Assignment 5 for Statistical Computing and Empirical Methods: Exploratory data analysis

Henry Reeve

Introduction

This document describes your fifth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommended that you first watch video lecture 5 entitled “Exploratory data analysis”.

Begin by creating an Rmarkdown document with html output. You will need to load the Tidyverse library and the Hawks data set from the Stat2Data package.

```
library(tidyverse)
library(Stat2Data)
data("Hawks")
```

1 Combining location estimators with the summarise function

Use a combination of the **summarise()**, **mean()** and **median()** to compute the sample mean, sample median and trimmed sample mean (with $q = 0.1$) of the Hawk’s wing length and Hawk’s weight. Your result should look something like this:

```
##   Wing_mean Wing_t_mean Wing_med Weight_mean Weight_t_mean Weight_med
## 1  315.6375    322.2297      370    772.0802      779.3681        970
```

Combine with the **group_by()** function to obtain a break down by species. Your result should look something like this:

```
## # A tibble: 3 x 7
##   Species Wing_mean Wing_t_mean Wing_med Weight_mean Weight_t_mean Weight_med
##   <fct>       <dbl>       <dbl>    <dbl>       <dbl>         <dbl>      <dbl>
## 1 CH           244.        243.      240        420.          410.       378.
## 2 RT           383.        385.      384       1094.         1089.      1070
## 3 SS           185.        184.      191        148.          140.       155
```

2 Location and dispersion estimatiors under linear transformations

Suppose that a variable of interest X_i has values $X_1, \ldots, X_n$. Suppose that $X_1, \ldots, X_n$ has sample mean $\hat{\mu}$. Let $a, b \in \mathbb{R}$ be real numbers and define a new variable $\tilde{X}_i$ with values $\tilde{X}_1, \ldots, \tilde{X}_n$ defined by $\tilde{X}_i = a \cdot X_i + b$ for $i = 1, \cdots, n$. Show that $\tilde{X}_1, \ldots, \tilde{X}_n$ has sample mean $a \cdot \hat{\mu} + b$.

Suppose further that $X_1, \ldots, X_n$ has sample variance S_X^2. What is the sample variance of $\tilde{X}_1, \ldots, \tilde{X}_n$? What is the sample standard deviation of $\tilde{X}_1, \ldots, \tilde{X}_n$?

3 Robustness of location estimators

In this exercise we shall investigate the robustness of several location estimators: The sample mean, sample median and trimmed mean.

We begin by extracting a vector called “hal” consisting of the talon lengths of all the hawks with any missing values removed.

```
hal<-Hawks$Hallux # Extract the vector of hallux lengths
hal<-hal[!is.na(hal)] # Remove any nans
```

To investigate the effect of outliers on estimates of location we generate a new vector called “corrupted_hall” with 10 outliers each of value 100 created as follows:

```
outlier_val<-100
num_outliers<-10
corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
```

We can then compute the mean of the original sample and the corrupted sample as follows.

```
mean(hal)
```

```
## [1] 26.41086
```

```
mean(corrupted_hal)
```

```
## [1] 27.21776
```

Now let’s investigate what happens as the number of outliers changes from 0 to 1000. The code below generates a vector called “means_vect” which gives the sample means of corrupted samples with different numbers of outliers. More precisely, means_vect is a vector of length 1001 with the i-th entry equal to the mean of a sample with $i-1$ outliers.

```
num_outliers_vect<-seq(0,1000)
means_vect<-c()

for(num_outliers in num_outliers_vect){
  corrupted_hal<-c(hal,rep(outlier_val,times=num_outliers))
  means_vect<-c(means_vect,mean(corrupted_hal))
}
```

Copy and modify the above code to create an additional vector called “medians_vect” of length 1001 with the i-th entry equal to the median of a sample “corrupted_hal” with $i-1$ outliers.

Ammend the code further to add an additional vector called “t_means_vect” of length 1001 with the i-th entry equal to the trimmed mean of a sample with $i-1$ outliers, where the trimmed mean has a trim fraction $q=0.1$.

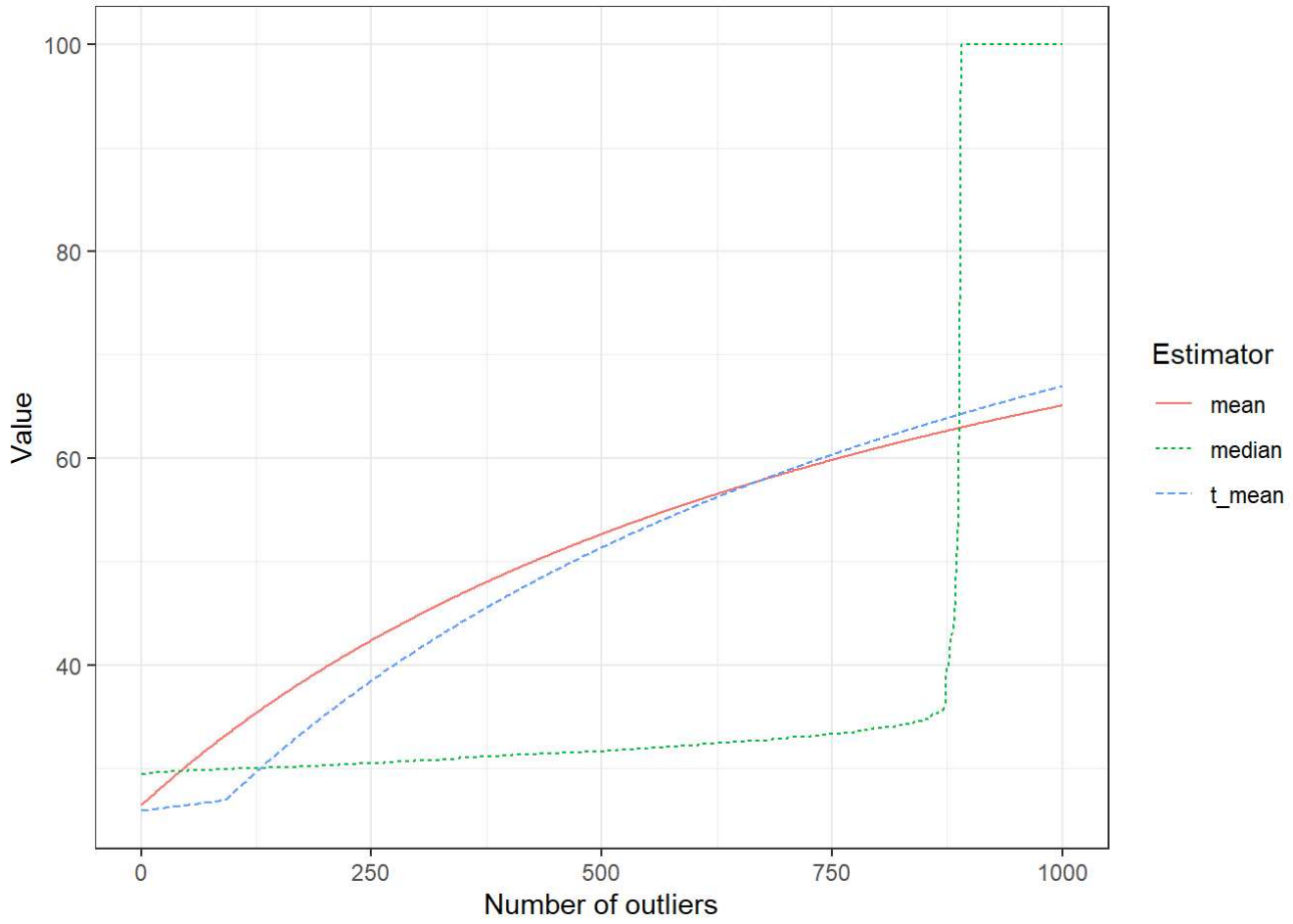
You should now have four vectors: “num_outliers_vect”, “means_vect”, “medians_vect” and “t_means_vect”. Combine these vectors into a data frame with the following code.

```
df_means_medians<-data.frame(num_outliers=num_outliers_vect,
                             mean=means_vect,t_mean=t_means_vect,
                             median=medians_vect)
```

Now use the code below to reshape and plot the data. The function **pivot_longer()** below is used to reshape the data. Don’t worry if this operation is unclear at this stage. Its use will be explained in Lecture 6.

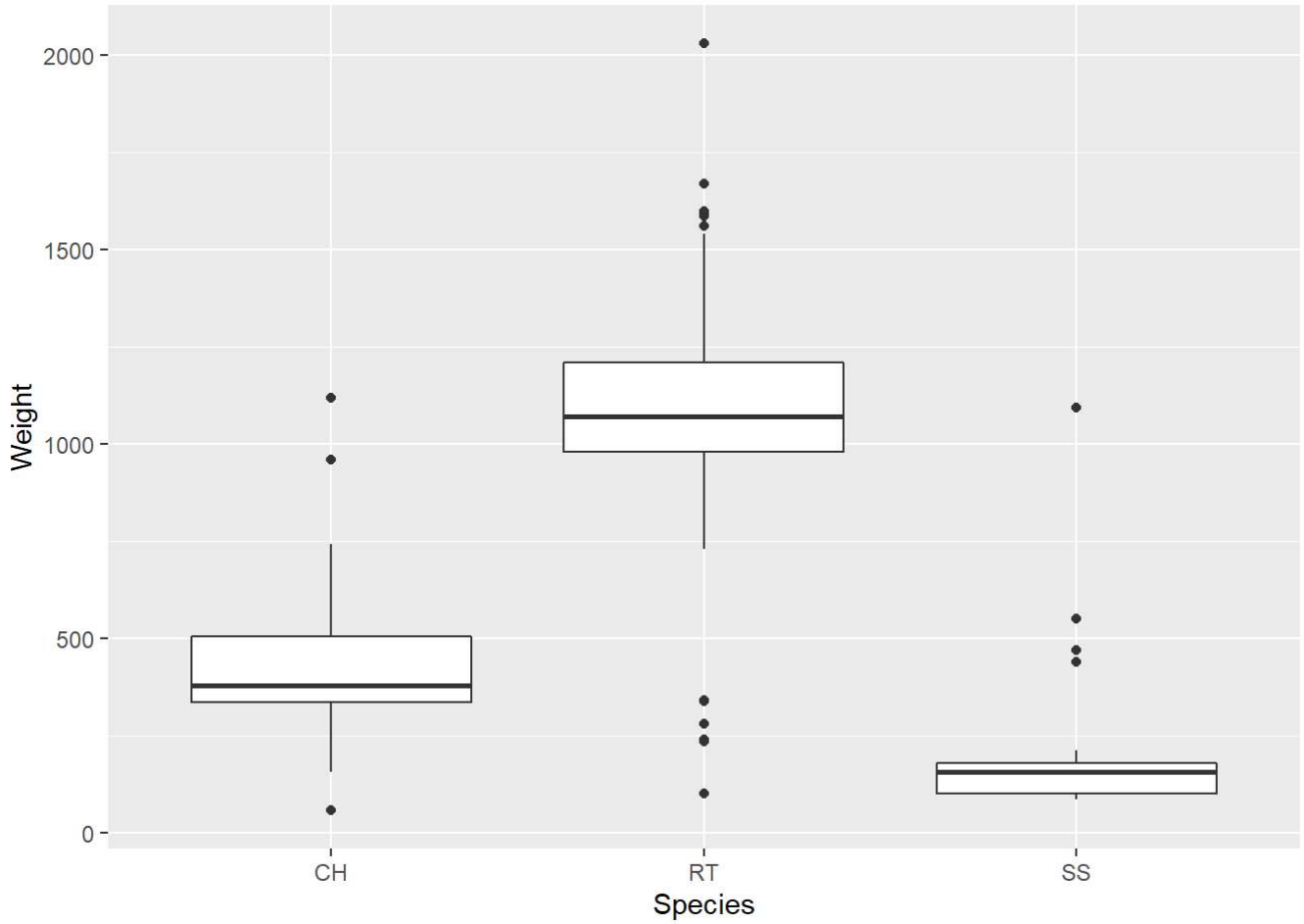
```
df_means_medians%>%
  pivot_longer(!num_outliers, names_to = "Estimator", values_to = "Value")%>%
  ggplot(aes(x=num_outliers,color=Estimator,
                  linetype=Estimator,y=Value))+
  geom_line()+xlab("Number of outliers")
```

The output of your code should look as follows:



4 Box plots and outliers

Use the functions **ggplot()** and **geom_boxplot()** to create a box plot which summarises the distribution of hawk weights broken down by species. Your plot should look as follows:



Note the outliers displayed as individual dots.

Suppose we have a sample $X_1, \cdots, X_n$. Let $q25$ denote the 0.25-quantile of the sample and let $q75$ denote the 0.75-quantile of the sample. We can then define the interquartile range, denoted IQR by $IQR := q75 - q25$. In the context of boxplots and outlier X_i is any numerical value such that the following holds if either of the following holds:

$$X_i < q25 - 1.5 \times IQR$$

$$X_i > q75 + 1.5 \times IQR.$$

Create a function called “num_outliers” which computes the number of outliers within a sample (with missing values excluded).

Now combine your function **num_outliers()** with the functions **group_by()** and **summarise()** to compute the number of outlier for the three samples of hawk weights broken down by specied. Your result should look as follows:

```
## # A tibble: 3 x 2
##   Species num_outliers_weight
##   <fct>                 <int>
## 1 CH                        3
## 2 RT                       13
## 3 SS                        4
```

5 Covariance and correlation under linear transformations

Suppose that we have a pair of variables: X_i with values $X_1, \ldots, X_n$ and Y_i with values $Y_1, \ldots, Y_n$. Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ have sample covariance $\Sigma_{X,Y}$. Let a, $b \in \mathbb{R}$ be real numbers and define a new variable $\tilde{X}_i$ with values $\tilde{X}_1, \ldots, \tilde{X}_n$ defined by $\tilde{X}_i = a \cdot X_i + b$ for $i = 1, \cdots, n$. In addition, let c, $d \in \mathbb{R}$ be real numbers and define a new variable $\tilde{Y}_i$ with values $\tilde{Y}_1, \ldots, \tilde{Y}_n$ defined by $\tilde{Y}_i = c \cdot Y_i + d$ for $i = 1, \cdots, n$. What is the covariance between the pair of variables $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$?

Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ have correlation $\rho_{X,Y}$. What is the correlation between the pair of variables $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$?