Psy/Educ 6600: Unit 2 Homework

Groundwork for Inference

Dr. Sarah Schwartz - Answer Key

Spring 2019

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
library(car)           # Companion to "Applied Regression"
```

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 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 Ihno's students took* ***averaged 28*** *with a* ***standard deviation of 8.5****. What is the two-tailed p value for Ihno's students with respect to that past population? (Don't forget that the N for mathquiz is not 100.) Would you say that Ihno'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. Recall that the not all participants have scores recorded for the math quiz. When you have such missing values, care must be taken on how to handle the partial-data cases.

First, use the `furniture::table1()` function to compute the mean of each othe two variables using the default settings. Notice that the only sample size given is thenumber of cases with complete data. Participants with incomplete data are ignored when computing the mean and standard deviations.

```
# Find the mean and n for: mathquiz, statquiz  <-- default settings: na.rm = TRUE
```

DIRECTIONS: Second, use the `furniture::table1()` function to compute the mean of each othe two variables changing settings with the option to include incomplete cases. Notice that the only sample size given is the total of all cases with complete and incomplete data, although the means and standard deviations are based on different sub-samples (option: `na.rm = FALSE`).

```
# Find the mean and n for: mathquiz, statquiz  <-- override the default settings: na.rm = FALSE
```

Third, use the `psych::describe()` function to compute the mean of each othe two variables. Notice that two different samples sizes are given. Compare the means to the two tables above.

> **NOTE:** 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
```

See results above.

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
```

Histogram

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

```r
# Histogram for: mathquiz
```

```r
# Histogram for: statquiz
```

QQ Plot

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 Ihno'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 specify 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_pre
```

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

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 Ihno'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 Ihno'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 MEN only (`genderF == "Male"`) 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 Ihno'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 Ihno's* ***female*** *students.*

DIRECTIONS: This time, subset out WOMEN (`genderF == "Female"`) and choose the post-quiz heart rate. Also, use a different population null value (μ).

```
# 1-sample t-test for FEMALES: hr_post
```

Chapter 7. Independent Samples t-Test for Means

7C-1. Independent Samples t-Test for Mean `hr_base` by `genderF`

TEXTBOOK QUESTION: *Perform a two-sample t test to determine whether there is a statistically significant difference in* ***baseline heart rate*** *between the* ***men and the women*** *of Ihno's class. Do you have* ***homogeneity of variance****? Report your results as they might appear in a journal article. Include the 95% CI for this gender difference.*

Assumtion Check: Homogeneity of Variance

DIRECTIONS: Before performing the test, check to see if the assumption of homogeneity of variance is met using **Levene's Test**. For a independent samples `t`-test for means, the men and women need to have the same amount of spread (SD) in their baseline hear rates.

> **NOTE:** Use the `car::leveneTest()` function to do this. Inside the funtion you need to specify at least three things (sepearated by commas):
>
> - the formula: `continuous_var ~ grouping_var` (replace with your variable names)
> - the dataset: `data = .` to pipe it from above
> - the center: `center = "mean"` since we are comparing means

```
# Levene's F-test for HOV: hr_base by genderF
```

Perform the t-Test for Means in 2 Indep Groups

DIRECTIONS: Test if men and women have different baseline heart rates using the `t.test()` function.

> Use the same `t.test()` funtion we have used in the prior chapters. This time you need to specify a two mandatory options:
>
> - the formula: `continuous_var ~ grouping_var` (replace with your variable names)
> - the dataset: `data = .` to pipe it from above
>
> There are also more optional options, the first of which is most commonly used:
>
> - is homogeneity satified: `var.equal = TRUE` (NOT the default)
> - independent vs. paired: `paired = FALSE` (this is the default)
> - confidence level: `conf.level = #` (defults to .95)

```
# indep groups t-test for means: hr_base by genderF
```

7C-2. Independent Samples `t`-Test for Mean `phobia` by `genderF`

TEXTBOOK QUESTION: *Repeat Exercise 1 for the phobia variable.*

Assumtion Check: Homogeneity of Variance

DIRECTIONS: Before performing the test, check to see if the assumption of homogeneity of variance is met using **Levene's Test**. For a independent samples `t`-test for means, the men and women need to have the same amount of spread (SD) in their phobia self-ratings.

```
# Levene's F-test for HOV: phobia by genderF
```

Perform the `t`-Test for Means in 2 Indep Groups

DIRECTIONS: Test if men and women have different phobia self-ratings using the `t.test()` function.

```
# indep groups t-test for means: phobia by genderF
```

7C-3. Independent Samples `t`-Test for Mean `hr_post` by `exp_condF` (Restricted to just the Easy vs.Imposible groups)

TEXTBOOK QUESTION: *Perform a two-sample `t` test to determine whether the students in the `impossible to solve` condition exhibited significantly higher postquiz heart rates than the students in the `easy to solve` condition at the .05 level. Is this t test significantly at the .01 level? Find the 99% CI for the difference of the two population means.*

Assumtion Check: Homogeneity of Variance

DIRECTIONS: Before performing the test, check to see if the assumption of homogeneity of variance is met using **Levene's Test**. For a independent samples `t`-test for means, the **"Easy"** and **"Imposible"** groups have the same amount of spread (SD) in their post quiz.

> Prior to running Levene's test, make sure to reduce your dataset by throwing out the students who were assigned the middle two difficulty levels of experimental quiz. You can do this by prefacing levene's test with `dplyr::filter(exp_condF %in% c("Easy", "Impossible"))`.

```
# Levene's F-test for HOV: hr_post by exp_condF
```

Perform the `t`-Test for Means in 2 Indep Groups

DIRECTIONS: Test if **"Easy"** and **"Imposible"** groups have different phobia self-ratings using the `t.test()` function.

> Prior to running the t-test, make sure to reduce your dataset by throwing out the students who were assigned the middle two difficulty levels of experimental quiz. You can do this by prefacing levene's test with `dplyr::filter(exp_condF %in% c("Easy", "Impossible"))`.

```
# indep groups t-test for means: hr_post by exp_condF
```

7C-4. Independent Samples t-Test for Mean `anx_post` by `exp_condF` (Restricted to just the Easy vs.Imposible groups)

TEXTBOOK QUESTIONS: *Repeat Exercise 3 for the postquiz anxiety variable.*

Assumtion Check: Homogeneity of Variance

DIRECTIONS: Before performing the test, check to see if the assumption of homogeneity of variance is met using **Levene's Test**. For a independent samples `t`-test for means, the **"Easy"** and **"Imposible"** groups have the same amount of spread (SD) in their post quiz anxiety levels.

> Prior to running Levene's test, make sure to reduce your dataset by throwing out the students who were assigned the middle two difficulty levels of experimental quiz. You can do this by prefacing levene's test with `dplyr::filter(exp_condF %in% c("Easy", "Impossible"))`.

```
# Levene's F-test for HOV: anx_post by exp_condF
```

Perform the `t`-Test for Means in 2 Indep Groups

DIRECTIONS: Test if **"Easy"** and **"Imposible"** groups have different post quiz anxiety levels using the `t.test()` function.

> Prior to running the t-test, make sure to reduce your dataset by throwing out the students who were assigned the middle two difficulty levels of experimental quiz. You can do this by prefacing levene's test with `dplyr::filter(exp_condF %in% c("Easy", "Impossible"))`.

```
# indep groups t-test for means: anx_post by exp_condF
```

7C-5. Independent Samples `t`-Test for Mean `hr_post` by `coffeeF`

TEXTBOOK QUESTIONS: *Perform a two-sample t test to determine whether* ***coffee drinkers*** *exhibited significantly higher* ***postquiz heart rates*** *than nondrinkers at the .05 level. Is this t test significant at the .01 level? Find the* ***99%*** *confidence interval for the difference of the two population means and explain its connection to your decision regarding the null hypothesis at the* ***.01 level****.*

Assumtion Check: Homogeneity of Variance

DIRECTIONS: Just like the last question, run **Levene's test** first.

```
# Levene's F-test for HOV: hr_post by coffeeF
```

Perform the `t`-Test for Means in 2 Indep Groups

DIRECTIONS: Make sure to change the confidence level to **99%**.

```
# indep groups t-test for means: hr_post by coffeeF
```