Answer Sketch for Homework H

431 Staff and Professor Love

*Due **2019-11-08** at 2 PM. Last Edited 2019-10-21 23:45:50*

Contents


```
knitr::opts_chunk$set(comment=NA)
options(width = 70)
```

Load necessary packages

Questions 1-4

We don't provide answer sketches for essay Questions, like Questions 1-4.

R Setup for Questions 5-10

```
library(here); library(janitor); library(magrittr);
library(broom); library(patchwork)
library(tidyverse)
```

```
hwH_data1 <- read.csv(here("data", "hwH_data1.csv")) %>%
  tbl_df
hwH_data2 <- read.csv(here("data", "hwH_data2.csv")) %>%
  tbl_df
```

Question 5

The same data appear in the `hwH_data1.csv` and the `hwH_data2.csv` files. What is the difference between the two files, and which of the two files is more useful for fitting an ANOVA to compare the PDS means across the three groups of study participants? Why?

Answer 5

To calculate the PDS means for each subject, we want the data in a form with one row per subject. The `hWH_data1` file presents the `age` information in a wider form than the `hwH_data2` file. Thus, `hwH_data2` has twice as many rows as `hwH_data1`, and has two rows for each subject. So we want to use `hwH_data1` in this case to calculate the PDS scores for each subject.

Question 6

Calculate and compare the sample PDS means across the three groups, and specify the rank order (highest to lowest) of the sampled PDS means.

Answer 6

```
hwH_data1 <- hwH_data1 %>%
  mutate(PDS = (subj_age - age)/age)

hwH_data1 %>% group_by(category) %>%
  summarize(mean_PDS = mean(PDS)) %>%
  arrange(desc(mean_PDS))
```

```
# A tibble: 3 x 2
  category mean_PDS
  <fct>       <dbl>
1 GROUP_B   0.00451
2 GROUP_C   0.00429
3 GROUP_A   0.00173
```

Group B has the largest sample mean (0.0045), then C (0.0043), then A (0.0017).

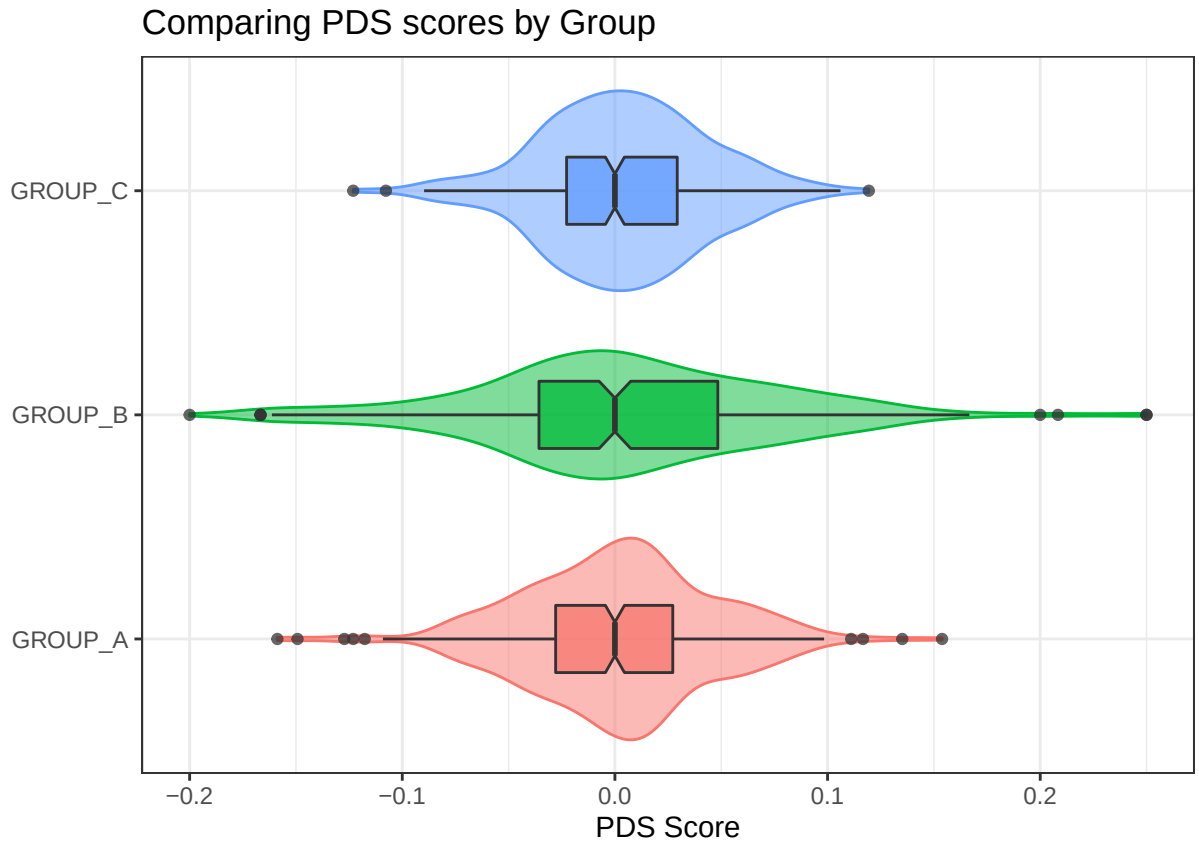
Question 7

Produce a graphical summary to compare the three groups that allows you to assess the Normality and Equal Variances assumptions of an ANOVA to compare the PDS means across the three groups. What conclusion do you draw about ANOVA assumptions in this setting?

Answer 7.

One good option was a boxplot, perhaps with violins.

```
ggplot(hwH_data1, aes(x = category, y = PDS, fill = category)) +
  geom_violin(aes(color = category), alpha = 0.5) +
  geom_boxplot(width = 0.3, notch = TRUE, alpha = 0.75) +
  theme_bw() +
  coord_flip() +
  guides(fill = FALSE, col = FALSE) +
  labs(title = "Comparing PDS scores by Group",
       y = "PDS Score", x = "")
```



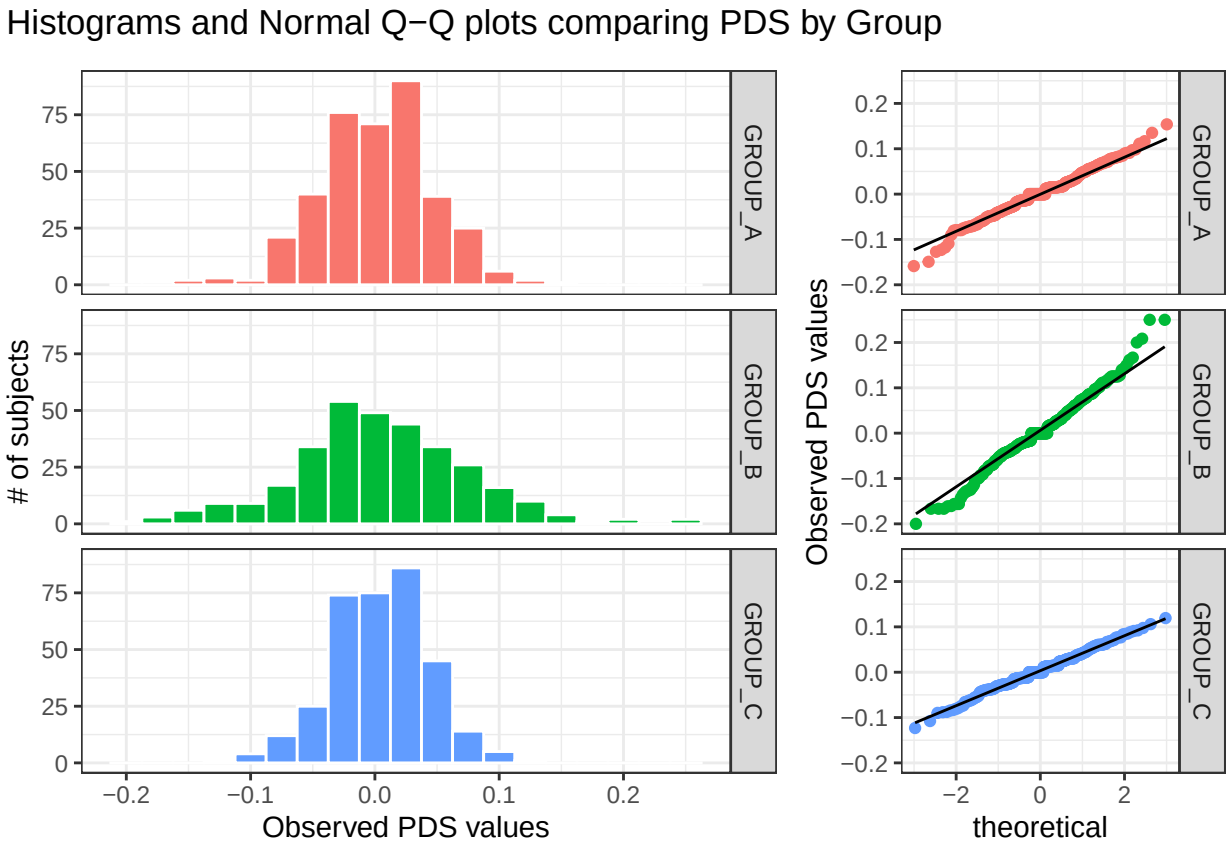
Another reasonable option would be a set of faceted histograms, perhaps next to some Normal Q-Q plots.

```
res <- mosaic::favstats(PDS ~ category, data = hwH_data1)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

```
bin_w = 0.025

p1 <- ggplot(hwH_data1, aes(x = PDS, fill = category)) +
  geom_histogram(binwidth = bin_w, col = "white") +
  guides(fill = FALSE) +
  theme_bw() +
  facet_grid(category ~ .) +
```

```
  labs(x = "Observed PDS values", y = "# of subjects")

p2 <- ggplot(hwH_data1, aes(sample = PDS, color = category)) +
  geom_qq() + geom_qq_line(col = "black") +
  guides(color = FALSE) +
  theme_bw() +
  facet_grid(category ~ .) +
  labs(y = "Observed PDS values")

p1 + p2 +
  plot_layout(nrow = 1, widths = c(7, 3)) +
  plot_annotation(title = "Histograms and Normal Q-Q plots comparing PDS by Group")
```



Here is a numerical summary, as well.

```
mosaic::favstats(PDS ~ category, data = hwH_data1)
```

```
  category        min          Q1 median         Q3       max
1  GROUP_A -0.1587302 -0.02788352      0 0.02721461 0.1538462
2  GROUP_B -0.2000000 -0.03571429      0 0.04838710 0.2500000
3  GROUP_C -0.1230769 -0.02272727      0 0.02930520 0.1194030
         mean         sd   n missing
1 0.001733341 0.04460588 378       0
2 0.004505656 0.07069228 321       0
3 0.004292067 0.03861400 342       0
```

Main conclusions: No apparent problems with the Normality assumption. Some indication of larger spread in

Group B than the other two groups, and Group B also has a somewhat smaller sample size than the other groups. ANOVA is pretty robust to problems with the equal variances assumption, so we are probably OK.

Question 8

Now do the actual comparison of the PDS means of the three groups (A, B and C) using an analysis of variance. What conclusion do you draw, using a **90%** confidence level?

Answer 8.

```
summary(aov(PDS ~ category, data = hwH_data1))
```

```
             Df Sum Sq   Mean Sq F value Pr(>F)
category      2 0.0017 0.0008568   0.311  0.733
Residuals  1038 2.8577 0.0027531
```

The ANOVA F test finds no statistically detectable differences between group means, as the p value far exceeds the required significance level of $\alpha = 0.10$.

$$\eta^2 = \frac{SS(category)}{SS(Total)} = \frac{0.0017}{0.0017 + 2.8577} = 0.00059$$

The group (category) accounts for about 0.06% of the variation in the PDS values.

Question 9

This is a pre-planned comparison, but the sample sizes differ across the groups being compared. Obtain the results from a Tukey HSD method and then a Bonferroni approach for pairwise comparisons of the population PDS means, in each case again using a 90% confidence level[^2]. Do your conclusions differ using these two approaches?

Answer 9.

Given the results from the ANOVA F test, neither of these results should show statistically detectable differences, and as we'll see, neither one does.

Tukey HSD Approach

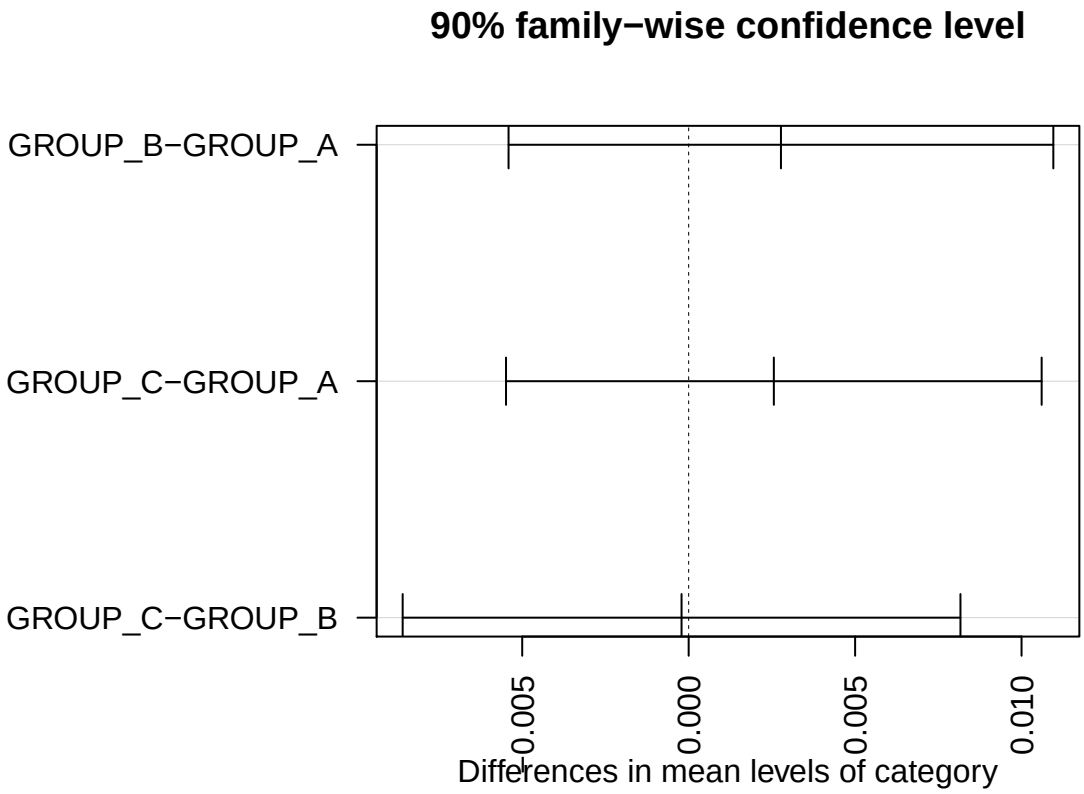
```
hwH_data1 %$% TukeyHSD(aov(PDS ~ category), conf.level = 0.90)
```

```
  Tukey multiple comparisons of means
    90% family-wise confidence level

Fit: aov(formula = PDS ~ category)

$category
                       diff          lwr        upr     p adj
GROUP_B-GROUP_A  0.002772314 -0.005409924 0.01095455 0.7658017
```

```
GROUP_C-GROUP_A  0.002558726 -0.005486518 0.01060397 0.7904578
GROUP_C-GROUP_B -0.000213588 -0.008591256 0.00816408 0.9984884
```

The confidence intervals each easily cover zero, as we can also see from the plot, below.

```
mar.default <- c(5,6,4,2) + 0.1 # save default plotting margins

par(mar = mar.default + c(0, 6, 0, 0))
hwH_data1 %$% plot(TukeyHSD(aov(PDS ~ category),
                            conf.level = 0.90), las = 2)
```




```
par(mar = mar.default) # return to normal plotting margins
```

Bonferroni Approach

```
hwH_data1 %$% pairwise.t.test(PDS, category,
                              p.adjust = "bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  PDS and category

        GROUP_A GROUP_B
GROUP_B 1       -
GROUP_C 1       1
```

```
P value adjustment method: bonferroni
```

Question 10

Specify the linear model regression equation used to predict our PDS outcome on the basis of group membership, but now also adjusting for whether or not the subject is `active`. What fraction of the variation in PDS levels is explained by this model? How much more of that variation is explained than by the model including group membership alone? How do you know?

Answer 10.

Here is a good choice of model...

```
m_10 <- lm(PDS ~ category + active, data = hwH_data1)

summary(m_10)
```

```

Call:
lm(formula = PDS ~ category + active, data = hwH_data1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.212583 -0.028646  0.001742  0.029077  0.237417

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.010185   0.003293   3.093  0.00203 **
categoryGROUP_B  0.002399   0.003949   0.607  0.54367
categoryGROUP_C  0.001731   0.003886   0.445  0.65611
active          -0.014326   0.003254  -4.403 1.18e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05201 on 1037 degrees of freedom
Multiple R-squared:  0.01894,   Adjusted R-squared:  0.0161
F-statistic: 6.673 on 3 and 1037 DF,  p-value: 0.0001835
```

This model accounts for 1.89% of the variation in `PDS`, according to the R^2 value.

We can compare this to the model without the `active` information, as follows:

```
m_08 <- lm(PDS ~ category, data = hwH_data1)

summary(m_08)
```

```

Call:
lm(formula = PDS ~ category, data = hwH_data1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.204506 -0.030719 -0.001733  0.026838  0.245494
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.001733   0.002699   0.642    0.521
categoryGROUP_B 0.002772   0.003982   0.696    0.487
categoryGROUP_C 0.002559   0.003916   0.653    0.514

Residual standard error: 0.05247 on 1038 degrees of freedom
Multiple R-squared:  0.0005993, Adjusted R-squared:  -0.001326
F-statistic: 0.3112 on 2 and 1038 DF,  p-value: 0.7326
```

Of course, the multiple R^2 for Model `m_08` is just the η^2 from our ANOVA comparison in Question 8. Again, we see that model accounts for less than 0.06% of the variation in PDS.

So the additional impact of `active` (even after Group is already in the model) is substantially larger than the impact of Group alone, even though Model `m_10` isn't strong, either.

Grading Rubric

The grading rubric will be prepared by the teaching assistants, and will be available when grades are posted. Each of the ten questions is worth 10 points.