

Author: Brenda Gunderson, Ph.D., 2015

License: Unless otherwise noted, this material is made available under the terms of the Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License: <http://creativecommons.org/licenses/by-nc-sa/3.0/>

The University of Michigan Open.Michigan initiative has reviewed this material in accordance with U.S. Copyright Law and have tried to maximize your ability to use, share, and adapt it. The attribution key provides information about how you may share and adapt this material.

Copyright holders of content included in this material should contact open.michigan@umich.edu with any questions, corrections, or clarification regarding the use of content.

For more information about how to attribute these materials visit: <http://open.umich.edu/education/about/terms-of-use>. Some materials are used with permission from the copyright holders. You may need to obtain new permission to use those materials for other uses. This includes all content from:

Attribution Key

For more information see: <http://open.umich.edu/wiki/AttributionPolicy>

Content the copyright holder, author, or law permits you to use, share and adapt:



Creative Commons Attribution-NonCommercial-Share Alike License



Public Domain – Self Dedicated: Works that a copyright holder has dedicated to the public domain.

Make Your Own Assessment

Content Open.Michigan believes can be used, shared, and adapted because it is ineligible for copyright.



Public Domain – Ineligible. Works that are ineligible for copyright protection in the U.S. (17 USC §102(b)) *laws in your jurisdiction may differ.



Content Open.Michigan has used under a Fair Use determination
Fair Use: Use of works that is determined to be Fair consistent with the U.S. Copyright Act (17 USC § 107)
*laws in your jurisdiction may differ.

Our determination DOES NOT mean that all uses of this third-party content are Fair Uses and we DO NOT guarantee that your use of the content is Fair. To use this content you should conduct your own independent analysis to determine whether or not your use will be Fair.

Lab 9: One-Way Analysis of Variance (ANOVA)

Objective: In this lab you will perform a **one-way analysis of variance**, often abbreviated **ANOVA**. We have already seen that the two independent samples *t*-test can be used to compare the means of two populations (when the samples are independent). However, when we want to compare the means of *three or more populations*, we turn to ANOVA. You can think of ANOVA as an extension of the two independent sample pooled *t*-test since it compares several population means and **requires** the assumption that the populations have equal variances.

Application: Andy is conducting a test for UM Hospital to explore the effects of a new antibiotic drug. They wish to explore if this drug has the same effects on the mean white blood cell count for the four different age populations: 1 = 0 to 19 years, 2 = 20 to 29 years, 3 = 31 to 40 years, and 4 = 40 years and older. They need a test that would allow them to compare the mean white blood cell count for these four populations. In this example, the number of populations under study is $k=4$, and the total sample size is 24 (6 people in each age group).

Overview: ANOVA is a statistical tool for analyzing how the mean value of a quantitative response (or dependent) variable is affected by one or more categorical variables, known as treatment variables or factors. While ANOVA allows us to compare the means of more than two populations, it can only tell us whether differences appear to exist, not specifically which population means are different. Consequently, the appropriate hypotheses for ANOVA are $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ (that the population mean responses are equal, where k is the number of populations or treatment groups) and H_a : at least one of the population mean responses, μ_i , is different.

Going back to the application, for Andy's test we have:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_a : at least one of the population mean white blood cell counts, μ_i , is different.

As with our other hypothesis tests, several assumptions are required for ANOVA. Andy will need to assume that the white blood cell count measurements in each of the four age populations follows a normal model. To check this assumption, he can create four QQ plots, one for each sample of measurements.

He will also need to assume that the four populations of white blood cell counts have equal variance. The equal population variance assumption can be checked as done in the two independent samples *t*-test – side-by-side boxplots, comparing sample standard deviations, and Levene's test. Further, the data are assumed to consist of independent random samples.

The analysis of variance that Andy will run involves decomposing the total variation of the white blood cell count into two parts: (1) that due to variation among the four sample white blood cell counts (*between groups* variation) and (2) that due to the natural variation of the white blood cell count in each of the four age groups (variation due to error, or *within groups* variation): **SS Total = SS Groups + SS Error**.

If Andy finds that the sum of squares between groups (SS Groups) is large relative to the sum of squares within groups (SS Error), it implies that the model of different white blood cell means explains a significant portion of the observed variability.

In order to determine what is "relatively large," the sum of squares values are divided by their respective degrees of freedom, creating what are called mean square terms. The degrees of freedom for SS Groups is the number of treatment groups, k , minus one ($k - 1$).

$$\text{MS Groups} = \text{SS Groups} / (k - 1)$$

$$\text{MS Error} = \text{SS Error} / (N - k)$$

In Andy's test, the degrees of freedom for SS Groups is equal to 3. For SS Error, the degrees of freedom is the total number of observations, N , minus the number of treatment groups ($N - k$). For Andy's ANOVA procedure, the degrees of freedom for the SS Error is equal to 20.

The ratio of these two mean squares forms the **F-statistic**, which has numerator degrees of freedom ($k - 1$) and denominator degrees of freedom ($N - k$). Note in the equation that MSE stands for MS Error.

$$F = \frac{\text{Variation among sample means}}{\text{Natural variation within groups}} = \frac{\text{MS Groups}}{\text{MSE}}$$

We can view this **F-statistic** as the ratio of two estimators of the common population variance, σ^2 , where the denominator (MSE) is a good (unbiased) estimator, and the numerator (MS Groups) is only good when the H_0 is true (otherwise, it tends to overestimate σ^2). If Andy's data results in a large F value then there is some evidence against the null hypothesis of equal population white blood cell count means.

If Andy rejects the null hypothesis, indicating that at least one of the population mean white blood cell count is different, then we can turn to a **multiple comparisons procedure** for determining which population mean(s) appear to be different and how they differ. The most common method to analyze this is by looking at the set of all *pairwise comparisons*. Two equivalent techniques can be used by Andy for each pair of means: either perform a hypothesis test to see if two population mean white blood cell counts are significantly different, or construct a confidence interval for the difference in population mean white blood cell counts to see whether the value of 0 is contained in the interval. Specifically, a multiple comparisons procedure called **Tukey's procedure**, which is available in most computer packages, controls for the overall Type I error rate (overall significance level) or the overall confidence level.

Formula Card

One-Way ANOVA				
SS Groups = $\text{SSG} = \sum_{\text{groups}} n_i (\bar{x}_i - \bar{x})^2$	MS Groups = $\text{MSG} = \frac{\text{SSG}}{k - 1}$	ANOVA Table		
SS Error = $\text{SSE} = \sum_{\text{groups}} (n_i - 1) s_i^2$	MS Error = $\text{MSE} = s_p^2 = \frac{\text{SSE}}{N - k}$			
SS Total = $\text{SSTO} = \sum_{\text{values}} (x_{ij} - \bar{x})^2$	$F = \frac{\text{MS Groups}}{\text{MS Error}}$			
Confidence Interval	$\bar{x}_i \pm t^* \frac{s_p}{\sqrt{n_i}}$	df = $N - k$	Under H_0 , the F statistic follows an $F(k - 1, N - k)$ distribution.	

Warm-Up: Content of TV Shows

A study examined whether the content of TV shows influence the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned 90 adults to watch one of three programs (30 to each). One program had violent content, another sexual content, and the third neutral content. Each program contained the same nine commercials. After the shows ended, the subjects were asked to recall the brands of products that were advertised. The ANOVA table based on the data is provided below.

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Content	2	11.8	5.90	7.47	0.001
Residuals	87	68.6	0.79		

- a. The researcher would like to assess if there is any effect of the program type on the average number of brands recalled, namely, test $H_0: \mu_1 = \mu_2 = \mu_3$. Provide the appropriate alternative hypothesis in the context of the problem.

H_a : _____

- b. One assumption is the model for number of brands recalled for each population is normal. What graph(s) would you would make to assess this assumption?
- c. At the 5% level, was there a significant effect of the program type on the average number of brands recalled? Explain your answer.

ILP: Is There a Difference Among the Mean Freshman GPAs for Three Different Socioeconomic Classes?

Background: Sociologists often conduct experiments to investigate the relationship between socioeconomic status and college performance. Socioeconomic status is generally partitioned into three classes: lower, middle, and upper. Consider the problem of comparing the mean grade point average (GPA) of college freshmen across the three socioeconomic populations. The GPAs for random samples of seven college freshmen from each of the three socioeconomic classes were selected from a

university's files at the end of the first academic year. The data are in the **gpa.Rdata** data set. (Source: Mendenhall and Sincich, 1996, page 589)

Task: Perform a test to assess whether the population mean freshman GPAs among the three socioeconomic classes differ. If there is sufficient evidence to indicate significant differences, determine which groups differ and how.

Procedure: We want to compare three populations with respect to a quantitative response (GPA). The appropriate inference procedure for this scenario is ANOVA and the value of k for this problem is _____.

Hypothesis Test:

1. **State the Hypotheses:** H_0 : _____

versus H_a : _____ ,

Clearly define the one of the parameters in the null hypothesis in context:

μ_1 represents _____

Determine Alpha: We were told the significance level was 5%.

Remember: Your hypotheses and parameter definition should always be a statement about the **population(s)** under study.

2. **Checking the Assumptions**

We need to assume:

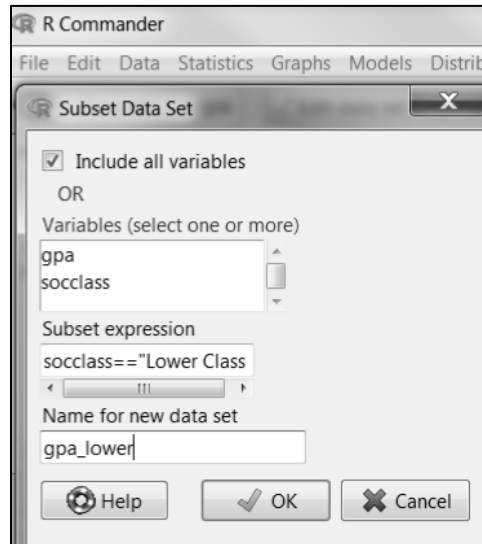
- The k samples are _____ from each other.
 - Each sample is a random sample. To check this assumption, we would make a _____ plot (if there was time order) for each sample.
 - Each sample needs to come from a normally distributed population. To check this assumption, we would make a _____ plot for each sample.
 - All k populations have equal _____ .
- a. Based on the description about how these samples were collected, can we assume we have random and independent samples?

Note that there is no time order for this data. *If there were*, since you need EACH sample to be a random sample, how many time plots would you need to make to check this assumption?

Answer = Need to make _____ time plot(s)

- b. Construct the QQ plots necessary to check the assumption about normally distributed populations. To do this, we have to subset the file for the three subgroups (or classes).

To do this, go to **Data -> Active data set -> Subset active data set**



Enter **socclass=="Lower"** as the subset expression, give an appropriate new data set name **gpa_lower** and click OK. Note that this corresponds to the line command:

```
gpa_lower <- subset(gpa, subset=socclass=="Lower")
```

This should result in the message: NOTE: The dataset gpa_lower has 7 rows and 2 columns.

With this new data set as the active data set, **create the qqplot** by going to **Graphs > Quantile-comparison plot** and select the **gpa** variable.

Alternatively you could type in, then highlight, and submit these commands:

```
qqnorm(gpa_lower$gpa,main="Normal QQ Plot by yournamehere")  
qqline(gpa_lower$gpa)
```

Based on your qqplot, does it appear that the population of GPAs for all students in the lower socioeconomic class is (approximately) normal? Why or why not?

Note: you would need to repeat the above steps for the “Middle Class” and “Upper Class” to complete the normal model for each population assumption checking. You will not need to make the other qqplots here, but when you do repeat these steps, you would need to be sure you go back and select the original gpa data set each time. The resulting three commands lines would look like:

```
gpa_lower <- subset(gpa, subset=socclass=="Lower")
```

```
gpa_middle <- subset(gpa, subset=socclass=="Middle")
gpa_upper <- subset(gpa, subset=socclass=="Upper")
```

3. Compute the Test-Statistic and Calculate the p -value

Test-Statistic

- a. First, test to see if the variances are equal using Levene's test, using **Statistics > Variances > Levene's test...** making sure to select "mean" as the measure of center, not "median."
If the p -value > 0.10 , then we fail to reject the null hypothesis that the variances are equal.

The assumption of equal population variances seems: **valid** **not valid**

- b. Next, generate the ANOVA output using **Statistics > Means > One-way ANOVA**. (Make sure you are using the full **gpa** dataset, not one of the subsets)

The symbol for the estimate of the common population standard deviation is _____,

which for this problem is found to be _____.

- c. What is the value of the test statistic? _____ = _____

- d. What is the distribution of the test statistic if the null hypothesis is true?

Note: This is not the same as the distribution of the population that the data were drawn from, and will be the model used to find the p -value.

Calculate the p -Value:

- e. What is the reported p -value? _____
- f. Draw a picture of the p -value, with labels for the distribution and x-axis.
Use the `pval()` function in R to check your work.

4. Evaluate the p -value and Conclusion

Evaluate the p -value:

What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember: Reject H_0 \Leftrightarrow Results statistically significant
Fail to Reject H_0 \Leftrightarrow Results not statistically significant

Conclusion:

What is your conclusion in the context of the problem?

Note: Conclusions should always include a reference to the population parameter(s) of interest. They should not be too strong; you can say that you have sufficient evidence, but do NOT say that we have *proven* anything true or false.

5. Follow-up Analyses: If ANOVA has indicated that there appears to be significant differences between two or more of groups, we can use a multiple comparison test to tell us which groups appear to be different and by how much.

- a. Obtain the multiple comparisons output using using **Statistics > Means > One-way ANOVA**. This time, click on the “Pairwise comparison of mean” box, which has a default significance level of 0.05. The multiple comparisons output contains both *p*-values and confidence intervals for every possible pairwise comparison of groups; either can be used to determine where differences exist. The *p*-values that are less than or equal to 0.05 or confidence intervals that do NOT contain 0 indicate a difference between those two population means.

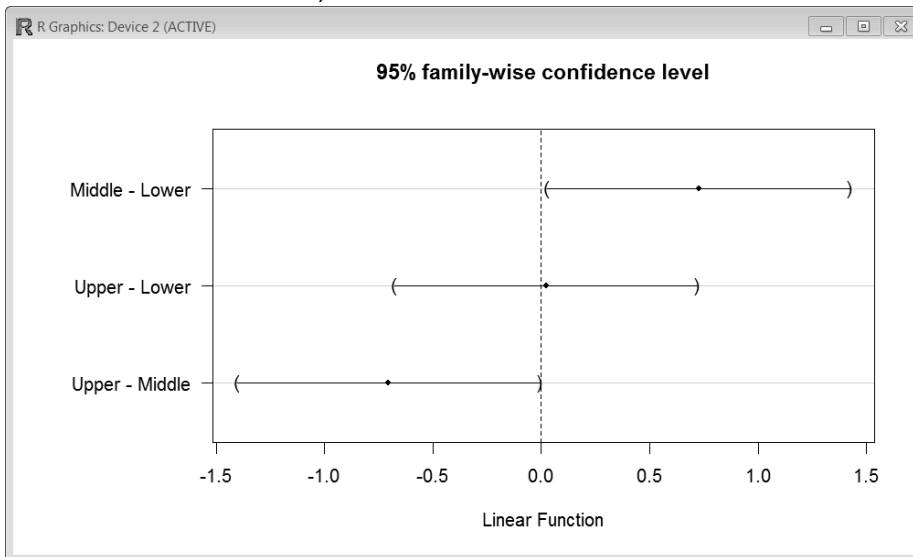
```

95% family-wise confidence level

Linear Hypotheses:
      Estimate   lwr      upr
Middle - Lower == 0  0.727143  0.029443  1.424843
Upper - Lower == 0  0.021429 -0.676271  0.719128
Upper - Middle == 0 -0.705714 -1.403414 -0.008014
  
```

The “estimate” is the estimated difference of means, and the “lwr” and “upr” are the lower and upper bounds of the confidence intervals. If these CI’s contain zero, we fail to reject the null hypothesis that the two means are equal.

You also can see this chart, which illustrates the same information:



- b. Summarize the findings about the differences in population means for the GPAs of freshmen in the different socioeconomic classes. Which pairs are significantly different at the 5% level?

- c. Calculate a 95% confidence interval for the mean GPA for the middle class group. The sample mean GPA for the 7 subjects in the group was 3.25.

$$\text{Confidence Interval} \quad \bar{x}_i \pm t^* \frac{S_p}{\sqrt{n_i}} \quad df = N - k$$

Cool-Down: Check Your Understanding

Complete the following sentences by circling words or filling in blanks as necessary.

- The p -value of 0.025 from this activity implies that if this study were repeated many times, we would see an F test statistic of 4.579 or **greater** **less** in about _____ % of repetitions if the population means were really all equal.
- ANOVA procedures can be thought of as an extension of the two independent samples **pooled** **unpooled** t -test, and hence requires the assumption of equal **population** **sample** variances.
- One way to check this assumption is to use Levene's test and see if the p -value is **greater than** **less than or equal to** 0.10.

