Psy/Educ 6600: Unit 1 Homework

Exploratory Data Analysis

Your Name

Spring 2018

Contents

Chapter 1. DATA PREPARATION

Load Packages

- Make sure the packages are **installed** *(Package tab)*

```
library(tidyverse)     # Loads several very helpful 'tidy' packages
library(readxl)        # Read in Excel datasets
library(furniture)     # Nice tables (by our own Tyson Barrett)
library(psych)         # Lots of nice tid-bits
```

Import Data, Define Factors, and Compute New Variables

- Make sure the **dataset** is saved in the same *folder* as this file
- Make sure the that *folder* is the **working directory**

  NOTE: I added the second line to convert all the variables names to lower case. I still kept the `F` as a capital letter at the end of the five factor variables.

```
data_clean <- read_excel("Ihno_dataset.xls") %>%
  dplyr::rename_all(tolower) %>%
  dplyr::mutate(genderF = factor(gender,
                                 levels = c(1, 2),
                                 labels = c("Female",
                                            "Male"))) %>%
  dplyr::mutate(majorF = factor(major,
                                levels = c(1, 2, 3, 4,5),
                                labels = c("Psychology",
                                           "Premed",
                                           "Biology",
                                           "Sociology",
                                           "Economics"))) %>%
  dplyr::mutate(reasonF = factor(reason,
                                 levels = c(1, 2, 3),
                                 labels = c("Program requirement",
                                            "Personal interest",
                                            "Advisor recommendation"))) %>%
  dplyr::mutate(exp_condF = factor(exp_cond,
                                   levels = c(1, 2, 3, 4),
                                   labels = c("Easy",
                                              "Moderate",
                                              "Difficult",
                                              "Impossible"))) %>%
  dplyr::mutate(coffeeF = factor(coffee,
                                 levels = c(0, 1),
                                 labels = c("Not a regular coffee drinker",
                                            "Regularly drinks coffee")))  %>%
  dplyr::mutate(hr_base_bps = hr_base / 60) %>%
  dplyr::mutate(anx_plus = rowsums(anx_base, anx_pre, anx_post)) %>%
  dplyr::mutate(hr_avg = rowmeans(hr_base + hr_pre + hr_post)) %>%
  dplyr::mutate(statDiff = statquiz - exp_sqz)
```

Chapter 2. DISTRIBUTION and UNIVARIATE PLOTS

2C-1. Frequency Distribution and Bar Chart

Request a frequency distribution using the `furniture::tableF(continuous_var)` function

```
# Frequency distrubution: majorF
```

Create a bar chart using `geom_bar()` for the Undergraduate Major (`majorF`) variable for Ihno's students.

> Make sure to add the variable of interest into the asthetics: `ggplot(aes(continuous_var))` before adding the `geom_bar()` layer.

```
# Bar Plot: majorF
```

2C-2. Bar Charts

Repeat Exercise 1 for the variables `prevmath` and `phobia`.

> IN THE WRITEUP: Would it make sense to request a histogram instead of a bar chart for phobia ? Discuss.

```
# Bar Plot: prevmath
```

```
# Bar Plot: phobia
```

2C-3. Frequency Distribution and Histogram

Request a frequency distribution and a histogram for the variable `statquiz`. Use the option in the function `geom_histogram(bins = #)` to change the number of bins or `geom_histogram(binwidth = #)` to change the bin width to give a better figure.

IN THE WRITEUP: Describe the shape of this distribution.

```
# Frequency distrubution: statquiz
```

```
# Histogram: statquiz, with a different number/width of bins
```

2C-4. Frequency Distribution and Histogram

Request a frequency distribution and a histogram for the variables baseline anxiety (`anx_base`) and baseline heart rate (`hr_base`).

IN THE WRITEUP: Comment on R's choice of class intervals for each histogram.

```
# Frequency distrubution: anx_base
```

```
# Histogram: anx_base
```

```
# Frequency distrubution: hr_base
```

```
# Histogram: hr_base
```

2C-6. Histograms -by- a Factor

Request Histograms for the variables `anx_base` and `hr_base` divided by `genderF` using an additional `facet_grid(group_var ~ .)` layer to create two plots.

```
# Histogram: anx_base, by genderF
```

```
# Histogram: hr_base, by genderF
```

2C-9. Deciles and Quartiles

Using the `quantile(probs = c(#, #, ..., #))` function, request the deciles and quartiles for the `phobia` variable.

> Make sure to add a `dplyr::pull(varname)` step to pull out only the one variable you are interested in.

```
# Deciles: phobia
```

```
# Quartiles: phobia
```

2C-10. Various Percentiles

Request the following percentiles for the variables `hr_base` and `hr_pre`: 15, 30, 42.5, 81, and 96.

```
# Percentiles: hr_base
```

```
# Percentiles: hr_pre
```

Chapter 3. SUMMARY DESCRIPTIVE STATISTICS

3C-1/3. Descriptive Statistics -full-

Use the `psych::describe()` function to find the ~~mode~~, **median** and **mean**, as well as the ~~**range**, semi-interquartile range, *unbiased* **variance*z, and*~~ unbiased*** standard deviation** for each of the *quantitative variables* in Ihno's data set.

> Make sure to use a `dplyr::select(var1, var2, ..., var12)` step to select only the variables of interest.

```
# Descriptive Stats: all quant vars
```

3C-4 Boxplots

(a) Boxplot

Create a plot for the `statquiz` variable using a `geom_boxplot()` layer.

> Make sure to specify the astheticis in `ggplot(aes(...))`. Since you want to plot the entire sample together, set `x = "Full Sample"` and `y = continuous_var`

```
# Boxplot: statquiz
```

(b) Boxplots -by- a Factor

Create a plot for the `statquiz` variable by `majorF`.

> You may choose to (1) split the x-axis with the `x = grouping_var` option in the asthetics, (2) specify a variable to fill in the boxes with color with the `fill = grouping_var`, or (3) make seperate panels by adding a `facet_grid(. ~ grouping_var` layer.

```
# Boxplot: statquiz, by majorF
```

(c) Boxplot -for- a Subset

Use a `dplyr::filter()` step filter the subjects in the dataset to create a **Boxplot** for the `statquiz` variable for just the `female Biology` majors.

> Make sure to use `==` instead of `=` to test for equality within the filter step. Make sure the use a `&` symbol to require multiple conditions.

```
# Boxplot: statquiz, for a subset
```

(d) Boxplots -by- a Factor and -for- a Subset

Use `dplyr::filter()` to create a SIDE-by-SIDE Boxplots for the `statquiz` variable that compares the `female Psychology` majors to the `female Biology` majors.

> A helpful symbol-set is `%in%`, which tests if the thing before it (`majorF`) is included in the concatinated list of options (eg. `c("Biology", "Psychology")`) that comes after it. Make sure the use a `&` symbol to require multiple conditions.

```
# Boxplot: statquiz, by a factor, for a subset
```

3C-5. Boxplots -for- Repeated Measures

Create Boxplots for both baseline and prequiz **anxiety**, so that they appear side-by-side on the same graph.

> Some data manipulations is needed to "stack" the two variables (baseline and pre-test) into a single variable. This is done with with the `tidyr::gather(key = "new_key", value = "new_value", old_var_1, old_var_2, ...)` function.

```
# Boxplot: anxiety, compare two repeated measures
```

3C-6. Descriptive Statistics -by- a Factor

Use `furniture::table1()` to find the *mean* and *standard deviation* for each of the *quantitative variables* separately for the `male` and `female` econ majors.

> Make sure to use the `dplyr::group_by(grouping_var)` step before the `furniture::table1()` step.

```
# Descriptive Stats: all quant vars, by genderF
```

Chapter 4. STANDARDIZED SCORES

4C-1. Calculate z-Scores

Use the `dplyr::mutate(new_zscore_var = scale(old_orig_var))` function to create two new variables consisting of the *z scores* for the **anxiety** and **heart rate** measures at **baseline** in Ihno's data set.

Then request *means* and *SD's* of the *z-score* variables with the `furniture::table1()` function to demonstrate that the means and SD s are `0` and `1`, respectively, in each case.

```
# Descriptive Stats: baseline anx & hr, original and z-scores
```