

Module 2 solutions

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 b) and c), 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

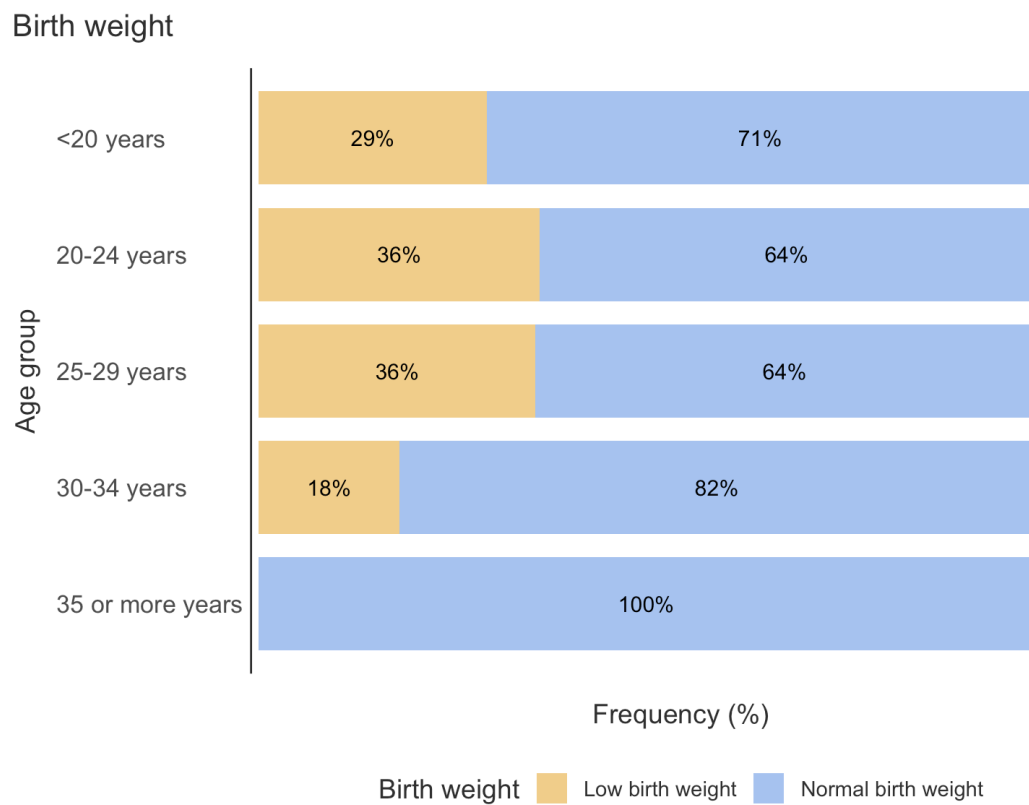
Smoking status during pregnancy	Low birth weight		
	Yes (%)	No (%)	Total (%)
Yes	30 (40.5)	44 (59.5)	74 (100)
No	39 (25.2)	86 (74.8)	115 (100)
Total	59 (31.2)	130 (68.8)	189 (100)

Note: this table has been constructed from jamovi 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).

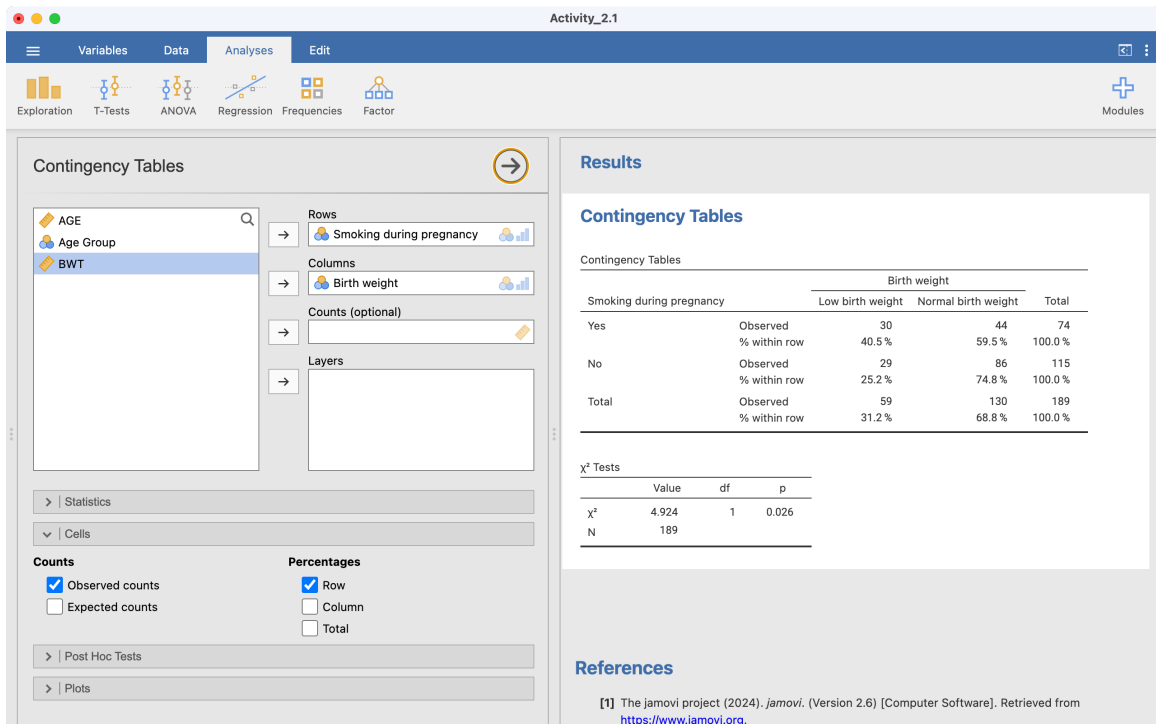
NB: You will revisit two-way tables in Module 7 where you will conduct statistical tests to determine if there is evidence of a difference in proportions.

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



Process

After reading in the data and creating more meaningful variable names, Table 1 was created as follows. Note that we request **Row Percentages**:



The screenshot shows the jamovi software interface with the 'Analyses' tab selected. The 'Contingency Tables' module is active, showing a list of variables on the left: AGE, Age Group, and BWT. The 'Rows' variable is 'Smoking during pregnancy' and the 'Columns' variable is 'Birth weight'. The 'Counts (optional)' section is empty. The 'Layers' section is empty. The 'Statistics' section is expanded, showing 'Counts' and 'Percentages' options. 'Observed counts' and 'Row' are selected. The 'Results' panel displays the contingency table and chi-square test results.

Contingency Tables

Smoking during pregnancy	Birth weight		Total
	Low birth weight	Normal birth weight	
Yes	Observed 30	44	74
	% within row 40.5 %	59.5 %	100.0 %
No	Observed 29	86	115
	% within row 25.2 %	74.8 %	100.0 %
Total	Observed 59	130	189
	% within row 31.2 %	68.8 %	100.0 %

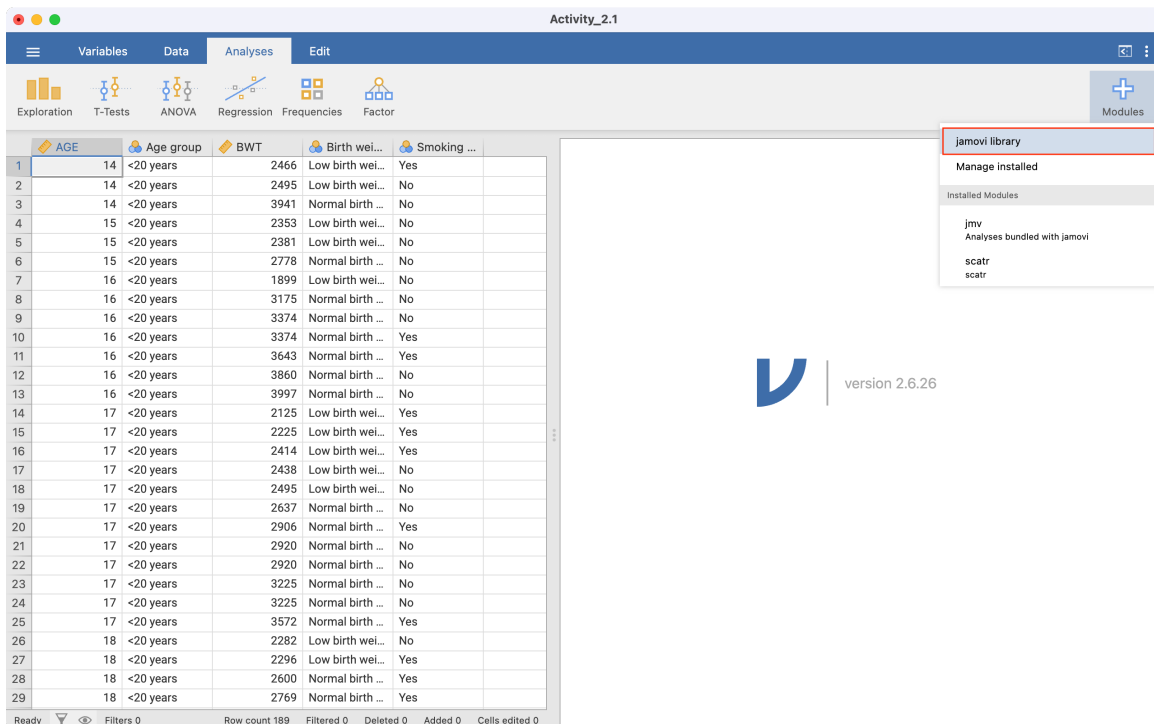
Chi-Square Tests

	Value	df	p
χ^2	4.924	1	0.026
N	189		

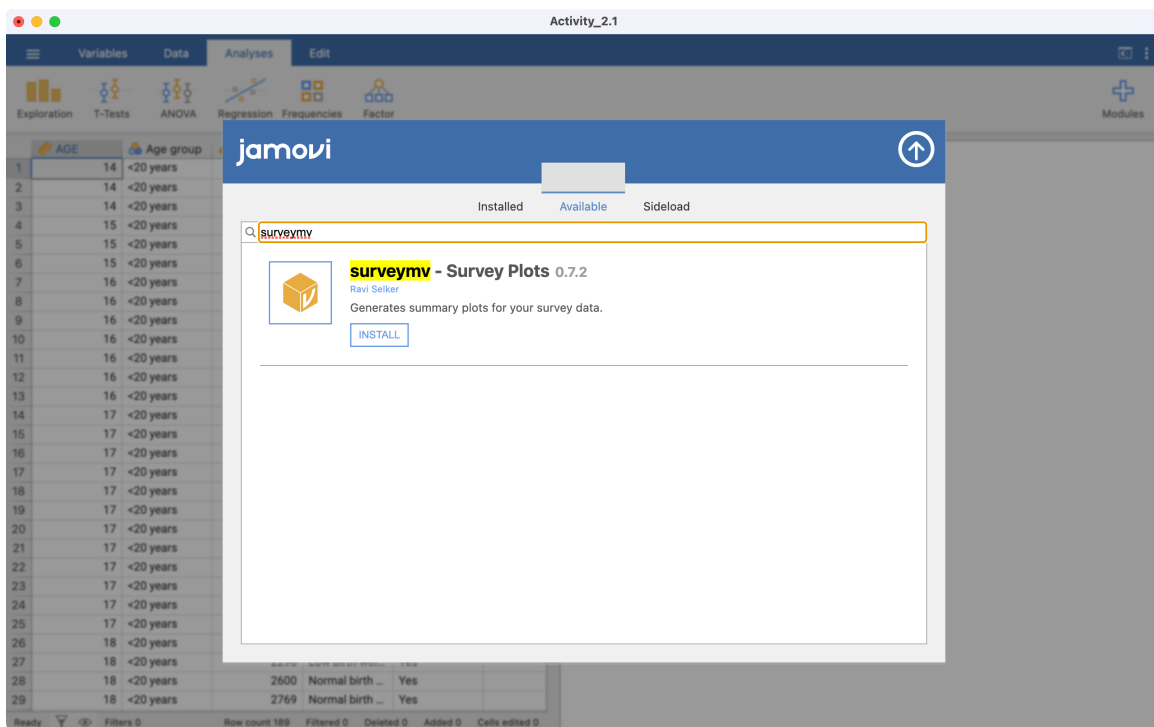
References

[1] The jamovi project (2024). *jamovi*. (Version 2.6) [Computer Software]. Retrieved from <https://www.jamovi.org>.

Figure 1 was created using the **survey** module. To install the module, click the **Analyses** tab, and click the large **+** at the top-right of the window. Choose **jamovi library**:



Click **Available** and search for **surveymv**, then click install:



The module has now been installed. To run the module, click the up-arrow to return to the **Analyses** tab, click the large **+** and choose **surveymv > Survey Plots**:

Activity_2.1

Variables Data Analyses Edit

Exploration T-Tests ANOVA Regression Frequencies Factor

Modules

jamovi library
Manage installed

Installed Modules

jmv
Analyses bundled with jamovi

scatr
scatr

survey
Survey Plots

Module - surveymv

☒ Show in main menu

Analyses

Survey Plots

	AGE	Age group	BWT	Birth wei...	Smoking ...
1	14	<20 years	2466	Low birth wei...	Yes
2	14	<20 years	2495	Low birth wei...	No
3	14	<20 years	3941	Normal birth ...	No
4	15	<20 years	2353	Low birth wei...	No
5	15	<20 years	2381	Low birth wei...	No
6	15	<20 years	2778	Normal birth ...	No
7	16	<20 years	1899	Low birth wei...	No
8	16	<20 years	3175	Normal birth ...	No
9	16	<20 years	3374	Normal birth ...	No
10	16	<20 years	3374	Normal birth ...	Yes
11	16	<20 years	3643	Normal birth ...	Yes
12	16	<20 years	3860	Normal birth ...	No
13	16	<20 years	3997	Normal birth ...	No
14	17	<20 years	2125	Low birth wei...	Yes
15	17	<20 years	2225	Low birth wei...	Yes
16	17	<20 years	2414	Low birth wei...	Yes
17	17	<20 years	2438	Low birth wei...	No
18	17	<20 years	2495	Low birth wei...	No
19	17	<20 years	2637	Normal birth ...	No
20	17	<20 years	2906	Normal birth ...	Yes
21	17	<20 years	2920	Normal birth ...	No
22	17	<20 years	2920	Normal birth ...	No
23	17	<20 years	3225	Normal birth ...	No
24	17	<20 years	3225	Normal birth ...	No
25	17	<20 years	3572	Normal birth ...	Yes
26	18	<20 years	2282	Low birth wei...	No
27	18	<20 years	2296	Low birth wei...	Yes
28	18	<20 years	2600	Normal birth ...	Yes
29	18	<20 years	2769	Normal birth ...	Yes

Ready Filters 0 Row count 189 Filtered 0 Deleted 0 Added 0 Cells edited 0

We want to plot the variable **Birth weight**, by **Age group**. We want a stacked bar plot, and we want to plot percentages. Putting it all together:

Activity_2.1

Variables Data Analyses Edit

Exploration T-Tests ANOVA Regression Frequencies Factor

Modules

Survey Plots

AGE
BWT
Smoking during pregnancy

Variables
Birth weight

Grouping Variable
Age group

☒ Variable description

Nominal / Ordinal Plots

Plot Type
☐ Grouped bar
☒ Stacked bar

Frequency Type
☐ Counts
☒ Percentages

Frequency Labels
☒ In plot
☐ On x-axis

Additional options
☒ Hide missing values

Continuous Plots

Results

Survey Plots

Birth weight

Age group

Frequency (%)

Birth weight Low birth weight Normal birth weight

Age group	Low birth weight	Normal birth weight
<20 years	29%	71%
20-24 years	36%	64%
25-29 years	36%	64%
30-34 years	18%	82%
35 or more years		100%

References

[1] The jamovi project (2024). *jamovi*. (Version 2.6) [Computer Software]. Retrieved from <https://www.jamovi.org>.

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:

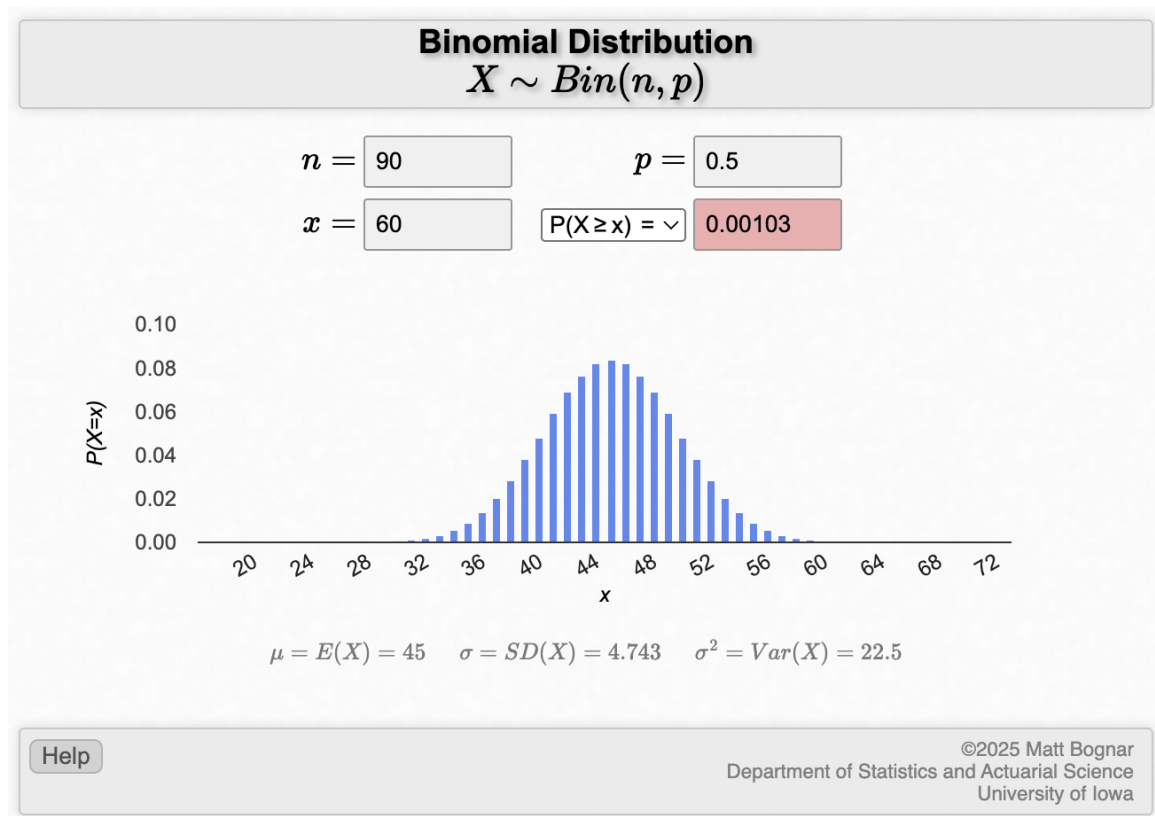
- The number of trials (**n** for the web applet, **size** for R)
- The number of successes we are interested in (**x** for web applet, **x** or **q** for R)
- The probability of success for each trial (**p** for the web applet, **prob** for R)
- The form of the binomial function
 - for the web applet: $P(X=x)$, $P(X\leq x)$ or $P(X\geq x)$;
 - for R: `dbinom`, `pbinom` or `pbinom(lower.tail=FALSE)`
- The final probability.

Answers

- Here, each participant represents a 'trial', so **n** is 90.
- We are interested in determining the probability that 60 or more participants prefer the new drug, so we define **x** as 60.
- We are told to assume that the two drugs are equally preferred, so **p** is 0.5.
- We need to calculate the probability that 60 or more participants prefer the new drug, so we change the drop-down to **$P(X\geq x)=$** .
- The result is computed as 0.00103. Therefore, the probability that 60 or more patients would prefer the new drug is 0.001 or 0.1%.

Process

The applet is completed as below.



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:

- Will not be identified?
- Will be identified in two of the samples?
- Will be identified in all the samples?
- Will be identified in at most 3 of the samples?

Answers

- The probability $P(X=0) = 0.116$ or 11.6%.
- The probability $P(X=2) = 0.336$ or 33.6%.
- The probability $P(X=5) = .005$ or 0.5%.
- The probability $P(X \leq 3) = .946$ or 94.6%.

Process

In all of these questions, n is 5 and p is 0.35. For (a) to (c), we need to calculate the probability of finding a certain number of infected samples, so we change the drop-down to $P(X=x)=$. For (a) we define x as 0, for (b) we define x as 2, and for (c) we define x as 5.

For (d) we change the drop-down to $P(X\leq x)=$ and define x as 3.

Activity 2.4

A health survey was conducted, and an extract of data has been provided in `Activity_2.4-health-survey.csv`. 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 opening the data, it is useful to plot a density plot to check the distribution of height. After confirming there are no biologically impossible values of height, and noting the minimum (122cm) and maximum (201cm) we create height groups.

First, click on the `height` column, then choose **Data > Transform**. Name the new variable, e.g. `Height_group`, and select **Create New Transform...**:

TRANSFORMED VARIABLE

Height group

Description

Source variable **height**

using transform **None** **Create New Transform...**

Retain unused levels in analyses ☐

	sex	height	Height gr...	weight
1	1	1.63	1.63	81.7
2	1	1.63	1.63	68.0
3	1	1.85	1.85	97.1
4	1	1.78	1.78	89.8
5	1	1.73	1.73	70.3
6	2	1.57	1.57	85.7
7	1	1.70	1.70	69.4
8	1	1.83	1.83	90.7
9	2	1.75	1.75	83.9
10	2	1.60	1.60	72.6
11	1	1.73	1.73	81.7
12	1	1.70	1.70	68.0
13	2	1.52	1.52	65.8
14	2	1.70	1.70	90.7
15	2	1.63	1.63	68.0

Ready Filters 0 Row count 1140 Filtered 0 Deleted 0 Added 0 Cells edited 0

version 2.6.26

Follow the process as outlined in Section 2.16 of the course notes. Your screen should look as follows, with the final two conditions showing:

TRANSFORMED VARIABLE

● TRANSFORM used by 1

Transform 1

Description Variable suffix

+ Add recode condition

if \$source < 2.0 use "180 to less than 200 cm"

else use "200 to less than 220cm"

Measure type **Auto**

	sex	height	Height gr...	weight
1	1	1.63	160 to less th...	81.7
2	1	1.63	160 to less th...	68.0
3	1	1.85	180 to less th...	97.1
4	1	1.78	160 to less th...	89.8
5	1	1.73	160 to less th...	70.3
6	2	1.57	140 to less th...	85.7
7	1	1.70	160 to less th...	69.4
8	1	1.83	180 to less th...	90.7
9	2	1.75	160 to less th...	83.9
10	2	1.60	160 to less th...	72.6
11	1	1.73	160 to less th...	81.7
12	1	1.70	160 to less th...	68.0
13	2	1.52	140 to less th...	65.8
14	2	1.70	160 to less th...	90.7
15	2	1.63	160 to less th...	68.0

Ready Filters 0 Row count 1140 Filtered 0 Deleted 0 Added 0 Cells edited 0

version 2.6.26

Finally, we use **Analyses > Exploration > Descriptives** to summarise the new Height group variable, and request a Frequency table:

Activity_2.4-health-survey

VariablesDataAnalysesEdit

Exploration

T-Tests

ANOVA

Regression

Frequencies

Factor

Modules

Descriptives

sexheightweight

Height group

Split by

Descriptives

Variables across columns

☒ Frequency tables

> | Statistics

> | Plots

Results

Descriptives

Descriptives

Height group	
N	1140
Missing	0
Mean	
Median	
Standard deviation	
Minimum	
Maximum	

Frequencies

Frequencies of Height group

Height group	Counts	% of Total	Cumulative %
120 to less than 140 cm	1	0.1 %	0.1 %
140 to less than 160 cm	160	14.0 %	14.1 %
160 to less than 180 cm	756	66.3 %	80.4 %
180 to less than 200 cm	222	19.5 %	99.9 %
200 to less than 220cm	1	0.1 %	100.0 %

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:

- Create a histogram, density plot and boxplot to inspect the distribution of birth weight and length of stay;
- 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

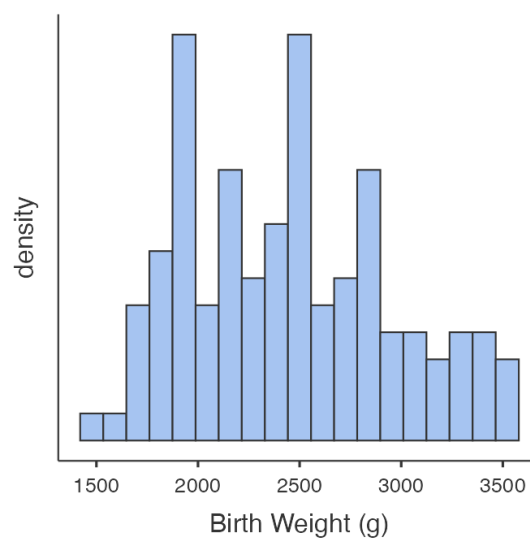


Figure 3: Density plot of birth weight

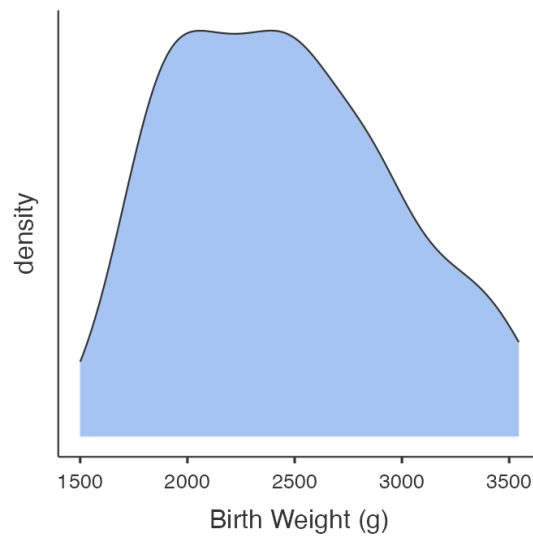


Figure 4: Box plot of birth weight

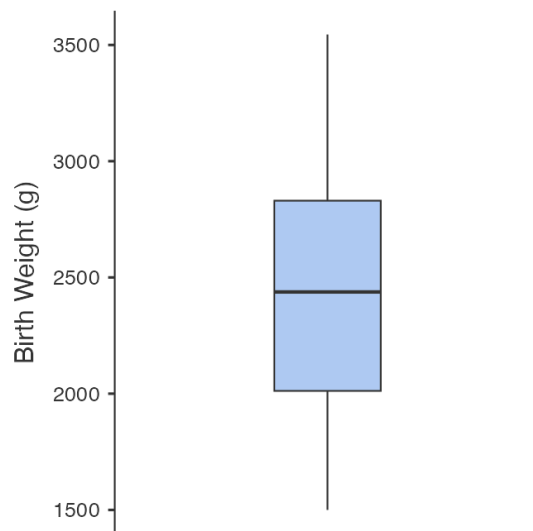


Figure 5: Histogram of length of stay

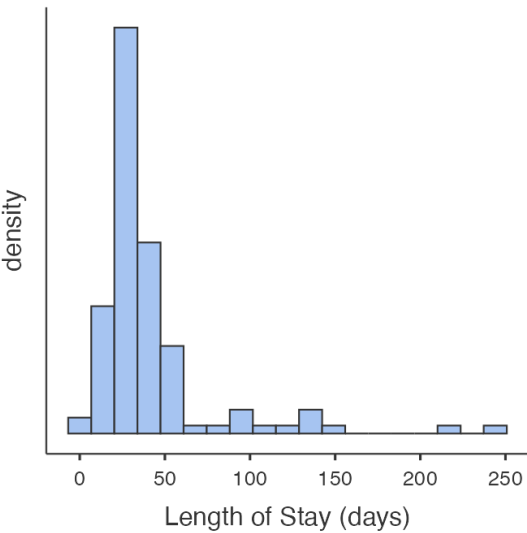


Figure 6: Density plot of length of stay

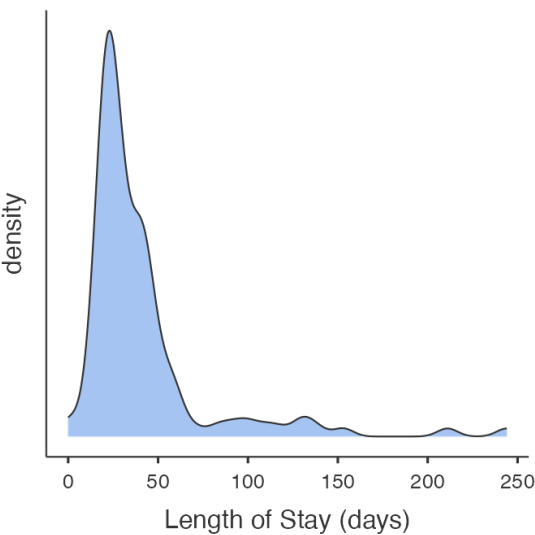
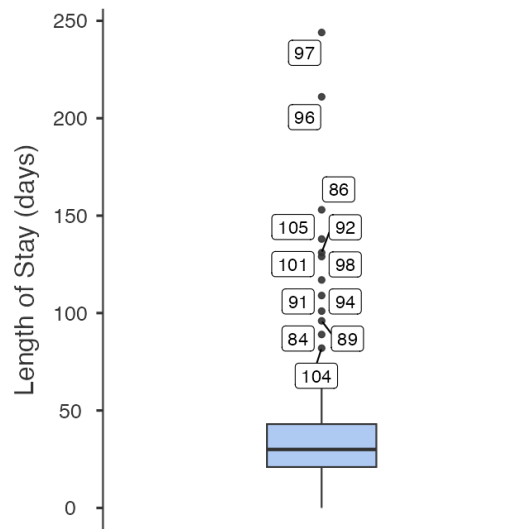


Figure 7: Box plot of length of stay



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)	2451.2 (504.82)	41.1 (36.93)
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.