PSYCH308A - Data Analysis 4 (DA4)

Brady C. Jackson

2024/10/04

Contents

```
# Load packages. Set messages and warnings to FALSE so I don't have to see the
# masking messages in the output.
library(psych)
library(jmv)       # for descriptive
library(ggplot2)
library(dplyr)
library(magrittr)
library(tidyr)     # for pivot_longer
```

Question 01 (Q01):

1. Interpret a p-value of .042. (This question is **not** asking for the decision this p-value results in, rather what does this value mean?)

Answer to Q01:

A p-value of .042 implies that the finding of some statistical test (z-test, t-test, etc.) that the probability of obtaining the test statistic found or something more extreme was .042 (a.k.a. 4.2%) assuming the null hypothesis was true

---

Question 02 (Q02):

2. ***In no more than two sentences,*** what is the relationship between sample size, effect size, and power?

Answer to Q02:

Power is positively correlated with both sample size and effect size in a manner that convolutes the two (sample and effect). That is to say, the power of a finding will increase if either sample size increases or effect size increases, and if one (e.g. sample size) is necessarily small, a higher power can be achieved by increasing the other (e.g. effect size).

Question 03 (Q03):

3. A researcher records the number of words recalled by students presented with a list of words for 1 minute. In one group, students were presented with the list of words in color; in a second group, the same words were presented in black and white. An equal number of students were in each group. The researcher reports the following: *Participants recalled significantly more words when the words were presented in color (M = 12.4 words) versus black and white (M = 10.9 words), t(48) = 2.01, p = .035, d = 0.18.* Based on the previous statement, what is the sample size in each group?

Answer to Q03:

Given the wording of the prompt, I'm assuming this test was conducted as an independent sample t-test ("In one group … in a second group" instead of phrasing like, "…the group of students looked at colored words and then later looked at black and white words…").

The degrees of freedom for the t-test ran is given as 48 (*t(48)*).
As this is an indepdendent t-test, two parameters had to be estimated, the mean of the population from which group 1 was drawn, and the mean of the population from which group 2 was drawn. Therefore, we have the relationship

$$
\begin{aligned}
df &= [Things\ we\ know\ (n_{total})] - [Estimated\ Parameters\ (n_{P,est})] \\
df &= n_{total} - n_{P,est} \\
48 &= n_{total} - 2 \\
48 + 2 &= n_{total} \\
50 &= n_{total}
\end{aligned}
$$

Therefore, there were a total of 50 students in the study, with 25 in each group given that the prompt states the same number of students were in each group.

Research Prompt for Q04 through Q09:

You are teaching your first Intro to Psychology course! After the midterm, you are disappointed with your students' overall test scores. You decide to implement two different required study techniques. There are 100 students in the class; 50 of them will be required to meet in groups to study right before the final (Group A) and the other 50 will be required to create flashcards to aid in memorization (Group B). You are interested in two primary research questions:

Research Question 1 (RQ1): Q04 - Q06:

Did student test scores improve significantly from the midterm to the final?
Data: *308A.RQ1 Data.DA4.csv*

Question 04 (Q04):

4. Visualize your data for this research question. Include your visualization here.

Answer to Q04:

Loading Data for Q04 The first step to visualizing the data is to load it. See code below:

```
# In order to visualize the data we must first load it. To do so we lool
# in the present working directory for CSV files, and take the one that
# has RQ1 in its filename. Append it to the current working directory
# to create the fullpath for loading
here <- getwd()
rq1_name <- list.files(here, pattern = ".*RQ1.*csv")

# Use the file.path function to create a platform appropriate fullpath / filename
# to the Research Question 1 data.
rq1_file <- file.path(here, rq1_name)

# Read the CSV data in as rq1_dat
rq1_dat <- read.csv(rq1_file, header = TRUE)

# Lower-case all column names for convenience
colnames(rq1_dat) <- tolower(colnames(rq1_dat))
```

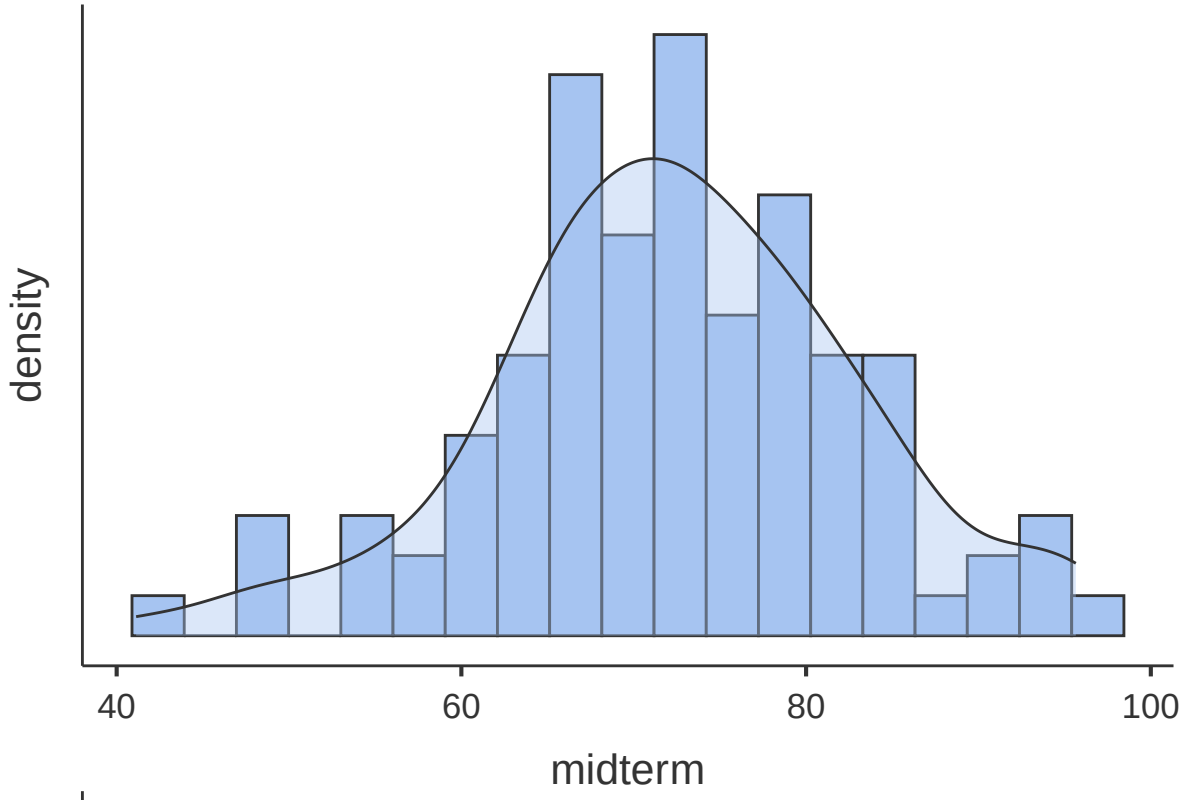
Data Prep Q04 Ultimately we're going to answer Research Question 1 as a dependent T-test since we're evaluating change in student test scores over time (regardless) of study method. So we're going to prepare the data further with some math. See code below:

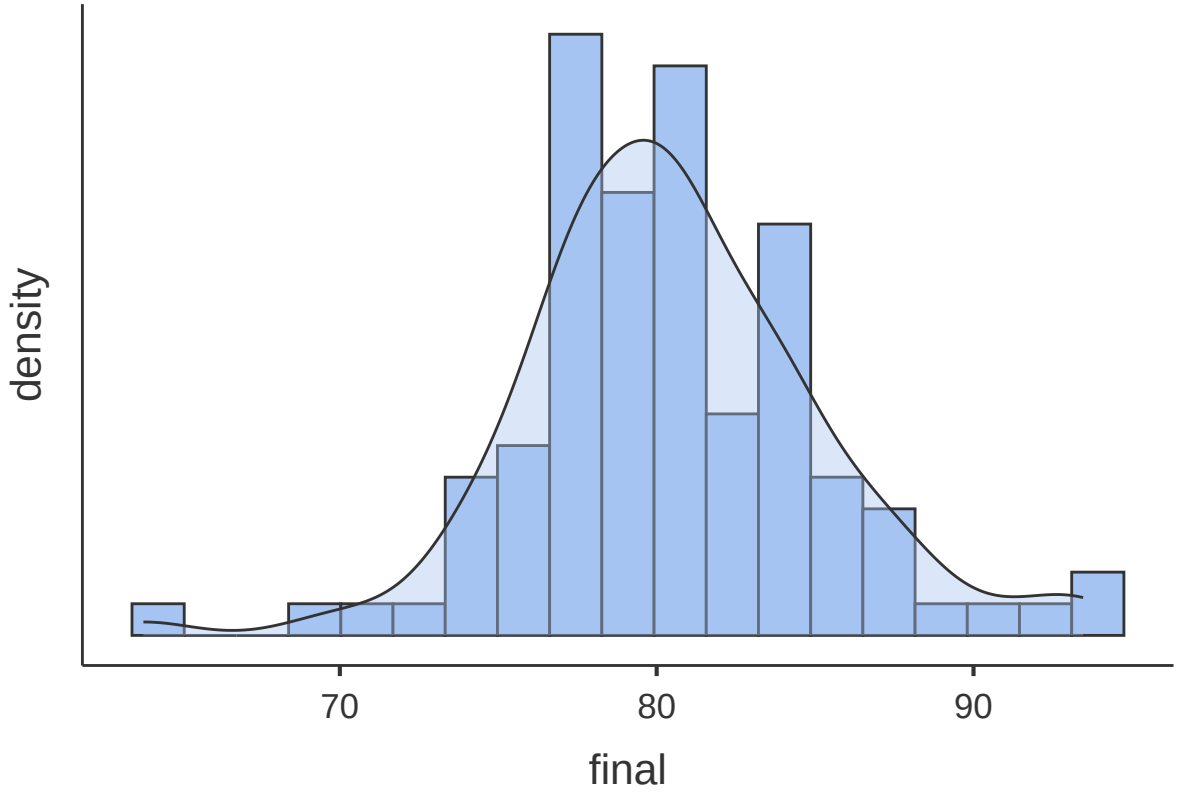
```
# Since we want to know if scores improved between the midterm and final we'll
# look at midterm scores, final scores, and the delta between the two.
# To do this, we need to add one more column to our math, the difference
# between the two. We need to maintain chirality (direction) as it will
# be informative if scores rise or decrease so we will do final - midterm
# without taking the absolute value. If scores decreased, we should see a
# positive mean of the diff. If they decreased we'll see a negative mean
# of the diff.
rq1_dat$fin_mid_diff <- rq1_dat$final - rq1_dat$midterm
```

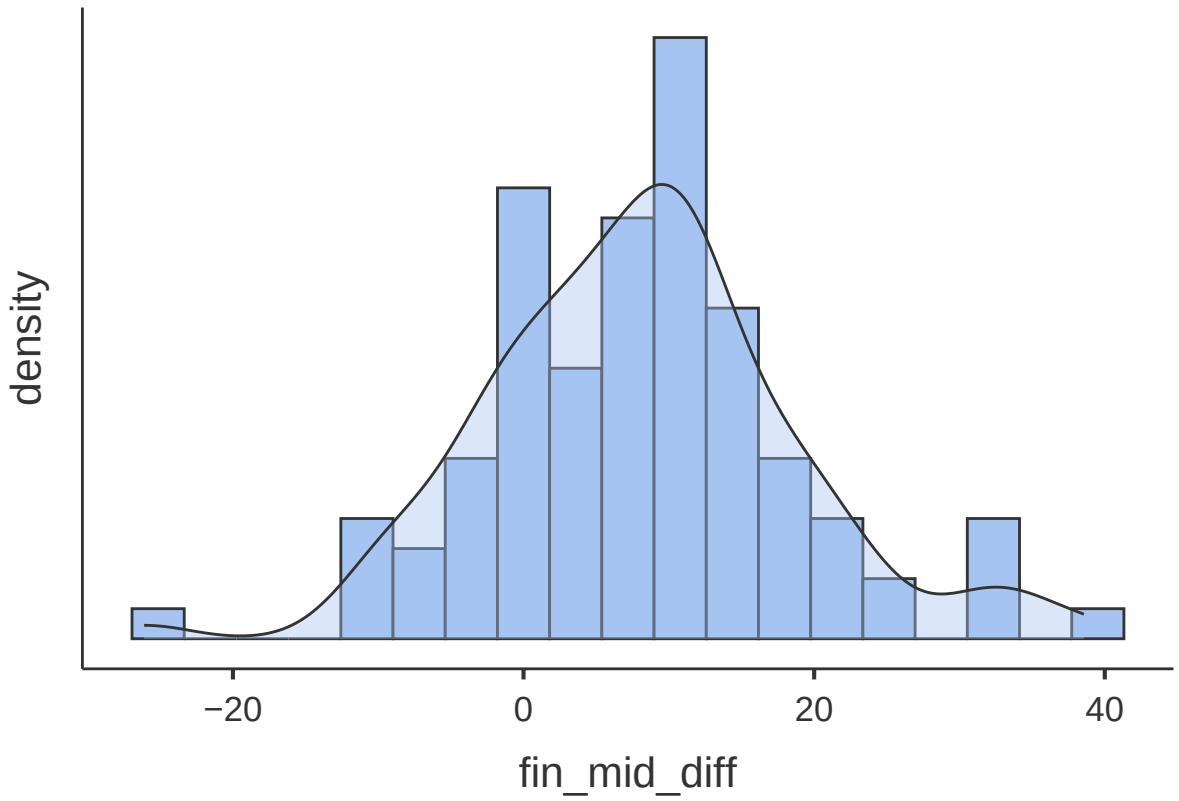
Visualize Prepared Data Q04 - Descriptives and Bar Chart See the histograms below for visualization of the data pertinent to this research question. Bar graphs of the midterm scores, the final scores, and the diff of final - midterm are also shown for better understanding

```
# We're going to create a descriptives object so we can checkout the histogram
# and descriptive stats of the midterm scores, the final scores, and the
# diff of both.
rq1_desc <- jmv::descriptives( rq1_dat[2:4],
                               hist = TRUE,
                               dens = TRUE,
                               sd = TRUE,
                               variance = TRUE,
                               se = TRUE,
                               skew = TRUE,
                               kurt = TRUE
                             )

# Render the rq1 descriptives object as output.
rq1_desc
```

```
##
##  DESCRIPTIVES
##
##  Descriptives
##
##                            midterm       final         fin_mid_diff
##
##    N                             100         100              100
##    Missing                         0           0                0
##    Mean                      72.39620    80.45750         8.061300
##    Std. error mean           1.060126   0.4742058         1.064727
##    Median                    72.16500    80.22500         8.545000
##    Standard deviation        10.60126    4.742058         10.64727
##    Variance                  112.3867    22.48712         113.3644
##    Minimum                   41.13000    63.80000        -26.11000
##    Maximum                   95.65000    93.46000         38.56000
##    Skewness                 -0.1713860  0.08208768        0.1736520
##    Std. error skewness       0.2413798   0.2413798        0.2413798
##    Kurtosis                  0.4235590    1.572569        0.9956757
##    Std. error kurtosis       0.4783311   0.4783311        0.4783311
##
```

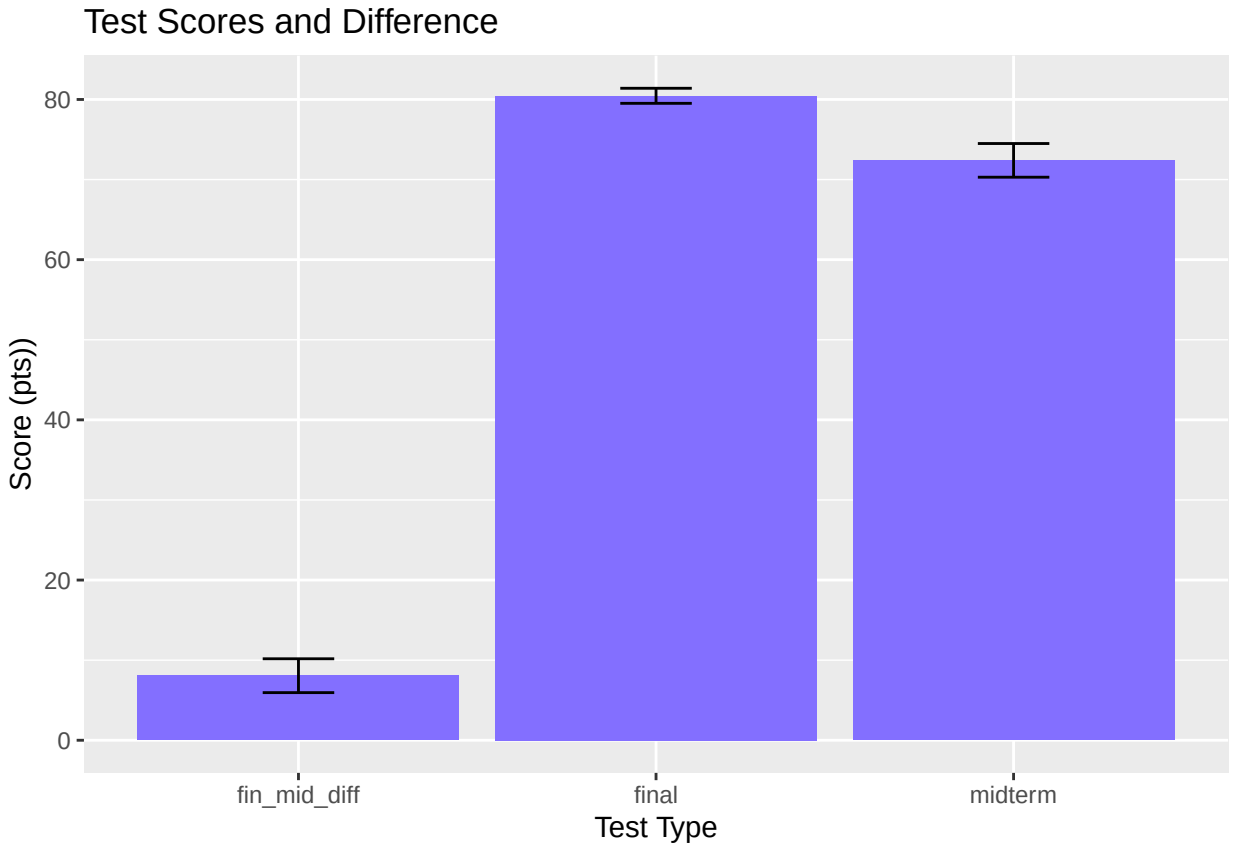

density
40
60
80
100
midterm


density
70
80
90
final




```
# Now we're going to create a bar chart of the midterm scores, final scores,
# and diffs .. though the diffs data willbe oddly scaled (much smaller) than the
# raw scores

# In order to plot the scores as bar charts, we need to permute the data so all
# scores are in one vector and there's n indy axis vector of the same size that
# defines whether the score is a midterm score, final score, or difference.
# We use pivot_longer to concatenate the three dcore vectors to each other
# (interleaved), with the column names saved in the new "name" vector.
rq1_long <- rq1_dat[2:4] %>%
            pivot_longer(cols = c(midterm, final, fin_mid_diff) )

# We're going to lay all three datasets out on a single bar graph
# even though that may jackup the scaling to be bad for the diff.
bar_rq1 <- ggplot( rq1_long, aes(name, value) )



#make a bar graph of relationship between fieldtrip and hard tangrams
# START HERE FIX BUGGY CODE BELOW
bar_rq1 + stat_summary( fun = mean,
                        geom = "bar",
                        position = "dodge",
                        fill="slateblue1"
                      ) +
          stat_summary( fun.data = mean_cl_normal,
                        geom = "errorbar",
                        position = position_dodge(width = 0.90),
```

```
                    width = 0.2
                  ) +
        labs(x = "Test Type", y = "Score (pts))") +
        ggtitle('Test Scores and Difference')
```



Discussion of Visualization for RQ1 It is notable that the error bar on the final scores is so much tighter than
that of the midterm scores. This is reflected in the final scores histogram which looks less "spread out" (smaller variance) than the same of the midterm scores. The larger spread in the differences of scores is also accounted for by the larger spread in the midterms as the tighter final scores minus the more spread out midterm scores will result in more spread out differences.

Question 05 (Q05):

5. Did student test scores improve significantly from the midterm to the final? Using RStudio to analyze, conduct a hypothesis test to evaluate this question. Organize your answer according to the 4 steps of hypothesis testing.

Answer to Q05: This is my answer

Question 06 (Q06):

6. Report your findings in APA format. (Hint: make sure to answer the research question!)

Answer to Q06: This is my answer

Research Question 2 (RQ2): Q07 - Q09:

Does the study technique used predict scores on the final exam?
Data: *308A.RQ2 Data.DA4.csv*

Question 07 (Q07):

7. Visualize your data for this research question. Include your visualization here.

Answer to Q07: This is my answer

Question 08 (Q08):

8. Does the study technique used predict scores on the final exam? Using RStudio to analyze, conduct a hypothesis test to evaluate this question. Organize your answer according to the 4 steps of hypothesis testing.

Answer to Q08: This is my answer

Question 09 (Q09):

9. ...
    a. Report your findings in APA format. (Hint: make sure to answer the research question!)
    b. The Dean of the university was also interested in your results, as this may help to raise scores in other departments. Unfortunately, she does not understand statistical language. Please interpret your findings for the Dean. Did scores improve? Which technique is better?

Answer to Q09: This is my answer

Question 10 (Q10):

10. A developmental psychologist is interested in the effect of a positive psychology intervention on the well-being of aging adults. She administers the intervention, collects well-being scores from a sample of 100 participants, and tests whether their well-being differs significantly from

the national average. Using G*Power, she determines the power for her test is .80.

a. Interpret this value

b. What suggestion would you give her if she wants a higher probability of detecting a true effect?

Answer to Q10:

This is my answer

Question 11 (Q11):

11. What would it mean if your analysis returned the following values? Consider the meaning of t - not the decision associated with it.

    a. *t(24) = 0.35*

    b. *t(24) = 1.00*

    c. *t(24) = 3.2*

Answer to Q11:

This is my answer

Question 12 (Q12):

12. Draw and annotate all the properties of the null and alternative curves: power, beta, alpha, type 1 error, type 2 error.

Answer to Q12:

This is my answer