

## Lab 7: ANOVA

### Objectives:

1. Understand how to perform ANOVA  $F$  test for comparing three or more population means.
2. Understand how to use the `aov()` function in R to construct ANOVA tables.

### Definitions:

- ANOVA (analysis of variance)
- treatment groups
- grand mean
- MSTR (mean sum of squares for treatment)
- MSE (mean sum of squares for error)
- $F$ -distribution
- $F$  statistic,  $F$  test

### Introduction:

In Labs 2 and 6, we considered methods for testing claims about two population means, namely, permutation tests and  $t$  tests. In this lab, we consider a technique for testing claims about three or more population means known as *analysis of variance* (ANOVA). The ANOVA procedure compares the variation in the means of samples taken from the populations. The idea is to partition the variability in all the samples into the variability *between* each sample and the variability *within* each sample. If the population means are indeed equal, then the variability between and within each sample should be roughly the same. The ratio of the between and within variability provides a test statistic, and the classical approach, which we consider in this lab, uses a theoretical sampling distribution. In the next lab, we will develop a permutation test approach for performing ANOVA.

### Activities:

**Getting Organized:** *If you are already organized, and remember the basic protocol from previous labs, you can skip this section.*

Navigate to your class folder structure. Within your "Labs" folder make a subfolder called "Lab7". Next, download the lab notebook .Rmd file for this lab from Blackboard and save it in your "Lab7" folder. There are no datasets used in this lab. You will be working with the `Zombies.csv` data set on this lab. You should download the data file into your "Lab7" folder from Blackboard.

Within RStudio, navigate to your "Lab7" folder via the file browser in the lower right pane and then click "More > Set as working directory". Get set to write your observations and R commands in an R Markdown file by opening the "lab7\_notebook.Rmd" file in RStudio. Remember to add your name as the author in line 3 of the document. For this lab, enter all of your commands into code chunks in the lab notebook. You can still experiment with code in an R script, if you want. To set up an R Script in RStudio, in the upper left corner click "File > New File > R script". A new tab should open up in the upper left pane of RStudio.

**Notation:** Before we see how to perform an ANOVA test in R/RStudio, let's formally set up the procedure, starting with defining the notation:

- $G$  denotes the number of populations/samples
- $n_i$  denotes the number of observations in the  $i^{\text{th}}$  sample,  $i = 1, \dots, G$
- $n = n_1 + \dots + n_G$  denotes the total number of observations
- $X_{ij}$  denotes the  $j^{\text{th}}$  observation in the  $i^{\text{th}}$  sample,  $j = 1, \dots, n_i$
- $\bar{X}_i$  denotes the mean of the  $i^{\text{th}}$  sample
- $\bar{X}_{..}$  denotes the *grand mean*, i.e., the mean of all  $n$  observations in each sample

If we let  $\mu_i$  denote the mean of the  $i^{\text{th}}$  population, then we are testing the following hypotheses with the above sample data:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_G \quad \text{vs.} \quad H_A : \text{at least one } \mu_i \text{ is different, i.e., } \mu_i \neq \mu_j \text{ for some } i \neq j$$

As discussed in the introduction above, we test these hypotheses by comparing the variability between the sample means to the variability within each sample. For the variability between the samples, we use the *mean sum of squares for treatment* (MSTR), which is given by

$$\text{MSTR} = \frac{1}{G-1} \sum_{i=1}^G n_i (\bar{X}_i - \bar{X}_{..})^2.$$

For the variability within the samples, we use the *mean sum of squares for error* (MSE), which is given by

$$\text{MSE} = \frac{1}{n-G} \sum_{i=1}^G \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

### Pause for Reflection #1:

Four chemical plants, producing the same products and owned by the same company, discharge liquid waste into streams in the vicinity of their locations. To monitor the extent of pollution created by the liquid waste and determine whether this differs from plant to plant, the company collected random samples of liquid waste from each plant, resulting in the following data.

Plant	Polluting Effluents (lb/gal of waste)	Sample Mean
A	1.65 1.72 1.50 1.37 1.60	1.568
B	1.70 1.85 1.46 2.05	1.765
C	1.40 1.75 1.38 1.65 1.55	1.546
D	2.10 1.95 1.65 1.88	1.895

- State the hypotheses we will test to determine if there is a difference in the mean weight of polluting effluents per gallon in the liquid waste discharged from the four plants. Be sure to define your notation.
- Identify what the values of  $G$  and  $n$  are, and for each sample identify the values of  $n_i$  and  $\bar{X}_i$  are.
- Finally, find the values of the grand mean  $\bar{X}_{..}$  and the variability between the samples' MSTR.

**The ANOVA  $F$  Test:** If  $H_0$  is true, i.e., the population means are all equal, then the variability between the samples should be roughly the same as the variability within the samples (assuming also that the populations have equal variance). If  $H_0$  is false, then the variability between the samples will be larger than the variability within the samples. Thus, we use the ratio of the between and within variability measures as the test statistic,

$$F = \frac{\text{MSTR}}{\text{MSE}},$$

which has a  $F$  distribution with  $(G-1)$  and  $(n-G)$  df. The observed test statistic based on the sample data obtained is denoted  $f$ , and then its associated  $P$ -value is calculated using the  $F$  distribution as follows:

$$P\text{-value} = P(F \geq f)$$

Note that the  $P$ -value is given by the probability of obtaining a test statistic as large or larger than what was observed, i.e., the  $P$ -value for an ANOVA  $F$  test is always a right-tail probability. This is because "more extreme" in this context would be sample data that produced more between sample variability resulting in a larger ratio of MSTR to MSE.

### Pause for Reflection #2:

Return to the pollution example and compute the observed  $F$  statistic using the value of MSTR you found in Reflection #1 and given that  $\text{MSE} = 0.03336$ . Then use the following code (with the corresponding values of  $f$ ,  $G-1$ , and  $n-G$  substituted in) to

calculate the corresponding  $P$ -value:

```
pf(f, G-1, n-G, lower.tail = FALSE)
```

Based on the  $P$ -value, do the data provide sufficient evidence to indicate a difference in the mean weight of polluting effluents per gallon in the liquid waste discharged from the four plants?

**The ANOVA Table:** As we can see, there are a lot of calculations that go into performing ANOVA. The ANOVA table given below is a tool that summarizes and organizes these calculations in an easy to use format.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Factor	G - 1	$\sum n_i (\bar{X}_{i.} - \bar{X}_{..})^2$	MSTR	MSTR/MSE	P-value
Error	n - G	$\sum \sum (X_{ij} - \bar{X}_{i.})^2$	MSE		
Total	n - 1	$\sum \sum (X_{ij} - \bar{X}_{..})^2$			

Notice how the ANOVA table is arranged:

- The last two columns are the most useful, **SINCE** they give the test statistic and its associated  $P$ -value:
  - the last column with heading  $\text{Pr}(>F)$  gives the  $P$ -value, so it is easy to read off;
  - the second to last column with heading **F value** gives the observed  $F$  statistic, which the  $P$ -value is based on.
- The other columns give the supporting calculations used to find the test statistic and  $P$ -value:
  - the first column provides labels for the source of variability, where **Factor** corresponds to the between samples variability and **Error** corresponds to the within sample variability;
  - the second column gives the corresponding degrees of freedom within each row, note that the sum of the **Factor** and **Error** df equals the **Total** df;
  - the third column with heading **Sum Sq** gives the sum of squares corresponding to each source, note that the sum of squares for the **Factor** and **Error** add up to the **Total** sum of squares (this is where the partitioning of the variability occurs that makes ANOVA possible);
  - the fourth column with heading **Mean Sq** gives the mean sum of squares corresponding to each source, note that these are found by dividing the sum of squares in each row by the corresponding df.

### Pause for Reflection #3:

Copy the following partial ANOVA table for the pollution example into your lab notebook and fill in the missing values, denoted by ---. Upload an image of your work into your lab notebook.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Factor	---	---	---	---	---
Error	---	---	0.03336		
Total	---	---			

**Performing the ANOVA  $F$  Test in R:** Thankfully, there is a function in R that performs the extensive calculations needed to perform ANOVA, given by `aov()`. Calling the `aov()` function on the data performs the calculations, and then using the `summary()` function on the results constructs the ANOVA table. The following code demonstrates how this works in the pollution example. The first step is to format the data in R. Notice that an object `Plant` is created to store labels for the observed waste amounts so that we can sort the observations into the appropriate *treatment groups* corresponding to the four populations given by the four plants.

```
# format the data in R
Waste = c(1.65, 1.72, 1.50, 1.37, 1.60,
          1.70, 1.85, 1.46, 2.05,
          1.40, 1.75, 1.38, 1.65, 1.55,
          2.10, 1.95, 1.65, 1.88)
Plant = rep(c("A", "B", "C", "D"), c(5, 4, 5, 4))

# Perform the ANOVA:
results = aov(Waste ~ Plant)      # store the ANOVA calculations in results
summary(results)                 # construct the ANOVA table
```

By running the above code for yourself (already provided in the Lab 7 Notebook), you can check your answers to Reflection #3.

#### Pause for Reflection #4:

In the above code, explain what the following line does:

```
Plant = rep(c("A", "B", "C", "D"), c(5, 4, 5, 4))
```

In particular, what does the function `rep()` do?

**Zombies:** Let's look at another example to see how to use the `aov()` function given a data set. The `Zombies.csv` file contains data about the number of zombies killed (killed) and by what household weapon (weapon) for a sample of 31 apocalypse survivors. Load the data and view it:

```
Zombies = read.csv("Zombies.csv")
View(Zombies)
```

#### Pause for Reflection #5:

Conduct some EDA:

- What are the mean and standard deviation of zombies killed across weapons (hint: the `tapply()` function will be useful)?
- How many observations of zombies killed are there for each of the weapons (hint: the `table()` function will be useful)?
- Create side-by-side boxplot to compare the distributions of zombies killed across weapons.

From the EDA, it sure looks as though there are differences in the number of zombies killed by each weapon, but are these differences due to sampling error, or do they represent real differences in zombie-killing effectiveness? To answer that question, we need to run an ANOVA test.

```
aov = aov(killed ~ weapon, Zombies)      # store results of ANOVA test
summary(aov)                             # view the ANOVA table
```

### Pause for Reflection #6:

State the hypotheses being tested by the above ANOVA calculations. Report the  $P$ -value you find and state the conclusion.

---

**Assumptions for the ANOVA  $F$  Test:** In performing the ANOVA  $F$  test, the following assumptions are made:

- the samples are independent
- the populations are normally distributed
- the populations have equal variance

The independence assumption is critical, if the samples are related in some way then a different procedure is needed. Violations of the assumptions of normality and equal variances are less important.

The big problem with non-normality in  $t$  tests is the effect of skewness on one-sided tests. But ANOVA tests are inherently two sided (we are testing for any differences between means, not differences in one direction) so non-normal distributions generally have little effect as long as the sample sizes are reasonably large.

If the sample sizes  $n_i$  are roughly equal, then unequal variances do not have a great impact, but if the population variances differ, then the actual sampling distribution of the  $F$  statistic could be very different from an  $F$  distribution. In particular, if there is a small sample from a population with large variance, then the  $F$  statistic can explode.

We will run simulations to explore the assumptions for ANOVA: in particular, how does "un-balancedness" (sample sizes not the same) and unequal population variances affect the outcome? We consider the hypotheses

$$H_0 : \mu_A = \mu_B = \mu_C \quad \text{vs.} \quad H_A : \text{at least one mean is different.}$$

The code below simulates drawing three random samples from populations (called  $A, B, C$ ) with the same mean ( $\mu = 20$ ) and standard deviation ( $\sigma = 3$ ) and then performs an ANOVA test. Using a significance level of 0.05, the object `counter` keeps track of how many times the null hypothesis is incorrectly rejected (false positive) and then corresponding proportion is computed.

```
n.A = 50                                # set sample sizes
n.B = 50
n.C = 50

Group = rep(c("A","B","C"), c(n.A, n.B, n.C)) # create group labels

counter = 0
N = 10^4

for (i in 1:N)
{
  a = rnorm(n.A, 20, 3)                  # Draw samples from N(20,3) pop
  b = rnorm(n.B, 20, 3)
  c = rnorm(n.C, 20, 3)
  X = c(a, b, c)                         # Combine into one vector

  Pvalue = summary(aov(X ~ Group))[1,5]  # Extract P-value from ANOVA table
  if (Pvalue < 0.05)                     # Reject H0, at 0.05 sig level?
    counter = counter + 1                # If yes, increase counter
}

counter/N                                # proportion of times H0 rejected
```

### Pause for Reflection #7:

What type of error is `counter` keeping track of? Is the proportion given by `counter/N` close to what you would expect the probability of making that type of error to be?

---

### Pause for Reflection #8:

Alter the code so that the sample size from  $A$  is 10 (`n.A = 10`) and redo the simulation. What happens to the proportion of times  $H_0$  is rejected?

---

### Pause for Reflection #9:

Alter the code again by increasing the standard deviation of population  $A$  to 9 and trying samples of size 50 and 10 (keeping the other sample sizes to 50). What proportion of times do you reject the null hypothesis in each case?

---

### Pause for Reflection #10:

Explore other scenarios: What if the population means are all different, but the population variances are the same? How do sample sizes affect the outcome? Try with all sample sizes the same and then unequal. Now try different variances and again, with balanced and unbalanced samples.

Record in your lab notebook what scenarios you tried and what results you found, i.e., how the proportion of times  $H_0$  was rejected is impacted.

---

---

Lab 7: ANOVA is shared under a [not declared](#) license and was authored, remixed, and/or curated by LibreTexts.