One-Way ANOVA

Taylor Shrode

4/5/2020

MSDS 660

One-Way ANOVA, or One-Way Analysis of Variance, "is a statistical technique that allows experimenters to compare mean differences of one outcome (dependent) variable across two or more groups (levels) of one independent variable (factor) (Powell, 2017)." One-Way ANOVA is essentially an extension of an independent two-sample t-test for comparing means where there are two or more groups (One-way anova test in r, n.d.). The summarized steps of the ANOVA procedure are listed below (One-way anova, n.d.).

1. Set up hypotheses and determine level of significance.
2. Calculate test statistic, which is the F-statistic.
3. Determine p-value.
4. Make a decision.
5. Conclude findings.

For this assignment, we will be analyzing a dataset that contains pain relief data from three different treatments (A, B, C) with One-Way ANOVA using a 5% significance level. The investigator for this study randomly assigned 12 participants to the treatments: four participants per treatment. The participants were instructed to take the given treatments when they feel pain and record the time (in minutes) until the pain eases off.

To begin, we need to upload the data into R. Our dataset contains twelve rows of data with two columns, treatment and time to relief (minutes). The dataset is located below.

Treatment	Time to relief (mins)
A	14
	24
	12
	25
B	20
	14
	17
	18
C	22
	29
	36
	20

To upload the data into R, we can enter the data into an Excel file and then read from the file or we can enter the data directly into R, which is the method chosen for this dataset.

```
> treatment = c(rep("A",4), rep("B", 4), rep("C",4))
> time_to_relief = c(14,24,12,25,20,14,17,18,22,29,36,20)
> pain_data <- data.frame(treatment, time_to_relief)
```

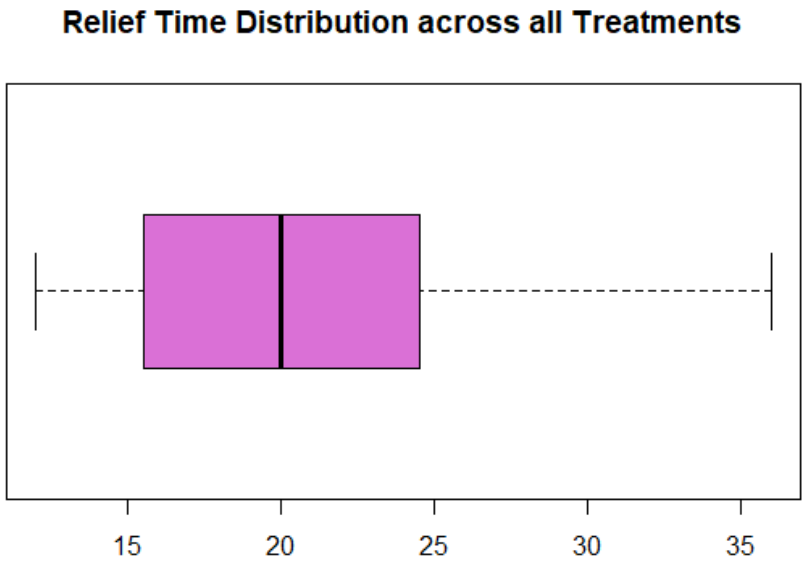
To ensure the data was entered into R correctly, we can use the **View(<dataset>)** command.

	treatment	time_to_relief
1	A	14
2	A	24
3	A	12
4	A	25
5	B	20
6	B	14
7	B	17
8	B	18
9	C	22
10	C	29
11	C	36
12	C	20

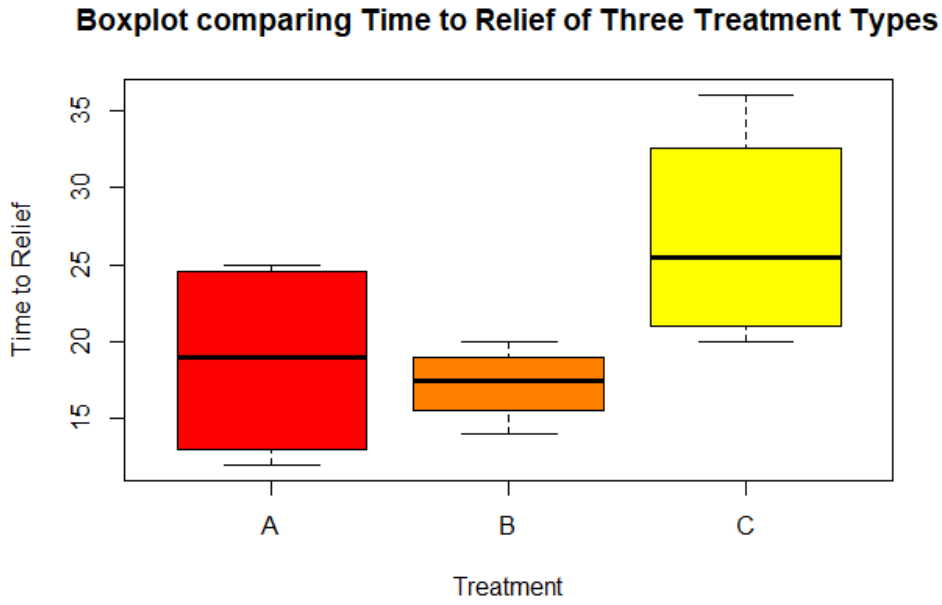
As we can see above, our data was correctly entered into R. Now, we can use the **str(<dataset>)** command to display the structure of the data and use the **summary(<dataset>)** to produce a summary of our dataset.

```
> str(pain_data)
'data.frame':   12 obs. of  2 variables:
 $ treatment     : Factor w/ 3 levels "A","B","C": 1 1 1 1 2 2 2 2 3 3 ...
 $ time_to_relief: num  14 24 12 25 20 14 17 18 22 29 ...
> summary(pain_data)
 treatment time_to_relief
 A:4       Min.   :12.00
 B:4       1st Qu.:16.25
 C:4       Median :20.00
           Mean   :20.92
           3rd Qu.:24.25
           Max.   :36.00
>
```

We can also graphically display our data using a boxplot, which is a graphical summary of a distribution (Liu et al., n.d.). The boxplot allows us to visualize the median of the dataset, the first (Q1) and third quartiles (Q3), the maximum (Q3 + 1.5*IQR) and minimum points (Q1 - 1.5*IQR), and any outliers (R - boxplots, n.d.). First, we can view the boxplot for relief time distributed across all treatments, using the command **boxplot(pain_data$time_to_relief, horizontal = TRUE, main="Relief Time Distribution across all Treatments", col = "orchid")**.



This boxplot is a visualization of our **summary** we created above. Next, we can create a boxplot to compare relief times of each treatment type using the command **boxplot(pain_data$time_to_relief~pain_data$treatment, main="Boxplot comparing Time to Relief of Three Treatment Types", xlab = "Treatment", ylab = "Time to Relief", col= heat.colors(3)).** It should be noted that the color palette chosen is a built-in function in R (Kassambara, 2018).



This boxplot allows us to compare the range and distribution of the time to relief of each treatment (Galarnyk, 2019). We see that Treatment B has the tightest range of time to relief while Treatment C has the widest range and the widest distribution. We can also see that the medians in the Treatment A and B boxplots slightly overlap, while there is not any overlap in the Treatment C boxplot. This indicates that, with 95% confidence, the true medians may not differ (Galarnyk, 2019).

Before we create our ANOVA model, we need to define our null and alternative hypotheses. The null hypothesis for ANOVA testing states that the means of each group are the same. The alternative hypothesis states that at least one sample mean is not equal to the others (One-way anova test in r, n.d.). Generically (One-way anova, n.d.),

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k, \text{ where } k \text{ is the number of groups}$$

$$H_A: The\ means\ are\ not\ all\ equal.$$

In our case, we have three groups, treatments A, B, and C. Thus, our hypotheses are

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_A: The\ means\ of\ Treatment\ A, B, C\ are\ not\ all\ equal.$$

Now that we have defined our hypotheses, we need to find our critical F-value. We can do this by referencing a F-table that corresponds to our level of significance (0.05), which can be found here (Critical values of the f distribution, n.d.). To use the F-table, we need to find our two types of degrees of freedom: the numerator (v_1) and the denominator (v_2) (Anderson, n.d.). The numerator is calculated by subtracting 1 from the total number of groups and the denominator is calculated by subtracting the number of groups from the total number of observations (Anderson, n.d.). In other words,

$$v_1 = 3\ groups\ -\ 1\ =\ 2$$
$$v_2 = 12\ -\ 3 = 9.$$

To read the F-table, the numerator (v_1)is represented by the numbers across the top of the table, and the denominator (v_2) is represented by the numbers down the column. Thus, referencing our F-table, our critical F-value is

$$F_{0.05}(v_1, v_2) = F_{0.05}(2,9) = 4.256.$$

We can now fit our ANOVA model since we have defined our null and alternative hypotheses. To build the ANOVA model, we use the **aov()** function, which takes in a formula argument and a data argument, similar to the **lm()** function used for linear regression. Thus, our ANOVA function looks like **aov(X~Y, dataset)** where **X** is our explanatory variable (**time_to_relief**), **Y**

is the response variable (**treatment**), and our **dataset** is **pain_data**. Saving this model in a variable called **pain_aov**, we can apply the **summary()** function and interpret the results.

```
> pain_aov <- aov(time_to_relief~treatment, pain_data)
> names(pain_aov)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "contrasts"     "xlevels"       "call"          "terms"
[13] "model"
> summary(pain_aov)
            Df Sum Sq Mean Sq F value Pr(>F)
treatment    2  208.7  104.33   3.007    0.1 .
Residuals    9  312.2   34.69
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Notice, the table gives us the degrees of freedom that we calculated above for our F-table, where **Residuals** is the denominator and **treatment** is the numerator.

The first value we look at is the p-value (Pr(>F)). If our p-value is less than our alpha level (0.05), we can reject the null hypothesis. Above, we see that our p-value is 0.1, which is not less than our alpha level. In other words, we cannot reject the null hypothesis. Now, we consider the f-value (F value). The F-value is calculated as

$$F = \frac{MS_{between}}{MS_{within}}$$

where $MS_{between}$is the variation between sample means and $MS_{within}$is the variation within groups (Powell, 2017). These values can be found in the **Mean Sq** (Mean Squares) column in our output above. The $MS_{between}$value is the mean square for **treatment** and MS_{within}value is the mean square for **Residuals** (Sullivan, n.d.). Thus,

$$F = \frac{MS_{between}}{MS_{within}} = \frac{104.33}{34.69} = 3.007.$$

Notice that our F-value is 3.007, which is less than our critical F-value (4.256). This also means that we cannot reject our null hypothesis. In other words, there is no statistical significance between the means of the three treatment groups.

Now, if the ANOVA is significant, the next step would be to perform post hoc tests to confirm where those differences are. One of the most used post hoc tests is Tukey's Honest Significant Difference (HSD), which assesses the significance of differences between pairs of group means (Post hoc tests, 2020). The command to run this post hoc test in R is **TukeyHSD(aov(response~predictor, conf.level))**, where **aov(response~predictor)** is our ANOVA model above (**pain_aov**) and **conf.level** is the confidence level (Post hoc tests, 2020).

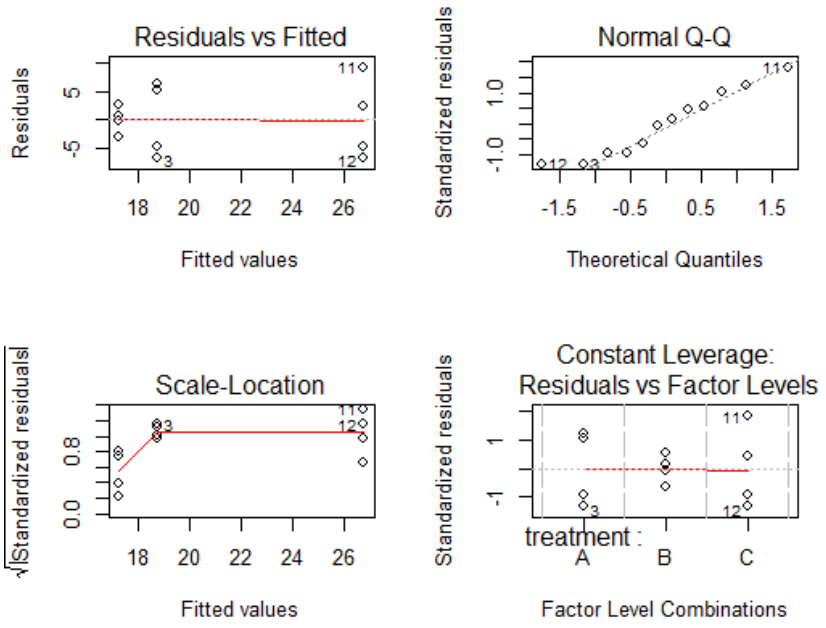
```
> tukey_test <- TukeyHSD(pain_aov, conf.level=0.95)
> tukey_test
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time_to_relief ~ treatment, data = pain_data)

$treatment
    diff        lwr     upr     p adj
B-A -1.5 -13.128702 10.1287 0.9315326
C-A  8.0  -3.628702 19.6287 0.1883467
C-B  9.5  -2.128702 21.1287 0.1102074

> |
```

The columns we are most interested in is **diff,** the difference of means, and **p adj,** the adjusted p-value for all pairs (Post hoc tests, 2020). As we can see above, none of the adjusted p-values are less than our level of significance (0.05), which indicates that there are no significant differences between any of the three pairs. Thus, further confirming the ANOVA model above. The results can also be plotted using the command **plot(tukey_test).** Significant pairs are the ones that do not cross the zero value (Pasin, 2014).

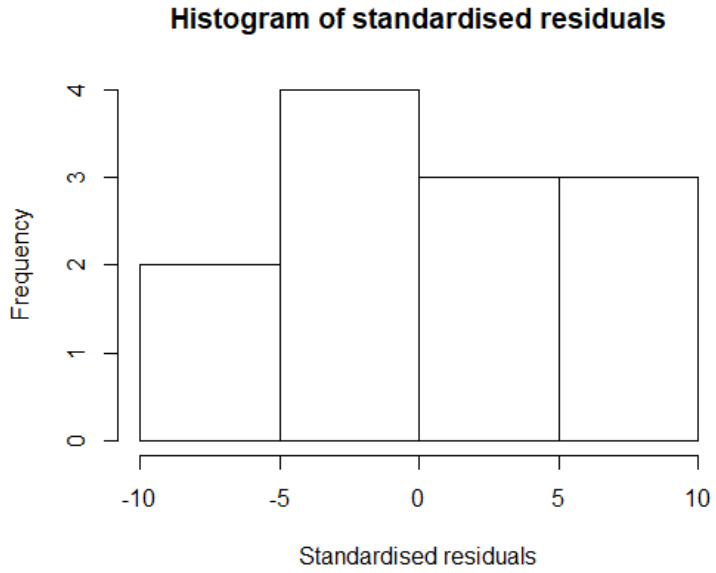
Even though our ANOVA model doesn't indicate that there is any statistical significance between our groups, we still need to check the assumptions of our model. The assumptions of One-Way ANOVA include the following (Mackenzie, 2018):

- Normality: Sample is taken from a normally distributed population.
    - Can be detected by plotting a histogram or using statistical tests such as Anderson-Darling test, Shapiro-Wilk test, Kolmogorov-Smirnov, etc (WorldClass FTE).
    - Can also use a normal probability plot or a quantile-normal plot of the residuals (WorldClass FTE).
- Variance Equality: The variance of data in the different groups should be the same.
    - This can be examined by Levene statistic test (WorldClass FTE)
    - Can also use a scatterplot between residuals and predicted values (WorldClass FTE).
- Sample Independence: Each sample has been drawn independently of the other samples.
    - Can be investigated by how that data was collected (WorldClass FTE).

First, we need to plot our ANOVA model. This is done by using the commands **par(mfrow=c(2,2))** and **plot(pain_aov).**



First, to check for normality, we look at our Normal Q-Q plot. Our Normal Q-Q plot indicates normality because of the straight slope. We can also create a histogram of the residuals. This can be done using the commands **residuals<-pain_aov$residuals** and **hist(residuals, main="Histogram of standardised residuals",xlab="Standardised residuals").**



The histogram indicates normality by the normal distribution of the plot. Next, to check for Variance Equality (homogeneity), we can use Levene's statistic test. We need to load the **car** library and then we can use the function **leveneTest** to test for equality of variances.

```
> library(car)
Loading required package: carData
> leveneTest(time_to_relief~treatment)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  2  4.5444 0.04323 *
       9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If the p-value is greater than 0.05, equal variances can be assumed (Karadimitriou & Marshall, n.d.). As we can see above, our p-value is 0.04323, which is not greater than 0.05. Thus, equal variances cannot be assumed. We can also verify this by using a scatter plot between residuals and predicted values. As seen in our plot above, our scatter plot seems to have a "pattern" to it, which helps verify our Levene's test above.

As we saw above, a One-Way ANOVA test was conducted to identify effectiveness of different pain treatments. Before we created the ANOVA model, we used a boxplot to visualize the time to pain relief (in minutes) distributed across all treatments, and then we used a boxplot to visualize the time to pain relief for each treatment. After we conducted our ANOVA test, we found that there were no significant differences in the means of time to pain relief for each treatment. We also conducted normality tests and homogeneity tests for One-Way ANOVA

model assumptions. We found that the normality assumption was met but there was no equality of variances. Although our ANOVA model found no significant differences between the three treatment groups, a post hoc test, the Tukey HSD test, was carried out. The Tukey test also found that there were no significant differences between the treatment groups.

Resources

Anderson, A. (n.d.). Critical values for an anova hypothesis using the f-table. *Dummies*. Retrieved April 2, 2020, from https://www.dummies.com/education/math/business-statistics/how-to-find-the-critical-values-for-an-anova-hypothesis-using-the-f-table/

Critical values of the f distribution. (n.d.). Retrieved April 2, 2020, from https://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm#ONE-05-1-10

Galarnyk, M. (2019, November 7). *Understanding boxplots*. Medium. https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Karadimitriou, S., & Marshall, E. (n.d.). https://www.sheffield.ac.uk/polopoly_fs/1.536445!/file/MASH_ANOVA_in_R.pdf

Kassambara. (2018, November 13). Top r color palettes to know for great data visualization. *Datanovia*. https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/

Liu, C.-T., Milton, J., & McIntosh, A. (n.d.). *Summary statistics and graphs with r*. Retrieved April 1, 2020, from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R2_SummaryStats-Graphs/R2_SummaryStats-Graphs_print.html

Mackenzie, Ruairi J . (2018, July 20). *One-way vs two-way anova: Differences, assumptions and hypotheses*. Informatics from Technology Networks. https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553

One-way anova. (n.d.). Retrieved March 31, 2020, from https://online.stat.psu.edu/stat200/book/export/html/212

One-way anova test in r. (n.d.). Retrieved March 31, 2020, from http://www.sthda.com/english/wiki/one-way-anova-test-in-r

Pasin, M. (2014, June 23). Performing anova test in r. *R-Bloggers*. https://www.r-bloggers.com/performing-anova-test-in-r-results-and-interpretation/

Post hoc tests. (2020, March). https://biostats.w.uib.no/post-hoc-tests-tukey-hsd/

Powell, C. (2017, September 18). *Oneway anova explanation and example in r*. http://rstudio-pubs-static.s3.amazonaws.com/308410_2ece93ee71a847af9cd12fa750ed8e51.html

R - boxplots. (n.d.). Retrieved April 1, 2020, from https://www.tutorialspoint.com/r/r_boxplots.htm

Sullivan, L. (n.d.). *Hypothesis testing—Analysis of variance(Anova)*. Retrieved April 2, 2020, from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ANOVA/BS704_HypothesisTesting-Anova_print.html