

Module 1: Solutions to learning activities

jamovi version

Activity 1.1

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

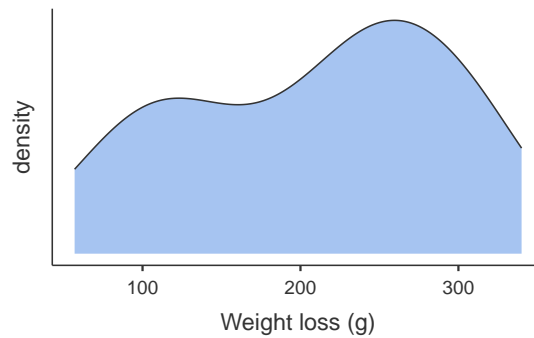
255	198	283	312	283
57	85	312	142	113
227	283	255	340	142
113	312	227	85	170
255	198	113	227	255

- a) These data have been saved as `Activity_1.1.rds`. Read the data into your software package.
- b) What type of data are these?
- c) Construct an appropriate graph to display the distribution of participants' weight loss. Provide appropriate labels for the axes and give the graph an appropriate title.

Answers

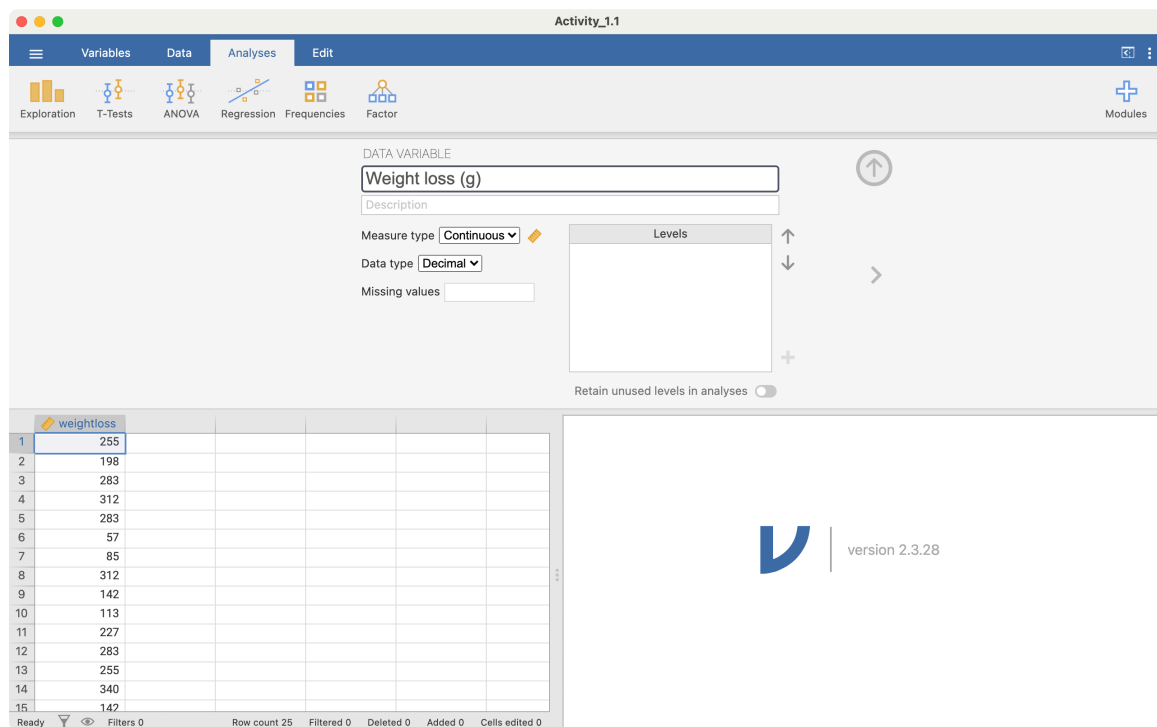
- b) These are continuous numeric data.
- c) See Figure 1.

Figure 1: Weight loss for 25 participants

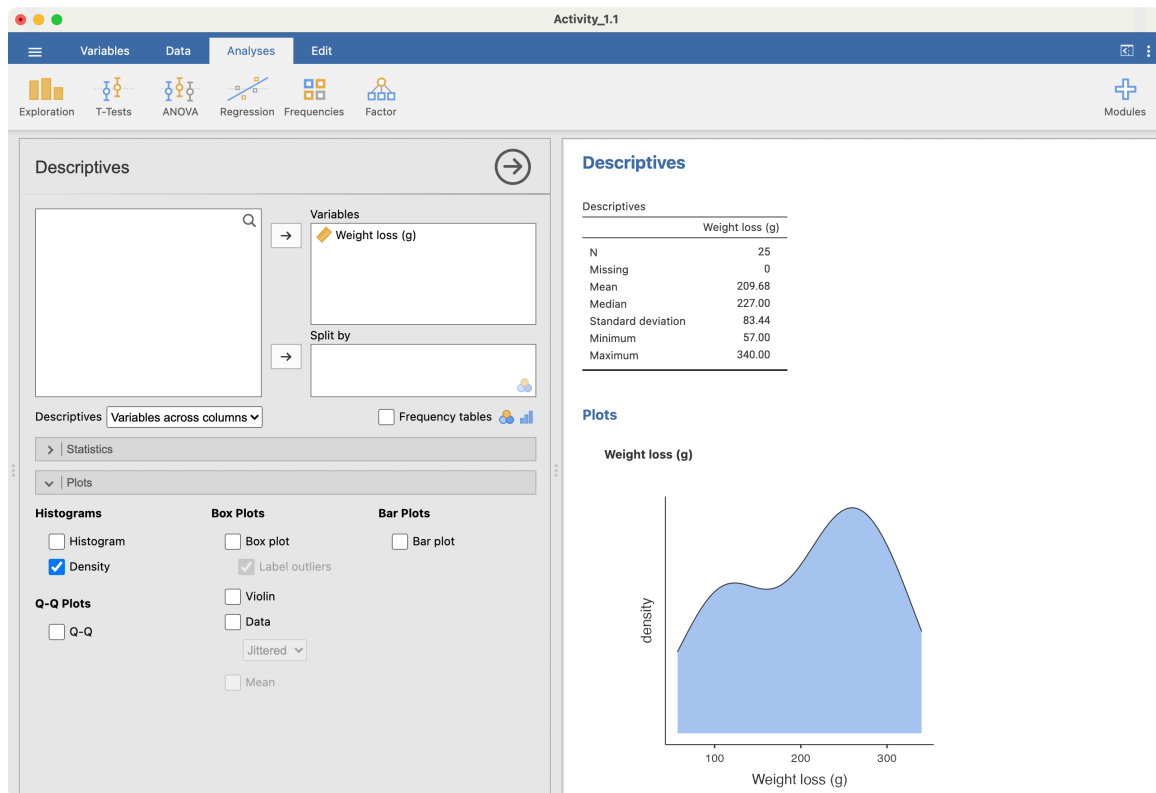


Process

After opening the data in jamovi, we first change the name of the variable `weightloss` to `Weight loss (g)`, so the plot is appropriately labelled. This can be done by double-clicking the column name of `weightloss` in the spreadsheet view:



After the name is changed, we choose **Analyses > Exploration > Descriptives**. Choose `Weight loss (g)` as the analysis variable, and make sure that **Plots > Density** is selected:



Note that the title of the figure can be added as text after copying the figure into a word processing package.

Activity 1.2

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

- a) The smaller the mean value
- b) The larger the standard deviation
- c) The smaller the variance

Answers

- a) is not true because the mean is not influenced by the spread (if the distribution is symmetric)
- b) is true because the larger the spread, the deviations from the mean will also be larger, and so the standard deviation will be larger.
- c) is not true because the variance will be larger if the deviations from the mean are larger.

Activity 1.3

Estimate the mean, median, standard deviation, range and interquartile range for the data `Activity_1.3.rds`, available on Moodle.

Answers

The mean is 1.50 and the median is 1.5.

The range of the data is from 0.1 to 3.2.

The standard deviation is estimated as 0.843, and the inter-quartile range is from 1.0 to 2.0.

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.

Process

A similar process is used as for Activity 1.1. To obtain the interquartile range, ensure that **Statistics > Percentiles** is selected, to obtain the first and third quartiles (the 25th and 75th percentiles respectively):

The screenshot shows the Jamovi software interface with the 'Analyses' tab selected. The 'Descriptives' analysis is configured for the variable 'Lead_concn'. The 'Statistics' section is expanded, showing various options for sample size, percentile values, dispersion, mean dispersion, central tendency, distribution, normality, and outliers. The 'Results' panel on the right displays the output for 'Lead_concn'.

Descriptives

Lead_concn	
N	15
Missing	0
Mean	1.50
Median	1.50
Standard deviation	0.84
Minimum	0.10
Maximum	3.20
25th percentile	0.95
50th percentile	1.50
75th percentile	1.95

References

- [1] The jamovi project (2022). *jamovi*. (Version 2.3) [Computer Software]. Retrieved from <https://www.jamovi.org>.
- [2] R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from <https://cran.r-project.org>. (R packages retrieved from MRAN snapshot 2022-01-01).

Activity 1.4

Data of diastolic blood pressure (BP) of a sample of study participants are provided in the datasets `Activity_1.4.rds`. Compute the mean, median, range and SD of diastolic BP.

Answers

The mean is 82.2 mmHg and the median is 83.0 mmHg. The range is 56.0 to 118.0 mmHg and the SD 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

The ages of 100 study participants have been saved as `Activity_1.5.rds`. Estimate the:

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

Plot the data using a density plot and boxplot. Is there anything unusual about the ages? What do you think is a possible explanation for this?

A clean version of the data have been saved as `Activity_1.5_clean.rds`. Recalculate the summary statistics and recreate the plots using the clean data.

Based on this exercise, what is your advice on coding unusual or missing values in data?

Answers

The summary statistics for the original dataset are estimated as follows. The mean age is 92.7 years, and the median is 45 years. The standard deviation is 209.07 years, and the interquartile range is 42 to 49 years. The range is 15 to 999 years.

A density plot and boxplot are presented in Figure 2 and Figure 3.

Figure 2: Density plot of age (in years) for 100 study participants

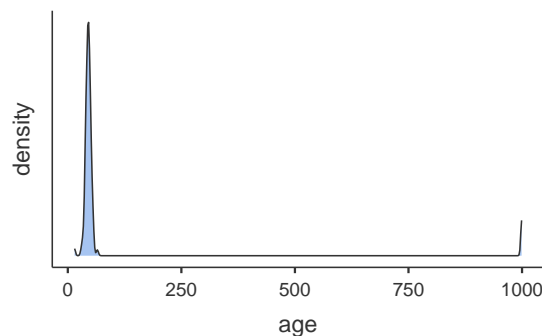
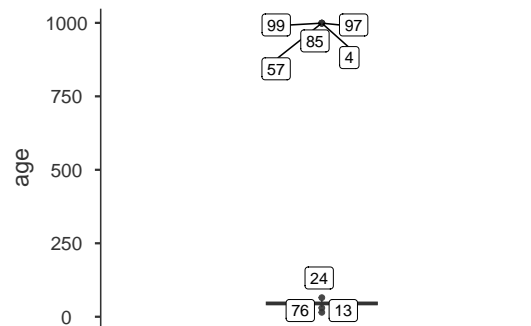


Figure 3: Boxplot of age (in years) for 100 study participants



In both plots, there are some very large, biologically impossible ages (around 1000 years). From the summary statistics, the highest age is recorded as 999 years. These values are either (a) a typographical error, or (b) more likely, a code representing a missing value of age.

Using the clean version of the data (with 95 observations), the mean age is 45.0 years, and the median is 45 years. The standard deviation is 6.34 years, and the interquartile range is 41 to 49 years. The range is 15 to 65 years.

Using the clean data, the mean, standard deviation and range have reduced compared to those obtained the original data. The median and interquartile range have not changed much, demonstrating the fact that these estimates are relatively robust in the presence of outlying observations.

The density plot (Figure 4) and boxplot (Figure 5) display a relatively symmetric distribution. While there are some large and small observations, these are not biologically impossible.

Figure 4: Density plot of age (in years) for 95 study participants with biologically plausible ages

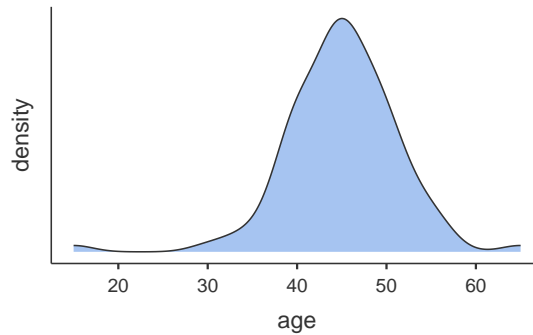
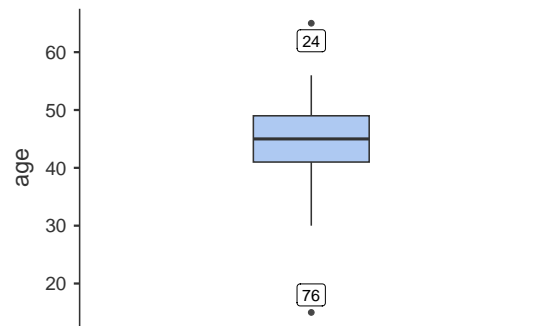


Figure 5: Boxplot of age (in years) for 95 study participants with biologically plausible ages



Based on this exercise, the best advice for coding unusual or missing values in data would be to never set the values as a numerical value (here 999). Numerical values can always be inadvertently analysed as if they were true, observed values resulting in inflated means and standard deviations. Further, if a code like 99 was used, it would be unclear whether this was a true age, or a code for a missing value.

Rather, values should be set to missing: using an empty cell in the spreadsheet of jamovi.

Finally, this question highlights the importance of always examining your data before analysing - either by plotting a density plot and/or a boxplot.