# HW 02: Analysis of Variance

due Wed, Feb 12 at 11:59p

## Getting Started

A brief outline of getting started is shown below. See the Lab 01 Instructions (https://www2.stat.duke.edu/courses/Spring20/sta210.001/labs/lab-01-review-r.html) for more details about the steps.

- Go to the sta210-sp20 organization on GitHub (https://github.com/sta210-sp20 (https://github.com/sta210-sp20)). Click on the repo with the prefix **hw-02**. It contains the starter documents you need to complete the lab.
- Clone the repo in RStudio Cloud

## Tips

Here are some tips as you complete HW 02:

- Periodically knit your document and commit changes (the more often the better, for example, once per each new task)
- Push all your changes back to your GitHub repo. The Git pane in RStudio should be empty after you push. You can also check your assignment repo on github.com to make sure it has updated.
- Once you have completed your work, you will submit your assignment in Gradescope. You are welcome to resubmit multiple times until the assignment deadline. We will grade the most recent version of the .pdf file in Gradescope.

## Packages

We will use the following packages in this assignment:

```
library(tidyverse)
library(broom)
library(knitr)
```

If you need to install any of the packages, type `install.packages("package_name")` in the console, where `package_name` is the package you need to install.
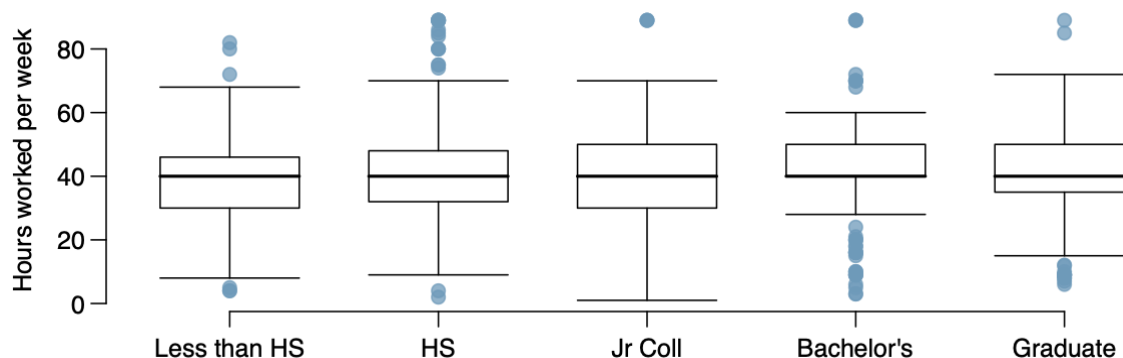
# Questions

## Part 1: Conceptual

The *Conceptual* section of homework contains short answer questions about the concepts discussed in class. Some of these questions may also require short chunks of code to produce the output needed to answer the question. Answers should be written in complete sentences.

The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

| | Less than High School | High School | Jr. College | Bachelor's | Graduate | Total |
|---|---|---|---|---|---|---|
| Mean | 38.67 | 39.6 | 41.39 | 42.55 | 40.85 | 40.45 |
| Standard deviation | 15.81 | 14.97 | 18.1 | 13.62 | 15.51 | 15.17 |
| n | 121 | 546 | 97 | 253 | 155 | 1172 |



We wish to test the following hypotheses:

$$H_0 : \mu_{<HS} = \mu_{HS} = \mu_{JC} = \mu_B = \mu_G$$
$$H_a : \text{ at least one } \mu_i \text{ is not equal to the others}$$

**Question 1.** Part of the ANOVA table associated with this test is shown below. Use the code to fill in the missing values and display the completed table.

| | Df | Sum Sq | Mean Sq | F-Stat | Pr(F > F-Stat) |
|---|---|---|---|---|---|
| degree | * | * | 501.54 | * | 0.0682 |
| Residuals | * | 267382 | * | * | |
| Total | * | * | | | |

```
ssw <- 267382 # sum of squares within (residuals)
msb <- 501.54 # mean square between (model)
p_val <- 0.0682
```

```
dfb <- _____  # degrees of freedom between (model)
dfw <- _____  # degrees of freedom within (residuals)
dft <- _____  # total degrees of freedom
ssb <- _____  # sum of squares between (model)
sst <- _____  # total sum of squares
msw <- _____   # mean square within (residuals)
f_stat <- _____   #F -statistic
```

```
source <- c("Degree (model)", "Residuals", "Total")
df <- c(dfb, dfw,dft)
ss <- c(ssb, ssw, sst)
ms <- c(msb, msw,NA)
f.statistic <- c(f_stat, NA, NA)
p.value <- c(p_val,NA,NA)

# combine the columns to make a table called "anova"
anova <- bind_cols("Source" = source, "df" = df, "Sum of squares" = s
s,
          "Mean square" = ms, "F-statistic" = f.statistic, "p-valu
e" = p.value)

# print the table
kable(anova)
```

**Question 2.** Check the assumptions for this test. Include your conclusion about whether the assumption is satisfied and a brief explanation supporting your conclusion.

**Question 3.** State the conclusion of the test in the context of the data. You can use a significance level of $\alpha = 0.05$ to make your conclusion.

# Part 2: Data Analysis

The *Data Analysis* section of homework contains open-ended data analysis questions. Your response should be neatly organized and read as a complete narrative. This means that in addition to addressing the question(s) stated below, you should include exploratory data analysis and check the appropriate model assumptions. In short, these questions should be treated as "mini-projects".

In a 1991 study, Allen et. al sought to answer whether the presence of a close friend or pet affected women's stress levels as they completed challenging tasks. To test this, they conducted an experiment in which 45 women were tasked with counting backwards by 13s or 17s under one of three test conditions ( group ): - C : Control group, Alone - F : Close friend present - P : Pet present

To quantify stress level, they measured each woman's heart rate and blood pressure after she completed the task. For today's analysis, we will focus on the heart rate ( heart_rate ).

Use Analysis of Variance to test whether there is an association between the presence of a friend or pet and stress level when completing challenging tasks. Your analysis should include

- Complete exploratory data analysis
- Code and output for the Analysis of Variance table
- Check of assumptions including all supporting graphs, code, output, narrative, etc.
- *If needed based on results from ANOVA*: Use a model with `group` as the predictor to compare the predicted mean `heart_rate` for each group and discuss which groups (if any) have a significantly different mean from one another.

The data is available in `stress-experiment.csv` in the `data` folder.

# Submitting Your Assignment

Once your work is finalized in your GitHub repo, you will submit it to Gradescope. **Your assignment must be submitted on Gradescope by the deadline to be considered "on time".**

To submit your assignment:

- Go to http://www.gradescope.com (http://www.gradescope.com) and click *Log in* in the top right corner.

- Click *School Credentials* ➡ *Duke NetID* and log in using your NetID credentials.

- Click on the *STA 210 Regression Analysis* course.

- Click on the assignment, and you'll be prompted to submit it.

    - If asked, login to your GitHub account.
    - If asked, click to request access to the "sta210-sp20" GitHub organization.
- Select your assignment repo and choose "master" for the branch.

- Click *Upload*. You should receive an email to confirm that the assignment has been submitted.

# Grading

| Total | 50 |
| --- | --- |
| Part 1: Conceptual problems | 18 |
| Part 2: Data analysis | 20 |
| rstudio::conf assignment | 10 |
| At least 3 informative commit messages | 2 |

*Note there is a 20% penalty if the .pdf file is incomplete and the .Rmd file has to be knitted for grading.*

## Acknowledgements & References

- Questions in Part 1 are from Ex. 7.42 in *OpenIntro Statistics, 4th ed.* (https://leanpub.com/openintro-statistics). Some parts have been reworded.

- National Opinion Research Center, General Social Survey, 2018. (http://gss.norc.org/get-the-data)

- Allen, K., Blascovich, J., Tomaka, J., and Kelsey, R., 1991, Presence of Human Friends and Pet Dogs as Moderators of Autonomic Responses to Stress in Women, *Journal of Personality and Social Psychology*, 61: 582-589. (https://search-ebscohost-com.proxy.lib.duke.edu/login.aspx?direct=true&db=pdh&AN=1992-04617-001&site=ehost-live&scope=site)