Answer Sketch for Homework 7

431 Staff and Professor Love

'Due 2018-11-16, version 2018-11-05

Contents

Initial R Setup

Here's the R setup we used.

```
knitr::opts_chunk$set(comment=NA)

library(tidyverse)
```

We'll read in each of the data sets, in case we need them.

```
coexpose1 <- read.csv("coexpose1.csv") %>% tbl_df
coexpose2 <- read.csv("coexpose2.csv") %>% tbl_df
```

1 Question 1

The same data appear in the `coexpose1.csv` and the `coexpose2.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 FEV_1 means across the three medical centers?

1.1 Answer for Question 1

```
glimpse(coexpose1)
```

```
Observations: 26
Variables: 3
$ johns.hopkins     <dbl> 3.23, 3.47, 1.86, 2.47, 3.01, 1.69, 2.10, 2....
$ rancho.los.amigos <dbl> 3.22, 2.88, 1.71, 2.89, 3.77, 3.29, 3.39, 3....
$ st.louis          <dbl> 2.79, 3.22, 2.25, 2.98, 2.47, 2.77, 2.95, 3....
```

```
glimpse(coexpose2)
```

```
Observations: 60
Variables: 3
$ pt.id  <int> 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, ...
$ fev1   <dbl> 3.23, 3.47, 1.86, 2.47, 3.01, 1.69, 2.10, 2.81, 3.28, 3...
$ center <fct> johns.hopkins, johns.hopkins, johns.hopkins, johns.hopk...
```

The `coexpose1` file contains 26 rows and 3 columns, labeled by the names of the three centers. Each column contains the response (FEV_1) for the subjects at that center. This is what is called data in the **wide** format, and is most appropriate when planning a matched samples analysis.

The `coexpose2` file contains 60 rows and 3 columns, labeled `pt.id`, `fev1` and `center`. We have each patient's ID, their FEV1 value, and their center, laid out in **long** format, which is most appropriate for an independent samples analysis. These are tidy data.

The ANOVA expects us to have a variable for our outcome (`fev1`) and the treatment/group identifier (`center` in this case), so the `coexpose2` data will be more useful for our purposes.

2 Question 2

Produce a numerical summary to compare the means across the three centers, and specify the rank order (highest to lowest) of the sampled FEV_1 levels.

2.1 Answer for Question 2

There are several ways to accomplish this, but I expect most of you used `favstats` from the `mosaic` package ...

```
mosaic::favstats(fev1 ~ center, data = coexpose2)
```

```
             center  min   Q1 median    Q3  max     mean        sd  n
1     johns.hopkins 1.69 2.47   2.61 2.910 3.47 2.626190 0.4961701 21
2 rancho.los.amigos 1.71 2.71   3.03 3.390 3.86 3.032500 0.5232399 16
3          st.louis 1.98 2.55   2.85 3.185 4.06 2.878696 0.4977157 23
  missing
1       0
2       0
3       0
```

By the mean (or, in fact, the median) `fev1` value in each `center`, we'd rank `rancho.los.amigos` highest, followed by `st. louis` and finally `johns.hopkins`, with the smallest `fev1` mean/median.

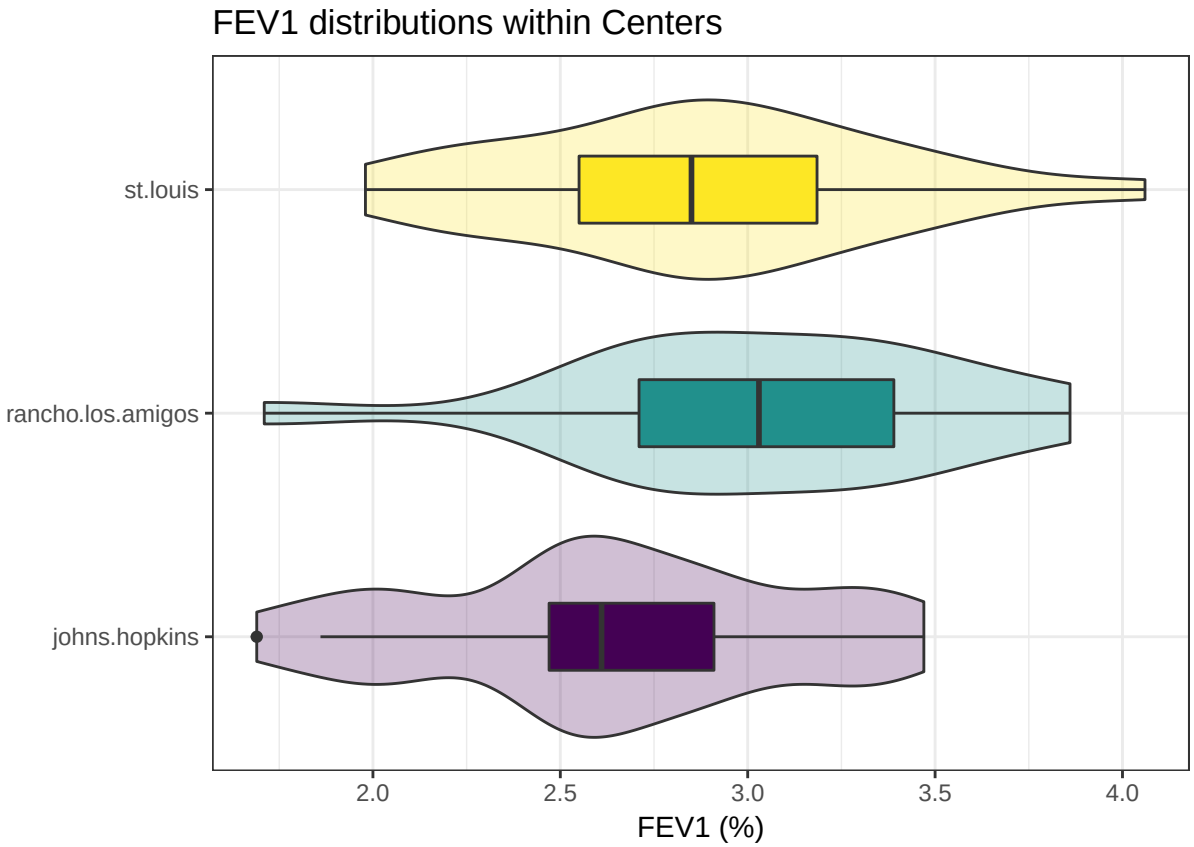
3 Question 3

Produce a graphical summary to compare the three centers that allows you to assess the Normality and Equal Variances assumptions of an ANOVA to compare the FEV_1 means across the three medical centers. What conclusion do you draw about the assumptions in this setting? Then do the actual comparison of the FEV_1 means of the three different medical centers using an analysis of variance. What conclusion do you draw, using a **90%** confidence level?

3.1 Answer for Question 3

I would be likely to use a combination boxplot and violin plot, although there are certainly other alternatives. There is one identified outlier in the Johns Hopkins data, but I see no serious concerns with either the Normality or the Constant Variance assumption.

```
ggplot(coexpose2, aes(y = fev1, x = center, fill = center)) +
  geom_violin(alpha = 0.25) +
  geom_boxplot(width = 0.3) +
  coord_flip() +
  labs(title = "FEV1 distributions within Centers",
       y = "FEV1 (%)",
       x = "") +
  guides(fill = FALSE) +
  scale_fill_viridis_d() +
  theme_bw()
```



Here's the ANOVA table, comparing the mean FEV_1 levels, by Center.

```
anova(lm(fev1 ~ center, data = coexpose2))
```

```
Analysis of Variance Table

Response: fev1
          Df  Sum Sq Mean Sq F value Pr(>F)
center     2  1.5828 0.79142  3.1153  0.052 .
Residuals 57 14.4803 0.25404
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since $p < 0.10$, at the 90% confidence level, we conclude that there is a statistically detectable difference in the means at the three centers.

4 Question 4

Specify the linear model regression equation used to predict our FEV_1 outcome on the basis of medical center. What fraction of the variation in FEV_1 levels is explained by the medical center?

4.1 Answer for Question 4

As we can see in the coefficients column of the output below, the equation is `fev1` = 2.63 + 0.41 (`rancho.los.amigos`) + 0.25 (`st.louis`).

```
summary(lm(fev1 ~ center, data = coexpose2))
```

```
Call:
lm(formula = fev1 ~ center, data = coexpose2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32250 -0.32250 -0.02244  0.32630  1.18130

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.6262     0.1100  23.877   <2e-16 ***
centerrancho.los.amigos  0.4063     0.1673   2.429   0.0183 *
centerst.louis           0.2525     0.1521   1.660   0.1024
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.504 on 57 degrees of freedom
Multiple R-squared:  0.09854,   Adjusted R-squared:  0.06691
F-statistic: 3.115 on 2 and 57 DF,  p-value: 0.052
```

The Johns Hopkins center patients are used as the baseline category in this case. Rancho Los Amigos patients had `fev1` values (on average) 0.41 points higher than did the Hopkins patients. St. Louis patients were (on average) 0.25 points higher than Hopkins.

The multiple $R^2 = 0.099$, alsofrom the output above, so 9.9% of the variation in `fev1` is accounted for by the linear model using `center`.

5 Question 5

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 FEV_1 means, in each case again using a 90% confidence level. Do your conclusions differ?

5.1 Answer for Question 5

Using an α of 0.10, the only significant difference we see in either the Bonferroni or the Tukey results is that Rancho Los Amigos has a statistically significantly larger mean than does Johns Hopkins.

```
pairwise.t.test(coexpose2$fev1, coexpose2$center,
                p.adjust.method="bonferroni")
```

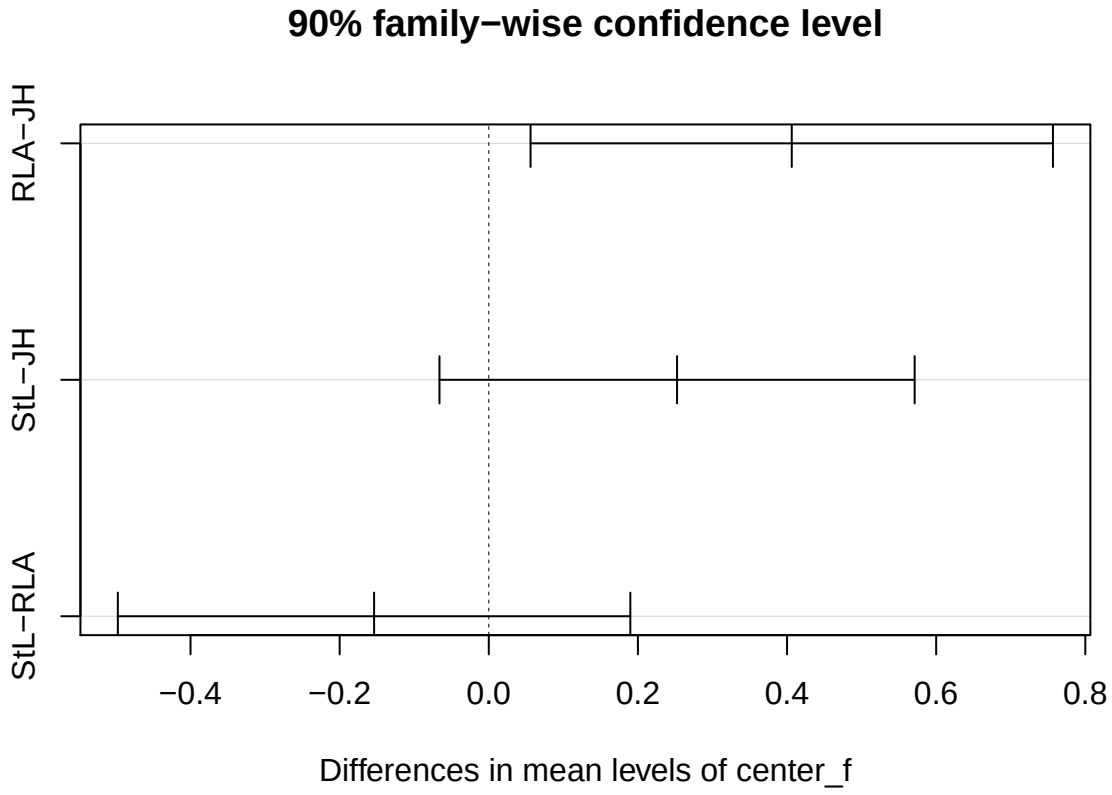
```
    Pairwise comparisons using t tests with pooled SD

data:  coexpose2$fev1 and coexpose2$center

                 johns.hopkins rancho.los.amigos
rancho.los.amigos 0.055         -
st.louis          0.307         1.000

P value adjustment method: bonferroni
```

```
TukeyHSD(aov(coexpose2$fev1 ~ coexpose2$center),
         conf.level=.90)
```

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

Fit: aov(formula = coexpose2$fev1 ~ coexpose2$center)

$`coexpose2$center`
                                      diff         lwr       upr     p adj
rancho.los.amigos-johns.hopkins  0.4063095  0.05601022 0.7566088 0.0473852
st.louis-johns.hopkins           0.2525052 -0.06610587 0.5711162 0.2294901
st.louis-rancho.los.amigos      -0.1538043 -0.49745443 0.1898457 0.6191128
```

We might also want to plot the confidence intervals from the Tukey procedure, but to do that, I will first create a new `center_f` factor that compresses the names of the centers, so they can fit more easily in the plot.

```
coexpose2 <- coexpose2 %>%
  mutate(center_f = fct_recode(center,
           "RLA" = "rancho.los.amigos",
           "JH" = "johns.hopkins",
           "StL" = "st.louis"))

plot(TukeyHSD(aov(fev1 ~ center_f, data = coexpose2),
              conf.level = 0.90))
```



6 Questions 6-9

6. Find an example of a visualization designed to support a comparison of at least two population means or medians using either paired or independent samples in a published work (online or not) for which you can find the complete sourcing information, and which was built no earlier than January 1, 2013. Provide the complete reference and a copy of the image itself (including any captions or titles) and surrounding material for the visualization.

7. In a few sentences, describe the purpose of the comparison being made in your example from Question 6. Explain its context and why it is important. Specify the research question that this comparison (and the accompanying p value or confidence interval based inference, if available) is providing to the reader.

8. In a few sentences, describe the visualization that you found which relates to the comparison being made in your example from Question 6. Explain what you believe the visualization is trying to do. Specify why it is or is not effective, in your view.

9. Provide your best suggestion as to how either the visualization or the comparison that you found in Question 6 might be improved, and explain why your change (or changes) would be an improvement.

6.1 We don't provide answer sketches for essay Questions, like 6-9