The data set of hospital stay data for 1323 hypothetical patients is available on Moodle in csv format (`Activity2.6.csv`). Import this dataset into R There are two variables in this dataset:

- female: female=1; male=0
- los: length of stay in days

a) Use R to examine the distribution of length of stay: overall; and separately for females and males. Comment on the distributions.

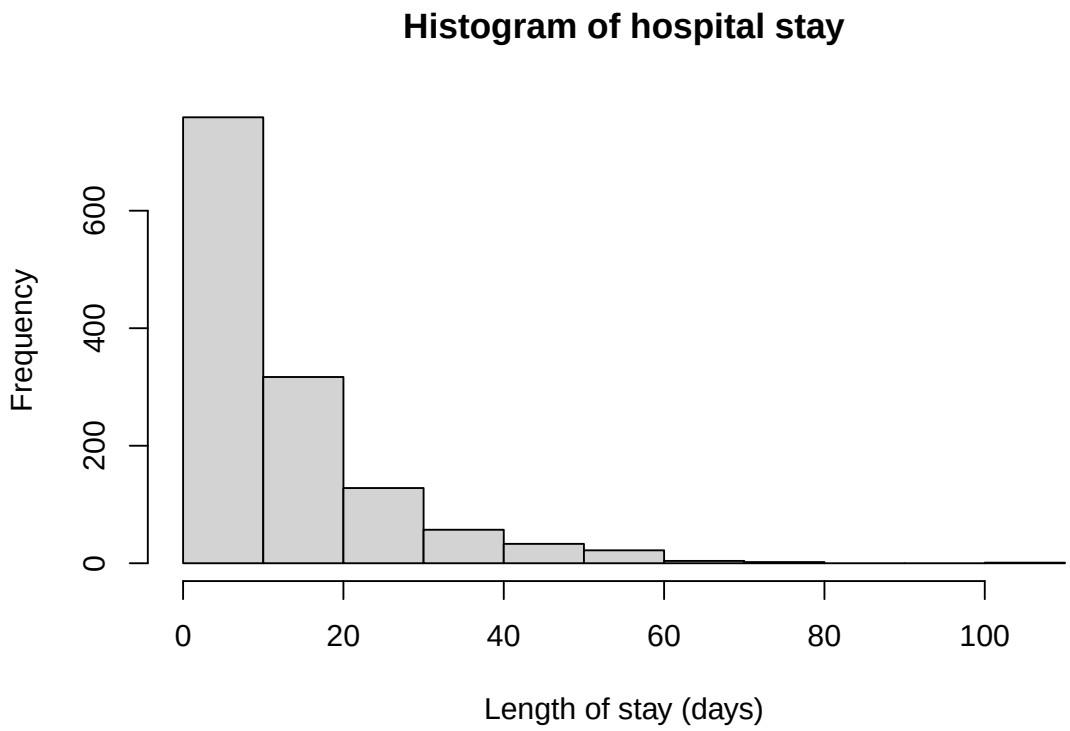
```
hospstay <- read.csv("data/activities/Activity_S2.5.csv")

summary(hospstay)
```

```
##      female            los
##  Min.   :0.0000   Min.   :  0.00
##  1st Qu.:0.0000   1st Qu.:  4.00
##  Median :0.0000   Median :  9.00
##  Mean   :0.1104   Mean   : 12.52
##  3rd Qu.:0.0000   3rd Qu.: 17.00
##  Max.   :1.0000   Max.   :106.00
```

```
# Define female as a factor
hospstay$female <- factor(hospstay$female, levels=c(0,1), labels=c("Male", "Female"))
summary(hospstay$female)
```

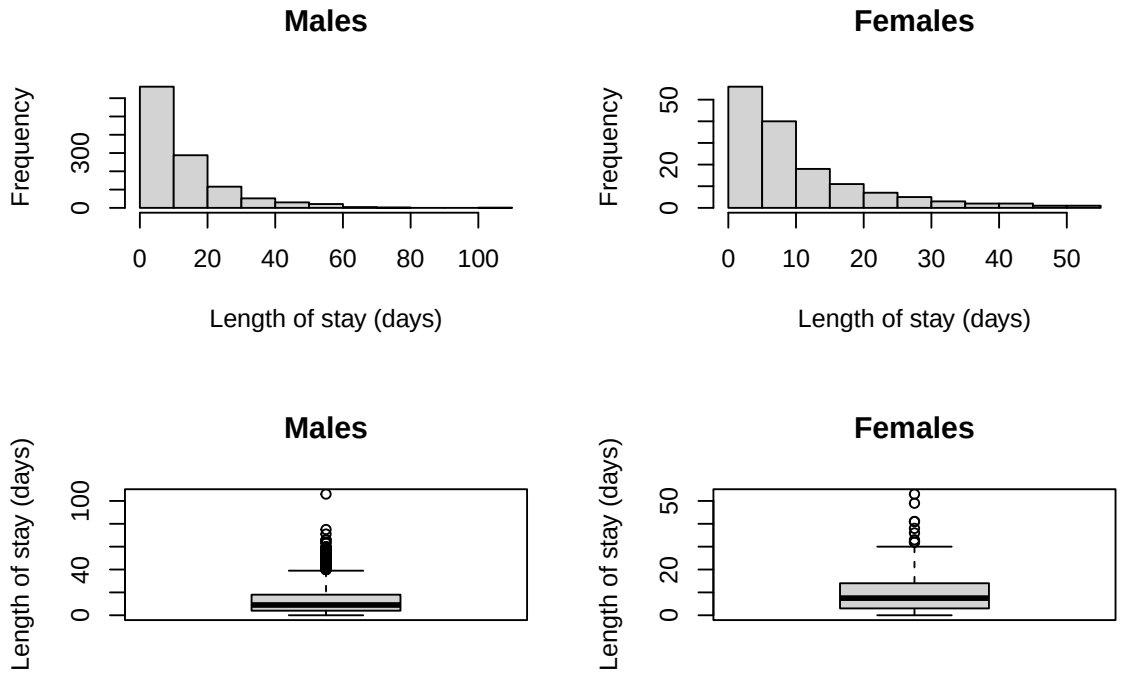
```
##   Male Female
##   1177    146
```

```
hist(hospstay$los, main="Histogram of hospital stay", xlab="Length of stay (days)")
```




```
boxplot(hospstay$los, main="Boxplot of hospital stay", ylab="Length of stay (days)")
```

```
hospstay_males <- subset(hospstay, female=="Male")
hospstay_females <- subset(hospstay, female=="Female")

# Set the graphics parameters to plot 2 rows and 2 columns:
par(mfrow=c(2,2))

# Specify each plot separately
hist(hospstay_males$los, xlab="Length of stay (days)", main="Males")
hist(hospstay_females$los, xlab="Length of stay (days)", main="Females")

boxplot(hospstay_males$los, ylab="Length of stay (days)", main="Males")
boxplot(hospstay_females$los, ylab="Length of stay (days)", main="Females")
```




```
# Reset graphics parameters
par(mfrow=c(1,1))
```

The histograms for overall length of stay and length of stay by gender all show that length of stay is heavily skewed (skewed to the right).

b) Use R to calculate measures of central tendency for hospital stay to obtain information about the average duration of hospital stay. Which summary statistics should you report and why? Report the appropriate statistics of the spread and measure of central tendency chosen.

```
descriptives(data = hospstay,
             vars = los,
             pc = TRUE,
             skew = TRUE,
             kurt = TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                           los
##
##    N                          1323
##    Missing                       0
##    Mean                   12.51550
##    Median                        9
##    Standard deviation     12.59933
##    Minimum                       0
##    Maximum                     106
##    Skewness                1.947803
##    Std. error skewness   0.06726732
##    Kurtosis                5.166837
##    Std. error kurtosis    0.1344336
##    25th percentile         4.000000
##    50th percentile         9.000000
##    75th percentile         17.00000
##
```

As the distribution of length of stay is highly skewed, the median and interquartile range should be presented. These can be calculated in the usual way, using the `descriptives()` function. The median length of stay is 9 days, with an interquartile range of 4 to 17 days.

c) Calculate the measures of central tendency for hospital duration separately for males and females. What can you conclude from comparing these measures for males and females?

```
descriptives(data = hospstay,
             vars = los,
             splitBy = female,
             pc = TRUE,
             skew = TRUE,
             kurt = TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                           female    los
##
##    N                      Male           1177
##                           Female          146
##    Missing                Male              0
##                           Female            0
##    Mean                   Male       12.75531
##                           Female     10.58219
##    Median                 Male              9
##                           Female     7.500000
##    Standard deviation     Male       12.83475
##                           Female     10.34625
```

```
##    Minimum               Male              0
##                          Female            0
##    Maximum               Male            106
##                          Female           53
##    Skewness              Male       1.943967
##                          Female     1.697009
##    Std. error skewness   Male      0.07130745
##                          Female    0.2006795
##    Kurtosis              Male       5.128450
##                          Female     3.067601
##    Std. error kurtosis   Male      0.1424946
##                          Female    0.3987670
##    25th percentile       Male       4.000000
##                          Female     3.000000
##    50th percentile       Male       9.000000
##                          Female     7.500000
##    75th percentile       Male       18.00000
##                          Female     14.00000
##
```

Lengths of stay are similar for men (median: 9 days, interquartile range: 4 to 18 days) and women (median: 8 days, interquartile range: 3 to 14 days).

Module 2: Full script

```
# Author: Timothy Dobbins
# Date: May, 2022
# Purpose: Learning activities for Module 2

library(jmv)
library(readxl)

### Activity 2.1

pbinom(q=59, size=90, prob=0.5, lower.tail = FALSE)


### Activity 2.2

dbinom(x=0, size=5, prob=0.35)
dbinom(x=2, size=5, prob=0.35)
dbinom(x=5, size=5, prob=0.35)
pbinom(q=3, size=5, prob=0.35)


### Activity 2.3
A <- pnorm(75, mean=87, sd=8, lower.tail=TRUE)
A

C <- pnorm(95, mean=87, sd=8, lower.tail=FALSE)
C

B <- 1 - A - C
B


### Activity 2.4

survey <- read_excel("data/examples/health-survey.xlsx")
summary(survey)

survey$sex <- factor(survey$sex, level=c(1,2), labels=c("Male", "Female"))

survey$bmi = survey$weight / (survey$height^2)
hist(survey$bmi, main="Histogram of BMI", xlab="BMI (kg/m2)")
boxplot(survey$bmi, main="Boxplot of BMI", ylab="BMI (kg/m2)")
```

```
subset(survey, bmi<15)
subset(survey, bmi>45)

survey$bmi_cat <- cut(survey$bmi, c(0, 18.5, 25, 30, 35, 40, 100), right=FALSE)
summary(survey$bmi_cat)

contTables(data=survey,
           rows = bmi_cat,
           cols = sex)

contTables(data=survey,
           rows = bmi_cat,
           cols = sex,
           pcCol = TRUE)

### Activity 2.5
babies <- readRDS("data/activities/Activity_S2.5-LengthOfStay.rds")
summary(babies)

hist(babies$BirthWt, main="Histogram of birth weights",
     xlab="Birth weight (kg)")

# We can specify our own cutpoints using the breaks command, with the seq() function:
hist(babies$BirthWt, main="Histogram of birth weights",
     xlab="Birth weight (kg)",
     breaks=seq(from=1500, to=4000, by=250))

hist(babies$LengthStay, main="Histogram of lengths of stay",
     xlab="Length of stay (days)")

hist(babies$LengthStay, main="Histogram of lengths of stay",
     xlab="Length of stay (days)",
     breaks=seq(from=0, to=250, by=25))

descriptives(data = babies,
             vars = c(BirthWt, LengthStay),
             pc = TRUE,
             skew = TRUE,
             kurt = TRUE)

hospstay <- read.csv("data/activities/Activity_S2.5.csv")

summary(hospstay)

# Define female as a factor
hospstay$female <- factor(hospstay$female, levels=c(0,1), labels=c("Male", "Female"))
summary(hospstay$female)

hist(hospstay$los, main="Histogram of hospital stay", xlab="Length of stay (days)")
boxplot(hospstay$los, main="Boxplot of hospital stay", ylab="Length of stay (days)")

hospstay_males <- subset(hospstay, female=="Male")
hospstay_females <- subset(hospstay, female=="Female")
```

```r
# Set the graphics parameters to plot 2 rows and 2 columns:
par(mfrow=c(2,2))

# Specify each plot separately
hist(hospstay_males$los, xlab="Length of stay (days)", main="Males")
hist(hospstay_females$los, xlab="Length of stay (days)", main="Females")

boxplot(hospstay_males$los, ylab="Length of stay (days)", main="Males")
boxplot(hospstay_females$los, ylab="Length of stay (days)", main="Females")

# Reset graphics parameters
par(mfrow=c(1,1))

descriptives(data = hospstay,
             vars = los,
             pc = TRUE,
             skew = TRUE,
             kurt = TRUE)

descriptives(data = hospstay,
             vars = los,
             splitBy = female,
             pc = TRUE,
             skew = TRUE,
             kurt = TRUE)
```

Module 3: Solutions to Learning Activities

Activity 3.1

An investigator wishes to study people living with agoraphobia (fear of open spaces). The investigator places an advertisement in a newspaper asking for volunteer participants. A total of 100 replies are received of which the investigator randomly selects 30. However, only 15 volunteers turn up for their interview.

1. Which of the following statements is true?

    a) The final 15 participants are likely to be a representative sample of the population available to the investigator
    b) The final 15 participants are likely to be a representative sample of the population of people with agoraphobia
    c) The randomly selected 30 participants are likely to be a representative sample of people with agoraphobia who replied to the newspaper advertisement
    d) None of the above

    ANSWER: C

2. The basic problem confronted by the investigator is that:

    a) The accessible population might be different from the target population
    b) The sample has been chosen using an unethical method
    c) The sample size was too small
    d) It is difficult to obtain a sample of people with agoraphobia in a scientific way

    ANSWER: A

Activity 3.2

A dental epidemiologist wishes to estimate the mean weekly consumption of sweets among children of a given age in her area. After devising a method which enables her to determine the weekly consumption of sweets by a child, she conducted a pilot survey and found that the standard deviation of sweet consumption by the children per week is 85 gm (assuming this is the σ). She considers taking a random sample for the main survey of:

i) 25 children, or
ii) 100 children, or
iii) 625 children or
iv) 3,000 children.

a) Estimate the standard error and maximum likely (95% confidence) error of the sample mean for each of these four sample sizes.

```r
# i: n=25
n <- 25
se <- 85 / sqrt(n)
se
```

```
## [1] 17
```

```r
mle <- 1.96 * se
mle
```

```
## [1] 33.32
```

i) The standard error of the mean for a sample of 25 = 85/√25 = 17 gm, and the maximum likely error = 1.96 × 17 = 33.32 gm.

```r
# ii: n=100
n <- 100
se <- 85 / sqrt(n)
se
```

```
## [1] 8.5
```

```r
mle <- 1.96 * se
mle
```

```
## [1] 16.66
```

ii) The standard error of the mean for a sample of 100 = 85/√100 = 8.5 gm, and the maximum likely error =1.96 × 8.5 = 16.66 gm.

```r
# iii: n=625
n <- 625
se <- 85 / sqrt(n)
se
```

```
## [1] 3.4
```

```r
mle <- 1.96 * se
mle
```

```
## [1] 6.664
```

iii) The standard error of the mean for a sample of 625 = 85/√625 = 3.4 gm, and the maximum likely error =1.96 × 3.4 = 6.66 gm.

```r
# iv: n=3000
n <- 3000
se <- 85 / sqrt(n)
se
```

```
## [1] 1.551881
```

```
mle <- 1.96 * se
mle
```

```
## [1] 3.041686
```

iv) The standard error of the mean for a sample of 3,000 = 85/√3000 = 1.55 gm, and the maximum likely error =1.96 × 1.551881 = 3.04 gm.

b) What happens to the standard error as the sample size increases? What can you say about the precision of the sample mean as the sample size increases?

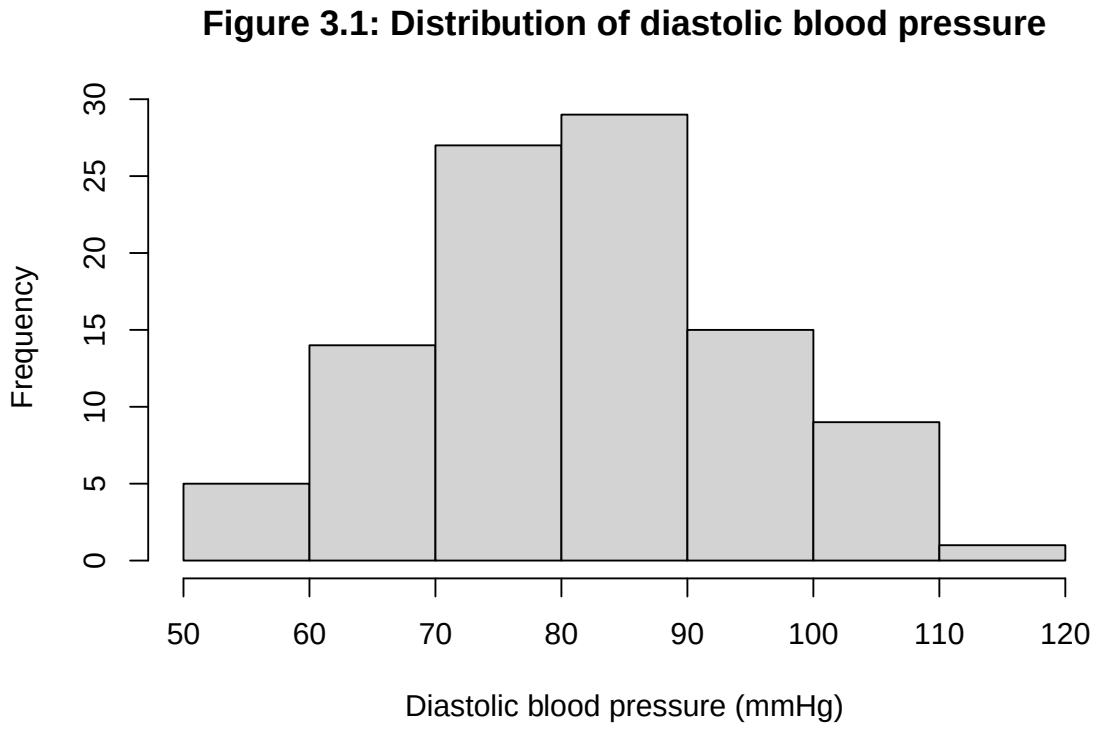
When the sample size increases, the standard error of the mean (and hence the maximum likely error) decreases. Thus, sample means from larger samples are more precise than from smaller samples.

Activity 3.3

The dataset for this activity is the same as the one used in Activity 1.4 in Module 1. The file is Activity1.4.rds on Moodle.

a) Plot a histogram of diastolic BP and describe the distribution.

```
library(jmv)

dbp <- readRDS("data/activities/Activity_S1.4.rds")

hist(dbp$diabp,
     main="Figure 3.1: Distribution of diastolic blood pressure",
     xlab="Diastolic blood pressure (mmHg)")
```



The distribution is approximately symmetrical, centered about the mean.

b) Use R to obtain an estimate of the mean, standard error of the mean and the 95% confidence interval for the mean diastolic blood pressure.

```
descriptives(data=dbp, vars=diabp, se=TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                          diabp
##
##    N                          100
##    Missing                      0
##    Mean                  82.23000
##    Std. error mean       1.301522
##    Median                83.00000
##    Standard deviation    13.01522
##    Minimum               56.00000
##    Maximum               118.0000
##
```

```
t.test(dbp$diabp)
```

```
##
##  One Sample t-test
##
## data:  dbp$diabp
## t = 63.18, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  79.6475 84.8125
## sample estimates:
## mean of x
##     82.23
```

The sample mean is estimated as 82.2 mmHg, and the standard error (SE) of the mean is 1.30 mmHg. The 95% confidence interval is from 79.6 to 84.8 mmHg.

Note that the original data have one decimal place. While we could present the mean to two decimal places when reporting the mean, it seems a bit excessive to present a mean blood pressure to two decimal places. Thus we report the mean and 95% confidence interval for the mean with 1 decimal place.

c) What can you say about the population mean from these results? (Include in you answer what is meant by the confidence interval of a mean).

We are 95% confident that true mean of the population from which we sampled lies between 79.6 mmHg and 84.8 mmHg.

Activity 3.4

Suppose that a random sample of 81 newborn babies delivered in a hospital located in a poor neighbourhood during the last year had a mean birth weight of 2.7 kg and a standard deviation of 0.9 kg. Calculate the 95% confidence interval for the unknown population mean. Interpret the 95% confidence interval.

> This question asks for a confidence interval to be calculated from summarised data. R does not have an in-built function to do this, but we can use the code presented in the R notes to complete this activitiy.

```
ci_mean <- function(n, mean, sd, width=0.95, digits=3){
  lcl <- mean - qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)
  ucl <- mean + qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)

  print(paste0(width*100, "%", " CI: ",
               format(round(lcl, digits=digits), nsmall = digits),
               " to ", format(round(ucl, digits=digits), nsmall = digits) ))

}

ci_mean(n=81, mean=2.7, sd=0.9, width=0.95)
```

```
## [1] "95% CI: 2.501 to 2.899"
```

> We are 95% confident that the true mean birthweight in the hospital located in a poor neighbourhood lies between 2.5 kg and 2.9 kg.

Module 3: Full script

```r
# Author: Timothy Dobbins
# Date: June, 2022
# Purpose: Learning activities for Module 3

library(jmv)

# Activity 3.2
# i: n=25
n <- 25
se <- 85 / sqrt(n)
se

mle <- 1.96 * se
mle

# ii: n=100
n <- 100
se <- 85 / sqrt(n)
se

mle <- 1.96 * se
mle

# iii: n=625
n <- 625
se <- 85 / sqrt(n)
se

mle <- 1.96 * se
mle

# iv: n=3000
n <- 3000
se <- 85 / sqrt(n)
se

mle <- 1.96 * se
mle


# Activity 3.3
dbp <- readRDS("data/activities/Activity_S1.4.rds")
```

```r
hist(dbp$diabp,
     main="Figure 3.1: Distribution of diastolic blood pressure",
     xlab="Diastolic blood pressure (mmHg)")

descriptives(data=dbp, vars=diabp, se=TRUE)
t.test(dbp$diabp)

# Activity 3.4
ci_mean <- function(n, mean, sd, width=0.95, digits=3){
  lcl <- mean - qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)
  ucl <- mean + qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)

  print(paste0(width*100, "%", " CI: ",
               format(round(lcl, digits=digits), nsmall = digits),
               " to ", format(round(ucl, digits=digits), nsmall = digits) ))

}

ci_mean(n=81, mean=2.7, sd=0.9, width=0.95)
```

Module 4: Solutions to Learning Activities

Activity 4.1

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

a) $P < 0.01$

   There is strong evidence against the null hypothesis.

b) $P > 0.05$

   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) 'ns' indicating not significant

   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) significant differences exist

   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; $P = 0.008$).

   H_0: There is no difference in the mean systolic blood pressure between males aged 40-60 years 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. The P value is 0.008 which provides strong evidence against the null hypothesis. Therefore, we can conclude that there is a difference in the mean SBP of males and females aged 40-60 years.

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).

H_0: There is no association between obesity and breast cancer. [A more formal 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. [A more formal 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, indicating a positive association between the study factor of obesity and the outcome of breast cancer. The 95% CI is 1.15 to 8.47 and excludes the null value of no association (i.e. OR = 1). The P value is 0.03 which provides evidence against the null hypothesis. Therefore, we can conclude that there is a positive association between obesity and breast cancer.

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).

H_0: There is no association between influenza infection and a healthy diet among adults aged 20-60 years. [A more formal 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. [A more formal 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, indicating a protective association between the study factor of healthy diet and the outcome factor of influenza infection among adults aged 20 to 60 years. However, the 95% confidence interval includes the null value of 1.0 (no association). The P value is 0.2, which means there is no evidence against the null hypothesis. Thus, we can conclude that there is no evidence of an association between a healthy diet and influenza infection among adults aged 20 to 60 years.

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 Excel file Activity_4.3.xls. Import the file into R for this activity.

a) State the research question

Is the mean daily energy intake of women aged 25-30 years different to the recommended daily intake of 7750 kJ/day?

b) Formulate the null hypothesis

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) Formulate the alternative hypothesis

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.

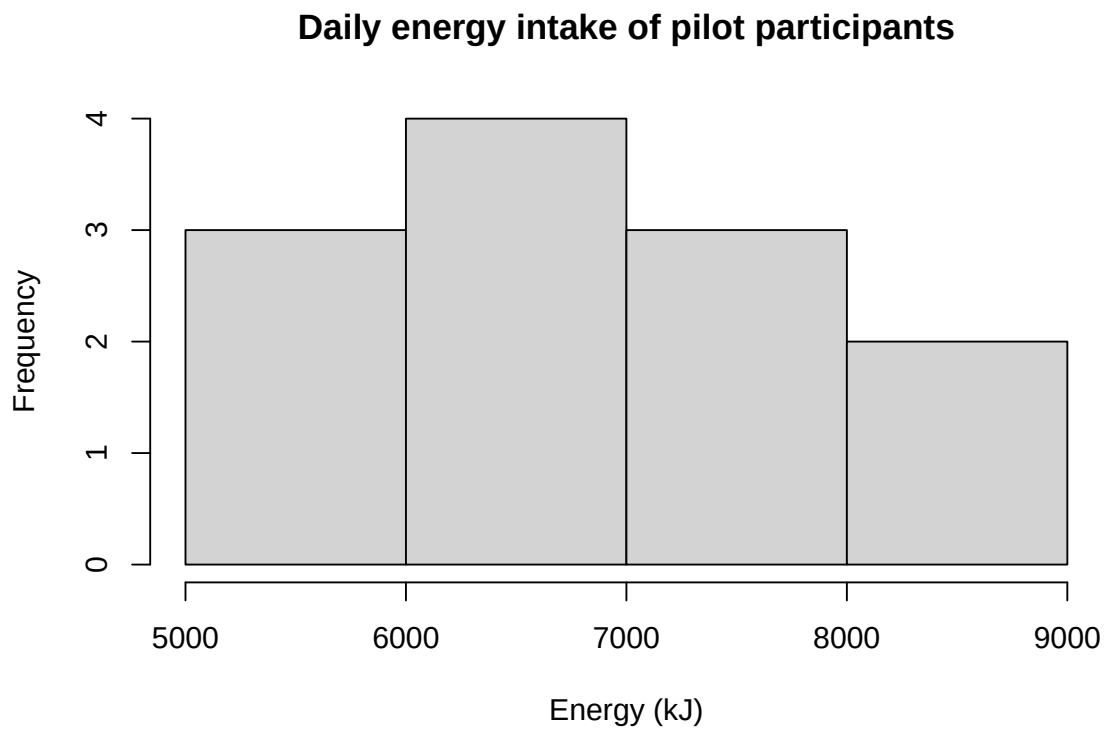
d) Analyse the data and report your conclusions

```
library(readxl)
library(jmv)

pilot <- read_excel("data/activities/Activity_S4.3.xls")
summary(pilot)
```

```
##      Energy
##  Min.   :5260
##  1st Qu.:6045
##  Median :6674
##  Mean   :6856
##  3rd Qu.:7642
##  Max.   :8770
```

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.

```
hist(pilot$Energy, main="Daily energy intake of pilot participants", 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.

The one-sample t-test is conducted as below, to compare the variable Energy to the hypothesised mean of 7750 kJ/day:

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

```
##
##  One Sample t-test
##
## data:  pilot$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
```

The one-sample t-test output shows that the mean daily energy intake of the 12 women is 6856 kJ (95% CI: 6131 to 7581 kJ). There is evidence (t = −2.71 with 11 DF, P = 0.02) that the mean daily energy intake of women aged 25-30 years is lower than the recommended daily intake of 7750 kJ/day.

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.

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. 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 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 cut-off 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.

Module 4: Full script

```
# Author: Timothy Dobbins
# Date: June, 2022
# Purpose: Learning activities for Module 4

library(readxl)
library(jmv)

pilot <- read_excel("data/activities/Activity_S4.3.xls")
hist(pilot$Energy, main="Daily energy intake of pilot participants", xlab="Energy (kJ)")

descriptives(pilot)

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

Module 5: Solutions to Learning Activities

Activity 5.1

Indicate what type of t-test could be used to analyse the data from the following studies and provide reasons:

a) A total of 60 university students are randomly assigned to undergo either behaviour therapy or Gestalt therapy. After twenty therapeutic sessions, each student earns a score on a mental health questionnaire.

   An independent samples t-test could be used because the two groups (behaviour therapy vs Gestalt therapy) are independent. The mental health scores would need to be normally distributed in each group.

b) A researcher wishes to determine whether attendance at a day care centre increases the scores of three year old twins on a motor skills test. Random assignment is used to decide which member from each of 30 pairs of twins attends the day care centre and which member stays at home.

   This is a twin pair study where one member of a twin is allocated to day care and the other member to stay at home. This is an example of an individually matched study and so a paired t-test is appropriate.

c) A child psychologist assigns aggression scores to each of 10 children during two 60 minute observation periods separated by an intervening exposure to a series of violent TV cartoons.

   The same children are scored twice (before and after the intervention), thus it is a paired design and a paired t-test is appropriate.

d) A marketing researcher measures 100 doctors' reports of the number of their patients asking them about a particular drug during the month before and the month after a major advertising campaign.

   The doctors' reports are paired because they report before and after an intervention. Therefore, a paired t-test is appropriate.

Activity 5.2

A study was conducted to compare haemoglobin levels in the blood of children with and without cystic fibrosis. It is known that haemoglobin levels are normally distributed in children. The study results are given below:

Table 0.1: Table 1: Summary of haemoglobin (g/dL)

Statistic	Children without CF	Children with CF
n	12	15
Mean	19.9	13.9
SD (SE)	5.9 (1.7)	6.2 (1.6)

a) State the appropriate null hypothesis and alternate hypothesis

   The null hypothesis: The mean haemoglobin level of children with cystic fibrosis is the same as the mean haemoglobin level of children without cystic fibrosis.

   The alternative hypothesis: The mean haemoglobin level of children with cystic fibrosis is different to the mean haemoglobin level of children without cystic fibrosis.

b) Use R to conduct an appropriate statistical test to evaluate the null hypothesis. Are the assumptions for the test met for this analysis to be valid?

   An independent samples t-test could be carried out to evaluate the study hypothesis because the data have been collected from 2 independent groups of children (children with and children without cystic fibrosis).

   The assumption of independence is met. The outcome variable is continuous and the data are approximately normally distributed in the underlying population (as mentioned in the question).

   We are provided with summarised data (i.e. means and standard deviations in each group), and not individual data. Therefore, we cannot use the standard `t.test()` function. The `BSDA` package has the `tsum.test()` function that can perform a t-test using summarised data.

```
# If necessary, install the BSDA package:
# install.packages("BSDA")
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
# Calculate difference in means by hand:
19.9 - 13.9
```

```
## [1] 6
```

```
# t-test assuming equal variance
tsum.test(mean.x=19.9, s.x=5.9, n.x=12,
          mean.y=13.9, s.y=6.2, n.y=15,
          mu=0, alternative="two.sided", var.equal = TRUE)
```

```
##
##  Standard Two-Sample t-Test
##
## data:  Summarized x and y
## t = 2.5523, df = 25, p-value = 0.01719
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.158367 10.841633
## sample estimates:
## mean of x mean of y
##      19.9      13.9
```

As the two standard deviations are similar, we can assume equal variances. There is evidence that the mean haemoglobin level is lower in children with cystic fibrosis (13.9 g/dL) than children without cystic fibrosis (19.9 g/dL; t=2.55 with 25 df, P=0.02). The difference in means is estimated as 6.0 g/dL (95% CI: 1.2 to 10.8).

Activity 5.3

A randomised controlled trial (RCT) was carried out to investigate the effect of a new tablet supplement in increasing the hematocrit (%) value in anaemic participants. In the study, hematocrit was measured as the proportion of blood that is made up of red blood cells. Hematocrit levels are often lower in anaemic people who do not have sufficient healthy red blood cells. In the RCT, 33 people in the intervention group received the new supplement and 31 people in the control group received standard care (i.e. the usual supplement was given). After 4 weeks, hematocrit values were measured as shown in the R file `ActivityS5.3.rds`. In the community, hematocrit levels are normally distributed.

a) State the research question and frame it as a null hypothesis.

   Research question: Do anaemic patients randomised to take a new supplement have different hematocrit values compared to the anaemic patients randomised to receive the usual care?

   Null hypothesis: There is no difference in the mean hematocrit value in patients randomised to take the new supplement and patients randomised to the control group.

b) Use R to conduct an appropriate statistical test to answer the research question. Before using the test, check the data to see if the assumptions required for the test are met. Obtain a box plot to obtain an estimate of the centre and spread of the data for each group.

   The appropriate test is an independent sample t-test. The assumptions for independent sample t-test are:

   - The two groups are independent
   - The measurements are independent
   - The outcome variable must be continuous and must be normally distributed in each group

   Based on the study design (RCT with 33 people in the intervention, 31 in the control group and the hematocrit level was measured only once per person) we can say that the first two assumptions are met.

   The outcome variable is the proportion of blood that is made up of red blood cells which can be assumed to be continuous.

The histograms and box-plots in Figure 2 (below) show that the data are approximately normally distributed in the intervention group but there is a slight deviation from normality observed in the control group. This is indicated by some deviation from symmetry of the histogram, although there are no influential outliers.

It is mentioned in the question that the outcome variable is normally distributed in the general population. Because the t-test is robust to some degree of non-normality with absence of influential outliers, we could say that the third assumption is also met.

We obtained descriptive statistics for both the intervention and control groups using `descriptives()` function from the `jmv` package. From the descriptive statistics we can see that standard deviation of the intervention group (1.57) is slightly larger than in the control group (0.99). Inspection of Figures 0.2 and 0.3 also indicates more variability in the intervention group. Therefore, it may not be reasonable to assume that the variances are equal. In this case, we will use independent sample t-test based on unequal variance assumption.

```
# Activity 5.3
library(jmv)

anaemia <- readRDS("data/activities/Activity_S5.3.rds")

descriptives(data=anaemia, vars=hematocrit,
             splitBy = group,
             skew = TRUE)
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                            group            hematocrit
##
##    N                       Intervention             33
##                            Standard care            31
##    Missing                 Intervention              0
##                            Standard care             0
##    Mean                    Intervention       32.43636
##                            Standard care      31.64516
##    Median                  Intervention       32.30000
##                            Standard care      31.80000
##    Standard deviation      Intervention       1.570991
##                            Standard care     0.9871976
##    Minimum                 Intervention       29.60000
##                            Standard care      29.80000
##    Maximum                 Intervention       36.10000
##                            Standard care      33.20000
##    Skewness                Intervention      0.2816846
##                            Standard care    -0.1638483
##    Std. error skewness     Intervention      0.4086354
##                            Standard care     0.4205365
##
```

```
# Plotting by group using the method from Module 2:
anaemia_i <- subset(anaemia, group=="Intervention")
anaemia_sc <- subset(anaemia, group=="Standard care")

# Set the graphics parameters to plot 2 rows and 2 columns:
par(mfrow=c(2,2))

# Specify each plot separately
hist(anaemia_i$hematocrit, xlab="Hematocrit", main="Intervention")
hist(anaemia_sc$hematocrit, xlab="Hematocrit", main="Standard care")

boxplot(anaemia_i$hematocrit, ylab="Hematocrit", main="Intervention")
boxplot(anaemia_sc$hematocrit, ylab="Hematocrit", main="Standard care")
```

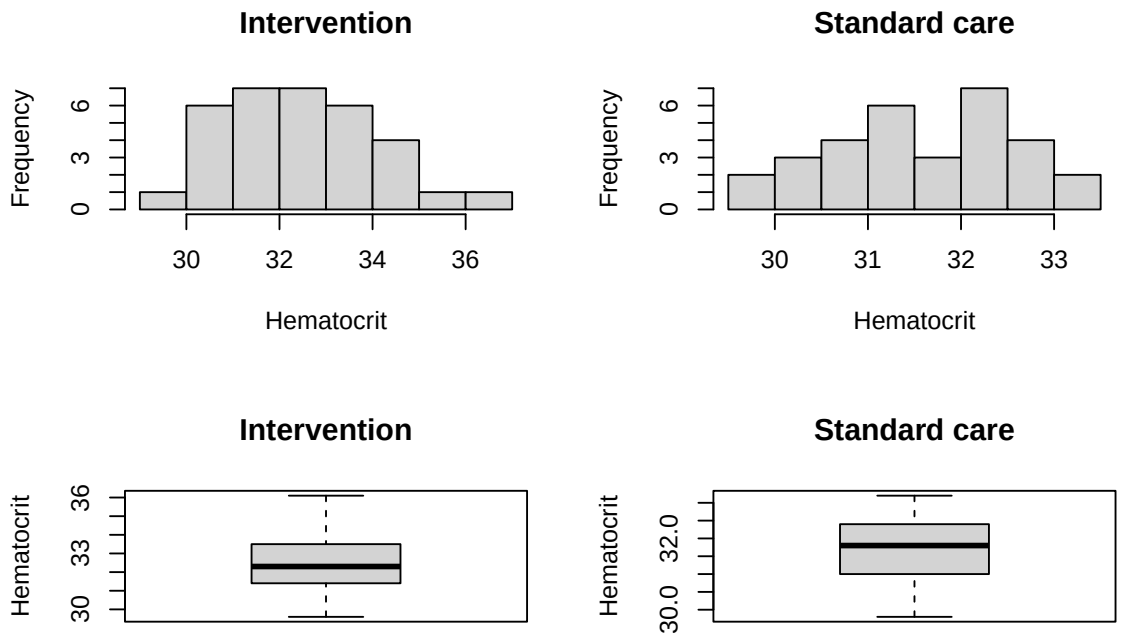


Figure 0.2: Graphical summaries of hematocrit by treatment group

```
# Reset graphics parameters
par(mfrow=c(1,1))
```

Note that the histograms and boxplots use different scales. We can standardise the scale limits using “xlim” and “ylim” by specifying the lower and upper bounds of the x- and y-axis:

```
par(mfrow=c(2,2))
```

```
hist(anaemia_i$hematocrit, xlab="Hematocrit", main="Intervention",
     xlim=c(28, 38))
hist(anaemia_sc$hematocrit, xlab="Hematocrit", main="Standard care",
     xlim=c(28, 38))

boxplot(anaemia_i$hematocrit, ylab="Hematocrit", main="Intervention",
        ylim=c(28, 38))
boxplot(anaemia_sc$hematocrit, ylab="Hematocrit", main="Standard care",
        ylim=c(28, 38))
```

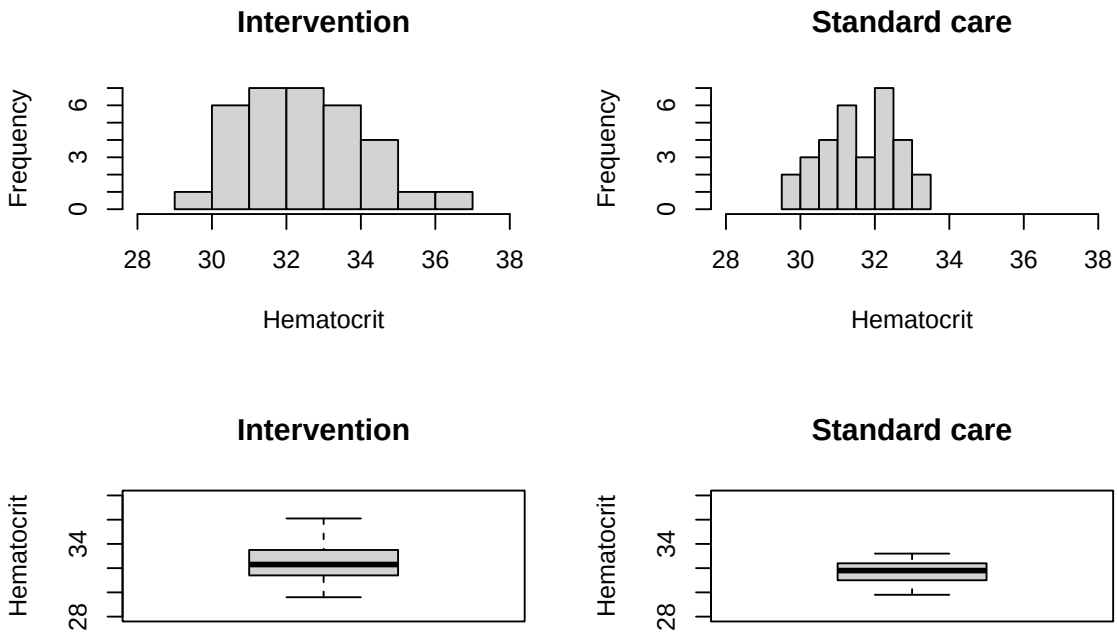


Figure 0.3: Graphical summaries of hematocrit by treatment group

```
# Reset graphics parameters
par(mfrow=c(1,1))
```

c) Run your statistical test.

```
# Welch's t-test
ttestIS(data=anaemia, vars=hematocrit, group=group, meanDiff=TRUE, ci=TRUE, welchs=TRUE)
```

```
##
##  INDEPENDENT SAMPLES T-TEST
##
##  Independent Samples T-Test
```

```
##
##                      Statistic   df       p        Mean difference   SE difference   Lower
##
##   hematocrit   Student's t    2.394370   62.00000   0.0196861       0.7912023      0.3304428   0
##              Welch's t     2.427577   54.31900   0.0185439       0.7912023      0.3259227   0.137
##
```

d) Construct a table to show how you would report your results and write a conclusion.

The results are summarised in Table 2.

Table 0.2: Table 2: Mean hematocrit levels by study group

	Intervention	Standard care	Difference in means (95% CI)	t, df	P value
	Mean (SD)	Mean (SD)			
Hematocrit level (%)	32.44 (1.57)	31.65 (0.99)	0.79 (0.14, 1.44)	2.43, 54.3df	0.019

Conclusion

The mean haematocrit level among the standard care group is 31.65 and among the intervention group is 32.44. There is evidence that the mean hematocrit level is different for the two study groups (P = 0.019, t= 2.43 with 54.3 df). The mean difference indicates that the mean hematocrit level was 0.79 units higher (95% CI: 0.14, 1.44) in the intervention group compared to the control group.

Activity 5.4

A total of 41 babies aged 6 months to 2 years with haemangioma (birth mark) were enrolled in a study to test the effect of a new topical medication in reducing the volume of their haemangioma. Parents were asked to apply the medication twice daily. The volume (in mm3) of the haemangioma was measured at enrolment and again after 12 weeks of using the medication.

a) What is the research question in this study? State the null and alternative hypotheses.

The research question is: does a 12 week application of new topical medication change the volume of haemangiomas among children aged 6 months to 2 years compared to the volume at baseline?

Null hypothesis: there is no change in the mean haemangioma volume among children aged 6 months to 2 years at baseline and after 12 weeks treatment with topical medication.

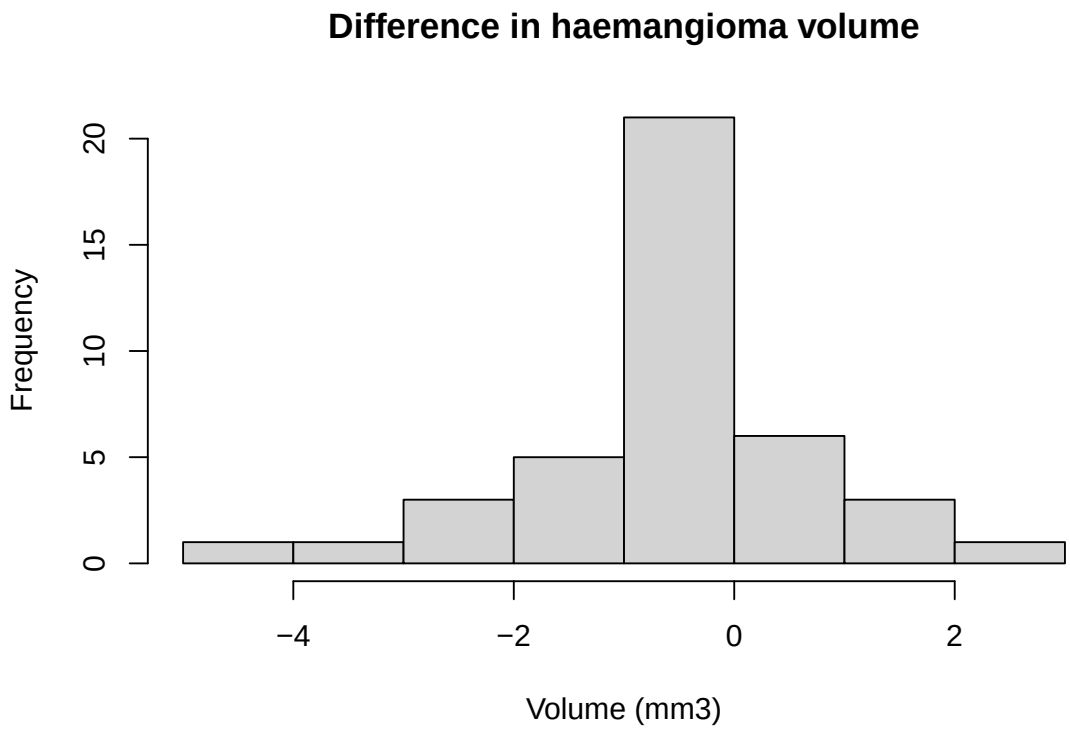
Alternative hypothesis: there is a change in the mean haemangioma volume among children aged 6 months to 2 years at baseline and after 12 weeks treatment with topical medication.

b) Use the data in the R file `ActivityS5.4.rds` to answer the research question. Which statistical test is appropriate to answer the research question and why? Conduct the test in R and write your conclusion.

A paired t-test is appropriate to test the null hypothesis. The measurement of haemangioma volume was made on each baby twice to compare the differences before and after the treatment, therefore, the study has a paired design. Because haemangioma volume is a continuous measurement (mm3), a paired t-test can be considered. The assumptions for a paired t-test are that the outcome variable is continuous, and differences of the measurements are normally distributed.

To check the distribution of the differences between the measurements, we first need to calculate the differences. We then examine the distribution of the differences using a histogram as shown in Figure 3.

```
babies <- readRDS("data/activities/Activity_S5.4.rds")

babies$diff = babies$week_12 - babies$baseline
hist(babies$diff, xlab="Volume (mm3)", main="Difference in haemangioma volume")
```



As we can see from the histogram, the differences in volume at the beginning and end of the study are reasonably symmetrically distributed. Although the distribution is very peaked, there are no influential outliers and the t-test is robust to the deviation of the normality assumption.

To conduct the paired t-test in R, we use the `t.test()` function, specifying the two columns of haemangioma volume and `paired=TRUE`:

```
# Using ttestPS from jmv
ttestPS(data=babies, pairs=list(list(i1 = 'week_12', i2 = 'baseline')), meanDiff=TRUE, ci=TRUE)
```

```
##
##  PAIRED SAMPLES T-TEST
##
##  Paired Samples T-Test
##
##                            statistic   df       p        Mean difference   SE difference   Lower
##
##   week_12   baseline   Student's t   -1.959437   40.00000   0.0570564      -0.4021951      0.2052605
##
```

The code for the `ttestPS()` function is quite cumbersome. You may want to use the `t.test()` function:

```
# Using t.test
t.test(babies$week_12, babies$baseline, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  babies$week_12 and babies$baseline
## t = -1.9594, df = 40, p-value = 0.05706
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.81704216  0.01265192
## sample estimates:
## mean difference
##      -0.4021951
```

The output shows that the mean volume at week 12 of 1.94 mm^3 was lower than the mean of 2.34 mm^3 at baseline. The mean decrease is 0.40 mm^3 (95% CI −0.01 to 0.82). From the paired t-test results, we can see that t = -1.96 with 40 df and P = 0.057. This P-value provides only weak evidence that the topical medication has an effect on haemangioma volume among children aged 6 months to 2 years. The P-value of 0.057 is consistent with the 95% CI just crossing the line of no difference (i.e. 0 mm3).

[Note that the estimated difference and its confidence interval (-0.40: 95% CI from -0.82 to 0.01) is presented as if it were an *increase* from baseline to 12-weeks. As the mean difference is negative, we can interpreted the estimates as *reductions* by multiplying each value by -1.]

c) What are the limitations of this study?

In a paired design, each subject serves as their own control (here by comparing the change in volume of the haemangioma at baseline and after 12 weeks of treatment). However, the reduction of −0.40 mm3 on average could have been due to the new medication or to the natural history of the condition. A better design would be a randomised controlled trial where subjects are randomised to the new treatment or the usual treatment and compare the volume between the 2 groups. More information on randomised controlled trials and other study designs is given in PHCM9794: Foundations of Epidemiology.

Module 5: Full script

```
# Author: Timothy Dobbins
# Date: June, 2022
# Purpose: Learning activities for Module 5

# If necessary, install the BSDA package:
# install.packages("BSDA")
library(BSDA)
library(jmv)

# Activity 5.2
# Calculate difference in means by hand:
19.9 - 13.9

# t-test assuming equal variance
tsum.test(mean.x=19.9, s.x=5.9, n.x=12,
          mean.y=13.9, s.y=6.2, n.y=15,
          mu=0, alternative="two.sided", var.equal = TRUE)

# Activity 5.3
anaemia <- readRDS("data/activities/Activity_S5.3.rds")

descriptives(data=anaemia, vars=hematocrit,
             splitBy = group,
             skew = TRUE)

# Plotting by group using the method from Module 2:
anaemia_i <- subset(anaemia, group=="Intervention")
anaemia_sc <- subset(anaemia, group=="Standard care")

# Set the graphics parameters to plot 2 rows and 2 columns:
par(mfrow=c(2,2))

# Specify each plot separately
hist(anaemia_i$hematocrit, xlab="Hematocrit", main="Intervention")
hist(anaemia_sc$hematocrit, xlab="Hematocrit", main="Standard care")

boxplot(anaemia_i$hematocrit, ylab="Hematocrit", main="Intervention")
boxplot(anaemia_sc$hematocrit, ylab="Hematocrit", main="Standard care")


# Create plots using common axis limits
hist(anaemia_i$hematocrit, xlab="Hematocrit", main="Intervention",
```

```
     xlim=c(28, 38))
hist(anaemia_sc$hematocrit, xlab="Hematocrit", main="Standard care",
     xlim=c(28, 38))

boxplot(anaemia_i$hematocrit, ylab="Hematocrit", main="Intervention",
        ylim=c(28, 38))
boxplot(anaemia_sc$hematocrit, ylab="Hematocrit", main="Standard care",
        ylim=c(28, 38))

# Reset graphics parameters
par(mfrow=c(1,1))

# Welch's t-test
ttestIS(data=anaemia, vars=hematocrit, group=group, meanDiff=TRUE, ci=TRUE, welchs=TRUE)

# Activity 5.4
babies <- readRDS("data/activities/Activity_S5.4.rds")

babies$diff = babies$week_12 - babies$baseline
hist(babies$diff, xlab="Volume (mm3)", main="Difference in haemangioma volume")

# Using ttestPS from jmv
ttestPS(data=babies, pairs=list(list(i1 = 'week_12', i2 = 'baseline')), meanDiff=TRUE, ci=TRUE)

# Using t.test
t.test(babies$week_12, babies$baseline, paired=TRUE)
```

Module 6: Solutions to Learning Activities

Activity 6.1

In a clinical trial involving a dietary intervention, 150 adult volunteers agreed to participate. The investigator wanted to know whether this sample was representative of the general population. One interesting finding was that 90 of the participants drink alcohol regularly compared to 70% of the general population.

a) State the null hypothesis

   Null hypothesis: The proportion of volunteers who consume alcohol regularly is the same as the proportion who consume alcohol regularly in the general population.

b) Calculate the 95% CI for the proportion of regular drinkers in the sample using R.

```
library(DescTools)

# BinomCI from DescTools calculates a Wilson confidence interval:
BinomCI(90, n=150, method="wilson")
```

```
##      est    lwr.ci    upr.ci
## [1,] 0.6 0.5200492 0.6749568
```

```
# The default binom.test calculates a Wald confidence interval
binom.test(90, n=150, p=0.7)
```

```
##
##  Exact binomial test
##
## data:  90 and 150
## number of successes = 90, number of trials = 150, p-value = 0.009553
## alternative hypothesis: true probability of success is not equal to 0.7
## 95 percent confidence interval:
##  0.5169313 0.6790370
## sample estimates:
## probability of success
##                    0.6
```

   The proportion of volunteers who drink alcohol regularly is estimated as 60%, with a 95% confidence interval from 52% to 67%. The 95% confidence interval can be interpreted as: we are 95% confident that the true prevalence of alcohol drinkers in the population from where the sample was drawn lies between 52% and 67%. Note that the prevalence of

regular drinkers in the general population is 70%, which does not fall within the 95% CI calculated from the sample. This may be because the participants were not randomly selected from the general population; they volunteered to participate in the study.

c) Use the R file Activity_S6.1.rds to decide if the sample of volunteers is representative of the population.

We have stated the null hypothesis in part (a):

Null hypothesis: The proportion of volunteers who consume alcohol regularly is the same as the proportion who consume alcohol regularly in the general population.

The one-sample proportion test has been calculated with the `binom.test` command in (a). If we only had individual data (i.e. did not have the summary results above), we would need to tabulate the data first:

```
alcohol <- readRDS("data/activities/Activity_S6.1.rds")
table(alcohol$Drinking_Status)
```

```
##
## Non-drinker     Drinker
##          60          90
```

```
binom.test(90, n=150, p=0.7)
```

```
##
##  Exact binomial test
##
## data:  90 and 150
## number of successes = 90, number of trials = 150, p-value = 0.009553
## alternative hypothesis: true probability of success is not equal to 0.7
## 95 percent confidence interval:
##  0.5169313 0.6790370
## sample estimates:
## probability of success
##                    0.6
```

The R output gives the P-value from a two-sided test of 0.0096. Thus, we can conclude that there is strong evidence that the proportion of drinkers among the sample population of the dietary intervention group (60%) is lower than that (70%) in the general population.

Activity 6.2

A survey was conducted of a random sample of upper primary school children to measure the prevalence of asthma using questionnaires completed by the parents. A total of 514 children were enrolled. Use the R dataset Activity_S6.2.rds for this activity.

a) Calculate the relative risk and odds ratio with 95% confidence interval using Stata for children to have asthma symptoms if they are male? Which risk estimate would be the correct statistic to report?

Before we begin analysing any binary data, we must ensure that the binary variables of interest are coded as factors, with the positive exposure and outcomes ordered first. We can check this using the `summary()` function:

```
library(jmv)
asthma <- readRDS("data/activities/Activity_S6.2.rds")

summary(asthma$Asthma)
```

```
## Yes  No
## 119 395
```

```
summary(asthma$Gender)
```

```
##   Male Female   NA's
##    258    242     14
```

Here, Asthma is coded with "Yes" as the first level, as required. Gender is coded with "Male" as the first level, which means that R will produce summaries of male vs female. Note that there are 14 missing values for gender, indicated as NA.

The relative risk of asthma can be calculated using the `contTables()` function within the `jmv` library. The relative risk is calculated using `relRisk = TRUE`, and the proportion of asthma within each sex calculated using `pcRow = TRUE`:

```
contTables(data=asthma, rows=Gender, cols=Asthma,
          pcRow = TRUE, relRisk = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    Gender                    Yes          No           Total
##
##    Male      Observed              70          188          258
##              % within row    27.13178    72.86822    100.00000
##
##    Female    Observed              46          196          242
##              % within row    19.00826    80.99174    100.00000
##
##    Total     Observed             116          384          500
##              % within row    23.20000    76.80000    100.00000
##
##
##
##  x² Tests
##
##          Value       df    p
##
##    x²    4.624920     1    0.0315107
##    N          500
##
##
##
##  Comparative Measures
##
```

```
##                     Value         Lower       Upper
##
##    Relative risk    1.427368      1.028159    1.981579
##
##      Rows compared
```

From the output we can see that the relative risk of Asthma for males compared to females is 1.43 (95% CI: 1.03 to 1.98).

To calculate the odds ratio, we use `odds = TRUE`:

```
contTables(data=asthma, rows=Gender, cols=Asthma,
          odds = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    Gender    Yes    No     Total
##
##    Male       70    188      258
##    Female     46    196      242
##    Total     116    384      500
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²    4.624920     1    0.0315107
##    N          500
##
##
##
##  Comparative Measures
##
##                  Value       Lower       Upper
##
##    Odds ratio    1.586494    1.039904    2.420381
##
```

The OR of Asthma 1.59 (95% CI: 1.04 to 2.42) for males compared to females.

The relative risk is the correct risk estimate to use as this a cross-sectional study. The relative risk is a direct comparison of the proportion with asthma symptoms in each exposure group. The odds ratio is only appropriate for a case-control study.

b) Use the tabulated data on the frequency of cases and exposure you obtained in R output in part a to calculate RR and OR with their 95% confidence interval using R.

This question assumes we are only given the values of the four cells in the cross-tabulation. We can re-write this table as follows to explain the process of entering summarised data:

Gender	Asthma	Number
Male	Yes	70
Male	No	188
Female	Yes	46
Female	No	196

We can enter these data in a dataframe, comprising three vectors, as follows:

```
asthma_summary <- data.frame(
  Gender = c("Male", "Male", "Female", "Female"),
  Asthma = c("Yes", "No", "Yes", "No"),
  Number = c(70, 188, 46, 196))
```

We need to define Gender and Asthma as factors. Here we must define the `levels` **in the order we want the categories to appear in the table**. Note that as Gender and Asthma are entered as text variables, we can omit `labels` command when defining the factors, and the factor will be labelled using the text entry:

```
asthma_summary$Gender <- factor(asthma_summary$Gender,
                                levels = c("Male", "Female"))

asthma_summary$Asthma <- factor(asthma_summary$Asthma,
                                levels = c("Yes", "No"))
```

We can calculate the relative risk using the summarised data in the same was done previously. However, we need to include the number of observations in each cell using the `counts` command:

```
contTables(data=asthma_summary,
           rows = Gender, cols = Asthma,
           counts = Number,
           relRisk = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    Gender    Yes    No     Total
##
##    Male       70    188      258
##    Female     46    196      242
##    Total     116    384      500
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²    4.624920     1    0.0315107
##    N          500
```

```
##
##
##
##  Comparative Measures
##
##                     Value         Lower       Upper
##
##    Relative risk    1.427368      1.028159    1.981579
##
##      Rows compared
```

The output from the summarised data is identical to the output of the individual-level data.

Activity 6.3

In a study to determine the cause of mortality, 89 people were followed up for 5 years. The participants are classified into two groups of those who did or did not have a heart attack. At the end of the follow-up 15 people died among them 10 had a heart attack. Among the 74 survivors 35 had a heart attack. Present the data on a 2x2 table and calculate relative risk of death from heart attack with 95% confidence interval using R.

The cross-tabulation of heart attack and mortality is given in Table 6.1.

HeartAttack	Death		Total
	Yes	No	
Yes	10	35	45
No	5	39	44
Total	15	74	89

To calculate relative risk using the information from Table 6.1, we enter first enter the summarised data into a new dataframe:

```
mortality <- data.frame(
  HeartAttack = c("Yes", "Yes", "No", "No"),
  Death = c("Yes", "No", "Yes", "No"),
  n = c(10, 35, 5, 39))

mortality$HeartAttack <- factor(mortality$HeartAttack,
                                levels = c("Yes", "No"))

mortality$Death <- factor(mortality$Death,
                                levels = c("Yes", "No"))
```

We then estimate the relative risk using the `contTables()` function:

```
contTables(data = mortality,
           rows = HeartAttack, cols = Death,
           counts = n,
           pcRow = TRUE, relRisk = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    HeartAttack                    Yes          No           Total
##
##    Yes            Observed              10           35           45
##                   % within row     22.22222     77.77778    100.00000
##
##    No             Observed               5           39           44
##                   % within row     11.36364     88.63636    100.00000
##
##    Total          Observed              15           74           89
##                   % within row     16.85393     83.14607    100.00000
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²     1.871883     1    0.1712595
##    N           89
##
##
##
##  Comparative Measures
##
##                      Value         Lower        Upper
##
##    Relative risk     1.955556      0.7267615    5.261971
##
##      Rows compared
```

From the output we can see that the relative risk of death from heart attack is 1.96 (95% CI: 0.73 to 5.26).

Activity 6.4

A study is conducted to test the hypothesis that the observed frequency of a certain health outcome is 30%. If the results yield a CI around the sample proportion that extends from 23.8 to 30.2, what can you say about the evidence against the null hypothesis?

> As the 95% confidence interval includes the hypothesised population proportion, we can infer that the P-value of the study will be greater than 0.05. Hence, this study provides weak to no evidence against the null hypothesis.

Activity 6.5

In an experiment to test the effect of vitamin C on IQ scores, the following confidence intervals were estimated around the percentage of people with improved scores for five different populations:

Population	% with improved IQ	95% confidence interval
1	35.0	32.0 to 38.0
2	29.5	25.0 to 34.0
3	43.5	42.0 to 45.0
4	30.5	20.0 to 41.0
5	24.5	21.0 to 28.0

a) Which CI is the most precise?

   Population 3. This has the smallest interval which is 3 (42 - 45).

b) Which CI implies the largest sample size?

   Population 3. The larger the sample size, the smaller the standard error and hence narrower the confidence interval. Therefore the largest sample size will have the narrowest confidence interval provided that the frequency is the same.

c) Which CI is the least precise? > Population 4. This has the widest interval which is 21 (41 - 20), thus is less precise than the others.

d) Which CI most strongly supports the conclusion that vitamin C increases IQ score and why?

   Population 3. This has the narrowest confidence interval where the lower bound is higher than the upper bound of all others.

e) Which would most likely to stimulate the investigator to conduct an additional experiment using a larger sample size?

   Population 4. This estimate is the least precise. By increasing sample size, the estimate of frequency as shown by the 95% CI would become narrower.

Module 6: Full script

```
# Author: Timothy Dobbins
# Date: June, 2022
# Purpose: Learning activities for Module 6

# Activity 6.1

library(DescTools)

# BinomCI from DescTools calculates a Wilson confidence interval:
BinomCI(90, n=150, method="wilson")

# The default binom.test calculates a Wald confidence interval
binom.test(90, n=150, p=0.7)

alcohol <- readRDS("data/activities/Activity_S6.1.rds")
table(alcohol$Drinking_Status)

binom.test(90, n=150, p=0.7)

# Activity 6.2

library(jmv)
asthma <- readRDS("data/activities/Activity_S6.2.rds")

summary(asthma$Asthma)
summary(asthma$Gender)

contTables(data=asthma, rows=Gender, cols=Asthma,
          pcRow = TRUE, relRisk = TRUE)

contTables(data=asthma, rows=Gender, cols=Asthma,
          odds = TRUE)

asthma_summary <- data.frame(
  Gender = c("Male", "Male", "Female", "Female"),
  Asthma = c("Yes", "No", "Yes", "No"),
  Number = c(70, 188, 46, 196))

asthma_summary$Gender <- factor(asthma_summary$Gender,
                                levels = c("Male", "Female"))

asthma_summary$Asthma <- factor(asthma_summary$Asthma,
```

```
                                levels = c("Yes", "No"))

contTables(data=asthma_summary,
           rows = Gender, cols = Asthma,
           counts = Number,
           relRisk = TRUE)


# Activity 6.3
mortality <- data.frame(
  HeartAttack = c("Yes", "Yes", "No", "No"),
  Death = c("Yes", "No", "Yes", "No"),
  n = c(10, 35, 5, 39))

mortality$HeartAttack <- factor(mortality$HeartAttack,
                                levels = c("Yes", "No"))

mortality$Death <- factor(mortality$Death,
                                levels = c("Yes", "No"))

contTables(data = mortality,
           rows = HeartAttack, cols = Death,
           counts = n,
           pcRow = TRUE, relRisk = TRUE)
```

Module 7: Solutions to Learning Activities

Activity 7.1

Use the file Activity_S7.1.rds to further investigate whether there is a gender difference in asthma in a random sample of 514 upper primary school children:

a) Use a contingency table (cross-tabulation) to determine the observed and expected frequencies. Which cell has the lowest expected cell count?

   Here we can use the `contTables()` function within `jmv` with the option `exp = TRUE` to present the expected frequencies.

```
library(jmv)

children <- readRDS("data/activities/Activity_S7.1.rds")

contTables(data=children, rows = gender, cols = asthma,
           exp = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    gender                No          Yes          Total
##
##    Female    Observed         196           46         242
##              Expected    185.8560     56.14400    242.0000
##
##    Male      Observed         188           70         258
##              Expected    198.1440     59.85600    258.0000
##
##    Total     Observed         384          116         500
##              Expected    384.0000    116.00000    500.0000
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²     4.624920     1    0.0315107
##    N           500
```

```
##
```

In the table, cell “b” (asthma symptoms among females) gives the lowest expected frequency which is 56.1 - much larger than 5. Therefore, we can conduct a Pearson’s chi-square test.

b) Use a chi-squared test to evaluate the hypothesis and interpret the result. Are the assumptions for a chi-squared test met? Calculate the 95% CI of the difference in proportions.

Note the table above has asthma symptoms with the “No” level appearing first. To estimate the difference in the proportion with asthma symptoms, we should re-order the `asthma` factor so that “Yes” appears first.

```
str(children$asthma)
```

```
##  Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 2 2 2 1 ...
##  - attr(*, "label")= chr "Asthma symptoms"
```

```
children$asthma <- relevel(children$asthma, ref="Yes")

str(children$asthma)
```

```
##  Factor w/ 2 levels "Yes","No": 1 2 2 1 2 2 1 1 1 2 ...
```

We can now request a contingency table with row-percents using `pcRow = TRUE` and `diffProp` to request the difference in proportions.

```
contTables(data=children, rows = gender, cols = asthma,
           pcRow = TRUE, diffProp = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    gender                    Yes          No           Total
##
##    Female    Observed              46          196          242
##              % within row     19.00826     80.99174    100.00000
##
##    Male      Observed              70          188          258
##              % within row     27.13178     72.86822    100.00000
##
##    Total     Observed             116          384          500
##              % within row     23.20000     76.80000    100.00000
##
##
##
##  x² Tests
##
##          Value       df    p
##
##    x²    4.624920     1    0.0315107
##    N          500
```

```
##
##
##
##  Comparative Measures
##
##                                    Value            Lower         Upper
##
##    Difference in 2 proportions    -0.08123518      -0.1546347    -0.007835685
##
##      Rows compared
```

The P value is 0.032 from the Chi-Square value of 4.62 with 1 df, which provides evidence of a difference in the proportions diagnosed with asthma between males and females.

There are four assumptions for a Pearson's chi-squared test:

- each observation must be independent
- each participant is represented in the table once only
- at least 80% of the expected cell counts should exceed a value of five
- all expected cell counts should exceed a value of one.

It is evident from the study design that the observations are independent, and each participant was presented in the table once only (asthma was measured once only). Therefore, the first two assumptions are met. From part (a), we can see that smallest expected cell count is 56.1 (>5). Thus, the last two assumptions are also met.

The proportion of asthma is 27.1% among males and 19.0% among females. There is evidence of a difference in the proportions diagnosed with asthma between males and females (chi-square=4.62 with 1 df, P=0.03). The proportion with asthma is 8.0% lower in females than males with a 95% confidence interval of 0.8% to 15.5% (read from the Risk difference).

Activity 7.2

The file Activity_S7.2.rds summarises the 5-year mortality data for 89 people who did or did not have a heart attack.

a) State the null hypothesis.

   Null hypothesis: There is no association between having a heart attack and risk of death in the next five years

b) Using R, carry out the appropriate significance test to evaluate the hypothesis. Do the data fulfil the assumptions of the statistical test you have used?

   A Pearson's Chi-Square test is appropriate to test the null hypothesis. To check whether our data fulfils the assumptions for a Pearson's Chi-Squared test we need to obtain the expected frequencies for the 2 × 2 table using cross-tabulation:

```
heart <- readRDS("data/activities/Activity_S7.2.rds")

head(heart)
```

```
##   survival heart_attack mort
## 1       No          yes  yes
## 2       No          yes  yes
## 3       No          yes  yes
## 4       No          yes  yes
## 5       No          yes  yes
## 6       No          yes  yes
```

```
str(heart)
```

```
## 'data.frame':    89 obs. of  3 variables:
##  $ survival    : Factor w/ 2 levels "No","yes": 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Survival"
##  $ heart_attack: Factor w/ 2 levels "No","yes": 2 2 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Heart_attack"
##  $ mort        : Factor w/ 2 levels "No","yes": 2 2 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Death"
```

We can see that each variable has been entered as a factor, but "No" is the first level. We should reorder our exposure (`heart_attack`) and outcome (`mort`) so that "yes" becomes the first level:

```
heart$heart_attack <- relevel(heart$heart_attack, ref="yes")
heart$mort <- relevel(heart$mort, ref="yes")

str(heart)
```

```
## 'data.frame':    89 obs. of  3 variables:
##  $ survival    : Factor w/ 2 levels "No","yes": 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Survival"
##  $ heart_attack: Factor w/ 2 levels "yes","No": 1 1 1 1 1 1 1 1 1 1 ...
##  $ mort        : Factor w/ 2 levels "yes","No": 1 1 1 1 1 1 1 1 1 1 ...
```

We can now use `contTables()` to produce the 2-by-2 table with expected values:

```
contTables(data=heart, rows=heart_attack, cols=mort,
           exp=TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    heart_attack                yes          No          Total
##
##    yes             Observed          10          35          45
##                    Expected    7.584270    37.41573    45.00000
##
##    No              Observed           5          39          44
##                    Expected    7.415730    36.58427    44.00000
##
##    Total           Observed          15          74          89
##                    Expected   15.000000    74.00000    89.00000
##
```

```
##
##
##  x² Tests
##
##          Value       df    p
##
##    x²    1.871883     1    0.1712595
##    N           89
##
```

From the study design, it is clear that the observations are independent, and participants are represented only once in the table. Therefore, the data fulfils the first two assumptions of independence. From the above table we can see that the lowest expected frequency is 7.4, which is greater than 5. Thus, the last two assumptions are also met, and we can use the results from the Pearson's Chi-Squared test.

The P-value from the Chi-Square test (chi-square = 1.87 with 1 df) test statistic is 0.17, that is, there is no evidence of an association between heart attack and mortality during the next 5 years.

c) Estimate the appropriate risk estimate for mortality. Are the confidence intervals of the risk estimates consistent with the P value?

Because this is a follow-up study from where we can estimate incidence, the relative risk is the appropriate risk estimate:

```
contTables(data=heart, rows=heart_attack, cols=mort,
           pcRow = TRUE, relRisk = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    heart_attack                    yes          No           Total
##
##    yes             Observed              10           35           45
##                    % within row     22.22222     77.77778    100.00000
##
##    No              Observed               5           39           44
##                    % within row     11.36364     88.63636    100.00000
##
##    Total           Observed              15           74           89
##                    % within row     16.85393     83.14607    100.00000
##
##
##
##  x² Tests
##
##          Value       df    p
##
##    x²    1.871883     1    0.1712595
##    N           89
##
##
```

```
##
##  Comparative Measures
##
##                     Value         Lower        Upper
##
##    Relative risk    1.955556      0.7267615    5.261971
##
##      Rows compared
```

The relative risk is 1.96 (95% CI: 0.73, 5.26). The confidence interval includes the null value (1) which is consistent with the P-value (P = 0.17).

Note that the risk difference would also be an acceptable measure for this study.

d) Summarise your results and state your conclusion.

There is no evidence of an association between heart attack and 5-year mortality (chi-square=1.87 with 1 df, P = 0.17). The risk of dying during next 5 years for those who experienced a heart attack is 1.96 (95% CI: 0.73, 5.26) times that for those who did not experience a heart attack, but the 95% confidence interval indicates that the relative risk may be as low as 0.73 and as high as 5.26 (with 95% confidence).

Activity 7.3

The effect of two penicillin allergens B and G was tested in a random sample of 500 people. All people were tested with both allergens. For each person, data were recorded for whether or not there was an allergic reaction to the allergen.
Use the data set Activity_S7.3.rds to test the null hypothesis that the proportion of participants who react to allergen G is the same as the proportion who react to allergen B. Are the 95% CI around the difference consistent with the P value?

As usual, after reading the data, we should check the ordering of the factor levels for the two variables.

```
study <- readRDS("data/activities/Activity_S7.3.rds")

head(study)
```

```
##   React_Allergen_G React_Allergen_B
## 1            React            React
## 2            React            React
## 3            React            React
## 4            React            React
## 5            React            React
## 6            React            React
```

```
str(study)
```

```
## 'data.frame':    500 obs. of  2 variables:
##  $ React_Allergen_G: Factor w/ 2 levels "Don't react",..: 2 2 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Reacts to allergen G"
##  $ React_Allergen_B: Factor w/ 2 levels "Don't react",..: 2 2 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Reacts to allergen B"
```

It's a little difficult to determine the order from this output, so let's look at a table:

```
table(study$React_Allergen_B)
```

```
##
## Don't react       React
##         432          68
```

```
table(study$React_Allergen_G)
```

```
##
## Don't react       React
##         448          52
```

In both cases, "Don't react" is the first level. This should be re-ordered:

```
study$React_Allergen_B <- relevel(study$React_Allergen_B, ref="React")
study$React_Allergen_G <- relevel(study$React_Allergen_G, ref="React")
```

Because the data are paired, McNemar's test should be used to test the null hypothesis using the `contTablesPaired()` function:

```
contTablesPaired(data=study,
                 rows=React_Allergen_B,
                 cols=React_Allergen_G)
```

```
##
##  PAIRED SAMPLES CONTINGENCY TABLES
##
##  Contingency Tables
##
##    React_Allergen_B    React    Don't react    Total
##
##    React                  40             28       68
##    Don't react            12            420      432
##    Total                  52            448      500
##
##
##
##  McNemar Test
##
##          Value       df    p
##
##    x²    6.400000     1    0.0114120
##    N          500
##
```

We use the `mcNemarDiff()` function (stored in Microsoft Teams and here) to estimate the proportions, the difference in proportions and its 95% confidence interval. We should copy the function and paste it into R. Note that when defining the variables to be summarised, the variable names must be surrounded by quotation marks.

```
### Copied from Microsoft Teams
mcNemarDiff <- function(data, var1, var2, digits = 3) {
  if (!requireNamespace("epibasix", quietly = TRUE)) {
    stop("This function requires epibasix to be installed")
  }

  tab <- table(data[[var1]], data[[var2]])
  p1 <- (tab[1, 1] + tab[1, 2]) / sum(tab)
  p2 <- (tab[1, 1] + tab[2, 1]) / sum(tab)
  pd <- epibasix::mcNemar(tab)$rd
  pd.cil <- epibasix::mcNemar(tab)$rd.CIL
  pd.ciu <- epibasix::mcNemar(tab)$rd.CIU
  print(paste0(
    "Proportion 1: ",
    format(round(p1, digits = digits), nsmall = digits),
    "; Proportion 2: ", format(round(p2, digits = digits), nsmall = digits)
  ))
  print(paste0(
    "Difference in paired proportions: ",
    format(round(pd, digits = digits), nsmall = digits),
    "; 95% CI: ", format(round(pd.cil, digits = digits), nsmall = digits),
    " to ", format(round(pd.ciu, digits = digits), nsmall = digits)
  ))
}
### End copy

mcNemarDiff(data = study, var1 = "React_Allergen_B", var2 = "React_Allergen_G", digits = 3)
```

```
## [1] "Proportion 1: 0.136; Proportion 2: 0.104"
## [1] "Difference in paired proportions: 0.032; 95% CI: 0.005 to 0.059"
```

The McNemar's test result shows evidence of a difference in reactions to allergens B and G (chi-square=6.40 with 1 df, P = 0.011).
The output from the `mcNemarDiff()` function shows that 13.6% of participants react to allergen B and 10.4% of participants react to allergen G, with the difference in reaction rate of 3.2% (95% CI: 0.5% to 5.9%). The 95% CI does not include the null value (which is 0), which is consistent with the P-value being less than 0.05 (P = 0.017).
Therefore, we can conclude that there is evidence (P = 0.017) of a difference in the effect of penicillin allergen B and G. A total of 3.2% more patients reacted to penicillin B (13.6%) compared to penicillin G (10.4%), and we are 95% confident that the true difference in the underlying population is between 0.5% and 5.9%.

Activity 7.4

We examined a survey of 200 live births in an urban region in which 2 babies were born prematurely. We also surveyed 80 live births in a rural region and found that 5 babies were born prematurely. Conduct an appropriate statistical analysis to find out whether the proportion of premature births is higher in the rural region.

> Two cross-sectional surveys were conducted in two different regions (urban and rural), thus the data are independent and each participant appeared in the dataset only once. We need to check the expected frequencies in a 2x2 table to determine if a Pearson's chi-square test or a Fisher's exact test is appropriate.

The data are aggregate, so we need to enter them into R. To produce the required information for the 2x2 table, we need to calculate total number of normal births in each survey by subtracting the number of premature births from the total births. Thus, in the urban area there are 200 − 2 = 198 normal births and in the rural area there are 80 − 5 = 75 normal births. The data should be entered in the following way:

```
babies <- data.frame(
  region = c("Urban", "Urban", "Rural", "Rural"),
  birth = c("Premature", "Not premature", "Premature", "Not premature"),
  n = c(2, 198, 5, 75)
)

babies$region <- factor(babies$region, levels=c("Urban", "Rural"))
babies$birth <- factor(babies$birth, levels=c("Premature", "Not premature"))

babies
```

```
##   region         birth   n
## 1  Urban     Premature   2
## 2  Urban Not premature 198
## 3  Rural     Premature   5
## 4  Rural Not premature  75
```

We can use `contTables()` to calculate the expected counts in each cell. As we are using aggregate data, we must use `counts = n` to specify that the column `n` contains the counts.

```
contTables(data=babies,
           rows = region, cols = birth, counts = n,
           exp = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    region                Premature    Not premature    Total
##
##    Urban     Observed            2              198          200
##              Expected     5.000000        195.00000    200.00000
##
##    Rural     Observed            5               75           80
##              Expected     2.000000         78.00000     80.00000
##
##    Total     Observed            7              273          280
##              Expected     7.000000        273.00000    280.00000
##
##
##
##  x² Tests
##
##           Value       df    p
##
##    x²    6.461538     1    0.0110234
```

```
##    N          280
##
```

We can see from the above output that the lowest expected frequency in cell “c” (expected number of premature birth in the urban region) is 2, which is less than 5. Thus, a Pearson’s chi-square test is not appropriate and we should conduct Fisher’s exact test. For doing this test, we should specify `fisher = TRUE`. Output from the test is shown below:
Stata output 7.8: Cross-tabulation to compare the proportion of premature birth from two surveys

```
contTables(data=babies,
           rows = region, cols = birth, counts = n,
           pcRow = TRUE, exp = TRUE, fisher = TRUE)
```

```
##
##  CONTINGENCY TABLES
##
##  Contingency Tables
##
##    region                    Premature    Not premature    Total
##
##    Urban     Observed              2             198          200
##              Expected          5.000000      195.00000    200.00000
##              % within row       1.00000       99.00000    100.00000
##
##    Rural     Observed              5              75           80
##              Expected          2.000000       78.00000     80.00000
##              % within row       6.25000       93.75000    100.00000
##
##    Total     Observed              7             273          280
##              Expected          7.000000      273.00000    280.00000
##              % within row       2.50000       97.50000    100.00000
##
##
##
##  x² Tests
##
##                             Value       df    p
##
##    x²                       6.461538     1    0.0110234
##    Fisher's exact test                       0.0218217
##    N                           280
##
```

Here we see the proportion of premature birth in the urban area is only 1% and in the rural area it is 6%. The P-value from Fisher’s exact test is 0.022. Thus, we can conclude that there is evidence that the proportion of premature birth in the rural area is higher than that in the urban area.

Module 7: Full script

```
# Author: Timothy Dobbins
# Date: July, 2022
# Purpose: Learning activities for Module 7

# Activity 7.1

library(jmv)

children <- readRDS("data/activities/Activity_S7.1.rds")

contTables(data=children, rows = gender, cols = asthma,
           exp = TRUE)

# Check levels of outcome variable
str(children$asthma)
children$asthma <- relevel(children$asthma, ref="Yes")
str(children$asthma)

contTables(data=children, rows = gender, cols = asthma,
           pcRow = TRUE, diffProp = TRUE)


# Activity 7.2
heart <- readRDS("data/activities/Activity_S7.2.rds")

head(heart)
str(heart)

heart$heart_attack <- relevel(heart$heart_attack, ref="yes")
heart$mort <- relevel(heart$mort, ref="yes")
str(heart)

contTables(data=heart, rows=heart_attack, cols=mort,
           exp=TRUE)

contTables(data=heart, rows=heart_attack, cols=mort,
           pcRow = TRUE, relRisk = TRUE)


# Activity 7.3
study <- readRDS("data/activities/Activity_S7.3.rds")
```

```r
head(study)
str(study)

table(study$React_Allergen_B)
table(study$React_Allergen_G)

study$React_Allergen_B <- relevel(study$React_Allergen_B, ref="React")
study$React_Allergen_G <- relevel(study$React_Allergen_G, ref="React")

contTablesPaired(data=study,
                 rows=React_Allergen_B,
                 cols=React_Allergen_G)

### Copied from Microsoft Teams
mcNemarDiff <- function(data, var1, var2, digits = 3) {
  if (!requireNamespace("epibasix", quietly = TRUE)) {
    stop("This function requires epibasix to be installed")
  }

  tab <- table(data[[var1]], data[[var2]])
  p1 <- (tab[1, 1] + tab[1, 2]) / sum(tab)
  p2 <- (tab[1, 1] + tab[2, 1]) / sum(tab)
  pd <- epibasix::mcNemar(tab)$rd
  pd.cil <- epibasix::mcNemar(tab)$rd.CIL
  pd.ciu <- epibasix::mcNemar(tab)$rd.CIU
  print(paste0(
    "Proportion 1: ",
    format(round(p1, digits = digits), nsmall = digits),
    "; Proportion 2: ", format(round(p2, digits = digits), nsmall = digits)
  ))
  print(paste0(
    "Difference in paired proportions: ",
    format(round(pd, digits = digits), nsmall = digits),
    "; 95% CI: ", format(round(pd.cil, digits = digits), nsmall = digits),
    " to ", format(round(pd.ciu, digits = digits), nsmall = digits)
  ))
}
### End copy

mcNemarDiff(data = study, var1 = "React_Allergen_B", var2 = "React_Allergen_G", digits = 3)


# Activity 7.4
babies <- data.frame(
  region = c("Urban", "Urban", "Rural", "Rural"),
  birth = c("Premature", "Not premature", "Premature", "Not premature"),
  n = c(2, 198, 5, 75)
)

babies$region <- factor(babies$region, levels=c("Urban", "Rural"))
babies$birth <- factor(babies$birth, levels=c("Premature", "Not premature"))

babies
```

```
contTables(data=babies,
           rows = region, cols = birth, counts = n,
           exp = TRUE)

contTables(data=babies,
           rows = region, cols = birth, counts = n,
           pcRow = TRUE, exp = TRUE, fisher = TRUE)
```

Module 8: Solutions to Learning Activities

Activity 8.1

To investigate the effect of body weight (kg) on blood plasma volume (mL), data were collected from 30 participants and a simple linear regression analysis was conducted. The slope of the regression was 68 (95% confidence interval 52 to 84) and the intercept was −1570 (95% confidence interval −2655 to −492).

a) What is the outcome variable and explanatory (exposure) variable?

   Because we want to know the extent to which body weight predicts blood plasma volume, weight is the explanatory variable and blood plasma volume is the outcome.

b) Interpret the regression slope and its 95% CI

   The regression slope is 68 which means that for every 1 kg increase in body weight, blood plasma volume is predicted to increase by 68 mL. The 95% confidence interval indicates that we are 95% confident that the true increase in blood plasma volume per 1 kg in body weight lies between 52 mL and 84 mL.

c) Write the regression equation

   The estimated regression equation is: Blood plasma volume (mL) = − 1570 + 68 × Body weight (kg)

d) If we randomly sampled a person from the population and found that their weight is 80kg, what would be the predicted value of plasma volume for this person?

   By substituting the value of 80 kg into the above equation we can predict that the person's blood plasma level will be: Blood plasma volume = − 1570 + (68 × 80) = 3870 mL.

   Therefore, the predicted plasma volume of a person who is 80 kg is 3870 mL.

Activity 8.2

To examine whether age predicts IQ, data were collected on 104 people. Use the data in the Stata file Activity_8.2.dta to answer the following questions.

a) What are the outcome variable and the explanatory variable?

   We are examining whether age predicts IQ, not the other way around. Therefore, age is the explanatory variable and IQ is the outcome variable.

b) Create a scatter plot with the two variables. What can you infer from the scatter plot?

The plot shows that as age increases, IQ decreases. In other words, there is a negative relationship between the two variables. The relationship appears roughly linear, but is not strong (the points are quite scattered around the line of best fit).

Figure 1: Scatter plot of IQ against age (years)

```
iq <- readRDS("data/activities/Activity_S8.2.rds")

plot(iq$age, iq$iq,
     main="Scatter plot of IQ against age",
     xlab = "Age (years)",
     ylab = "IQ")

abline(lm(iq$iq ~ iq$age))
```

c) Using R, obtain the correlation coefficient between age and IQ and interpret it.

To obtain the correlation coefficient, we use the `cor.test()` function:

R Output 1: Correlation coefficient between IQ and age

```
cor.test(iq$age, iq$iq)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iq$age and iq$iq
## t = -6.2429, df = 102, p-value = 9.952e-09
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6523298 -0.3707534
## sample estimates:
##       cor
## -0.5257982
```

The correlation coefficient is −0.526 indicating that as age increases, IQ decreases, which is consistent with the scatter plot. The P value is <0.0001, indicating that there is very strong evidence of a negative linear association between IQ and age, and the strength of that relationship is fair (based on the descriptions given in Section 8.2.1 of the course notes).

d) Conduct a simple linear regression using R and report the relationship between the two variables including the interpretation of the R-squared value. Are the assumptions for linear regression met in this model?

We use the `lm()` function to regress `iq` (as the outcome variable) on `age` (the explanatory variable). To obtain more descriptive output, we save the model as an object, called `model` here, and then use the `summary()` function. Finally, confidence intervals for the intercept and slopes are obtained using the `confint()` function.

R Output 2: Simple linear regression of IQ on Age

```
model <- lm(iq$iq ~ iq$age)
summary(model)
```

```
##
## Call:
## lm(formula = iq$iq ~ iq$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1164 -1.9364  0.2843  2.0367  9.3070
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.32431    2.57770  41.636  < 2e-16 ***
## iq$age       -0.42344    0.06783  -6.243 9.95e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.255 on 102 degrees of freedom
## Multiple R-squared:  0.2765, Adjusted R-squared:  0.2694
## F-statistic: 38.97 on 1 and 102 DF,  p-value: 9.952e-09
```

```
confint(model)
```

```
##                   2.5 %      97.5 %
## (Intercept) 102.2114610 112.4371606
## iq$age       -0.5579735  -0.2889048
```

The Model Summary table shows the R-squared value of 0.2765. This indicates that 27.6% of the variation in IQ in the sample can be explained by variability in age.

The coefficients section provides the regression coefficients: an estimated intercept of 107.324 and an estimated slope of -0.423.

The equation is estimated as: IQ = 107.324 + (−0.423 × age)

The assumptions for simple linear regression are:

1. the observations are independent of one another;
2. the relation between the explanatory variable and the outcome variable is linear;
3. the residuals are normally distributed.

Information on the first assumption come from the study design. It is not mentioned in the study description that the data were collected on more than one occasion from each participant or that the participants are related to one another in any ways. Therefore, the observations are independent of one another.

Evidence on the second assumption of a linear relationship between the outcome and explanatory variables is obtained from the scatterplot. Figure 1 demonstrates a linear relationship between age and IQ and so this assumption is also satisfied.

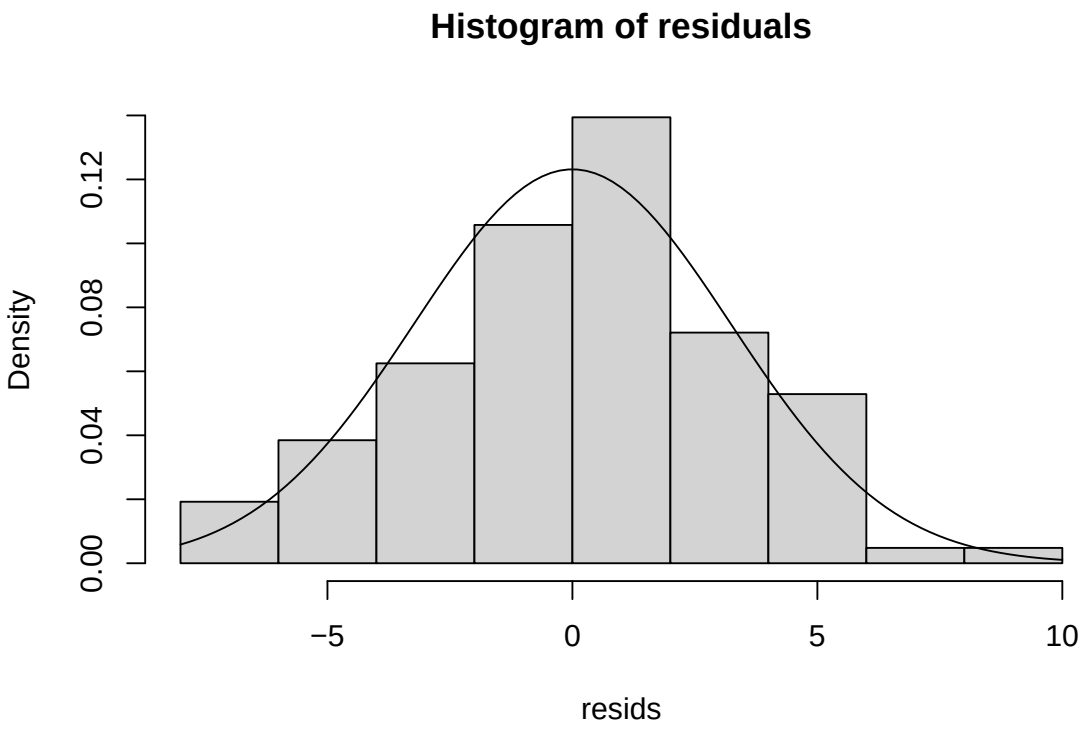
To check the third assumption, that the residuals are normally distributed, we need to first generate and save the residuals in a new object. In R, we can save the residuals using the `resid()` function:

```
resids <- resid(model)
```

To check the assumption, we need to examine the distribution of the residuals using a histogram. The histogram is shown in Figure 2. The histogram shows that the residuals are fairly normally distributed without any remarkable outliers. Therefore, the third assumption is also met.

Figure 2: Distribution of residuals from the regression of IQ on age

```
hist(resids, probability =TRUE,
     main = "Histogram of residuals")
curve(dnorm(x,
            mean=mean(resids),
            sd=sd(resids)), add = TRUE)
```



e) What could you infer about the association between age and IQ in the population, based on the results of the regression analysis in this sample?

This study provides very strong evidence that IQ is negatively associated with age (t=−6.24 with 102 df, P<0.001). For every year increase in age, we predict a 0.42 unit decrease in IQ and we are 95% confident that the true decrease in IQ lies between 0.29 to 0.56 units. Variability in age explains 27.6% of the variability in IQ.

Activity 8.3

Which of the following correlation coefficients indicates the weakest relationship and why? a) r=0.72 b) r=0.41 c) r=0.13 d) r = −0.33 e) r = −0.84

Answer: c.

r = 0.13 which is closest to 0. Note that this only relates to the strength of a linear relationship.

Activity 8.4

Are the following statements true or false?

a) If a correlation coefficient is closer to 1.00 than to 0.00, this indicates that the outcome is caused by the exposure.

False: Correlation cannot tell you about causation; it can only tell you if a relationship or association exists between 2 variables.

b) If a researcher has data on two variables, there will be a higher correlation if the two means are close together and a lower correlation if the two means are far apart.

False: Correlation is not determined by the means of the variables. Correlation only indicates whether the value of one variable increases as the value of the other variable increases or decreases (in a linear way).

Module 8: Full script

```
# Author: Timothy Dobbins
# Date: July, 2022
# Purpose: Learning activities for Module 8

# Activity 8.1
iq <- readRDS("data/activities/Activity_S8.2.rds")

plot(iq$age, iq$iq,
     main="Scatter plot of IQ against age",
     xlab = "Age (years)",
     ylab = "IQ")

abline(lm(iq$iq ~ iq$age))

cor.test(iq$age, iq$iq)

model <- lm(iq$iq ~ iq$age)
summary(model)
confint(model)

resids <- resid(model)

hist(resids, probability =TRUE,
     main = "Histogram of residuals")
curve(dnorm(x,
            mean=mean(resids),
            sd=sd(resids)), add = TRUE)
```