Psy/Educ 6600: Chapter 5

Intro to Hypothesis Testing: 1 Sample z-Test

Your Name

Spring 2019

Contents

PREPARATION

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"
```

SECTION C

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)
```

Question C-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 the number of cases with complete data. Participants with incomplete data are ignored when computing the mean and standard deviations.

```r
# Find the mean and n for: mathquiz, statquiz  <-- default settings: na.rm = TRUE
data_clean %>%
  furniture::table1(mathquiz, statquiz)
```

```
----------------------------
          Mean/Count (SD/%)
          n = 85
 mathquiz
          29.1 (9.5)
 statquiz
          6.8 (1.7)
----------------------------
```

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`).

```r
# Find the mean and n for: mathquiz, statquiz  <-- override the default settings: na.rm = FALSE
data_clean %>%
  furniture::table1(mathquiz, statquiz, na.rm = FALSE)
```

```
----------------------------
          Mean/Count (SD/%)
          n = 100
 mathquiz
          29.1 (9.5)
 statquiz
          6.9 (1.7)
----------------------------
```

Third, use the `psych::describe()` function to compute the mean of each other 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
data_clean %>%
  dplyr::select(mathquiz, statquiz) %>%
  psych::describe()
```

```
         vars   n  mean   sd median trimmed   mad min max range  skew kurtosis
mathquiz    1  85 29.07 9.48     30   29.26 10.38   9  49    40 -0.19    -0.58
statquiz    2 100  6.86 1.70      7    7.03  1.48   1  10     9 -0.97     0.77
           se
mathquiz 1.03
statquiz 0.17
```

Question C-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
data_clean %>%
  dplyr::select(mathquiz, statquiz) %>%
  psych::describe()
```

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
data_clean %>%
  dplyr::pull(mathquiz) %>%
  shapiro.test()
```

```

    Shapiro-Wilk normality test

data:  .
W = 0.98221, p-value = 0.2917
```

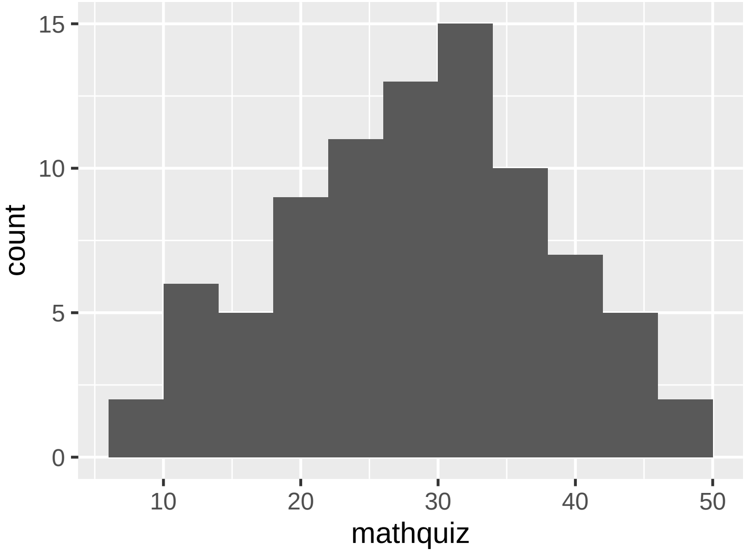
```
# Shapiro-Wilk's Normality Test for: statquiz
data_clean %>%
  dplyr::pull(statquiz) %>%
  shapiro.test()
```

```

    Shapiro-Wilk normality test

data:  .
W = 0.89865, p-value = 1.222e-06
```

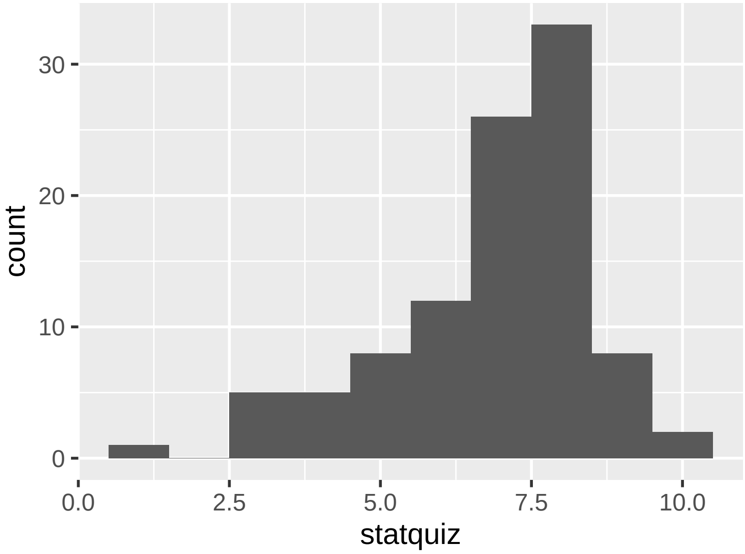
Histogram

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

```
# Histogram for: mathquiz
data_clean %>%
  ggplot(aes(mathquiz)) +
  geom_histogram(binwidth = 4)
```



```
# Histogram for: statquiz
data_clean %>%
  ggplot(aes(statquiz)) +
  geom_histogram(binwidth = 1)
```

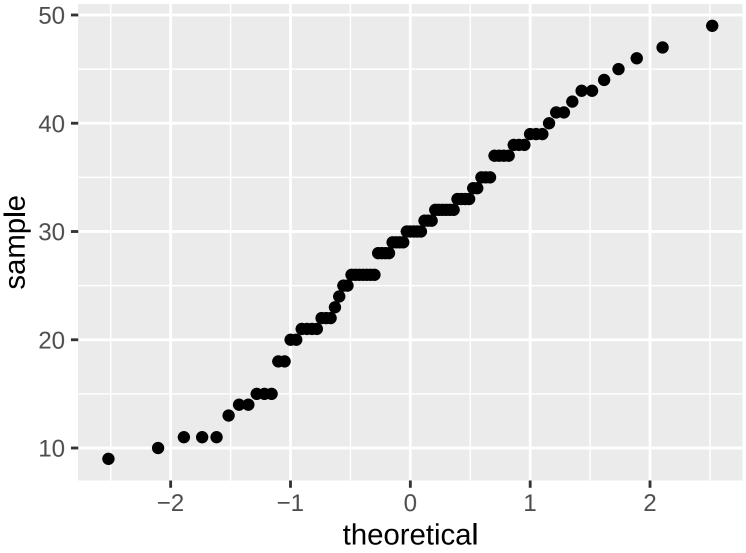


QQ Plot

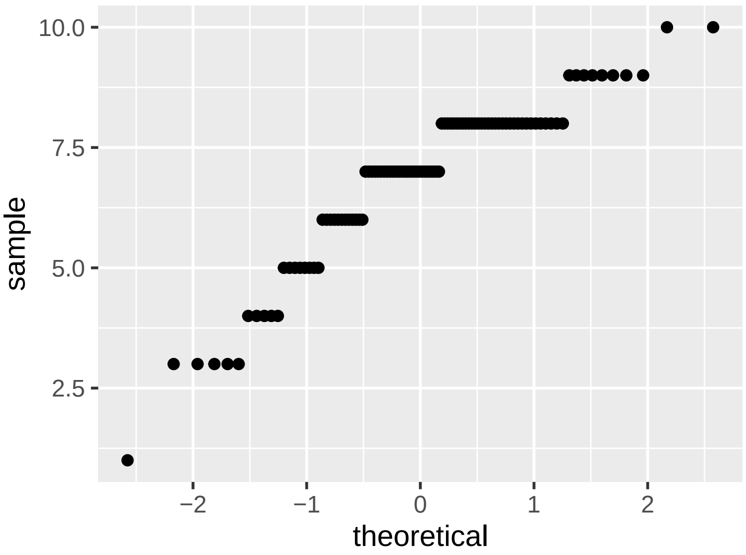
DIRECTIONS: Use `geom_qq()` after setting the `ggplot(aes())`.

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

```
# Histogram for: mathquiz
data_clean %>%
  ggplot(aes(sample = mathquiz)) +
  geom_qq()
```



```
# Histogram for: statquiz
data_clean %>%
  ggplot(aes(sample = statquiz)) +
  geom_qq()
```


10.0
7.5
5.0
2.5
sample
−2
−1
0
1
2
theoretical