2021025081 DESQUITADO, Yuichi
2021011911 GACASA, Ymanuel Josh
APM1111 Statistical Theory - Formative Assessment 8

Dataset and the Problem

The dataset *PlantGrowth* from R [*library(datasets)*] talks about a study conducted about plants that were given two different treatment conditions and their weights after being dried out.

This dataset is composed of three columns, such as their V1 (indexing), independent variable (group: Control, Treatment 1, Treatment 2), and the dependent variable (weight).

From this dataset, we may investigate to see if such treatments will be significant to incur differences among the tested specimens.

As for this study, the formulated hypotheses (null and alternate), which are the following:

H_0: All group population means are equal (i.e., $\mu_1 = \mu_2 = \mu_3$)

H_A: At least one group population mean is different (i.e., they are not all equal)

Checking Assumptions

Assumption #1: You have one dependent variable that is measured at the continuous level.

> **Remark** - This dependent variable was called "*weight*" in the dataset. This dependent variable is continuous.

Assumption #2: You have one independent variable that consists of three categorical, independent groups.

> **Remark** - The independent variable in the dataset is called "group", which is composed of Control, Treatment 1, and Treatment 2, which refers to the variation of treatment done by the researchers.

Assumption #3: You should have independence of observations.

> **Remark** - Each observation is independent of each other as there is no relationship between the observations in each group of the observation or between the groups themselves.

Assumption #4: There should be no significant outliers in the three or more groups of your independent variable in terms of the dependent variable.

There are many ways to check for outliers. For this paper, we would be employing two ways, visual using boxplot and using interquartile.

Boxplot

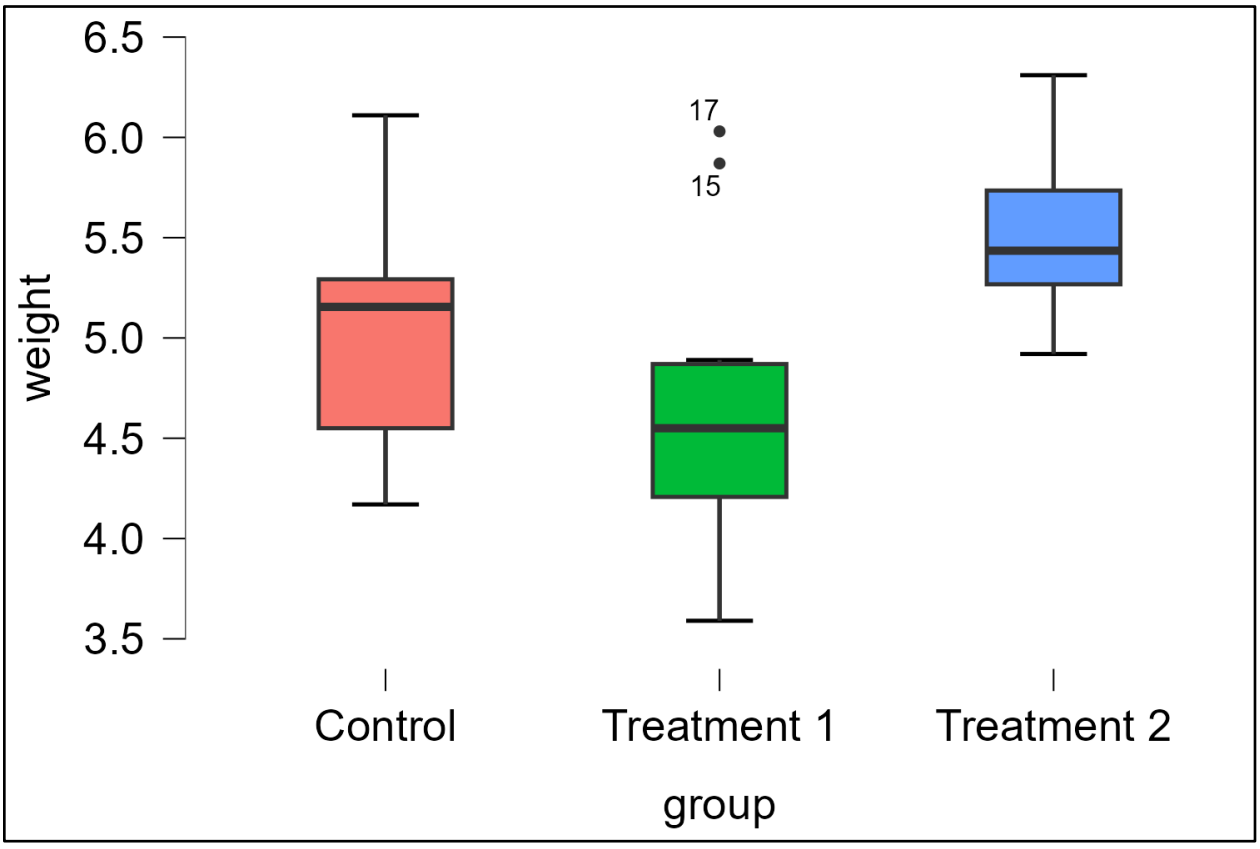

Figure 1. Boxplot of the dataset PlantGrowth

Interquartile

```
[1] import matplotlib.pyplot as plt
    import scipy.stats as stats
    import statistics
    import pandas as pd
    from google.colab import drive
    drive.mount('/content/drive')

    Mounted at /content/drive

[2] df = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/PlantGrowth.csv")

    # Using interquartiles to find outliers

    for i in ['ctrl', 'trt1', 'trt2']:
      quartile = df[df['group'] == i]['weight'].quantile([0.25, 0.75])
      iqr = quartile[0.75] - quartile[0.25]

      print("The IQR for group \"%s\"" % i, "is =", round(iqr, 4))

      highOutlier = round(quartile[0.75] + (1.5 * iqr), 4)
      lowOutlier = round(quartile[0.25] - (1.5 * iqr), 4)

      print(highOutlier, lowOutlier)

      sdf = df[df['group'] == i]['weight']

      for j in sdf.index:
        sdfValue = sdf[j]

        if (sdfValue >= highOutlier or sdfValue <= lowOutlier):
          print("[", j + 1, "]",sdfValue, "is an outlier.")

      print()

    The IQR for group "ctrl" is = 0.7425
    6.4062 3.4363

    The IQR for group "trt1" is = 0.6625
    5.8637 3.2138
    [ 15 ] 5.87 is an outlier.
    [ 17 ] 6.03 is an outlier.

    The IQR for group "trt2" is = 0.4675
    6.4362 4.5663
```

Image 1. Program for identifying outliers using IQR

Remark - There were significant outliers in the dataset, namely data entry 15 (*Treatment 1, 5.87*) and 17 (*Treatment 1, 6.03*) as assessed by visual inspection of boxplot and usage of interquartile. Such outliers will not be removed from the dataset in the continuation of the computation for ANOVA as they are not extreme outliers.

Assumption #5: Your dependent variable should be approximately normally distributed for each group of the independent variable.

Descriptive Statistics ▼

	weight		
	Control	Treatment 1	Treatment 2
Valid	10	10	10
Missing	0	0	0
Mean	5.032	4.661	5.526
Std. Deviation	0.583	0.794	0.443
IQR	0.743	0.662	0.467
Shapiro-Wilk	0.957	0.930	0.941
P-value of Shapiro-Wilk	0.747	0.452	0.564
Minimum	4.170	3.590	4.920
Maximum	6.110	6.030	6.310
25th percentile	4.550	4.207	5.268
50th percentile	5.155	4.550	5.435
75th percentile	5.293	4.870	5.735

Table 1. Descriptive statistics of PlantGrowth dataset with Shapiro-Wilk's Test P-Value

Remark - The weights are approximately normally distributed for each of the treatment conditions, as assessed by Shapiro-Wilk's Test, $p > 0.05$.

Assumption #6: You have homogeneity of variances (i.e., the variance of the dependent variable is equal in each group of your independent variable).

Test for Equality of Variances (Levene's)

F	df1	df2	p
1.237	2.000	27.000	0.306

Table 2. Levene's Test for Equality of Variance for PlantGrowth dataset

Remark - There was homogeneity of variances of the weights for all treatment conditions, as assessed by Levene's Test for Equality of Variance, $p > 0.05$.

Computation

ANOVA - weight

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
group	3.766	2	1.883	4.846	0.016	0.264
Residuals	10.492	27	0.389			

Note. Type III Sum of Squares

Table 3. Result of ANOVA for PlantGrowth dataset

Descriptives - weight

group	N	Mean	SD	SE	Coefficient of variation
Control	10	5.032	0.583	0.184	0.116
Treatment 1	10	4.661	0.794	0.251	0.170
Treatment 2	10	5.526	0.443	0.140	0.080

Table 4. Descriptives of the PlantGrowth dataset

Post Hoc Comparisons - group

			95% CI for Mean Difference				
		Mean Difference	Lower	Upper	SE	t	p_{tukey}
Control	Treatment 1	0.371	−0.320	1.062	0.279	1.331	0.391
	Treatment 2	−0.494	−1.185	0.197	0.279	−1.772	0.198
Treatment 1	Treatment 2	−0.865	−1.556	−0.174	0.279	−3.103	0.012*

* p < .05

Note. P-value and confidence intervals adjusted for comparing a family of 3 estimates (confidence intervals corrected using the tukey method).

Table 5. Post Hoc Comparison for PlantGrowth dataset

Reporting

A one-way ANOVA was conducted to determine if there is a significant difference between treatment conditions to the weight of the plants. Specimens were divided into three groups: Control (*n = 10*), Treatment 1 (*n = 10*), and Treatment 2 (*n = 10*).

There were outliers under Treatment 1, namely (*Row 15 = 5.87*) and (*Row 17 = 6.03*). The said outliers were identified using visual inspection via boxplot and interquartile computation (*Control IQR = 0.7425*, *Treatment 1 IQR = 0.6625*, *Treatment 2 IQR = 0.4675*).

The dataset is normally distributed for each group, as assessed by Shapiro-Wilk's Test ($p > 0.05$).

The dataset is homogeneous via Levene's Test ($p = 0.306 > 0.05$).

Data is presented as mean $\pm$ standard deviation. The weights were statistically significantly different between different treatment conditions, $F(2, 27) = 4.846$, $p = 0.016 < 0.05$, $\eta^2 = 0.264$.

Weights decreased from Control ($\mu = 5.032$, $\sigma = 0.583$) to Treatment 1 ($\mu = 4.661$, $\sigma = 0.794$), then increased to Treatment 2 ($\mu = 5.526$, $\sigma = 0.443$), in that order.

Turkey post hoc analysis showed that the mean increased from Treatment 1 to Treatment 2 (*0.865, 95% CI [0.174, 1.556]*), meaning a statistically significant ($p = 0.012 < 0.05$).