Module 1 solutions

Module 2: Solutions to Learning Activities

Activity 2.1

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 2500g or less at birth. Data were collected on age and smoking status of mothers and the birth weight of their babies. The file `Activity_2.1.rds` contain data on the participants in the study. The file is located on Moodle in the Learning Activities section.

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. Answer the following questions:

a) What was the total number of mothers who smoked during pregnancy?
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?
c) 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 the age group of the mother? Provide appropriate labels for the axes and give the graph an appropriate title. [Hint: plot the data using the `AgeGrp` variable]
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.

Answers

Table 1: Cross tabulation of smoking status during pregnancy by low birth weight of the babies among 189 mothers

Smoking status during pregnancy	Low birth weight		
	Yes (%)	No (%)	Total (%)
Yes	30 (40.5)	44 (59.5)	74 (100)
No	39 (25.2)	86 (74.8)	115 (100)
Total	59 (31.2)	130 (68.8)	189 (100)

Note: this table has been constructed from R output.

a) There were 74 mothers who smoked during pregnancy.
b) 41% of mothers who smoked and 25% of non-smoking mothers gave birth to low-birth-weight babies.
c) See Figure 2.1.
d) 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). 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 (Table 1).

NB: You will revisit two-way tables in Module 7 where you will conduct statistical tests to determine if the there is evidence of a difference in proportions.

Process

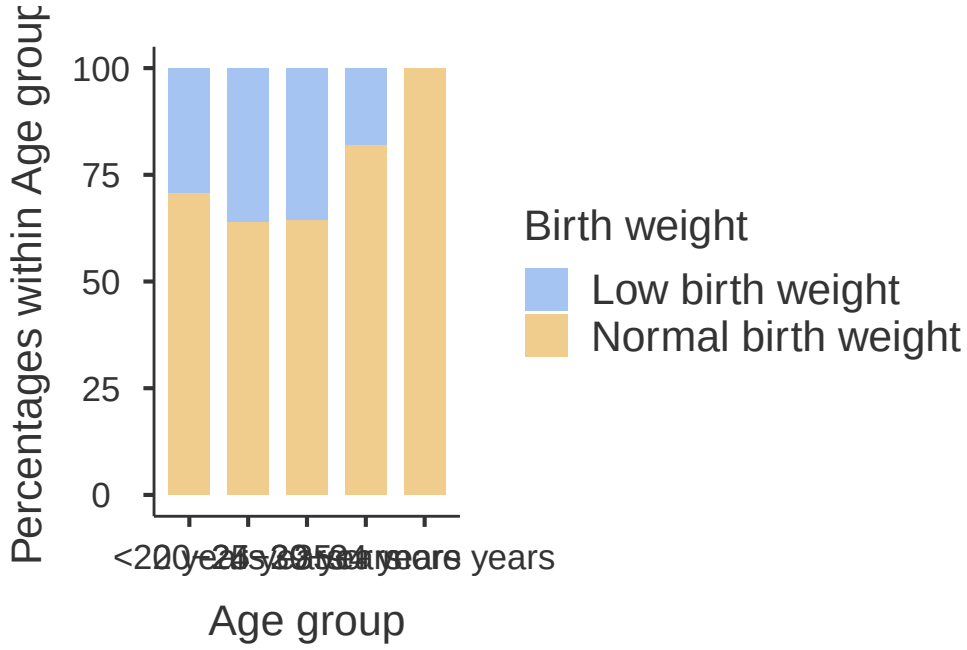
Table 1 was created using the following code:

```
library(jmv)
babies <-readRDS("data/activities/Activity_1.2.rds")

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

Figure 1 was also created using the `contTables` function, and is based on thinking of constructing a two-way table consisting of `Age group` and `Low birth weight`, and requesting percents within each age group:

Figure 1: Relative frequency of low birth weight by mother's age group




```
contTables(data=babies, rows=LOW, cols=AgeGrp, pcCol=TRUE)
```

```
CONTINGENCY TABLES

Contingency Tables

  LOW                                <20 years    20-24 years    25-29 years    30

  Low birth weight       Observed           15             25             15
                         % within column  29.41176       36.23188       35.71429

  Normal birth weight    Observed           36             44             27
                         % within column  70.58824       63.76812       64.28571

  Total                  Observed           51             69             42
                         % within column 100.00000      100.00000      100.00000


² Tests

        Value       df     p
```

```
    ²     5.291374      4     0.2586855
    N           189
```

We can request a bar-chart using `barplot=TRUE`, where the x-axis of the plot is `Age group` (i.e. the columns), and we want the bars to be stacked so we add `xaxis = "xcols"` and `bartype = "stack"`. We request the y-axis to be percentages (rather than counts: `yaxis="ypc"`), and the percentages to be the column percents from the previous table (`yaxisPc = "column_pc"`). Putting it all together:

```
contTables(data=babies, rows=LOW, cols=AgeGrp, pcCol=TRUE,
           barplot = TRUE, xaxis = "xcols", bartype = "stack",
           yaxis = "ypc", yaxisPc = "column_pc")
```

Note: an alternative, more flexible method for producing a stacked relative frequency bar chart is provided in your notes.

Activity 2.2

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 either the web applet, or 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` for the web applet, `size` for R)
b) The number of successes we are interested in (`x` for web applet, `x` or `q` for R)
c) The probability of success for each trial (`p` for the web applet, `prob` for R)
d) The form of the binomial function
    - for the web applet: `P(X=x)`, `P(X x)` or `P(X x)`;
    - 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.

Activity 2.3

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 the web applet 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 2.4

A health survey was conducted, and an extract of data has been provided in `Activity_2.4-health-survey.cs`
Categorise height into 20cm intervals, and present the height-groups appropriately.

Answer

Table 2: Heights of 1140 health survey participants

Height	Frequency	Relative frequency (%)
120 to less than 140cm	1	0.1
140 to less than 160cm	160	14.0
160 to less than 180cm	756	66.3
180 to less than 200cm	222	19.5
200 to less than 220cm	1	0.1

Process

After reading the data in using `read.csv`, it is useful to plot a density plot to check the distribution of height. After confirming there are no biologically impossible values of height, we use the `breaks` function to create height groups. Finally, we use `descriptives` with `freq = TRUE` to produce a one-way frequency table.

```
survey <- read.csv("data/activities/Activity_2.4-health-survey.csv")

descriptives(survey, height, dens=TRUE)

survey$"Height group" <- cut(survey$height,
                      breaks = c(0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2),
                      right=FALSE,
                      labels=c(
                        "120 to less than 140cm",
                        "140 to less than 160cm",
                        "160 to less than 180cm",
                        "180 to less than 200cm",
                        "200 to less than 220cm"
                      ))

descriptives(data=survey, vars="Height group", freq = TRUE)
```

Activity 2.5

The data in the file `Activity_2.5-LengthOfStay.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, density plot and boxplot to inspect the distribution of birth weight and length of stay;

b. Complete the following summary statistics for each variable:

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

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 2 to Figure 7.

Figure 2: Histogram of birth weight

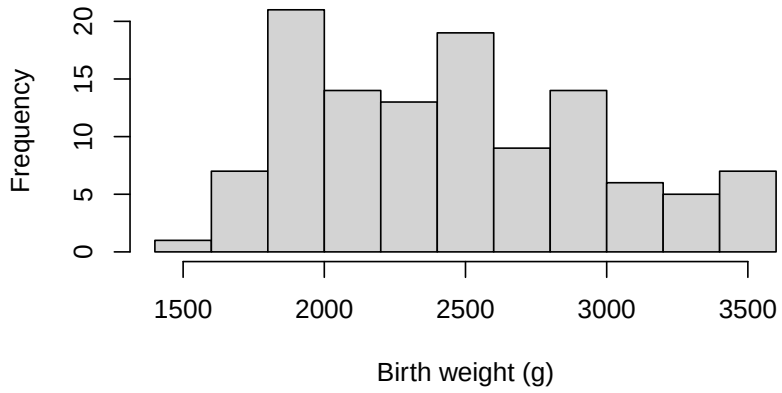


Figure 3: Density plot of birth weight

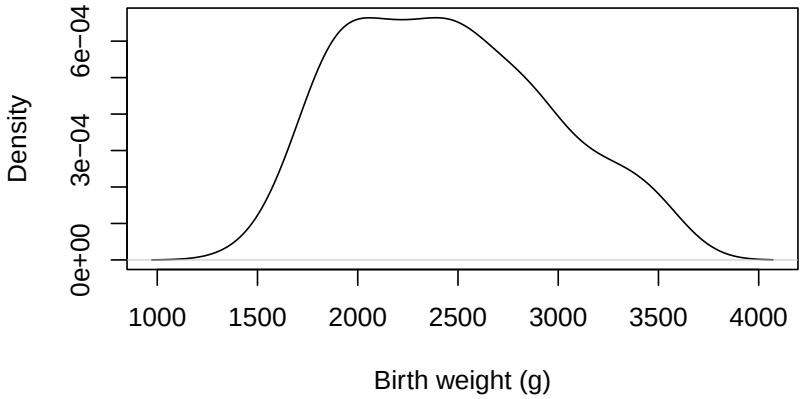


Figure 4: Boxplot of birth weight

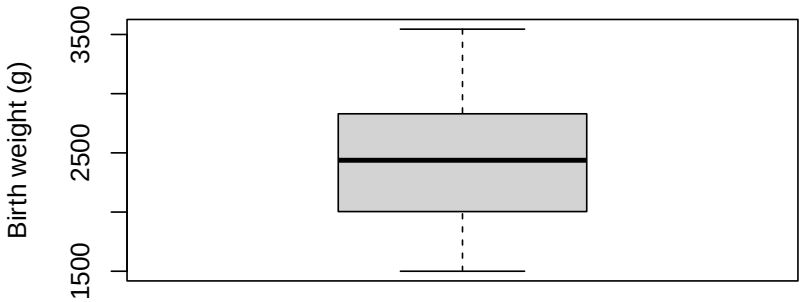


Figure 5: Histogram of length of stay

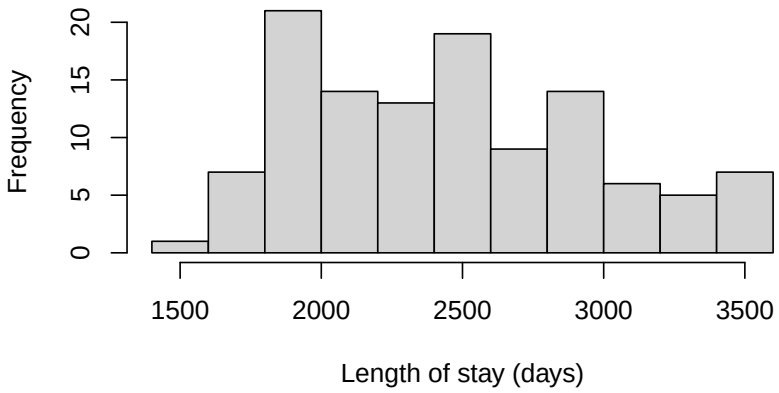


Figure 6: Density plot of length of stay

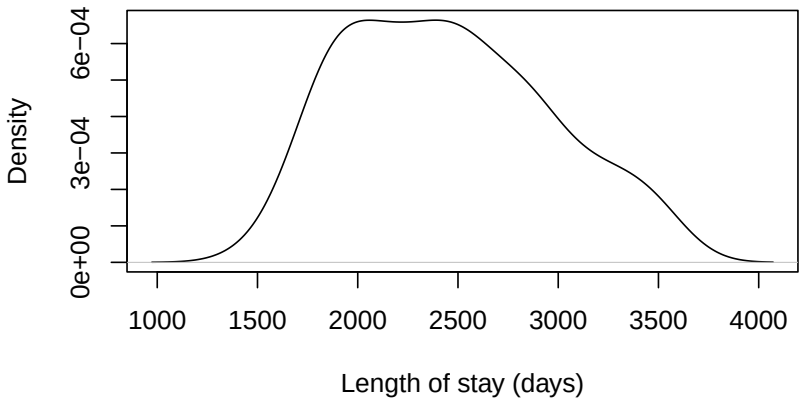
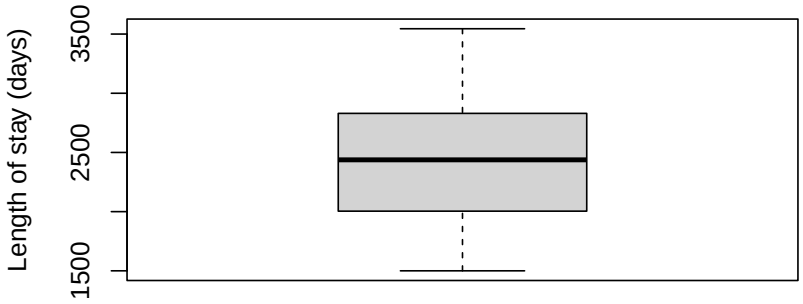


Figure 7: Boxplot of length of stay



b) See Table 3.

Table 3: 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 [2012 to 2830]	30 [21 to 43]

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.

Process

The figures were produced using the following code:

```
babies <- readRDS("data/activities/Activity_2.5-LengthOfStay.rds")

hist(los$BirthWt,
```

```
     xlab = "Birth weight (g)", main = "")

plot(density(los$BirthWt, na.rm = TRUE),
     xlab = "Birth weight (g)", main = "")

boxplot(los$BirthWt,
        ylab = "Birth weight (g)", main = "")

hist(los$LengthStay,
     xlab = "Length of stay (days)", main = "")

plot(density(los$LengthStay, na.rm = TRUE),
     xlab = "Length of stay (days)", main = "")

boxplot(los$LengthStay,
        ylab = "Length of stay (days)", main = "")
```

The summary statistics were produced with the following code:

```
library(jmv)

descriptives(data = babies, vars = c(BirthWt, LengthStay), pc = 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
   25th percentile       2012.000      21.00000
   50th percentile       2437.500      30.00000
   75th percentile       2830.000      43.00000
```