Module 1: Solutions to learning activities

R 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.

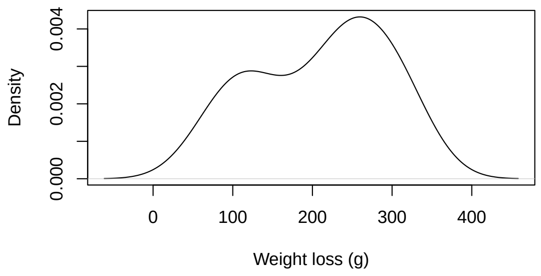
255	198	283	312	283
57	85	312	142	113
227	283	255	340	142
113	312	227	85	170
255	198	113	227	255

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

```
library(jmv)

# Read in the data
weightloss <- readRDS("data/activities/Activity_1.1.rds")

# Check the default density plot:
plot(density(weightloss$weightloss))

# Let's add labels and titles
plot(density(weightloss$weightloss),
     xlab="Weight loss (g)",
     main="Figure 1: Weight loss for 25 participants")

# Alternatively, use the descriptives function from jmv:
descriptives(data=weightloss, vars=weightloss, dens=TRUE)
```

Note that it is difficult to customise the plot when using the descriptives function. You could create a new column called ``Weight loss (g)" and plot it using the following code (note the use of the quotation marks -- these are **very important**):

```
weightloss$"Weight loss (g)" <- weightloss$weightloss
descriptives(data=weightloss, vars="Weight loss (g)", dens=TRUE)
```

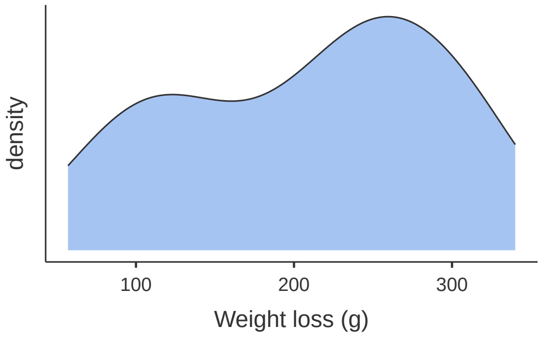
```
 DESCRIPTIVES

 Descriptives

                         Weight loss (g)

   N                                  25
   Missing                             0
   Mean                         209.6800
   Median                       227.0000
   Standard deviation           83.43796
   Minimum                      57.00000
   Maximum                      340.0000
```

Figure 2: Weight loss for 25 participants



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

```r
act1_3 <- readRDS("data/activities/Activity_1.3.rds")

descriptives(act1_3, pc=TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                         Lead_concn

   N                             15
   Missing                        0
   Mean                    1.500000
   Median                  1.500000
   Standard deviation     0.8434623
   Minimum                0.1000000
   Maximum                 3.200000
   25th percentile        0.9500000
   50th percentile         1.500000
   75th percentile         1.950000
```

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:

a. mean and median;
b. standard deviation and interquartile range;
c. range.

Plot the data using a density plot and boxpolot. 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 3 and Figure 4.

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

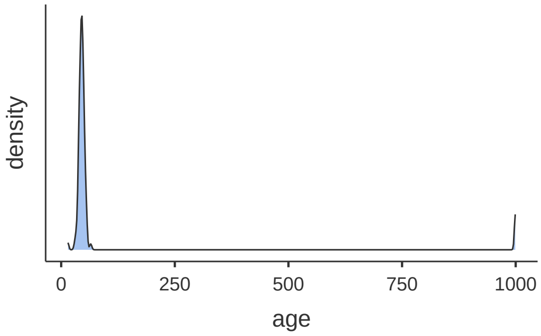
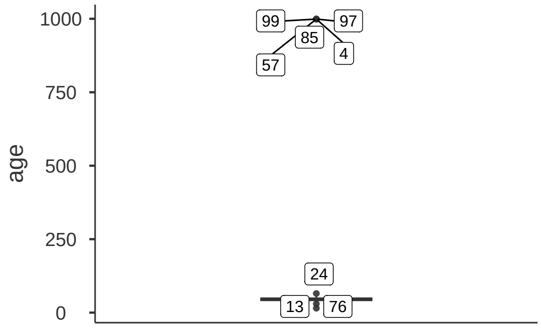


Figure 4: Boxplot of age (in years) for 100 study participants



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 5) and boxplot (Figure 6) display a relatively symmetric distribution. While there are some large and small observations, these are not biologically impossible.

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

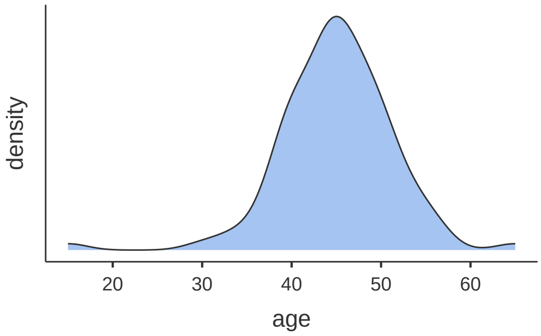
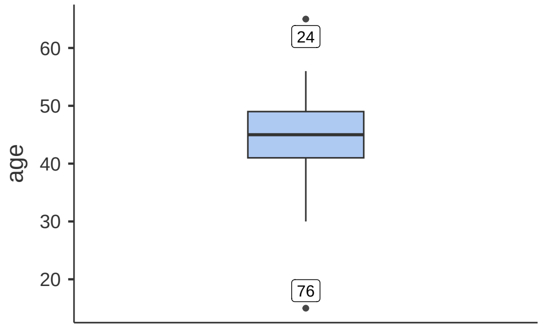


Figure 6: Boxplot of age (in years) for 95 study participants with biologically plausible ages



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 a "NA" in R.

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

Process

The following code was used to answer this question:

```
library(jmv)

# Load original data
```

```
study <- readRDS("data/activities/Activity_1.5.rds")

descriptives(data=study, vars=age, pc = TRUE, dens = TRUE, box = TRUE)

# Load cleaned data
study_clean <- readRDS("data/activities/Activity_1.5_clean.rds")

descriptives(data=study_clean, vars=age, pc = TRUE, dens = TRUE, box = TRUE)
```