Getting the (wrong) picture from the data

Olaniyan Folajimi

November 7, 2021

Foot length and spelling/grammatical errors

```
data = read.csv('fs.csv')
str(data)
```

‘data.frame’: 71 obs. of 2 variables: $ ft_size : num 17.5 17.5 17.5 17.5 18 18 18 18 18.5 18.5 … $ num_mistakes: num 15 18 19 20 16 17 18 19 14 16 …

Data exploration

Let us look at some of the statistics of the data. We start with the mean and standard deviation.

```
mean_sd(data$num_mistakes, denote_sd = "paren")
```

[1] “9.18 (5.38)” The data has high variance as highlighted above. We can also look at the confidence intervals:

```
mci <- mean_ci(data$num_mistakes)
print(mci, show_level = TRUE)
```

[1] “9.18 (95% CI: 7.93, 10.44)”

```
median_iqr(data$num_mistakes)
```

[1] “9.00 (5.00, 13.50)”

```
n_perc(data$num_mistakes == 10)
```

[1] “5 (7.04%)”

```
new_summary <-
  qsummary(data[, c("num_mistakes", "ft_size")],
           numeric_summaries = list("Minimum" = "~ min(%s)",
                                    "Maximum" = "~ max(%s)",
                                    "Mean (Std)" = "~ qwraps2::mean_sd(%s, denote_sd = 'paren')"),
           n_perc_args = list(digits = 1, show_symbol = TRUE, show_denom = "always"))
```

```
summary_table(data, new_summary)
```

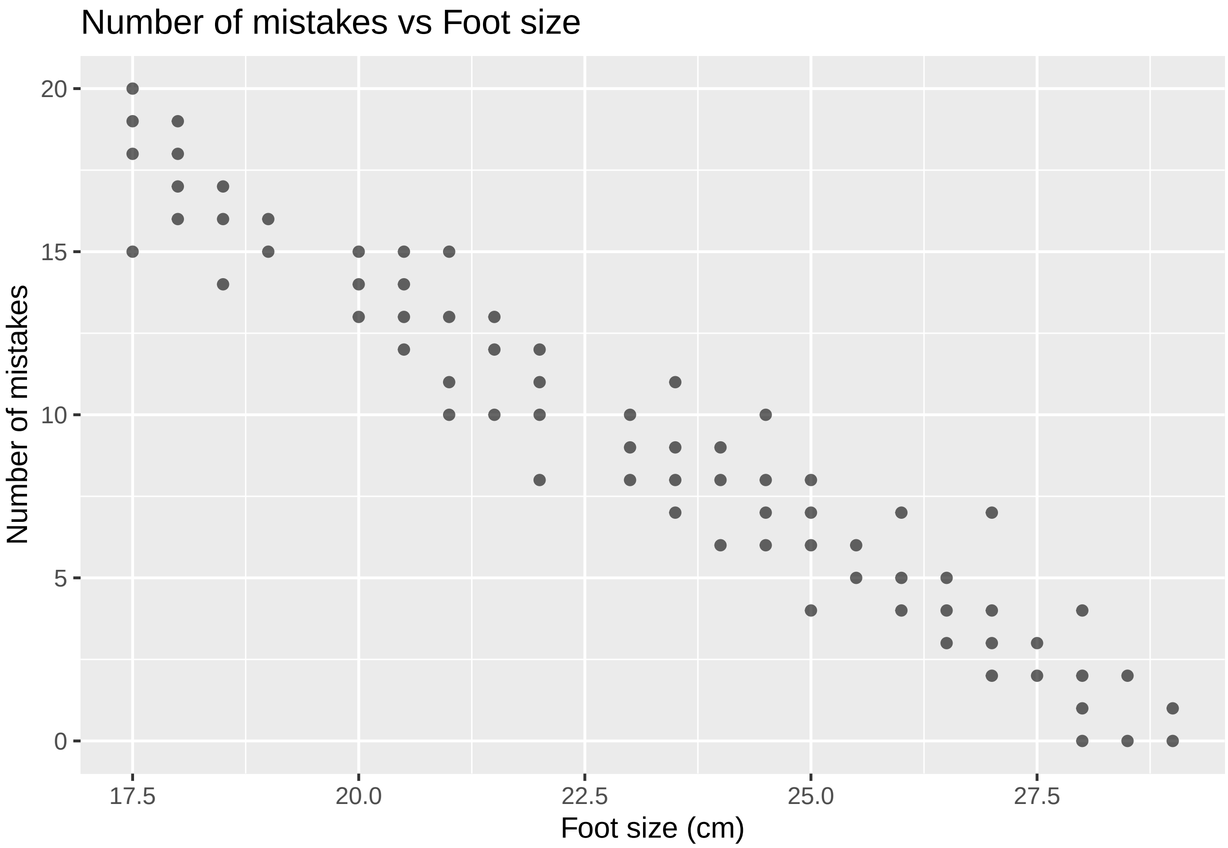
	data (N = 71)
num_mistakes	
Minimum	0
Maximum	20
Mean (Std)	9.18 (5.38)
ft_size	
Minimum	17.5

	data (N = 71)
Maximum	29
Mean (Std)	23.18 (3.45)

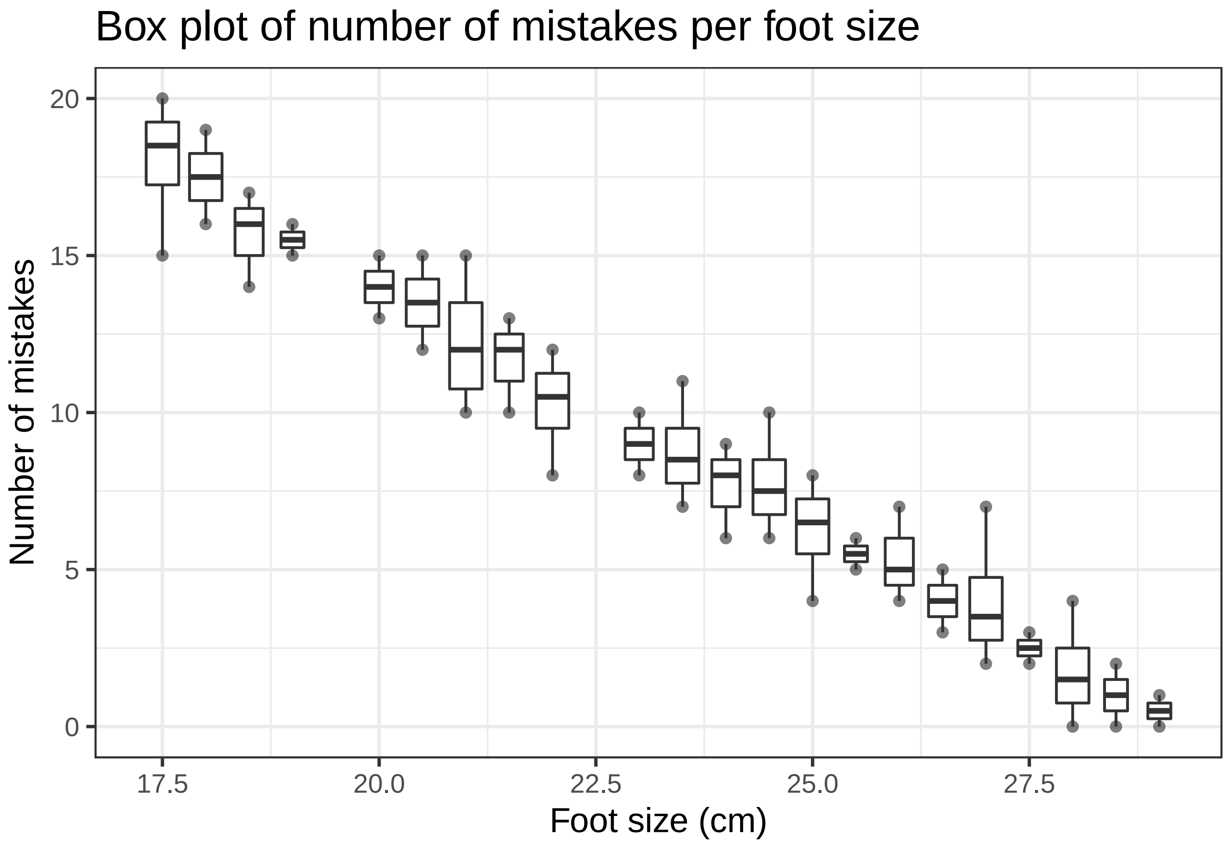
Steps

Graphical Representations of the data

```
ggplot(data=data, aes(x=ft_size, y=num_mistakes)) +geom_point(alpha=0.6) +ggtitle("Number of mistakes vs
  labs(x= "Foot size (cm)") +
  labs(y = "Number of mistakes")
```



```
ggplot(data, aes(x=ft_size, y=num_mistakes, group=ft_size)) +ggtitle("Box plot of number of mistakes per
 geom_point(alpha = .5) +
  geom_boxplot(varwidth=TRUE) +
  labs(x= "Foot size (cm)") +
  labs(y = "Number of mistakes") +
  theme_bw(base_size = 13)
```



Important points in making the graph?

In building the graph, we want to:

1. show the general pattern the data presents.
2. demonstrate the high variance in the data.

Why did we make that graph?

1. We made the first graph to highlight the possibility of a correlation between the number of mistakes made and the foot size.
2. We made the second graph to show that the box plots of mistakes varies from one foot size to another and to support our initial analysis of the variance in the data.

What can we calculate to make a summary of the variables?

We showed in the Data Exploration section the summary of statistics of the data such as mean, standard deviation range, inter-quartile range, etc. The summary of the table is presented below:

```
summary_table(data, new_summary)
```

	data (N = 71)
num_mistakes	
Minimum	0
Maximum	20
Mean (Std)	9.18 (5.38)
ft_size	
Minimum	17.5

	data (N = 71)
Maximum	29
Mean (Std)	23.18 (3.45)

What can be said about the size of students' feet and the number of mistakes they make?

The graphs show that the larger the feet size, the smaller the mistakes made by the students.

Is there a correlation between the two quantities? positive or negative? Is there a causal link?

The first graph clearly shows a downward trend in the number of mistakes as the foot size increases. This is negative correlation between the foot size and number of mistakes.

What is happening ?

In general, the size of student's feet should not determine the number of mistakes made. However, foot sizes are correlated with age and this especially true at elementary school level. Also, in many cases, the age of a student determines his/her class which may indicate how well the student will perform on the dictation. Therefore, a plausible explanation is that the performance of a student on the dictation is correlated with the age/class of the student and therefore the foot size.