# Homework 1

### Due date: Friday, September 16

### Logistical details

Complete the following exercises and submit your assignment via gradescope (linked on the course webpage). To do this, take the following steps:

- Complete your assignment.
- Scan your assignment as a PDF document.
- Upload your PDF to gradescope.
- Specify which problems are on each page.

### Problems to start after class Sept 12

#### Q1.

The `dplyr` R package provides a `starwars` data set containing data on 87 Star Wars characters. Load the data and necessary packages using the following code chunk:

```
library(ggformula)
library(mosaic)
data("starwars", package = "dplyr")
```

Explore the data by doing the following:

a. Use the `dplyr::glimpse()` to get an overview of the data set.

b. Pick a single numerical variable and make a histogram of it. Select a reasonable binwidth for it. Write a sentence summarizing what you see.

(A little bit of starter code is provided below, you can use it as a starting point and fill in the blanks.)

```
gf_histogram(~___, data = starwars, binwidth = ___)
```

c. Calculate summary statistics for the variable you selected in part (c) using `favstats()`.

d. Pick a numerical variable and a categorical variable and make a visualization (you pick the type!) to visualization the relationship between the two variables. Along with your code and output, provide an interpretation of the visualization.

e. Pick a two numerical variables and make a scatterplot to visualize the relationship between the two variables. Along with your code and output, provide an interpretation of the visualization.

f. Pick a single categorical variable from the data set and make a bar plot of its distribution.

g. Pick two categorical variables and make a visualization to visualize the relationship between the two variables. Along with your code and output, provide an interpretation of the visualization.

## Problems to start after class Sept 14

Student researchers tested to see whether listening to music would interfere with a person's ability to memorize words. Their subjects were randomly assigned to either listen to music or not. They were then shown 40 five-letter words for 90 seconds. The words were taken away and the subjects were asked to write down as many words as they could remember. The researchers wish to determine whether listening to music tends to hinder people's abilities to memorize words.

To load the data for questions 2-4, run the following:

```
words <- read.csv("https://aloy.rbind.io/data/words.csv")
```

### Q2.

(a) Graph the data and compare the center and spread of the number of words memorized for each of the groups. Does either group have outliers or skewed data?

(b) Use R to calculate a 95% confidence interval of the difference in the mean number of words recalled between the two groups. Interpret this interval without using the word "difference" (use words like "higher"/"more" or "lower"/"less" instead.)

**Q3.**

(a) State the null and alternative hypotheses comparing the mean number of words recalled for this research question. (You can typeset mathematics in R Markdown documents. Here is a quick tutorial, but you can also knit to a Word document and typeset from there.)

(b) Use R to compute the test statistic and p-value associated with the hypotheses you specified in the previous part and interpret the strength of evidence.

**Q4.**

(a) Write down the underlying theoretical model for a simple linear regression model that is equivalent to the two-sample t-test with equal variances. (You can typeset mathematics in R Markdown documents. Here is a quick tutorial, but you can also knit to a Word document and typeset from there.)

(b) What assumptions/conditions are necessary for your model to be valid? Are these assumptions reasonable for these data?