Lab 04: Analysis of Variance

due Tue, Feb 11 at 11:59p

The goal of this lab is to use Analysis of Variance (ANOVA) to understand the variation in price of diamonds that are 0.5 carats. Additionally, you will be introduced to new R fucntion used for wrangling and summarizing data.

Getting Started

- Go to the sta210-sp20 organization on GitHub (https://github.com/sta210-sp20). Click on the repo with the prefix **lab-04-anova-**. It contains the starter documents you need to complete the warmup exercise.

- Clone the repo and create a new project in RStudio Cloud.

- See Lab 01 for full instructions on getting started.

Packages

We will use the following packages in today's lab.

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

Data

In today's lab, we will analyze the `diamonds` dataset from the ggplot2 package . Type `?diamonds` in the console to see a dictionary of the variables in the data set. The primary focus of this analysis will be examining the relationship between a diamond's cut and price.

Before starting the exercises, take a moment to read more about the diamond attributes on the Gemological Institute of America webpage: https://www.gia.edu/diamond-quality-factor.

Exercises

The `diamonds` dataset contains the price and other characteristics for over 50,000 diamonds. For this analysis, we will only consider diamonds that have a carat weight of 0.5.

Exploratory data analysis

Exercise 1. Create a new data frame that is a subset of diamonds that weigh 0.5 carats. How many observations are in the new dataset?

You will use this subset for the remainder of lab.

Exercise 2. When using Analysis of Variance (ANOVA) to compare group means, it is ideal to have approximately the same number of observations for each group.

- Which two levels of `cut` have the fewest number of observations? Show the code and output used to support your answer. ⊕
- Recode the variable `cut`, so that the two levels with the fewest number of observations are combined into one level. Be sure to give the new level an informative name and save the results to the data frame. ***You will use the recoded version of `cut` for the remainder of the lab.***

Exercise 3. Confirm that the variable `cut` was recoded as expected. Show the code and output used to check the recoding.

Exercise 4. Create a plot to display the relationship between `cut` and `price`. Be sure to include informative axes labels and an informative title.

Exercise 5. Calculate the number of observations along with the mean and standard deviation of `price` for each level of `cut`.

Exercise 6. Based on the plots and summary statistics from the previous exercises , does there appear to be a relationship between the cut and price for diamonds that are 0.5 carats? Briefly explain your reasoning.

Analysis of Variance

Exercise 7. When using ANOVA to compare means across groups, we make the following assumptions (note how similar they are to the assumptions for regression):

- **Normality**: The distribution of the response, y, is approximately normal within each category of the predictor, x - in the i^{th} category, the y's follow a $N(\mu_i, \sigma^2)$ distribution.
- **Independence**: All observations are independent from one another, i.e. one observation does not affect another.
- **Constant Variance**: The distribution of the response within each category of predictor, x has a common variance, σ^2.

Are the assumptions for ANOVA satisfied? Comment on each assumption, including an explanation for your reasoning and any summary statistics and/or plots used to make the conclusion.

Exercise 8. Display the ANOVA table used to examine the relationship between `cut` and `price` for diamonds that are 0.5 carats.

Exercise 9. Use the ANOVA table from the previous question to calculate the sample variance of `price`. Show the code / formula used to calculate the sample variance.

Exercise 10. What is $\hat{\sigma}^2$, the estimated variance of `price` within each level of `cut`.

Exercise 11. State the null and alternative hypotheses for the test conducted using the ANOVA table in Exercise 8. State the hypotheses using both statistical notation and words in the context of the data.

Exercise 12. What is your conclusion for the test specified in the previous question? State the conclusion in the context of the data.

Additional Analysis

Exercise 13. Based on the conclusion of the ANOVA test, conduct further statistical analysis to provide more detail about which level(s) is(are) different and by how much. If further statistical analysis is not required, provide a brief explanation why it isn't based on the conclusion from the ANOVA test.

You're done and ready to submit your work! Knit, commit, and push all remaining changes. You can use the commit message "Done with Lab 4!", and make sure you have pushed all the files to GitHub (your Git pane in RStudio should be empty) and that all documents are updated in your repo on GitHub. Then submit the assignment on Gradescope following the instructions below.

Submitting the 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 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.
- **Make sure to include the names of all group members who participated in the assignment.** Click here for help on adding group members to an assignment.
- Click *Upload*. You should receive an email to confirm that the assignment has

been submitted.

Notes:

- You can see what has been submitted by click “Code” at the top of the page. **We will be grading the PDF file**, so please make sure that has all of your final code, output, and narrative.
- You are welcome to resubmit as many times as you’d like before the deadline (we will only grade the most recent version). Just click the “Resubmit” button at the bottom of the page and reselect your repo and master branch.

Grading

Exploratory Data Analysis	15
Analysis of Variance	21
Additional Analysis	4
Merge conflict exercise	3
Lab attendance & participation	3
Narrative in full sentences & document neatly organized	2
Commit messages from every member	2
Total	**50**