FORE224/STAT202 Assignment 6

One-Way analysis of variance (ANOVA) and post-hoc tests

Due at 12 noon, 29 September 2021

[1] As usual, create a new project in its own folder for this assignment.

Download the file **job_satisfaction1.csv** from Learn and save it in your project folder. The file contains data for 106 cases. Two variables are recorded:

- the categorical variable education_level, for the case's highest education level, with values "school", "college", "university" (it is a US dataset; college is roughly between school and university)
- the continuous variable score, which is the case's job satisfaction score.

[2] Load the `tidyverse` library.

Read the dataset from job_satisfaction1.csv and make sure that the education_level variable is treated as a factor with levels "school", "college", "university" where "school" is the baseline group. You can do this by using `read_csv` to read in the data and then using `mutate` with the following code to change the variable education_level to a factor with the required levels:

```
education_level <- factor(education_level,
  levels = c("school", "college", "university")
)
```

[*Optional: as we show in lecture 22, you can read in the data and make education_level a factor with the required levels all at the same time, with code like this:*

```
job_satisfaction1 <- read_csv("./rawdata/job_satisfaction1.csv",
  col_types = cols(education_level = col_factor(c("school", "college",
"university")))
)
```

]

Use `set.seed` to set the random seed to your student ID number and make a subset called my_js of 104 cases from the job satisfaction data using `sample_n` as in previous labs.

[3] Use R's `summarise` function (also works as `summarize`) to get a summary of the number of cases and mean (average) score by education level. Assuming that your sample of the data is called `my_js`, the following code will do this:

```
my_js %>%
  group_by(education_level) %>%
  summarise(
    count = n(),
    mean_score = mean(score)
  )
```

Show this summary in your assignment report (and take note of the mean values!).

[4] Make a boxplot of score by education level using `ggplot` (this is part of the tidyverse so you do not have to load it separately) .

The following code will do this assuming that your sample data is called `my_js`:

```
my_js %>% ggplot(aes(x = education_level, y = score)) +
  geom_boxplot() +
  labs(
    title = "Job satisfaction by education level",
    x = "Education level",
    y = "Job satisfaction score"
  )
```

Include the boxplot in your assignment report.

[5] Fit a model to predict score using education_level. Use `summary` to display the model summary.

Explain in up to 80 words what the coefficient values in the model tell you and relate these coefficient values to the summary of the data you obtained in part [3].

[6] Create at the usual 4 regression diagnostic plots (eg, as obtained using `plot(...)`. Include the plots in your report.

Explain in up to 100 words what assumptions the Scale-Location plot and Normal Q-Q plot give information about and whether these plots indicate anything unusual or wrong for this model.

[7] Use R's `anova` function to display the analysis of variance for your model. Include the output in your assignment report.

State in 30-40 words what null and alternative hypotheses the F-test statistic and its associated p-value shown in the `anova' output are testing, in the context of the variables used in this model.

Give your conclusion based on this F-test and state what evidence that conclusion is based on (about 20 words in total).

[8] How many pair-wise comparison tests would you need to make if you tested each group in the education_level factor compared to every other group? Show how the overall Type 1 error rate is calculated if a significance level of 0.05 is used for each test and give the result of this calculation (this is covered in Lecture 23). Answer in about 20 words in total.

[9] Use the `TukeyHSD` function to get adjusted pairwise confidence intervals from this model (this is covered in Lecture 23). Note that that you need to use the syntax `TukeyHSD(aov(model_name))` where `model_name` is the name of your regression model.

Show the output in your report. Explain in about 50 words what the confidence intervals shown in the `TukeyHSD` function output are confidence intervals for.

Explain in about 30 words whether the `TukeyHSD` confidence intervals show evidence of a difference between any of the education_level group population means and if so, which.

[10] For this and the following questions you will use a subset of the **PlantGrowth** dataset included in R. This dataset gives results from an experiment to compare yields (as measured by dried weight of plants), variable weight (unfortunately I cannot find what units this is in!) obtained under a control and two different treatment conditions, variable group. The dataset is from Dobson, A. J. (1983), *An Introduction to Statistical Modelling*.

Set the random number seed to your student ID number as usual and use the following code to create a subset called `my_plants` of 25 cases from PlantGrowth.

```
my_plants <- PlantGrowth %>% sample_n(25)
```

[11] Create a summary of the counts and mean weights per group. You can use and adapt the code from part [3] above.

Also make a plot of the weights for each group in your data. In this case there are not many observations in each group and a boxplot is not an ideal choice. A nice alternative for a smaller number of observations is a 'jittered scatter plot' (as shown in Lecture 21) which you could create using the following code:

```
my_plants %>% ggplot(aes(x = group, y = weight)) +
  geom_jitter(width = 0.1) + # width says how big the jitter is
  labs(
    x = "Group",
    y = "Weight"
  )
```

(Optional: You could add `aes(colour = group)` in the `geom_jitter` function to colour the points by group.)

Include the summary of counts and means per group and the boxplot in your assignment report.

On the basis of this information, do you think that there might be differences in the population mean weights by group for the sampled populations? Summarise your thoughts in about 50-70 words in your report.

[12] Create a regression model to predict weight using group. Then use the `TukeyHSD` function to get adjusted pairwise confidence intervals for this model (this is covered in Lecture 23).

Show the summary output for the regression and the TukeyHSD output in your report.

Comment in about 100 words on whether there is evidence of a difference between any of the pairs of group population mean weights and if so, which. You should explain clearly what values from the output you are using and how these justify your comments.

Include all the answers and code used to obtain them in a Word document and submit it through Learn. Remember to use your own words when answering the questions.

`