R Project 2: ANOVA and Tukey Comparisons

Data Analytics 2 – ADS523

James Sears

Applied Data Science Program, Graduate School, Longmeadow, MA

Submitted To: Professor Aja Shabana

Summer 2021

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

Introduction

The object of this project is to conduct one-way analysis of variance (ANOVA) and create Tukey confidence intervals. One-way ANOVA is used to test a hypothesis that three or more population means are equal in terms of a single independent variable; and alternatively that at least one population is different from the others. When the latter is determined, post-ANOVA analysis using the Tukey multiple-comparison method is used to determine which populations mean(s) are different.

These concepts are applied to happiness scores for nations generated by Gallup World Poll. The Mean Happiness Scores of multiple global regions are tested to determine if they are equal, and if not, which region or regions are different.

Data Description

The World Happiness Report is a landmark survey of the state of global happiness. The report continues to gain global recognition as governments, organizations and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness (World Happiness Report).

The happiness scores and rankings use data from the Gallup World Poll. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others (World Happiness Report).

For this project, the data is drawn from the 2021 World Happiness Report (Version 2), compiled by Ajayoal Singh for Kaggle Inc. and published online in April 2021. The variables of interest for the nations listed are the ten global regions to which they are assigned and the Happiness Scores for each of those nations. Four of global regions are to be selected and analyzed.

One-Way Analysis of Variance (ANOVA)

ANOVA is used to determine if sample data could come from populations with the same mean, or suggests that at least one sample comes from a population whose mean is different from the others. This is accomplished by comparing the variability among the samples' means (between-sample variability) and the variability within each sample (within-sample variability). If the between-sample variability is large relative to the within-sample variability, then there is evidence to suggest that the samples are of populations with different means (Sullivan, page 639-640).

The ratio of these variabilities is known as the ANOVA *F*-test statistic, expressed below as Formula 1:

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

$$F = \frac{(between\ sample\ variation)\ /\ (number\ of\ populations - 1)}{(within\ sample\ variation)/\ (total\ number\ of\ observations - number\ of\ populations)}$$

As the variations are based on mean squares, this can also be written as Formula 2:

$$F = \frac{SST\ /\ (k-1)}{SSE/\ (n-k)} = \frac{MST}{MSE}$$

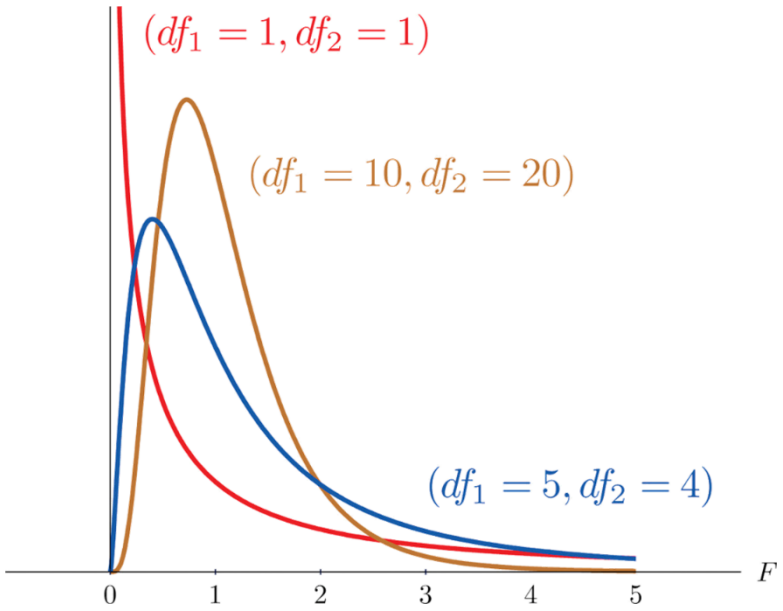
Where *n* is the total number of observations for all samples and *k* is the number of samples (or treatments) representing each population.

The between-sample variation is the mean square of errors of the samples (or treatments), or *MST*; it is calculated as the sum of squared errors of the samples (*SST*) divided by the samples' degrees of freedom (*k* number of samples minus one). The *SST* is calculated as the sum of squares of (*SS total*) of all observations about their mean minus the sum of squares error (*SSE*) of each sample.

The within-sample variation is the mean square of errors, or *MSE*; it is calculated as the sum of squared errors (*SSE*) of each sample about their means divided by the corresponding degrees of freedom (*n* total number of observations minus *k* number of samples).

Once again, the ratio of these variabilities (*MST* and *MSE*) is known as the ANOVA *F*-test statistic. *F*-test statistics make up a probability distribution known as the *F*-distribution. The *F*-distribution is right (positive) skewed can vary in shape (Figure 1) depending on the pair of degrees of freedom that correspond with the numerator (df_1; *k* number of samples minus one) and denominator (df_2; *n* total number of observations minus *k* number of samples) of the Formula 2 above. The values of *F* are always

Figure 1. Samples of F-distribution curves of various pairs of degrees of freedom.



greater than or equal to zero, and as with any probability distribution, the area beneath the curve is equal to one. Corresponding to the degrees of freedom in the numerator and denominator are various

areas (α; 0.1, 0.05, 0.025, 0.01 and 0.001) in the right tail of the *F*-distribution. These areas correspond to the desired level of significance (1 - α); 90%, 95%, 97.5%, 99% and 99.9%, respectively (Sullivan, page 576).

As noted, the null hypothesis with ANOVA that the sample means are equal, and alternatively that at least one sample is different from the others. If the *F*-test statistic calculated from the samples' observations, for the desired level of significance, is less than the corresponding theoretical *F*-distribution's critical value (from a lookup table or calculated by technology) then the null hypothesis is not rejected; the means and variances of each of the samples are within reasonable variance (that which may occur randomly) of a single population. In other words, the sample means are the same; the samples are of the same population. The *P*-values of the *F*-test statistic can also be calculated, estimating the probability of obtaining an *F*-test statistic that is equal or (more extreme (greater) than the critical value. This can be compared to the level of significance (α) and if it is greater, the null hypothesis is not rejected.

Conversely, if the *F*-test statistic is greater than the critical value, or the *P*-value is less than the level of significance α, the null hypothesis is rejected, indicating that at least one of the sample means is different from the others; not all of the samples are from the same population. However, ANOVA does not indicate which sample or samples are different from the others. Fortunately, this can be accomplished with post-ANOVA analysis using the Tukey multiple-comparison method.

Tukey Multiple-Comparison Method

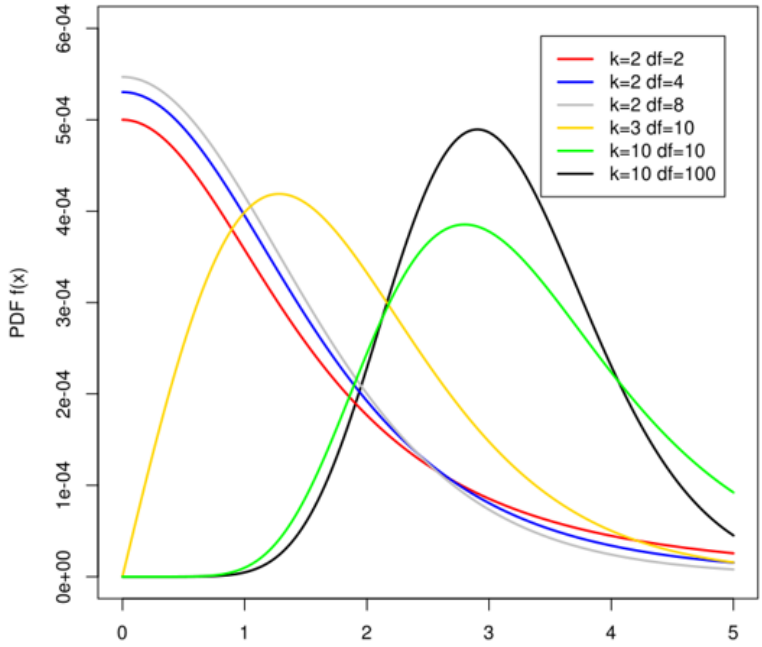
The Tukey test compares the means after the null hypothesis of equal means has been rejected. It identifies the pairs of means that differ significantly by constructing confidence intervals for each pair of means with a given family confidence level. An individual confidence level is the confidence that any *particular* confidence interval contains the difference between the corresponding population means. Whereas the family confidence level is the confidence that *all* the individual confidence intervals contain the differences between the corresponding population means. Therefore, the family confidence enables the simultaneous comparison of all of the pairwise confidence intervals (comparing the differences of population means). The formula for constructing the intervals of the family confidence level is expressed as Formula 3:

$$\left(\bar{x}_i - \bar{x}_j\right) \pm \frac{q_\alpha}{\sqrt{2}} \cdot s\sqrt{(1/n_i) + \left(1/n_j\right)}$$

$$s = \sqrt{SSE/(n-k)}$$

Where $\bar{x}_i - \bar{x}_j$ is the point estimate of the difference of the means, $\bar{x}_i$ and $\bar{x}_j$, of the two samples being compared; *s* is the square root of the mean square of error estimate (*MSE*, or the within-sample variation) from ANOVA, n_i is the sample size of the first population, and n_j is the sample size of the second population. The term q_α is the critical value of the Studentized Range (*q*) Distribution, which is another family of distribution curves that vary in shape depending on the κ number of populations, samples, or treatments ($\kappa = k$) and *v* degrees of freedom of random variation (*n* total observations from all samples minus *k*). Figure 2 demonstrates various Studentized Range (q) Distribution curves.

Figure 2. Samples of Studentized Range (q) Distribution curves of various pairs of *k* (number of populations, samples or treatments) and *v* degrees of freedom of random variation.



Theoretical *q*-distribution critical values of the desired level of significance are acquired from lookup tables or calculated by technology. The critical value and the Formula 3 above are used to construct confidence intervals for each pair of sample means. If the pair-wise confidence interval contains zero, then the samples' population means are possibly equal based on the desired level of significance (or confidence). The confidence interval gives an interval of possible values for the difference between the two populations. So, if the confidence interval contains zero then it contains both positive and negative values which means the difference could be positive or negative; either population could have the larger mean.

Conversely, if the pair-wise confidence interval does not contain zero, then there is evidence to support that there is a significant difference between the two population means being compared. In this case, the interval is either all positive or all negative, meaning the larger population mean is significantly larger.

Data Selection for One-Way Analysis of Variance (ANOVA)

The 2021 World Happiness Report contains a list of nations and their Happiness Scores. Each nation is also assigned to one of ten global regions. The goal is to select four regions and utilize one-way ANOVA to determine if their Regional Mean Happiness Scores (μ) are equal or not. These regions are listed in Table 1 along with their abbreviated label, the number of nations in each region (n), the regions' Mean Happiness Scores (μ) and standard deviation (σ).

It should be noted that the Regional Mean Happiness Scores are improperly calculated as means of means; they are simply the sum of each region's nations Happiness Scores divided by the number of

nations. This method assumes that the population of each nation in the study is the same. The proper method to obtain each Regional Mean Happiness Scores would be to weight each nation's Happiness Score based on each nation's population. As population data was not available, the project assumes the population of each nation in the study is the same.

Table 1. Regions and Regional Mean Happiness Scores. The regions analyzed are in bold font.

Region	Abbr.	Nations (n)	Mean Happiness Score (μ)	Std. Dev. (σ)
Central and Eastern Europe	**CaEE**	**17**	**5.985**	**0.493**
Commonwealth of Independent States	CoIS	12	5.467	0.438
East Asia	EstA	6	5.810	0.440
Latin America and Caribbean	**LAaC**	**20**	**5.908**	**0.693**
Middle East and North Africa	MEaNA	17	5.220	0.999
North America and ANZ	NAaA	4	7.129	0.138
South Asia	SthAs	7	4.442	0.993
Southeast Asia	SthsA	9	5.408	0.606
Sub-Saharan Africa	**S-SA**	**36**	**4.494**	**0.655**
Western Europe	**WstE**	**21**	**6.915**	**0.657**

The four regions of interest are **Central and Eastern Europe (CaEE), Latin America and Caribbean (LAaC), Sub-Saharan Africa (S-SA), and Western Europe (WstE)**; these regions were selected for multiple reasons. One, their sample sizes are relatively large and will yield more significant results. Two, their means are relatively close and testing the level of accuracy of ANOVA and the Tukey Test were of interest. And three, their standard deviations meet the conditions required by ANOVA:

- **The samples are from populations that follow the normal distribution:** The samples are population parameters based on populations of nations that are assumed to be large.
- **The populations are independent:** Each nation contributes to the sample independently and is assigned to only one region.
- **The populations have equal standard deviations, or the largest SD is not more than twice the smallest SD:** The largest standard deviation is 0.693 (LAaC), which is not more than twice the smallest of 0.493 (CaEE).

Application of One-Way Analysis of Variance (ANOVA)

As the conditions to perform ANOVA are met, the null and alternative hypotheses to test the Regional Mean Happiness Scores (μ) are stated as:

H_0: All the population means are equal, $\mu_{\text{CaEE}} = \mu_{LAaC} = \mu_{S-SA} = \mu_{WstE}$

H_1: *At least one* population mean is different.

Once again, ANOVA compares the variability among the samples' means (between-sample variability, or *MST*) and the variability within each sample (within-sample variability, or *MSE*), and calculates the F-test statistic by Formula 2. Technology (Excel and R) is used to perform ANOVA with a 95% level of significance ($\alpha = 0.05$), where k is the number of samples (4 regions) and n is the total number of observations for all samples (94 nations from the 4 regions). The results are displayed in Table 2:

Table 2. ANOVA results generated by technology (Excel).

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
CaEE	17	101.741	5.984765	0.24337
LAaC	20	118.161	5.90805	0.480896
S-SA	36	161.801	4.494472	0.428884
WstE	21	145.213	6.914905	0.431018

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Sample	*SST*		*MST*			
	84.71917	3	28.23972	69.32408	2.57E-23	2.705838
Within Sample	*SSE*		*MSE*			
	36.66223	90	0.407358			
Total	121.3814	93				

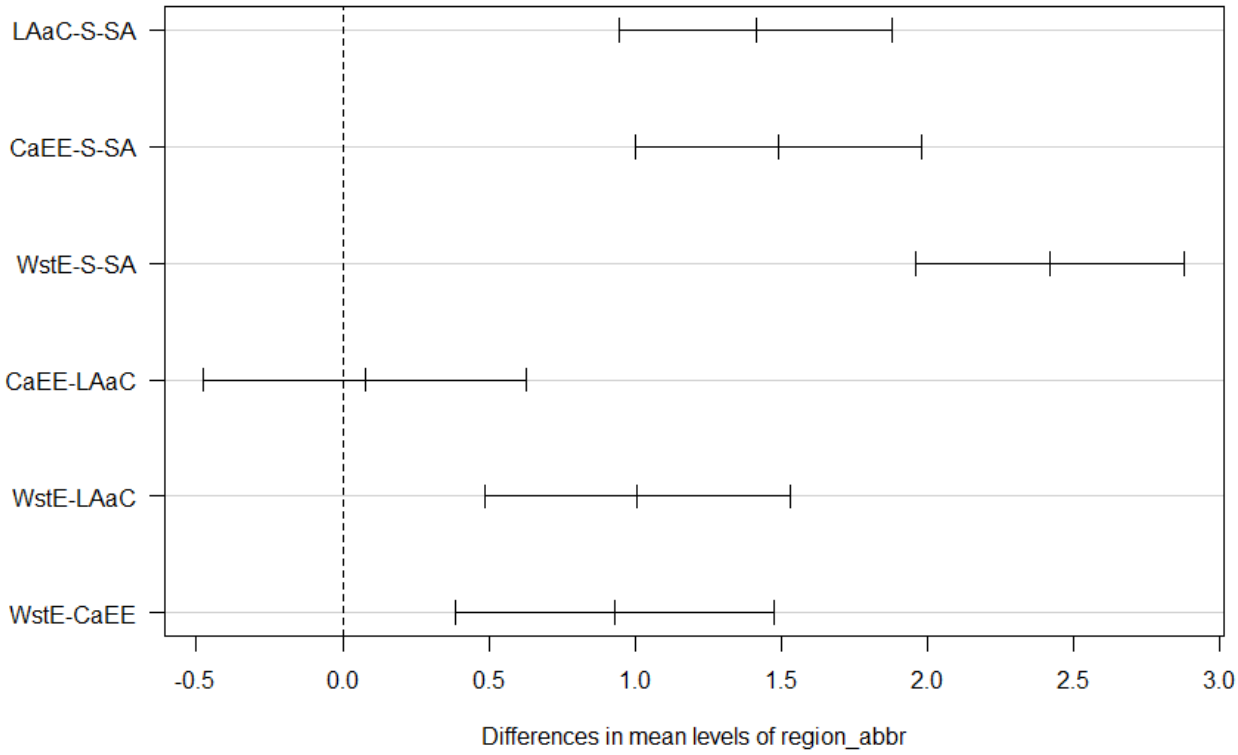
The calculated *F*-test statistic is 69.324, which is larger than the theoretical critical value of 2.706 for the appropriate *F*-distribution curve ($df_1 = k - 1 = 4 - 1 = 3$; $df_2 = n - k = 94 - 4 = 90$). The calculated *P*-value is very small at 2.57×10^{-23} compared to α of 0.05 for a 95% level of significance. As noted, if the *F*-test statistic is greater than the critical value, or the *P*-value is less than the level of significance α, the null hypothesis is rejected. This indicates that at least one of the sample means is significantly different from the others; with 95% confidence, the four Regional Mean Happiness Scores are not equal.

Application of the Tukey Multiple-Comparison Method

As the null hypothesis has been rejected, a Tukey test can be performed to determine which region's or regions' Regional Mean Happiness Score differs from the other. Technology (R) is used to construct confidence intervals for each of the six pair-wise comparisons of population means. Figure 3 displays the confidence intervals graphically.

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

Figure 3. Family-wise 95% confidence intervals of mean differences for Happiness Scores of selected global regions.



It is noted that CaEE – LAac (Central and Eastern Europe & Latin America and Caribbean) is the only pair of regions whose confidence interval contains zero. It can be concluded that the Mean Happiness Scores of these regions, 5.985 and 5.908, respectively, are possibly equal based on the desired level of significance (or confidence). And as all of the other pair-wise confidence intervals do not contain zero, there is evidence to support that there is a significant difference between the two Regional Mean Happiness Scores being compared in each case.

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

References

Sullivan III, Michael. Statistics: Informed Decisions Using Data; 6th Edition, ISBN-13: 9780135780121.

World Happiness Report, 2021 (Version 2). (April 2021). Compiled by Ajayoal Singh for Kaggle Inc. https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021/version/2?select=world-happiness-report.csv