Psy/Educ 6600: Unit 1 Homework

Exploratory Data Analysis

Original: Dr. Sarah Schwartz
Updated by: Tyson Barrett

Spring 2018

Contents

Chapter 1. DATA PREPARATION **2**
Load Packages . . . 2
Import Data, Define Factors, and Compute New Variables . . . 2

Chapter 2. DISTRIBUTION and UNIVARIATE PLOTS **3**
2C-1. Frequency Distribution and Bar Chart . . . 3
2C-2. Bar Charts . . . 4
2C-3. Frequency Distribution and Histogram . . . 5
2C-4. Frequency Distribution and Histogram . . . 6
2C-6. Histograms -by- a Factor . . . 8
2C-9. Deciles and Quartiles . . . 9
2C-10. Various Percentiles . . . 9

Chapter 3. SUMMARY DESCRIPTIVE STATISTICS **10**
3C-1/3. Descriptive Statistics -full- . . . 10
3C-4 Boxplots . . . 11
(a) Boxplot . . . 11
(b) Boxplots -by- a Factor . . . 12
(c) Boxplot -for- a Subset . . . 13
(d) Boxplots -by- a Factor and -for- a Subset . . . 14
3C-5. Boxplots -for- Repeated Measures . . . 15
3C-6. Descriptive Statistics -by- a Factor . . . 16

Chapter 4. STANDARDIZED SCORES **17**
4C-1. Calculate z-Scores . . . 17

Chapter 5. Intro to Hypothesis Testing: 1 Sample z-Test **17**
5C-3. 1 Sample `z`-Test compared to historic controls for `mathquiz` and `statquiz` . . . 17
5C-4. Test for Normaity for `mathquiz` and `statquiz` . . . 18
Skewness and Kurtosis . . . 18
Shapiro-Wilk's Test . . . 18
Create Histograms . . . 19
Create QQ Plots . . . 20

Chapter 6. Confidence Interval Estimation: The `t` Distribution **21**
6C-1. 1-sample `t`-tests for `anx_base`, `anx_pre`, and `anx_post` . . . 21
6C-2. 1-sample `t`-tests for `hr_base` among MEN . . . 22
6C-3. 1-sample `t`-tests for `hr_post` among FEMALE . . . 23

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
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 the ~~mode~~, **median**, and **mean**, as well as the **range**, ~~semi-interquartile range~~, *unbiased* **variance**, 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`.

Make sure to set `x = grouping_var` and `y = continuous_var` in the asthetics.

```
# 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. It will be helpful to set the asethics such that `x = one_grouping_var` and `fill = another_grouping_var`, while letting `y = continuous_var`.

```
# 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 test if the thing before it is included in the concatinated list of elements that comes after it.

```
# 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_var, value = new_value_var, 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 `splitby = ~ grouping_var` option.

```
# 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.

Request *means* and *SD's* of the *z-score* variables 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
```

Chapter 5. Intro to Hypothesis Testing: 1 Sample z-Test

5C-3. 1 Sample z-Test compared to historic controls for `mathquiz` and `statquiz`

TEXTBOOK QUESTION: *(A) In the past 10 years, previous stats classes who took the same math quiz that Inho's students took* ***averaged 28*** *with a* ***standard deviation of 8.5****. What is the two-tailed p value for Inho's students with respect to that past population? (Don't forget that the N for mathquiz is not 100.) Would you say that Inho's class performed significantly better than previous classes? Explain. (B) Redo part a assuming that the same previous classes had also taken the same statquiz and* ***averaged 6.1*** *with a* ***standard deviation of 2.5****.*

DIRECTIONS: Find the mean (`M`) and sample size (`n`) for `mathquiz` and `statquiz` and then work the rest of the statistical test by hand in the printed homework packet.

> **NOTE:** You may use the `furniture::table1()` funciton gives the mean, but it only gives the total `n` for all variables. Since some students were missing the math quiz, but not the stat quiz the sample sizes are different. So use the `psych::describe()` function to get the means and the sample size for each variable.

```
# Find the mean and n for: mathquiz, statquiz
```

5C-4. Test for Normaity for `mathquiz` and `statquiz`

TEXTBOOK QUESTION: *Test both the math quiz and stat quiz variables for their resemblance to normal distributions. Based on skewness, kurtosis, and the Shapiro-Wilk statistic, which variable has a sample distribution that is not very consistent with the assumption of normality in the population?*

Skewness and Kurtosis

DIRECTIONS: Find the skewness and kurtosis for `mathquiz` and `statquiz`

> **NOTE:** Yes, you just did this above using the `psych::describe()` function… so you may skip it here if you want.

```
# Find the skewness and kurtosis for: mathquiz, statquiz
```

Shapiro-Wilk's Test

DIRECTIONS: Use the `shapiro.test()` function to test for normality in a small'ish sample.

> **NOTE:** You must use a `dplyr::pull()` step to pull out one variable from the dataset before you can use the `shapiro.test()` function.

```
# Shapiro-Wilk's Normality Test for: mathquiz
```

```
# Shapiro-Wilk's Normality Test for: statquiz
```

Create Histograms

DIRECTIONS: Use `geom_histogram()` after setting the `ggplot(aes())`. Make sure to try different `bins = #` or `binwidth = #` to get a 'good looking' plot.

> **NOTE:** For histograms, you do need to specify the variable name as `x`in the `aes(x = variable)` option.

```
# Histogram for: mathquiz
```

```
# Histogram for: statquiz
```

Create QQ Plots

DIRECTIONS: Use `geom_qq()` after setting the `ggplot(aes())`.

> **NOTE:** For qq plots, you do need to specify the variable name as `sample`in the `aes(sample = variable)` option.

```
# Histogram for: mathquiz
```

```
# Histogram for: statquiz
```

Chapter 6. Confidence Interval Estimation: The `t` Distribution

6C-1. 1-sample `t`-tests for `anx_base`, `anx_pre`, and `anx_post`

TEXTBOOK QUESTION: *Perform one-sample t tests to determine whether the baseline, pre-, or postquiz anxiety scores of Inho's students differ significantly ($\alpha = .05$, two-tailed) from the mean ($\mu = 18$) found by a very large study of college students across the country. Find the 95% Cconfidence interval for the population mean for each of the* ***three*** *anxiety measures.*

DIRECTIONS: Use the `t.test(mu = #)` function to perform a 1 sample `t`-test. Make sure to sepify the Null hypothesis value for μ.

> **NOTE:** You must use a `dplyr::pull()` step to pull out one variable from the dataset before you can use the `t.test()` function.

```
# 1-sample t-test for: anx_base
```

```
# 1-sample t-test for: anx_base
```

```
# 1-sample t-test for: anx_base
```

6C-2. 1-sample `t`-tests for `hr_base` among MEN

TEXTBOOK QUESTION: *Perform a one-sample t test to determine whether the average baseline heart rate of Inho's* ***male*** *students differs significantly from the* ***mean*** *heart rate ($\mu = 70$) for college-aged men at the* ***.01 level****, two-tailed. Find the* ***99%*** *confidence intervals for the population mean represented by Inho's* ***male*** *students.*

DIRECTIONS: Similar to the last problem, use the `t.test(mu = #)` function to perform a 1 sample `t`-test. This time, make sure the subset out the males only with a `dplyr::filter()` step prior to the `dplyr::pull()` step.

> **note:** To change from the default 95% confidence intervals, make sure to specify `conf.level = 0.99` inside the `t.test()` function.

```
# 1-sample t-test for MALES: hr_base
```

6C-3. 1-sample `t`-tests for `hr_post` among FEMALE

TEXTBOOK QUESTION: *Perform a one-sample t test to determine whether the average postquiz heart rate of Inho's* ***female*** *students differs significantly ($\alpha = .05$, two-tailed) from the* ***mean*** *resting heart rate ($\mu = 72$) for college-aged women. Find the 95% confidence interval for the population mean represented by Inho's* ***female*** *students.*

DIRECTIONS: This time, subset out WOMEN and choose the post-quiz heart rate. Also, use a different population null value.

```
# 1-sample t-test for MALES: hr_base
```