PHCM9795 Foundations of Biostatistics

Learning activity solutions: R version

14 June, 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

2022-06-12 [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)
```
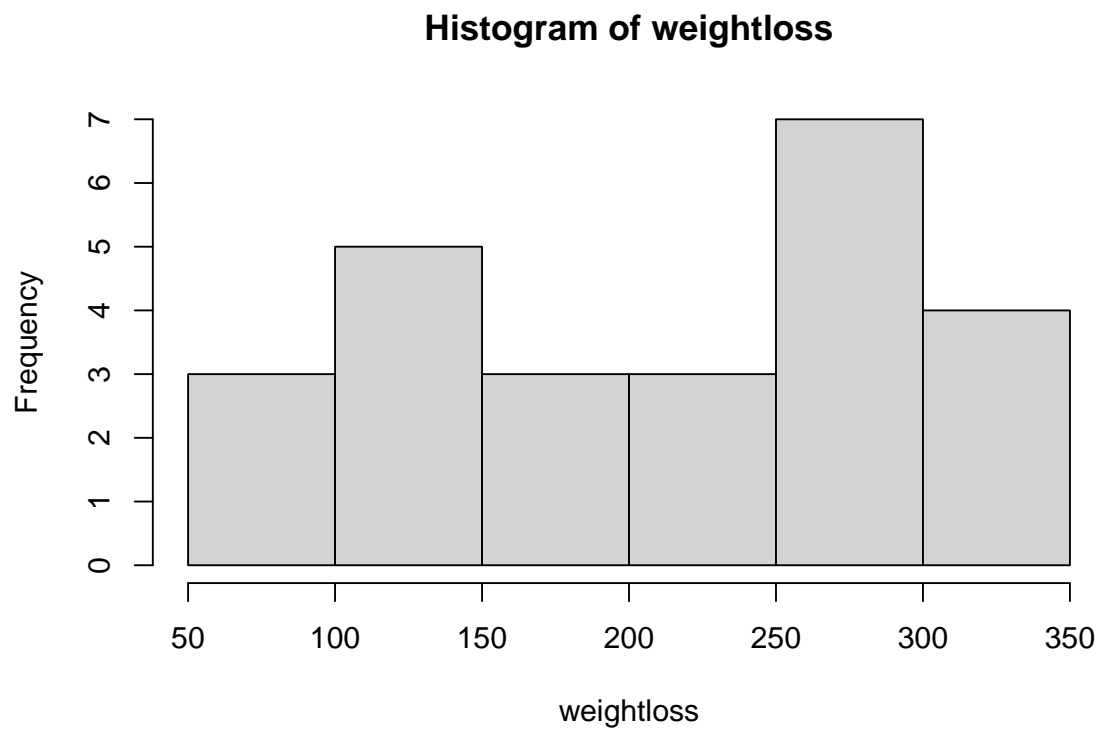
Histogram of weightloss


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

Weight loss for 25 participants
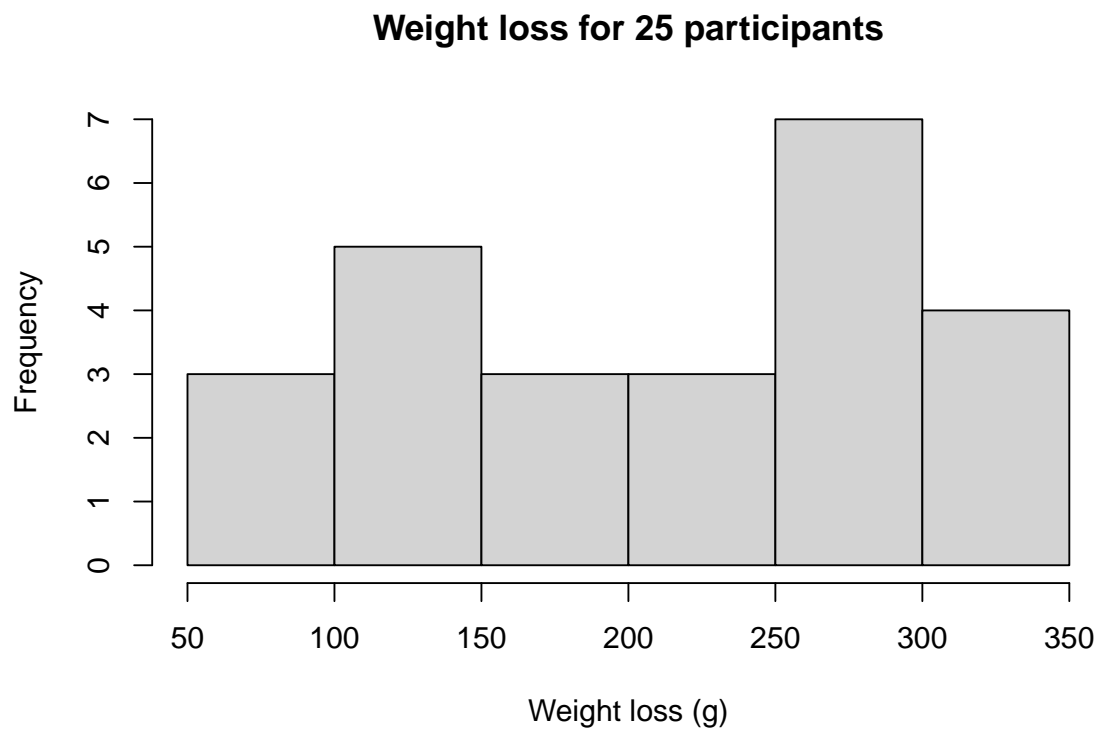
Note that the question requests **relative frequencies**, so we can use the code in Section 1.12 to amend this graph:

```
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
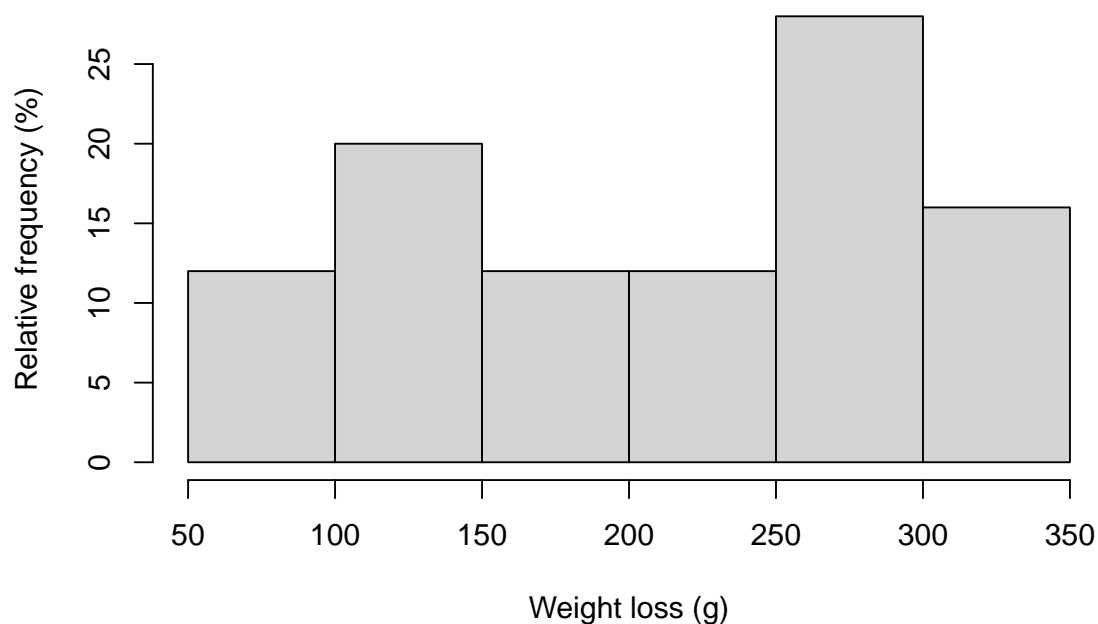
Activity 1.2

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.

```r
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
```

```r
# 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)
```

## Fig 1.2: Frequency of low birth weight by mother's age group
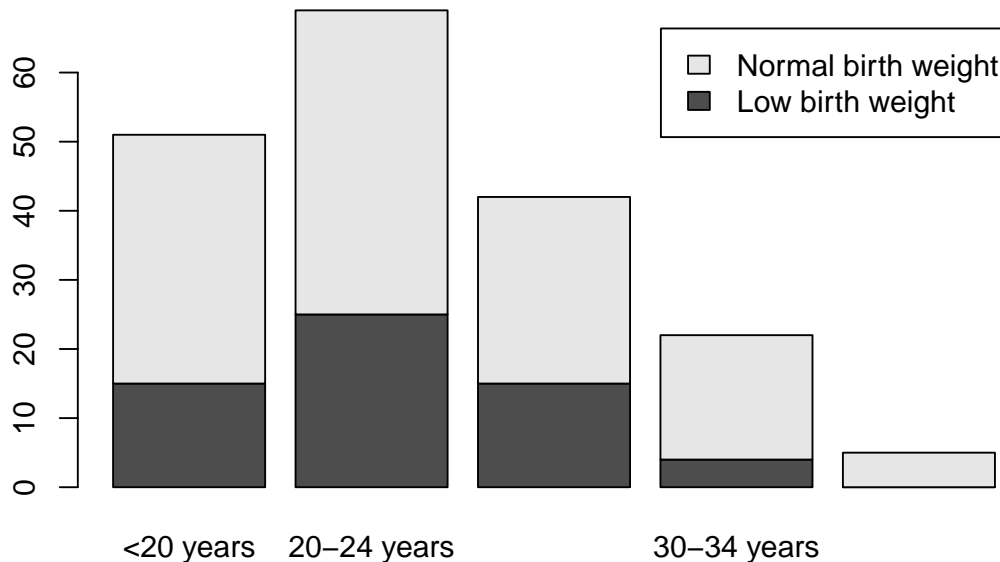
We then calculate the *relative frequency* of low-birth weight by mothers' age group

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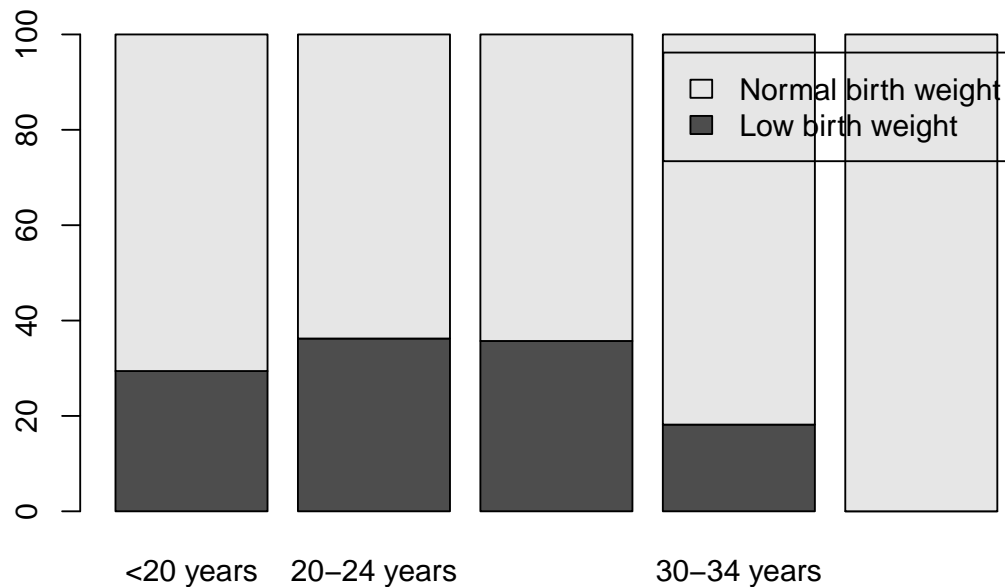
d) Using your answers to the question a) and b), write a brief conclusion about the relationship of low birth weight and mother's age and smoking status.

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?

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

Module 1: Full script

```r
# 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 probability 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.
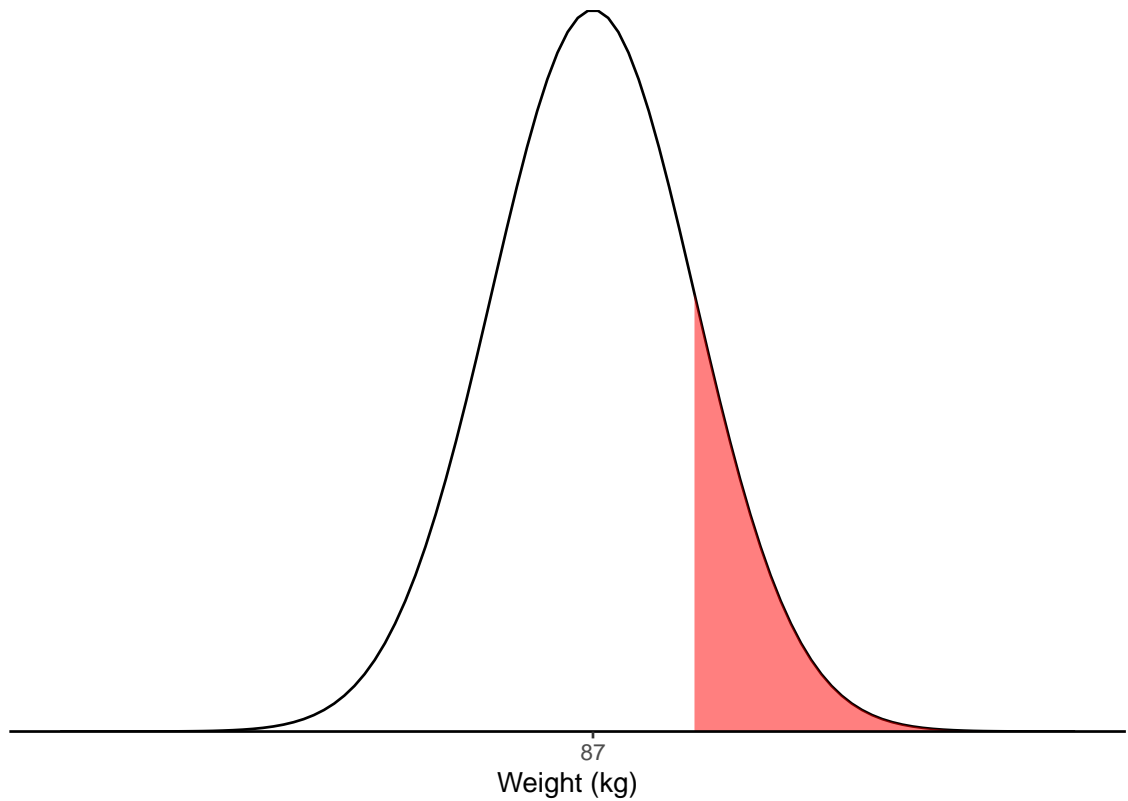
Figure 0.1: Probability that a man will weigh 95kg or more
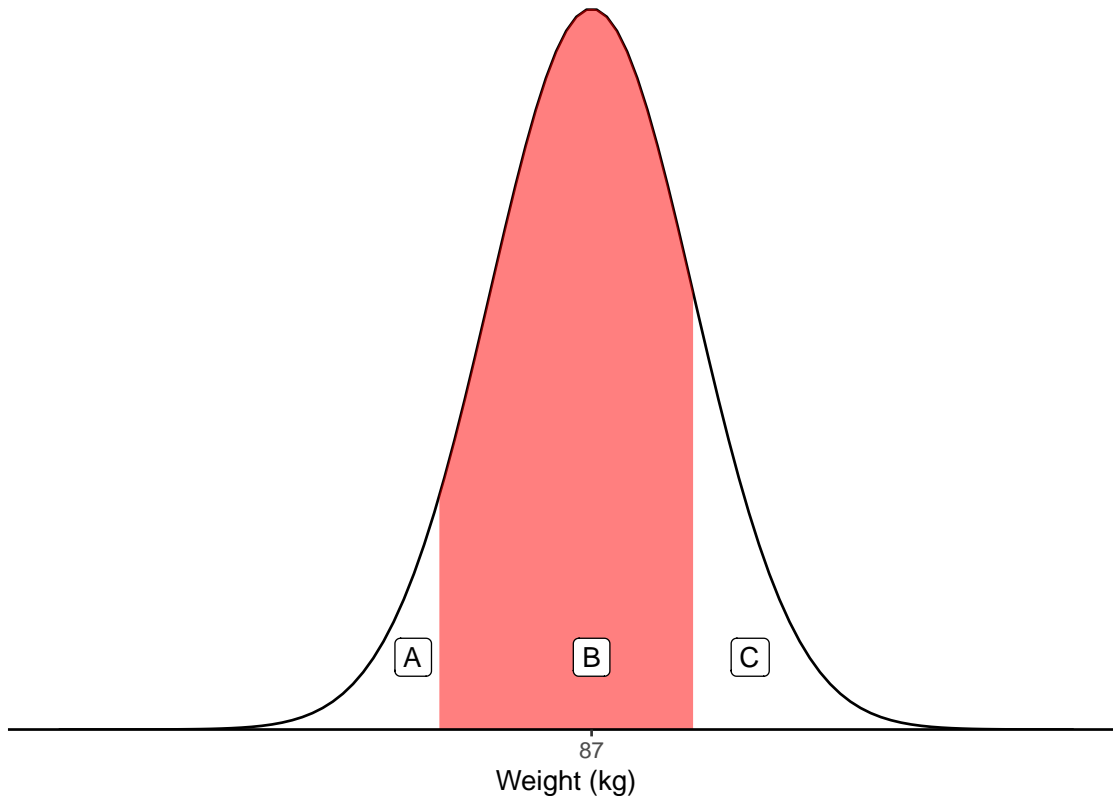
The probability is calculated as:

```
# 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.

87

Weight (kg)

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. $BMI = \frac{weight \ (kg)}{[height \ (m)]^2}$. 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)")
```

Histogram of BMI
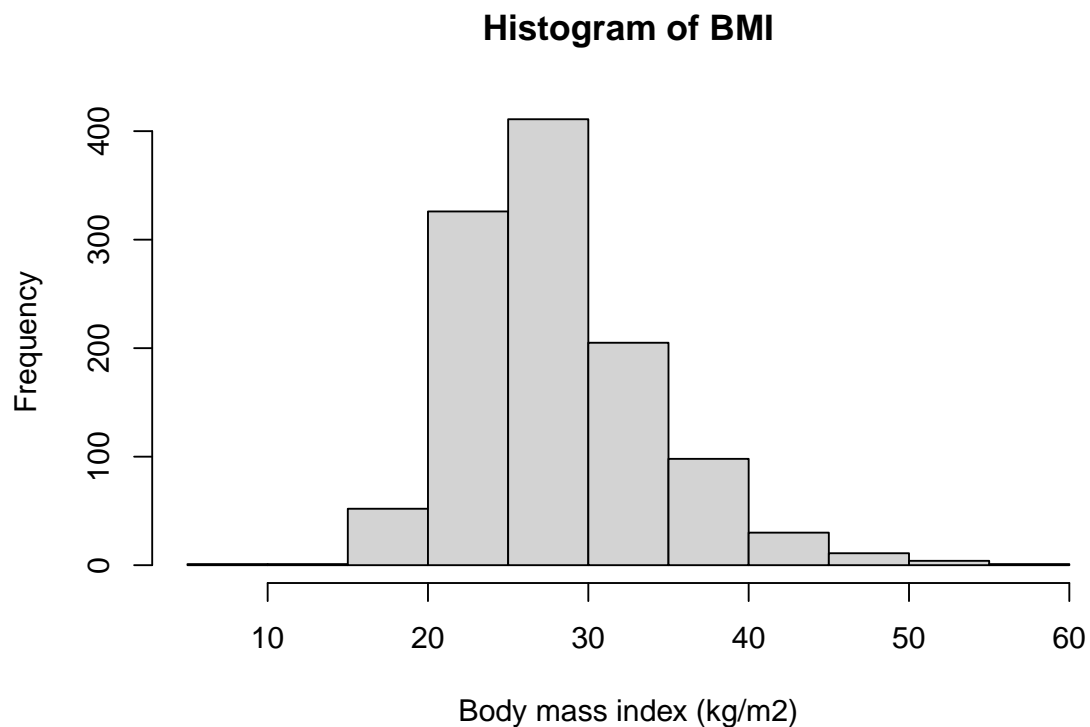

```
boxplot(survey$bmi, main="Boxplot of BMI", ylab="Body mass index (kg/m2)")
```

Boxplot of BMI

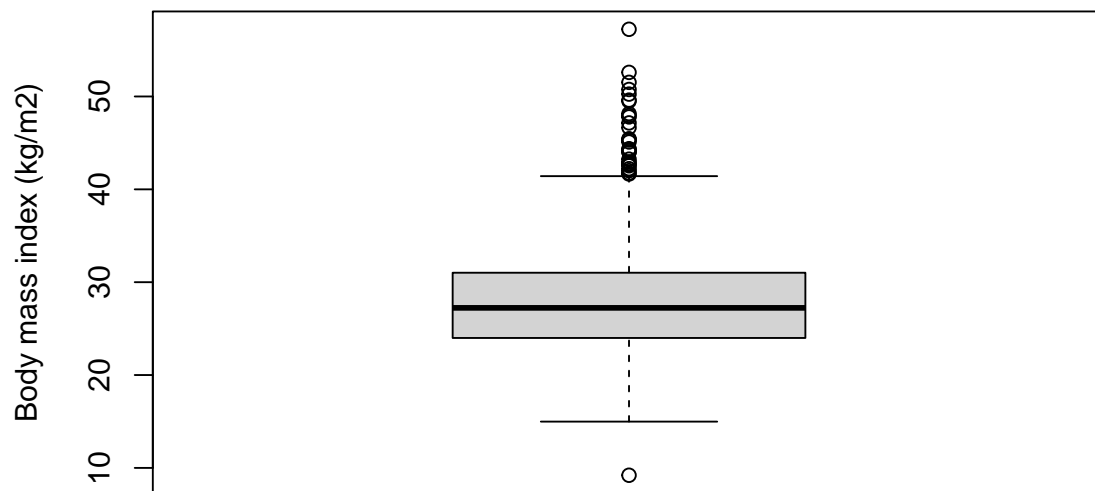The boxplot in particular shows that there are some extreme values of BMI. We can examine these records by viewing records with BMI less than, say 15, or greater than 45:

```
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
##
##
##
##  χ² Tests
##
##          Value         df      p
##
##   χ²    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;
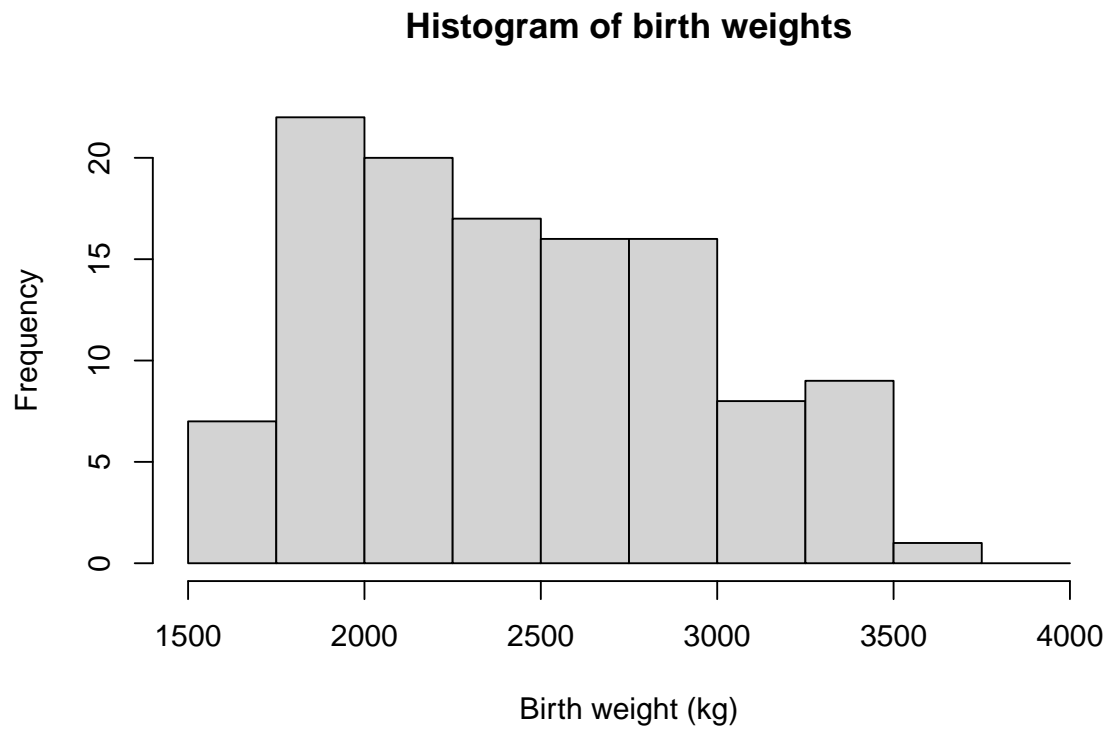
```
babies <- readRDS("data/activities/Activity_S2.5-LengthOfStay.rds")
summary(babies)
```

```
##       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)")
```
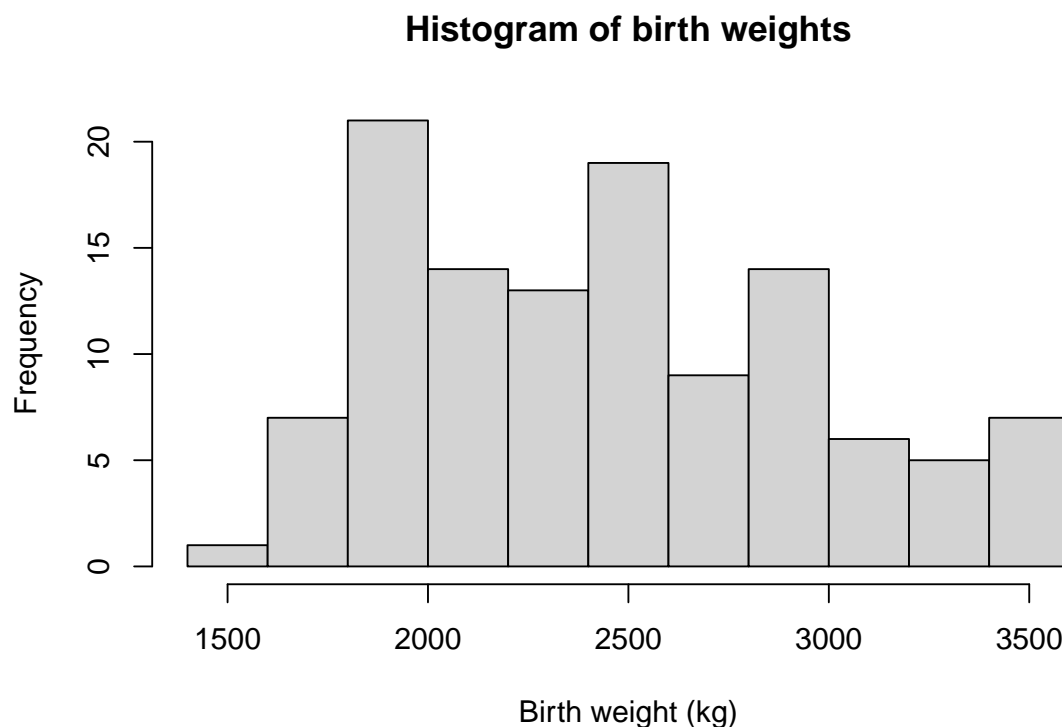


```
# We can specify our own cutpoints using the breaks command, with the seq() function:
hist(babies$BirthWt, main="Histogram of birth weights",
     xlab="Birth weight (kg)",
     breaks=seq(from=1500, to=4000, by=250))
```

Histogram of birth weights


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

Histogram of lengths of stay


```
hist(babies$LengthStay, main="Histogram of lengths of stay",
     xlab="Length of stay (days)",
     breaks=seq(from=0, to=250, by=25))
```
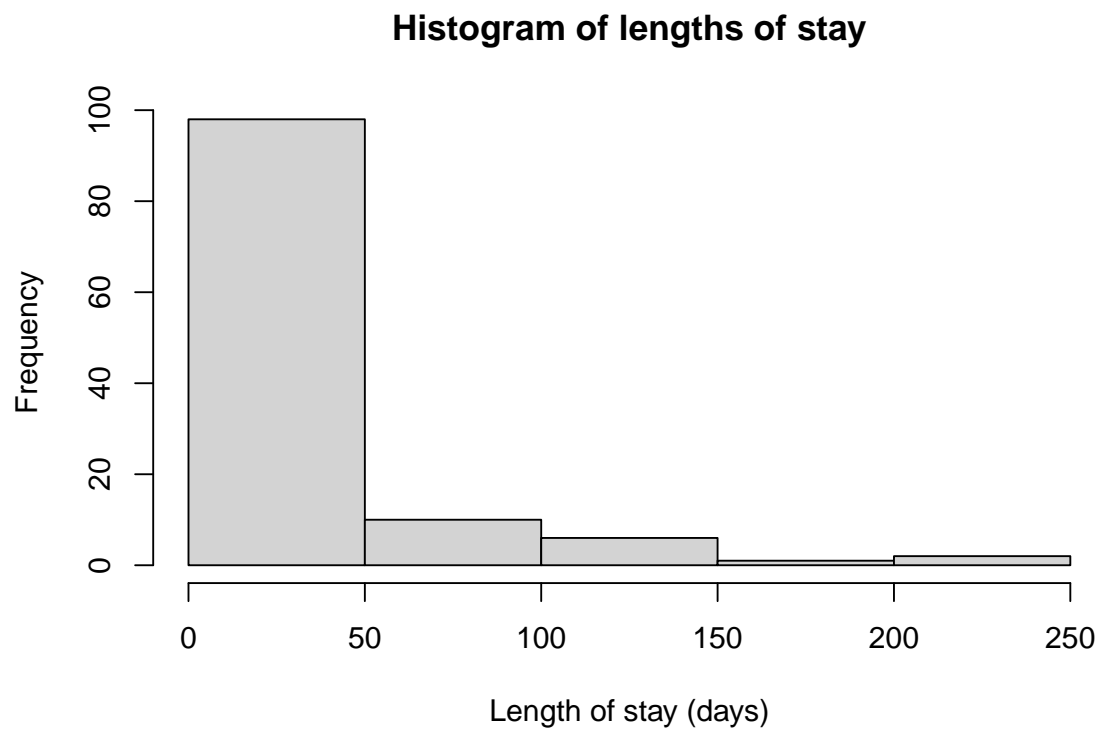
Histogram of lengths of stay
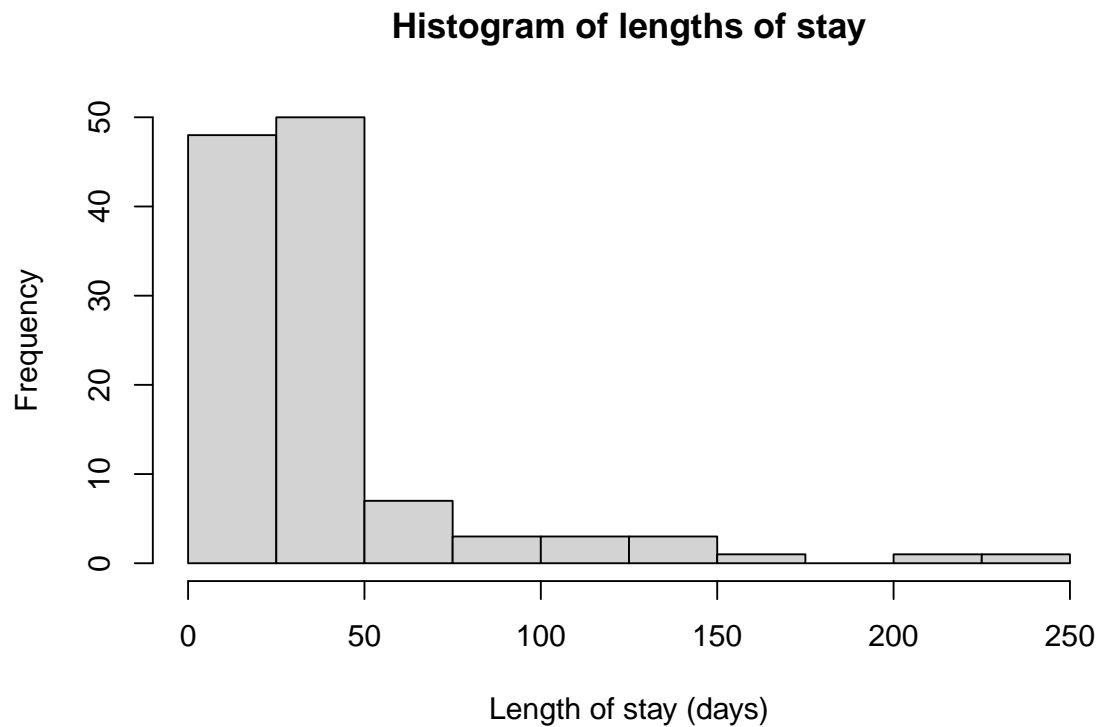
The histogram for birthweight shows a roughly symmetric distribution. The histogram for length of stay shows a highly skewed distribution (skewed to the right).

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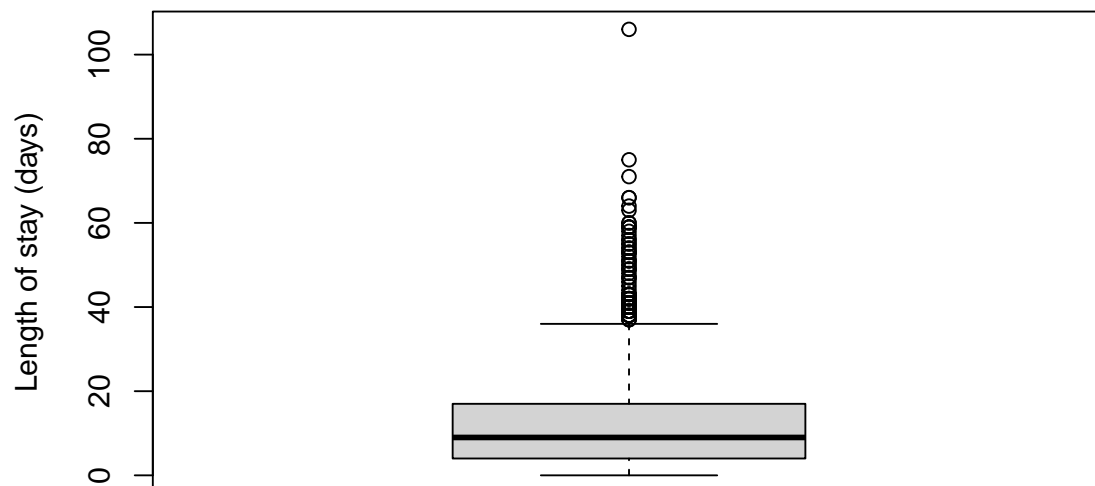
```
hist(hospstay$los, main="Histogram of hospital stay", xlab="Length of stay (days)")
```

Histogram of hospital stay
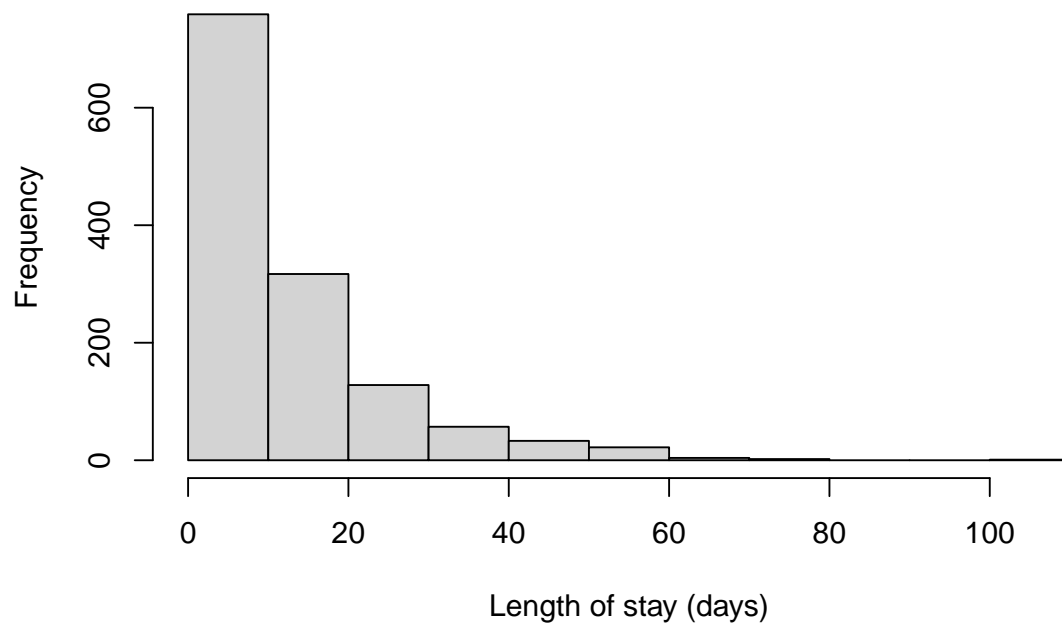

```
boxplot(hospstay$los, main="Boxplot of hospital stay", ylab="Length of stay (days)")
```

Boxplot of hospital stay

```
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")
```
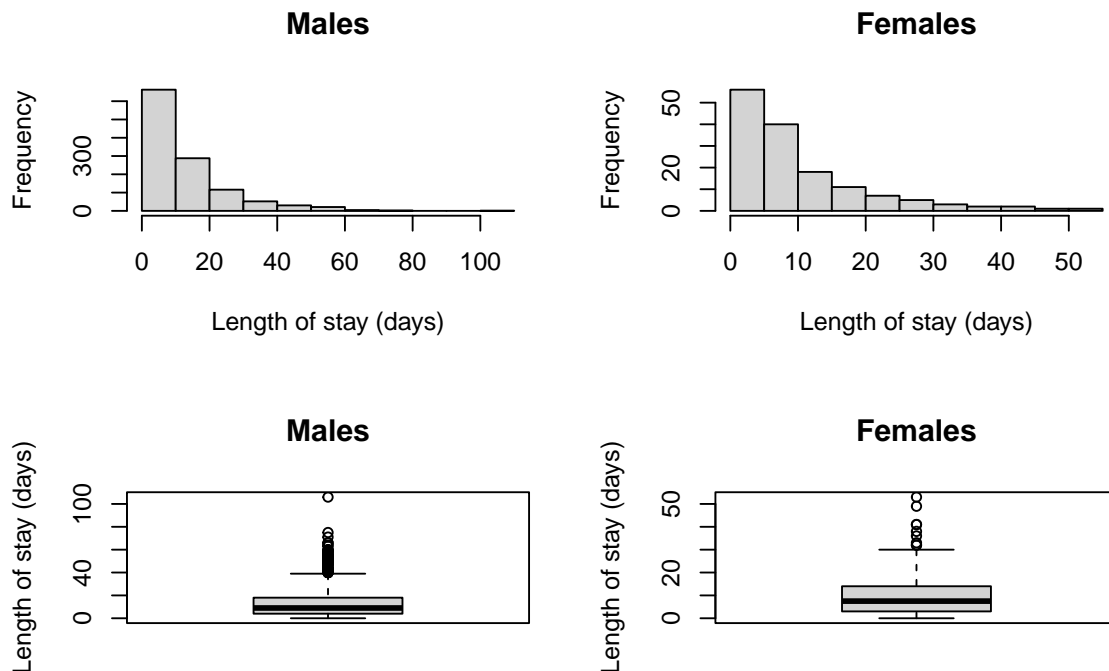


```
# Reset graphics parameters
par(mfrow=c(1,1))
```

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

```r
# 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)")
```

```r
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.

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

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

```
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.

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

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

```
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.

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

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

```
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.

```
# 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)")
```

Figure 3.1: Distribution of diastolic blood pressure
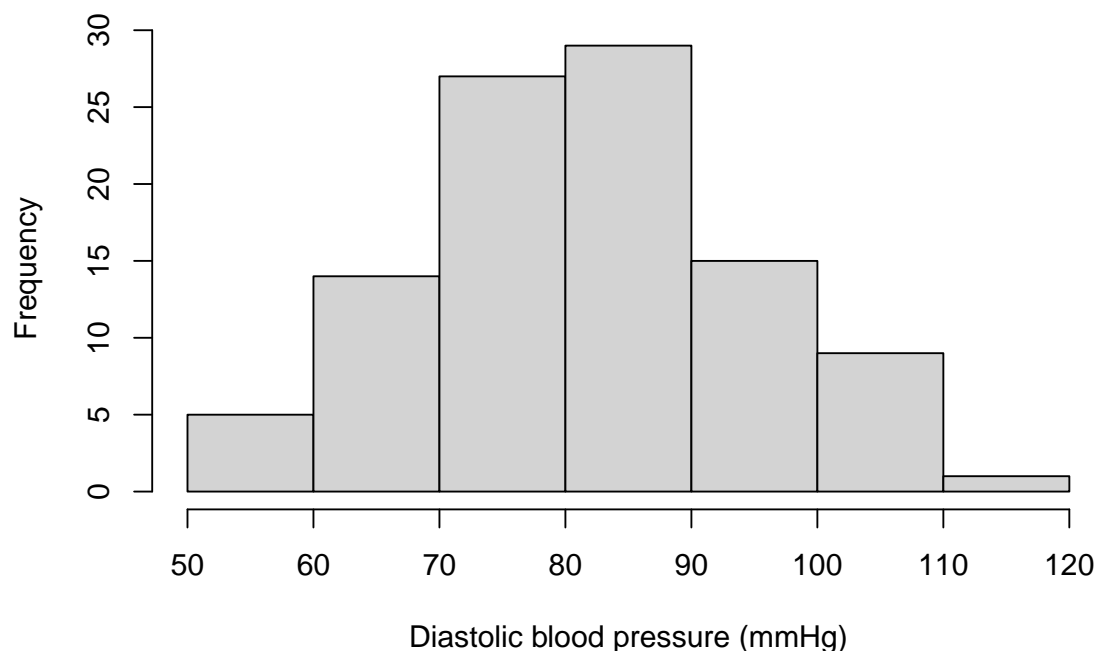
The distribution is approximately symmetrical, centered about the mean.

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)")
```

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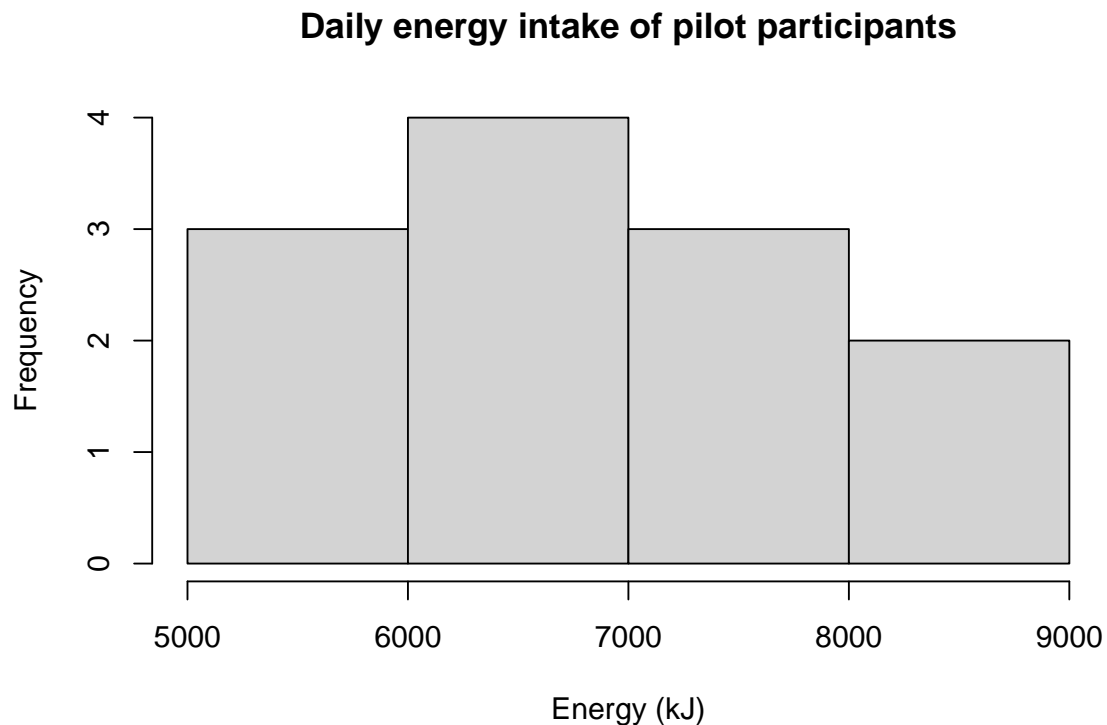
It is very difficult to assess the shape of a distribution with only 12 observations, but here we can see that the distribution looks roughly symmetric. In this case, we will assume Normality.

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

```r
# 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)
```