Psy/Educ 6600: Unit 3 Homework

Hypothesis Tests for TWO Measures Per Subject

Your Name

Spring 2019

Contents

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)
```

Ihno's Dataset for Section C's

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)
```

Other Datasets for Section B's

```
schizo <- data.frame(id       = c(1:10),
                     yr_hos   = c( 5,  7, 12,  5, 11,  3,  7,  2,  9,  6),
                     ori_test = c(22, 26, 16, 20, 18, 30, 14, 24, 15, 19))

GRE <- data.frame(id          = c(1:5),
                  verbalGRE_1 = c(540, 510, 580, 550, 520),
                  verbalGRE_2 = c(570, 520, 600, 530, 520))

test_scores <-  data.frame(id      = c(1:12),
                           spatial = c(13, 32, 41, 26, 28, 12, 19, 33, 24, 46, 22, 17),
                           math    = c(19, 25, 31, 18, 37, 16, 14, 28, 20, 39, 21, 15))

child_vars  <-  data.frame(child = c(1:8),
                           shoe  = c(5.2, 4.7, 7.0, 5.8, 7.2, 6.9, 7.7, 8.0),
                           read  = c(1.7, 1.5, 2.7, 3.1, 3.9, 4.5, 5.1, 7.4),
                           age   = c(  5,   6,   7,   8,   9,  10,  11,  12))

memory <- data.frame(id    = c(1:9),
                     sound = c(8, 5, 6, 10, 3, 4, 7, 11, 9),
                     look  = c(4, 5, 3, 11, 2, 6, 4,  6, 7))
```

Chapter 9. Linear Correlation

Section B - Other Datasets

9B-5 Calculating Pearson's r

TEXTBOOK QUESTION: *A psychiatrist has noticed that the schizophrenics who have been in the hospital the longest score the lowest on a mental orientation test. The data for 10 schizophrenics are listed in the following table. (a) Calculate Pearson's r for the data. (b) Test for statistical significance at the .05 level (two-tailed).*

```
schizo
```

```
   id yr_hos ori_test
1   1      5       22
2   2      7       26
3   3     12       16
4   4      5       20
5   5     11       18
6   6      3       30
7   7      7       14
8   8      2       24
9   9      9       15
10 10      6       19
```

DIRECTIONS: Calculate Pearson's r between `yr_hos` and `ori_test` in the `schizo` dataset. Also, test against the two-sided alternative.

The `cor.test()` function needs at least two arguments:

1. the formula: `~ continuous_var1 + continuous_var2`
2. the dataset: `data = .` *we use the period to signify that the datset is being piped from above*

   NOTE: The `cor.test()` function computes the Pearson correlation coefficient by default (`method = "pearson"`), but you may also specify the Kendall (`method = "kendall"`)or Spearman (`method = "spearman"`)methods. It also defauls to testing for the two-sided alternative and computing a 95% confidence interval (`conf.level = 0.95`). You will not need to change these options in this assignment.

```
# Pearson's r: yr_hos & ori_test
```

9B-6 One vs. Two Sided Alternative

TEXTBOOK QUESTION: *If a test is reliable, each participant will tend to get the same score each time he or she takes the test. Therefore, the correlation between two administrations of the test (test-retest reliability) should be high. The reliability of the verbal GRE score was tested using five participants, as shown below. (a) Calculate Pearson's r for the test-retest reliability of the verbal GRE score. (b) Test the significance of this correlation with* $\alpha = .05$ *(one-tailed). Would this correlation be significant with a twotailed test?*

```
GRE
```

```
  id verbalGRE_1 verbalGRE_2
1  1         540         570
2  2         510         520
3  3         580         600
4  4         550         530
5  5         520         520
```

DIRECTIONS: Calculate Pearson's r between `verbalGRE_1`and `verbalGRE_2` in the `GRE` dataset TWICE. The first time test for the **one-sided** alternative and the second time for the **two-sided** alternative.

The `cor.test()` function defaults to the `alternative = "two.sided"`. If you would like a one-sided alternative, you must choose which side you would like to test: `alternative = "greater"` or `alternative = "less"`

```
# Pearson's r: verbalGRE_1 & verbalGRE_2 --> ONE tail
```

```
# Pearson's r: verbalGRE_1 & verbalGRE_2 --> TWO tails
```

Section C - Ihno's Dataset

9C-1. Scatterplots - Eyeball method for estimating correlation

TEXTBOOK QUESTION: *(A) Create a scatter plot of phobia versus statquiz. From looking at the plot, do you think the Pearson's r will be positive or negative? Large, medium, or small? (B) Create a scatter plot of baseline anxiety versus postquiz anxiety. From looking at the plot, do you think the Pearson's r will be positive or negative? Large, medium, or small?*

DIRECTIONS: Create two scatter plots: the first with `phobia` on the horizontal axis (`x`) and `statquiz` on the vertical axis (`y`) and the second with `anx_base` on the x-axis and `anx_post` on the y-axis. Then answer the rest of the question in the printed homework packet.

> **NOTE:** You may use the `geom_count()` funciton instead of the `geom_point()` function due to the high number of points that are 'over plotted' or on top of each other, since the two measures are quite coursely captured.

```
# Scatterplot: phobia vs. statquiz
```

```
# Scatterplot: anx_base vs. anx_post
```

9C-2a. Calculating Pearson's r

TEXTBOOK QUESTION: *Compute the Pearson's r between* `phobia` *and* `statquiz` *for all students; also, find the Pearson's r between baseline and postquiz anxiety.*

DIRECTIONS: Compute Pearson's r: first for `phobia` and `statquiz`, followed by `anx_base` and `anx_post` using the `cor.test()` function.

```
# Pearson's r: phobia & statquiz
```

```
# Pearson's r: anx_base & anx_post
```

9C-2b. Effect of Excluding Extreme Values

TEXTBOOK QUESTION: *Use Select Cases to delete any student whose baseline anxiety is over 29, and repeat part (B) of the first exercise. Also, rerun the correlation of baseline and postquiz anxiety. What happened to the Pearson's r ? Use the change in the scatter plot to explain the change in the correlation coefficient.*

DIRECTIONS: Create a scatterplot for `anx_base` and `anx_post`, AFTER first using a `dplyr::filter()` funtion in a prepatroy step to restrict to the subsample of students with baseline anxiety of 29 and below.

```
# Scatterplot: anx_base vs. anx_post <-- restricting to baseline anxiety of 29 and lower
```

DIRECTIONS: Compute Pearson's r: for `anx_base` and `anx_post`, AFTER first using a `dplyr::filter()` funtion in a prepatroy step to restrict to the subsample of students with baseline anxiety of 29 and below.

```
# Pearson's r: anx_base & anx_post <-- restricting to baseline anxiety of 29 and lower
```

9C-3. Reporting APA Style

TEXTBOOK QUESTION: *(a) Compute Pearson's r's among the three measures of anxiety. Write up the results in APA style. (b) Compute the average of the three measures of anxiety, and then compute the correlation between each measure of anxiety and the average, ~~so that the output contains a single column of correlations (do this by creating and appropriately modifying a syntax file~~).*

DIRECTIONS: First, compute a new variable called `anx_mean` that is the average of all three of the anxiety measures using the `furniture::rowmeans()` function. Then use the `furniture::tableC()`function to create a correlation matrix for all FOUR anxiety meausres.

```
# Pearson's r: anx_mean, anx_base, anx_pre, & anx_post
```

9C-4. Missing Values

TEXTBOOK QUESTION: *(a) Compute Pearson's r for the following list of variables: mathquiz , statquiz , and phobia . (b) Repeat part a after selecting exclude cases listwise. Which correlation was changed? Explain why.*

Directions: Compute the correlation matrix between `mathquiz`, `statquiz`, and `phobia` using the `furniture::tableC()` function two times; first with all defaults and again with listwise deletion.

> **Note:** The `furniture::tableC()` funtion defaults to `na.rm = FALSE` which displays `NA` for any correlation between a pair of variables where even one subject is missing one value. To use listwise deletion, specify the option `na.rm = TRUE`.

```
# Pearson's r: (default: na.rm = FALSE)
```

```
# Pearson's r: "complete.obs" (list-wise deletion)
```

Chapter 10. Linear Regression

Section B - Other Datasets

10B-6 Swapping x and y

TEXTBOOK QUESTION: *A cognitive psychologist is interested in the relationship between spatial ability (e.g., ability to rotate objects mentally) and mathematical ability, so she measures 12 participants on both variables. The data appear in the following table. (a) Find the regression equation for predicting the math score from the spatial ability score. (b) Find the regression equation for predicting the spatial ability score from the math score. (c) According to your answer to part a, what math score is predicted from a spatial ability score of 20? (d) According to your answer to part b, what spatial ability score is predicted from a math score of 20?*

```
test_scores
```

```
   id spatial math
1   1      13   19
2   2      32   25
3   3      41   31
4   4      26   18
5   5      28   37
6   6      12   16
7   7      19   14
8   8      33   28
9   9      24   20
10 10      46   39
11 11      22   21
12 12      17   15
```

DIRECTIONS: Use the `lm()` function to fit a linear model or linear regression model TWICE for `math` and `spacial` in the `test_scores` dataset, specifing which is x and which is y.

The `lm()` function needs at least two arguments:

1. the formula: `continuous_y ~ continuous_x`
2. the dataset: `data = .` *we use the period to signify that the datset is being piped from above*

   NOTE: To view more complete information, add a `summary()` step using a pipe AFTER the `lm()` step

```
# Linear model: y = math & x = spatial
```

```
# Linear model: y = spatial & x = math
```

10B-9 Predictions and Residuals

TEXTBOOK QUESTION: *If you calculate the correlation between shoe size and reading level in a group of elementary school children, the correlation will turn out to be quite large, provided that you have a large range of ages in your sample. The fact that each variable is correlated with age means that they will be somewhat correlated with each other. The following table illustrates this point. Shoe size is measured in inches, for this example, reading level is by grade (4.0 is average for the fourth grade), and age is measured in years. (a) Find the regression equation for predicting shoe size from age. (b) Find the regression equation for predicting reading level from age. (c) Use the equations from parts a and b to make shoe size and reading level predictions for each child. Subtract each prediction from its actual value to find the residual.*

```
child_vars
```

```
  child shoe read age
1     1  5.2  1.7   5
2     2  4.7  1.5   6
3     3  7.0  2.7   7
4     4  5.8  3.1   8
5     5  7.2  3.9   9
6     6  6.9  4.5  10
7     7  7.7  5.1  11
8     8  8.0  7.4  12
```

DIRECTIONS: Use the `lm()` function to fit a linear model or linear regression model TWICE for `shoe` and `read` each perdicited in turn by `age` in the `child_vars` dataset, specifing which is x and which is y.

```
# Linear model: y = shoe & x = age
```

```r
# Linear model: y = read & x = age
```

DIRECTIONS: Starting with the `child_vars` dataset, create four new variables, each with a seperate `dplyr::mutate()` function step. Pipe it all together ans save it as new dataset with the `child_new <-` assignment operator to use in the next step.

1. `shoe_pred` Use the appropriate regression equation
2. `shoe_resid` Subtract: `shoe` (the original) minus `shoe_resid` (the residual)
3. `read_pred` Use the appropriate regression equation
4. `read_resid` Subtract: `read` (the original) minus `read_resid` (the residual)

```r
# create new variables --> save as: child_new
```

Note: Remove the hashtag symbol at the first of the code line below to show your new variables.

```r
#child_new
```

10B-10 Raw Correlation vs. Partial Correlation

TEXTBOOK QUESTION: *(a) Calculate Pearson's r for shoe size and reading level using the data from Exercise 9. (b) Calculate Pearson's r for the two sets of residuals you found in part c of Exercise 9. (c) Compare your answer in part b with your answer to part a. The correlation in part b is the partial correlation between shoe size and reading level after the confounding effect of age has been removed from each variable (see Chapter 17 for a much easier way to obtain partial correlations).*

DIRECTIONS: Calculate Pearson's r between `shoe`and `read` in the `child_new` dataset.

```
# Pearson's r: shoe & read
```

DIRECTIONS: Calculate Pearson's r between `shoe_resid`and `read_resid` in the `child_new` dataset.

```
# Pearson's r: shoe_resid & read_resid
```

Section C - Ihno's Dataset

10C-1. Linear Regression

TEXTBOOK QUESTION: *Perform a linear regression to predict statquiz from phobia, and write out the raw-score regression formula. Do the slope and Y intercept differ significantly from zero? Explain how you know. What stats quiz score would be predicted for a student with a phobia rating of 9? Approximately what phobia rating would a student need to have in order for her predicted statquiz score to be 7.2?*

Directions: Use the `lm()` function to fit a linear model or linear regression model predicting `statquiz` from `phobia`.

```
# Linear model: y = statquiz & x = phobia
```

10C-2. Subgroups Analysis

TEXTBOOK QUESTION: *(a) Perform a linear regression to predict pre-quiz anxiety from phobia, and write out the raw-score regression formula. (b) Repeat part a separately for men and women. For each gender, what prequiz anxiety rating would be predicted for someone reporting a phobia rating of 8? For which gender should you really not be making predictions at all? Explain.*

Directions: Use the `lm()` function to fit a linear model or linear regression model predicting `anx_pre` from `phobia`. Then repeat the same model TWICE more: first among just men and then for just women.

> **Note:** Use the `dplyr::filter()` function to subset the sample BEFORE fitting the model. Also, be aware of which type of variable you are using: `genderF == "Male"` or `gender == 2` works, but `gender == male` does NOT.

```
# Linear model: y = anx_pre & x = phobia  <-- full sample
```

```
# Linear model: y = anx_pre & x = phobia  <-- subset of men
```

```
# Linear model: y = anx_pre & x = phobia  <-- subset of women
```

Chapter 11. Matched pairs t-test

Section B - Other Datasets

11B-3 Matched Pairs vs. Direct Difference Methods

TEXTBOOK QUESTION: *Using the data from Exercise 9B6, which follows. (a) Determine whether there is a significant tendency for verbal GRE scores to improve on the second testing. Calculate the matched t in terms of the Pearson correlation coefficient already calculated for that exercise. (b) Recalculate the matched t test according to the direct-difference method and compare the result to your answer for part a.*

```
GRE
```

```
  id verbalGRE_1 verbalGRE_2
1  1         540         570
2  2         510         520
3  3         580         600
4  4         550         530
5  5         520         520
```

DIRECTIONS: Calculate the matched pairs t test between `verbalGRE_1`and `verbalGRE_2` in the `GRE` dataset.

In order to use this function, you MUST first restructure your dataset so that the TWO continous variables are stacked or **gathered** together. Use the `tidyr::gather()` function with the following FOUR options:

- A new variable name that will store the original variable names: `key = new_group_var`
- A new variable name that will store the original variable values: `value = new_continuous_var`
- List the original variable names: `continous_var1, continuous_var2`
- Do not get ride of blank values: `na.rm = FALSE`

After the dataset is fathered, ad the `t.test()` function, which needs at least THREE arguments:

- the formula: `continuous_var ~ group_var`
- the dataset: `data = .` *we use the period to signify that the datset is being piped from above*
- specify the data is paired: `paired = TRUE` *the default is independent groups*

    Note: I suggest using `key = time` and `value = verbalGRE`.

```
# Paired t-test: verbalGRE1 & verbalGRE2
```

DIRECTIONS: Calculate a NEW variable called `verbalGRE_diff` with the `dplyr::mutate()` function by subtracting the `verbalGRE_1`and `verbalGRE_2` variables in the `GRE` dataset. Pipe it all together ans save it as new dataset with the `GRE_new <-` assignment operator to use in the next step.

```
# Compute a new variable --> save as: child_new
```

> **Note:** Remember that before you do a one-sample t test for the mean, you have to use the `dplyr::pull()` function (see chapter 6)

```
# 1-sample t test: pop mean of verbalGRE_diff = 0 (no difference)
```

11B-8 Confidence Intervale for the Mean Difference

TEXTBOOK QUESTION: *A cognitive psychologist is testing the theory that short-term memory is mediated by subvocal rehearsal. This theory can be tested by reading aloud a string of letters to a participant, who must repeat the string correctly after a brief delay. If the theory is correct, there will be more errors when the list contains letters that sound alike (e.g., G and T) than when the list contains letters that look alike (e.g., P and R). Each participant gets both types of letter strings, which are randomly mixed in the same experimental session. The number of errors for each type of letter string for each participant are shown in the following table. (a) Perform a matched t test ($\alpha = .05$, one tailed) on the data above and state your conclusions. (b) Find the 95% confidence interval for the population difference for the two types of letters.*

```r
memory
```

```
  id sound look
1  1     8    4
2  2     5    5
3  3     6    3
4  4    10   11
5  5     3    2
6  6     4    6
7  7     7    4
8  8    11    6
9  9     9    7
```

DIRECTIONS: Calculate the matched pairs t test between `sound`and `look` in the `memory` dataset twice: first as a **one-tail** test and then again as a **two-tailed*** test.

> **Note:** I suggest using `key = type` and `value = errors`.

```r
# Paired t-test: sound and look --> ONE tail
```

```
# Paired t-test: sound and look --> TWO tails
```

11B-9 t-Test for Mean Difference vs. Correlation

TEXTBOOK QUESTION: *For the data in Exercise 10B6: (a) Calculate the matched t value to test whether there is a significant difference ($\alpha = .05$, two tailed) between the spatial ability and math scores. Use the correlation coefficient you calculated to find the regression slope in Exercise 10B6. (b) Explain how the Pearson r for paired data can be very high and statistically significant, while the matched t test for the same data fails to attain significance.*

```
test_scores
```

```
   id spatial math
1   1      13   19
2   2      32   25
3   3      41   31
4   4      26   18
5   5      28   37
6   6      12   16
7   7      19   14
8   8      33   28
9   9      24   20
10 10      46   39
11 11      22   21
12 12      17   15
```

DIRECTIONS: Calculate Pearson's r between `spatial` and `math` in the `schizo` test_scores

```
# Pearson's r: spatial & math
```

Note: I suggest using `key = type` and `value = score`.

```
# Paired t-test: spatial & math
```

Section C - Ihno's Dataset

11C-1a. Matched pairs t-test

TEXTBOOK QUESTION: *(a) Perform a matched-pairs t test to determine whether there is a significant increase in heart rate from baseline to the prequiz measurement. (b) Repeat these paired t tests separately for men and women.*

Directions: Calculate the matched pairs t test between `hr_base`and `hr_pre` Then repeat the calculation TWICE more: first among just men and then for just women.

> **Note:** Use the `dplyr::filter()` function to subset the sample BEFORE fitting the model. Also, be aware of which type of variable you are using: `genderF == "Male"` or `gender == 2` works, but `gender == male` does NOT.
>
> **Note:** I suggest using `key = time` and `value = hr`.

```
# Paired t-test: hr_base & hr_pre  <-- full sample
```

```
# Paired t-test: hr_base & hr_pre  <-- subset of men
```

```
# Paired t-test: hr_base & hr_pre  <-- subset of women
```

11C-2. More than Two Variables

TEXTBOOK QUESTION: *(a) Perform a matched-pairs t test to determine whether there is a significant increase in anxiety from baseline to the prequiz measurement. (b) Perform a matched-pairs t test to determine whether there is a significant decrease in anxiety from the prequiz to the postquiz measurement.*

Directions: Calculate the matched pairs t test first between `anx_base`and `anx_pre` and then between `anx_pre`and `anx_post`.

Note: I suggest using `key = time` and `value = anx`.

```
# Paired t-test: anx_base & anx_pre
```

```
# Paired t-test: anx_pre & anx_post
```

11C-3. Compared to Correlation

TEXTBOOK QUESTION: *Perform a matched-pairs t test to determine whether there is a significant difference in mean scores between the experimental stats quiz and the regular stats quiz. Is the correlation between the two quizzes statistically significant? Explain any discrepancy between the significance of the correlation and the significance of the matched t test.*

Directions: Calculate the matched pairs t test between `exp_sqz`and `statquiz`.

Note: I suggest using `key = type` and `value = score`.

```
# Paired t-test: exp_sqz & statquiz
```

Directions: Compute Pearson's r: for `exp_sqz` and `statquiz`

```
# Pearson's r: exp_sqz & statquiz
```