Module 2 solutions

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 Stata or 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` for Stata, `size` for R)
b) The number of successes we are interested in (`k` for Stata, `x` or `q` for R)
c) The probability of success for each trial (`p` for Stata, `prob` for R)
d) The form of the binomial function
    - for Stata: `binomialp`, `binomial` or `binomialtail`;
    - for R: `dbinom`, `pbinom` or `pbinom(lower.tail=FALSE)`
e) The final probability.

Answers

a) Here, each participant represents a 'trial', so `size` is 90.
b) We are interested in determining the probability that 60 or more participants prefer the new drug. This corresponds to *more than 59*, so we need to define `q` as 59.
c) We are told to assume that the two drugs are equally preferred, so `prob` is 0.5.
d) 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 59 successes. We therefore want to use `pbinom` function with `lower.tail=FALSE` here.

e) The result computed by R is 0.00103013. Therefore, the probability that 60 or more patients would prefer the new drug is 0.001 or 0.1%.

Process

We used the pbinom function, completed as follows:

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

```
[1] 0.001030133
```

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 Stata or R to compute the probability that a patient with a low level of infection:

a) Will not be identified?
b) Will be identified in two of the samples?
c) Will be identified in all the samples?
d) Will be identified in at most 3 of the samples?

Answers

a) The probability P(X=0) = 0.116 or 11.6%.
b) The probability P(X=2) = 0. 336 or 33.6%.
c) The probability P(X=5) = .005 or 0.5%.
d) The probability P(X≤3) = .946 or 94.6%.

Process

In all of these questions, size is 5 and prob is 0.35. For (a) to (c), we need to calculate the probability of finding a certain number of infected samples, and we can use the dbinom function:

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

```
[1] 0.1160291
```

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

```
[1] 0.3364156
```

```
# Part (c)
dbinom(x=5, size=5, prob=0.35)
```

```
[1] 0.005252187
```

For part (d), “at most 3 samples” is the same as 3 or fewer samples, so we can use the `pbinom` function.

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

```
[1] 0.9459775
```

Activity 2.3

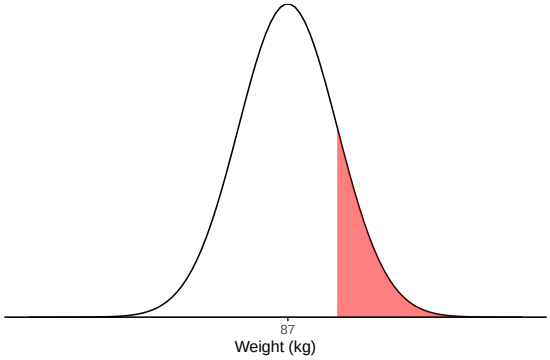
If weights of men are Normally distributed with a population mean μ = 87, and a population standard deviation, σ = 8 kg:

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).
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 Normal curve.

Answers

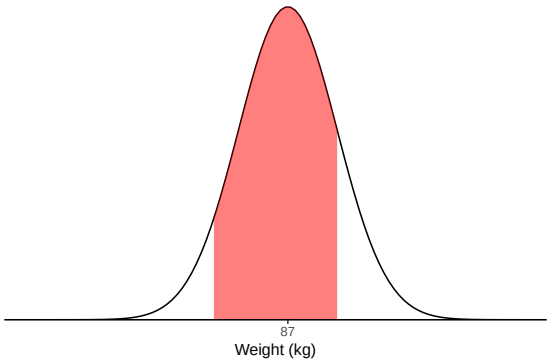
a) The probability that a man from this population weighs 95 kg or more is 0.16 or 16% (Figure 1).

Figure 1: Probability that a man will weigh 95kg or more



b) The probability that a man will weigh more than 75 kg but less than 95 kg is 0.77, or 77% (Figure 2).

Figure 2: Probability that a man will weigh more than 75kg by less than 95kg



Process

a) The curve representing the desired probability is shown in Figure 1, 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.

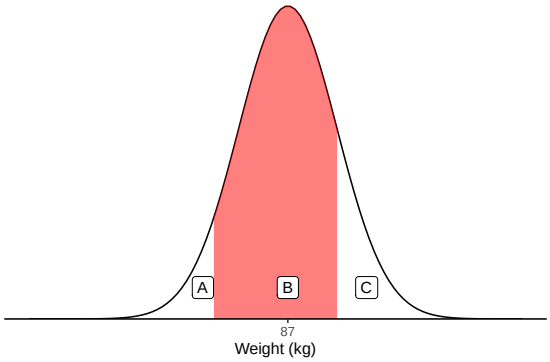
We can calculate the probability directly from R using the pnorm function. As we are calculating the probability above a certain value, we use lower.tail=FALSE

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

```
[1] 0.1586553
```

b) 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.

Figure 3: Probability that a man will weigh more than 75kg by less than 95kg



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
```

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

Alternatively, we could calculate:

```r
pnorm(95, mean=87, sd=8, lower.tail=TRUE) -
  pnorm(75, mean=87, sd=8, lower.tail=TRUE)
```

```
[1] 0.7745375
```

Activity 2.4

Using the health survey data (`Activity_S2.4.xlsx`) described in the computing 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:

- Underweight: BMI < 18.5
- Normal weight: 18.5 $\leq$ BMI < 25
- Pre-obesity: 25 $\leq$ BMI < 30
- Obesity Class I: 30 $\leq$ BMI < 35
- Obesity Class II: 35 $\leq$ BMI < 40
- Obesity Class III: BMI $\geq$ 40

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?

Answers

Table 1: Frequencies of BMI category by sex for 1140 participants in a health survey

BMI category	**Male**	**Female**	**Total**
Underweight	6 (1.2%)	12 (1.9%)	18 (1.6%)
Normal weight	134 (26.1%)	228 (36.4%)	362 (31.8%)
Pre-obesity	216 (42.1%)	195 (31.1%)	411 (36.1%)
Obesity Class I	95 (18.5%)	106 (16.9%)	201 (17.6%)
Obesity Class II	46 (9.0%)	55 (8.8%)	101 (8.9%)
Obesity Class III	16 (3.1%)	31 (4.9%)	47 (4.1%)
Total	**513 (100.0%)**	**627 (100.0%)**	**1,140 (100.0%)**

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%).

Process

We first read the Excel data into R, using the readxl package. It is useful to examine the dataset - here using the `summary()` function:

```
library(readxl)
library(jmv)

survey <- read_excel("data/activities/Activity_S2.4-health-survey.xlsx")
summary(survey)
```

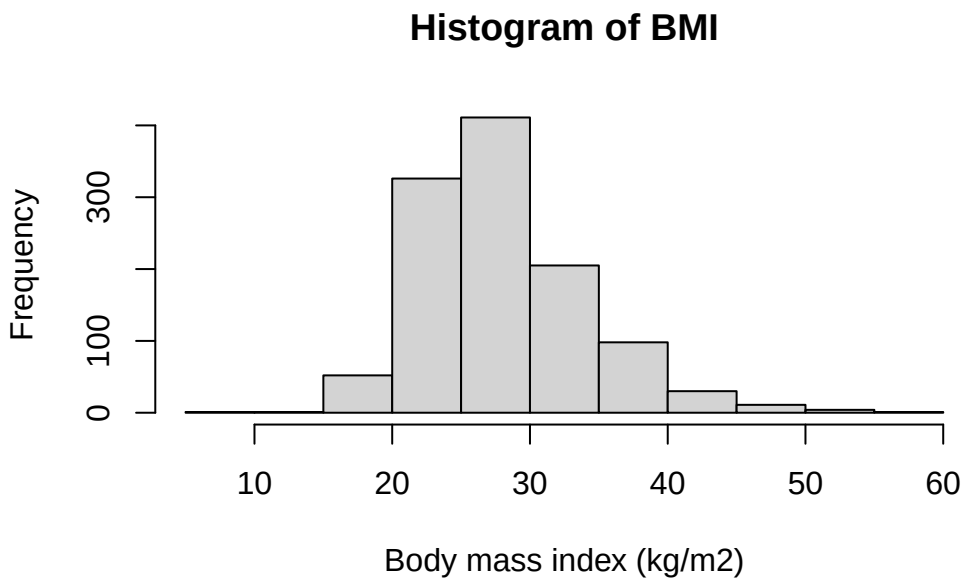
```
      sex           height          weight      
 Min.   :1.00   Min.   :1.220   Min.   : 22.70  
 1st Qu.:1.00   1st Qu.:1.630   1st Qu.: 68.00  
 Median :2.00   Median :1.700   Median : 79.40  
 Mean   :1.55   Mean   :1.698   Mean   : 81.19  
 3rd Qu.:2.00   3rd Qu.:1.780   3rd Qu.: 90.70  
 Max.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

Note that has been entered as a numeric variable. We should define sex as a factor, and then create BMI. After creating BMI, we should examine its distribution using a histogram and/or a boxplot:

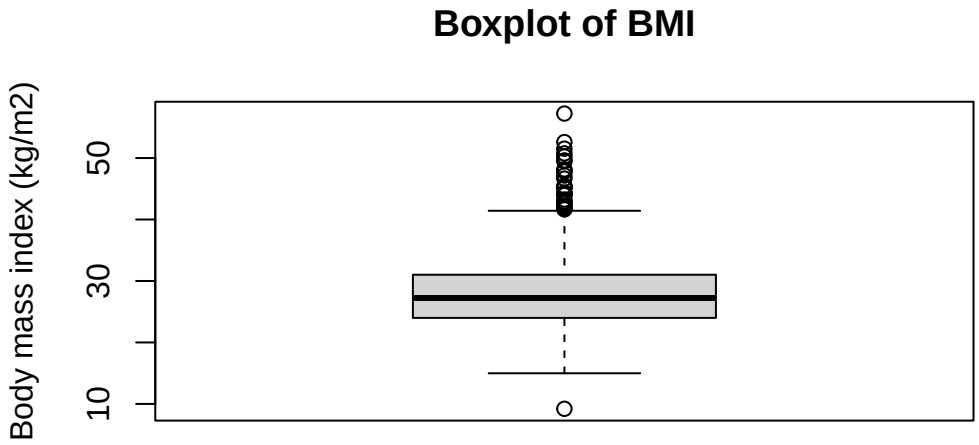
```
# Define sex as a factor
survey$sex <- factor(survey$sex, level=c(1,2), labels=c("Male", "Female"))

# Create BMI
survey$bmi = survey$weight / (survey$height^2)

# Examine the distribution of BMI
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 some records using the `subset()` function:

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

sex	height	weight	bmi
Female	1.57	22.7	9.21
Female	1.65	40.8	15

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

sex	height	weight	bmi
Female	1.52	105	45.4
Male	1.85	174	50.8
Female	1.22	74.8	50.3
Male	1.93	213	57.2
Female	1.63	127	47.8
Female	1.55	115	48
Female	1.65	131	48.2
Female	1.55	109	45.3
Male	1.78	143	45.1
Female	1.65	127	46.6
Female	1.63	132	49.5
Female	1.7	152	52.6
Female	1.6	127	49.6
Female	1.5	106	47.2
Female	1.73	154	51.5
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
```

Note the labelling of the categories. A square bracket indicates an inclusive range, while a round bracket indicates a range that doesn't include that value. For example, [18.5, 25) indicates a category of BMIs that are greater than or equal to 18.5, but less than 25.

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
```

```
² Tests

       Value       df     p

  ²    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
```

^{2} Tests

	Value	df	p
^{2}	22.49802	5	0.0004209
N	1140		

Activity 2.5

The data in the files `Activity_S2.5.dta` and `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; b) Complete the following summary statistics for each variable: - mean and median; - standard deviation and interquartile range; - skewness and kurtosis.

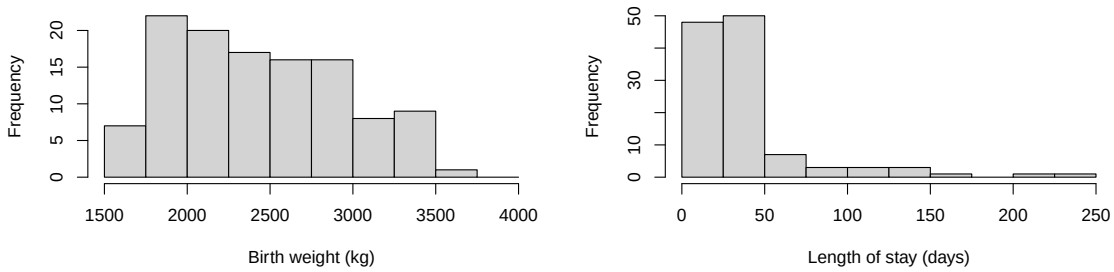
Make a decision about whether each variable is symmetric or not, and which measure of central tendency and variability should be reported.

Answers

a) See Figure 4.

Figure 4: Summary of data from 117 babies admitted to a hospital

(a) Birth weights (b) Length of stay



b) See Table 2.

Table 2: Summary of data from 117 babies admitted to a hospital

	Birthweight (grams)	Length of stay (days)
Mean (Standard deviation)	2451 (504.8)	41 (36.9)
Median [Interquartile range]	2438 [2004 to 2830]	30 [21 to 43]
Skewness	0.35	3.1
Kurtosis	-0.7	11.6

As the histogram for birthweight shows 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.

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

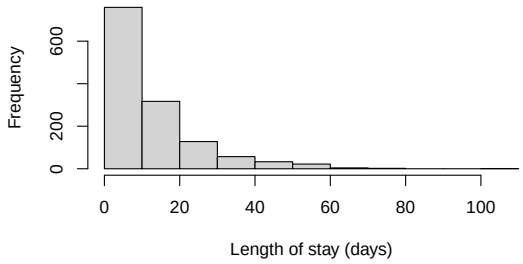
The data set of hospital stay data for 1323 hypothetical patients is available on Moodle in csv format (`Activity2.6.csv`). Import this dataset into Stata or R. There are two variables in this dataset: - female: female=1; male=0 - los: length of stay in days

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

a) The histograms for overall length of stay (Figure 4) and length of stay by gender (Figure 5) all show that length of stay is heavily skewed (skewed to the right).

Figure 5: Summary of length of hospital stay

(a) Histogram



(b) Boxplot

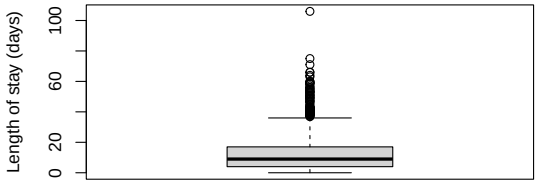
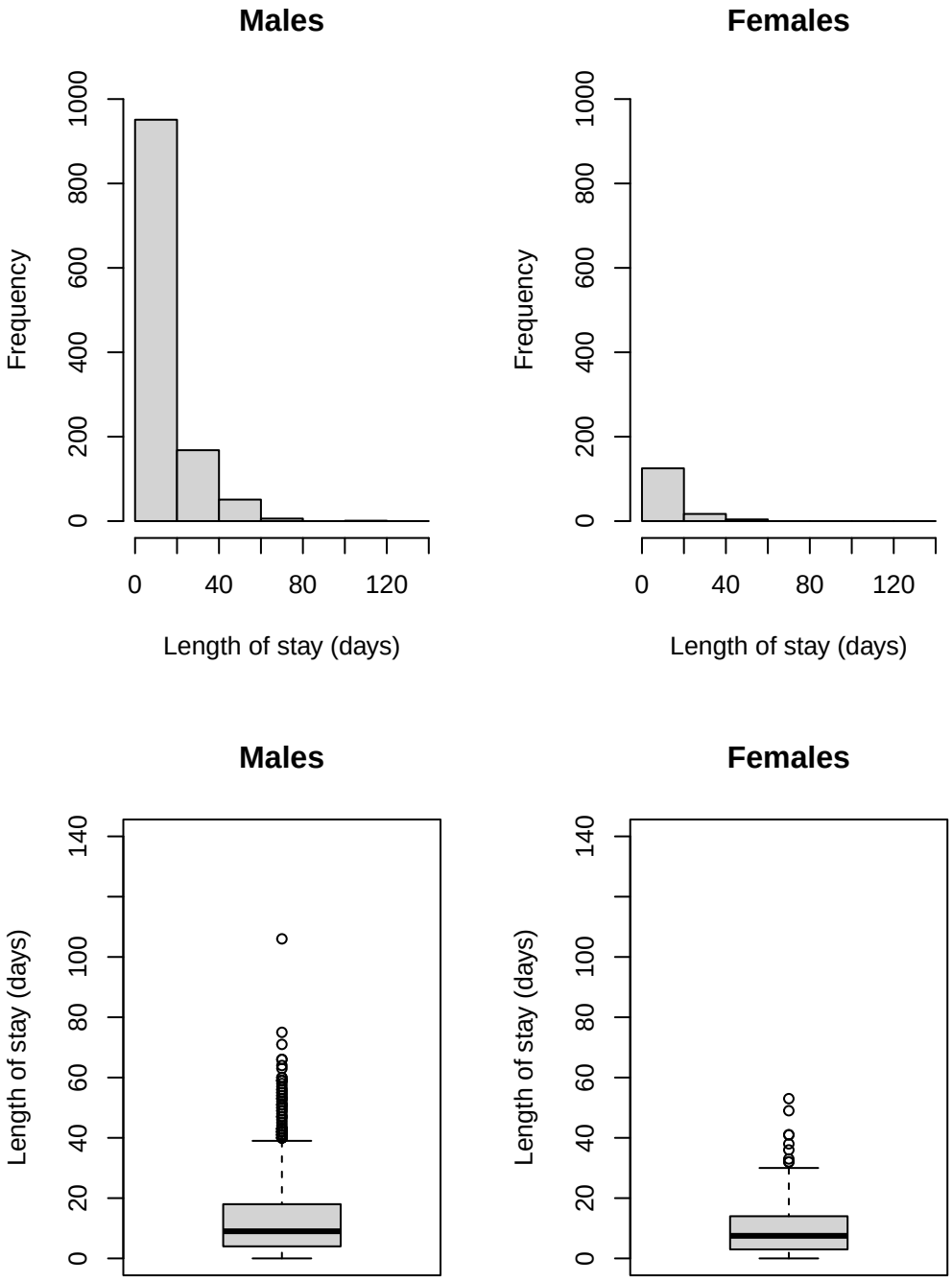


Figure 6: Summary of length of hospital stay



b) 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 summarize command. The median length of stay is 9 days, with an interquartile range of 4 to 17 days.
c) For men, the median length of stay was 9 days, with an interquartile range from 4 to 18 days. For women, the median length of stay was 8 days, with an interquartile range from 3 to 14 days. The lengths of stay for men and women are similar.

Process

The process of plotting histograms and boxplots by another variable (here, sex) is discussed in Section 2.30 in the course notes. Essentially, we can split the graphics window into 2 rows and 2 columns using the `par` function. It is important to keep both the x- and y-axes consistent for both plots, which we can do by specifying the limits and the breaks:

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

# Specify each plot separately
# First, the histograms
# Notice we specify the x-limits (from 0 to 140) and the breaks to keep
# the x-axes consistent

# We also specify the y-limits to keep the y-axes consistent

hist(hospstay_males$los,
     xlab="Length of stay (days)", main="Males",
     xlim=c(0, 140), breaks=c(0, 20, 40, 60, 80, 100, 120, 140),
     ylim=c(0, 1000))
hist(hospstay_females$los,
     xlab="Length of stay (days)", main="Females",
     xlim=c(0, 140), breaks=c(0, 20, 40, 60, 80, 100, 120, 140),
     ylim=c(0, 1000))

# Next, the boxplots
# Notice we specify the y limits and the breaks

boxplot(hospstay_males$los,
        ylab="Length of stay (days)", main="Males",
        ylim=c(0, 140), breaks=c(0, 20, 40, 60, 80, 100, 120, 140))
boxplot(hospstay_females$los,
        ylab="Length of stay (days)", main="Females",
        ylim=c(0, 140), breaks=c(0, 20, 40, 60, 80, 100, 120, 140))
```

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

The summary statistics for all participants can be calculated in the usual way:

```
descriptives(hospstay, vars=los, pc=TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                         los

   N                        1323
   Missing                     0
   Mean                 12.51550
   Median                      9
   Standard deviation   12.59933
   Minimum                     0
   Maximum                   106
   25th percentile      4.000000
   50th percentile      9.000000
   75th percentile      17.00000
```

Summary statistics by gender are calculated by defining a `splitBy` variable:

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

```
 DESCRIPTIVES

 Descriptives

                         female    los

   N                     0            1177
                         1             146
   Missing               0               0
                         1               0
   Mean                  0         12.75531
                         1         10.58219
```

Median	0	9
	1	7.500000
Standard deviation	0	12.83475
	1	10.34625
Minimum	0	0
	1	0
Maximum	0	106
	1	53
25th percentile	0	4.000000
	1	3.000000
50th percentile	0	9.000000
	1	7.500000
75th percentile	0	18.00000
	1	14.00000