SGTA Week 4 : One-Way Analysis of Variance 2023

Part 1

Question 1

In a paper entitled "The origin of Precambrian Iron Formation", Lindhurst reported on the total *Fe* content in four different types of iron formation, Carbonate, Silicate, Magnetite and Hematite. The *Fe* content was obtained from 6 different randomly selected samples for each of the four different iron types. The data are stored in text file called "iron.txt" on the unit iLearn. The data (as a percentage) are

Carbonate	Silicate	Magnetite	Hematite
34.78	46.43	35.67	22.89
36.08	42.76	32.43	27.69
41.72	45.10	26.54	31.32
39.33	36.65	39.92	33.64
42.50	35.91	35.25	24.88
39.10	41.96	29.84	30.21

Carry out an analysis of variance to determine if there are any differences in iron content for the four different iron formations. Below is a guide on how to do this

a. Load the data using the Tools menu or the command below

```
iron <- read.table("iron.txt", header = TRUE)
```

Notice the data has a column for the iron content and an indicator column for the formation type. Type head(iron) to see this.

b. Use the aov command to carry out the One-Way ANOVA.

```
iron.aov = aov(content ~ formation, data = iron)
```

c. Check if the data adheres to the conditions for ANOVA:

- Check for equal variances by looking at the boxplot of the data (**Hint**: boxplot(content ~ formation, data = iron))
- Assess the residuals for normality (**Hint**: Create QQ plot of residuals from qqnorm(iron.aov$residuals))

d. If the above diagnostic plots seem ok for ANOVA. Carry out the ANOVA and obtain the significance test results, testing for no difference between the four different iron types. As always, state the necessary parameters, hypotheses, test statistic and conclusion (both statistical and contextual). (**Hint**: See summary(iron.aov))

Question 2

In an experiment testing 6 treatments, each treatment was replicated 7 times, so that the 42 experimental units were arranged in a completely random fashion in the experimental area. The data are not given.

a. Set out the analysis of variance table that would be obtained, giving the Source of Variation and degrees of freedom. Leave them blank for the parts that you don't have nor couldn't work out without the data.
b. The analysis of variance, when it was carried out, gave an F observed value of 3.47. Obtain the p value or range of it from F tables, and determine whether there was a significant difference between the 6 treatments at a 5% level of significance. *Note*: If degrees of freedom (for numerator or denominator) is NOT listed in F tables, use the closest one (but below) in the tables. Eg, df = 39 is not in the table, you may look up tables for df=35 (NOT 40). You do similarly when using t tables.

Question 3

It is planned to carry out a similar experiment with the 4 different iron formations (i.e., $t = 4$). In obtaining an estimate of the inherent variability in the experiment, it would be preferred to have at least 30 degrees of freedom for the estimate of random variation σ^2 (i.e., Error M.S. or MSE). **Note**: Error M.S. (or MSE) $= \frac{\text{Error S.S.}}{\text{its degrees of freedom}} = \frac{\text{Error S.S.}}{n-t}$.

a. If we assume equal replication for each of the four treatments, what is the **minimum** number of replicates that should be used and still achieve our aim? Set out only the Source of variation and degrees of freedom for the ANOVA table.
b. Repeat the above exercise if we wish to obtain an estimate with at least 40 degrees of freedom.
c. If it is decided to expand the experiment to assess 8 different iron formations (t =8). Again, assuming equal replication, find the number of replicates necessary to obtain at least
   - 30 d.f.
   - 40 d.f for our estimate of random variation.

Part 2 [Previous exam question]

Question 1

During the Covid-19 pandemic outbreak an international health agency is assessing the impact of the number of cases and the choice of vaccines across 97 selected countries. Data from the second quarter of last year is to be analysed with a one-way ANOVA analysis in `R`. The three most popular vaccine brands are investigated: AstraZeneca, Moderna and Pfizer. The goal is to determine if the number of cases in 2020 has impacted the choice of the vaccine brand.

Cases	Number of recorded Covid-19 cases in the mid of 2020
Brand	The three most popular vaccination brands: AstraZeneca, Moderna and Pfizer

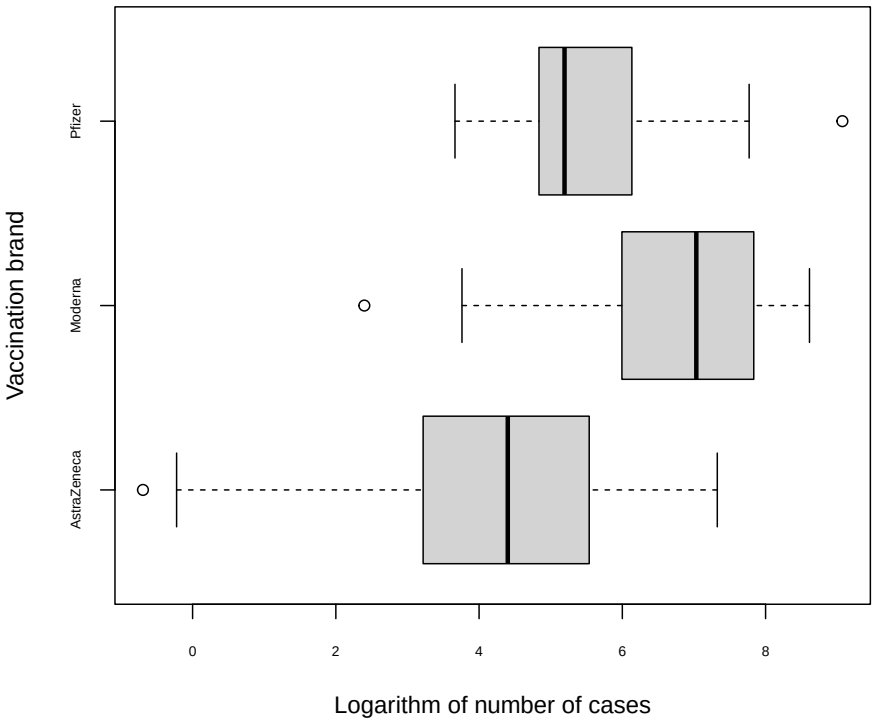

Figure 1: Boxplots of the number and the (natural) logarithm of the number of cases by vaccine brand

a. Look at the boxplots for the log-transformed number of cases. Do the log-transformed number of cases have

   - similar medians across the three groups?
   - similar spreads across the three groups?

b. Using the model diagnostics plots, explain why we cannot proceed with constructing an ANOVA model for the raw number of cases.

c. Suppose you wish to perform a One-Way ANOVA to compare the averages of the logarithm of the number of cases across the three brands. State the assumptions of the test and suggest how to check each of them.

d. Complete the ANOVA table below.

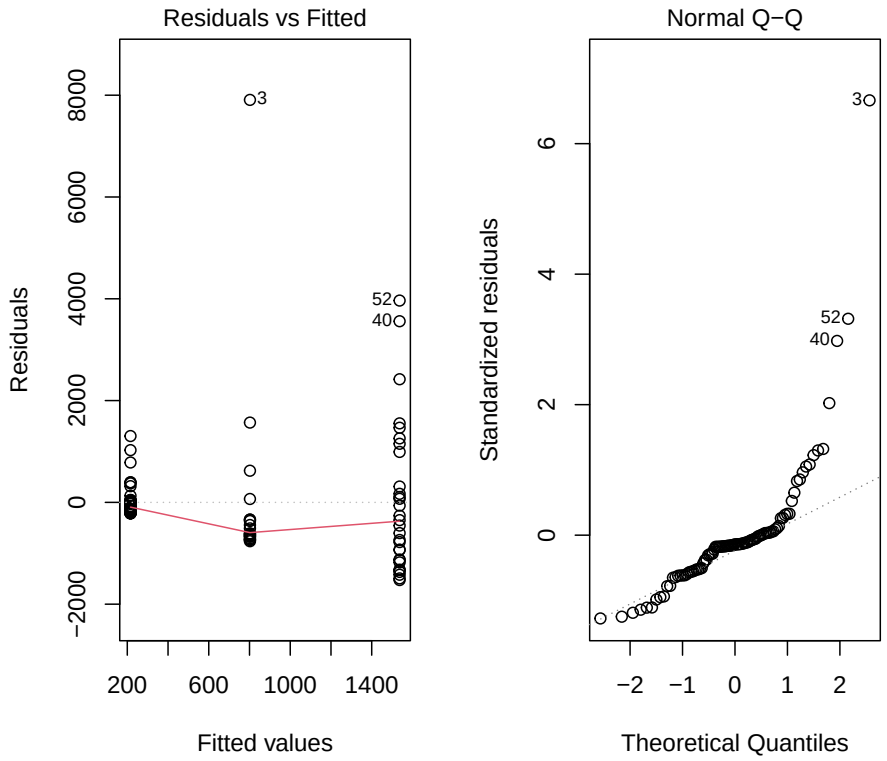

Figure 2: Model diagnostics output of one-way ANOVA for the number of cases.

Source	d.f.	S.S.	M.S.	F-value
Treatment	2			21.28
Residual	94	252.93	2.69	
Total				

e. Based on the ANOVA table, investigate difference in the mean of the logarithms of the number of cases across the three vaccine brands. In your answer:

i) State the null and alternative hypotheses.

ii) Give the appropriate test statistic.

iii) Under the null hypothesis name the distribution of the test statistic and give its parameters.

iv) Using the appropriate table in this exam script, find the rejection region for the test statistic.

v) Draw your conclusion.

Part 3

Recall that `setwd()` requires the specification of the directory. Since the directory may depend on the location of the file in the computer and the operating system of the computer, `setwd()` may no longer work when one shifts to another computer.

A much better way that puts you on the path to managing your `R` work like an expert is to use **projects**.

By clicking on the `RStudio` project file, a file with an extension of `.Rproj`, can automatically open an `RStudio` session and direct the working directory to the folder where the `Rproj` file is located. Hence, clicking on the `.Rproj` file can set `RStudio` to the desired directory even after one shifts to another computer. The exercises below will explore the idea of `project` in `RStudio` a bit further.

If you are using `RStudio Cloud` then using `project` is the default. You can skip ahead to part c) of Question 1.

Question 1: Create project in a new directory

Let's make a project while creating a new directory at the same time.

a) From the `RStudio` menu, click `File > New Project... > New Directory > New Project` to put a `project` file in a folder named `test` on your `Desktop`.

b) Once the process is completed, you'll see your current `RStudio` session switched to the new project. Confirm this by looking at the top right-hand of your RStudio window that you are working with a project named `test`. You can also look at the top of your `R` Console window to confirm `RStudio` is working from the correct working directory. When you are done, quit `RStudio`.

c) Inspect the new folder `test` you just created on your desktop, double-click on the `.Rproj` file to re-open the project. In the `RStudio` session that pops up, find out the current working directory with `getwd()`. Is this working directory identical to the directory of the `test` folder on your desktop?

d) Download the SGTA dataset `light.dat` from our iLearn space and move it to the `test` folder above. Could the codes below import `light.dat`?

```
read.table("light.dat", header = TRUE)
```

e) `here` package helps you to find your files, based on the current working directory at the time. This means that if you have been following the instructions provided so far, it will be relative to the folder with the `.Rproj` file in it.

- In the `RStudio` session in previous question, install `here` package with `install.packages()` and include this package to the library with `library()`.
- After that, type in

  ```
  here()
  ```

  in the `RStudio` Console. Is the resulting folder path identical to the `test` folder on your desktop?

f) What makes the `here` package such an important tool is that it provides a easy way to construct a path to any files relative to your working directory.

- Create a folder named `subfolder` inside the `test` folder on your desktop. Create another folder named `subsubfolder` inside the `subfolder` folder above.
- Type in the `RStudio` console the code below and check the paths it created.

```
here("subfolder", "light.dat")
here("subfolder", "subsubfolder", "light.dat")
```

Essentially you provided the path components from your project *root* to where you want `R` to locate your files and then the `here` package will do the rest and put the full path together.

g) Move the aforementioned dataset `light.dat` to the folder `subsubfolder` above. Correct all the errors in the codes below, which try to import this dataset `light.dat` to `RStudio`.

```
read.table(here("subsubfolder", "subfolder", light.dat), header = TRUE
```

- Hint: in the `RStudio` console, `+` at the start of a line may indicate missing parentheses in the codes. To escape from non-stopping `+` signs, pressing the `ESC` key may help.

h) We will now try to "simulate" the process of moving your analysis to a new computer.

- Quit `RStudio`.
- Move the `test` folder to another location on your computer. It can be within `Documents` or `Downloads`.
- Now open a new `RStudio` session by clicking on the `.Rproj` file under the new location.
- Load any required packages and then run the (fixed) code from part g) to import `light.dat`..
- Notice that you don't have to adjust your code at all even the dataset is in a new location on your computer as long as the (sub-)folder structure remains unchanged.

In summary, `RStudio` projects give you a solid workflow that will serve you well in the future:

- Create an RStudio project for each data analysis project.
- Keep data files there.
- Keep your `R` scripts there.
- Save your outputs (plots and cleaned data) there.
- Only use the `here` package to create relative paths, not absolute paths.

Everything you need is in one place, and cleanly separated from all the other projects that you are working on.

Question 2: Create a project with an existing directory

Complete this question if you haven't had a dedicated `R` folder for this unit and with a project file in it.

a) Designate/create a folder that will keep all the files associated with this unit — input data, R scripts, analytical results, figures.

b) From the `RStudio` menu: `File > New Project... > Existing Directory` and point `R` to the folder in a).

c) `RStudio` should switch to this new project. Find out the `RStudio` is working out from the correct working directory.

d) Now quit `RStudio`. To get back into the project, simply double-click on the `.Rproj` file.