

Ch5 - Comparing Two or More Population Means

Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East West University, Dhaka
Last Updated December 16, 2023

Outline

Outline

1. Comparing Two Means

- σ_1^2 and σ_2^2 known
- σ_1^2 and σ_2^2 unknown

2. Comparing Several Means (ANOVA)

- 1. The One-Way Layout
- 2. Two - Way Layout
- Two - Way ANOVA without replication

What's Next!

- ▶ This chapter will be about *comparing means of two or more populations* assuming the *population distribution is Normal*.
- ▶ When we compare two means, in this case we can perform either z test or t-test, this is called *two-sample test*, for example we can test

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

- ▶ However we will discuss that one of the key issues in t-test is, we need to impose assumptions on *variances of the two populations*. If we assume variances are known, the problem is not difficult, but if the variances are unknown it becomes a bit complicated.
- ▶ In that case we can assume two populations have variances with known ratio k , i.e., $\sigma_1^2 / \sigma_2^2 = k$, for example, $k = 1$, then we have $\sigma_1^2 = \sigma_2^2$ or equal variances. This is also easy to handle.
- ▶ But if we don't assume anything about the variances and we don't know their ratio, then the problem becomes very hard and still there is no known distribution of the test statistic in finite samples. This problem is famously known as *Behrens-Fisher problem*, there are some way to get some approximations we will see that in coming sections.

What's Next!

- Here are two pictures you should have in mind when we compare two populations.

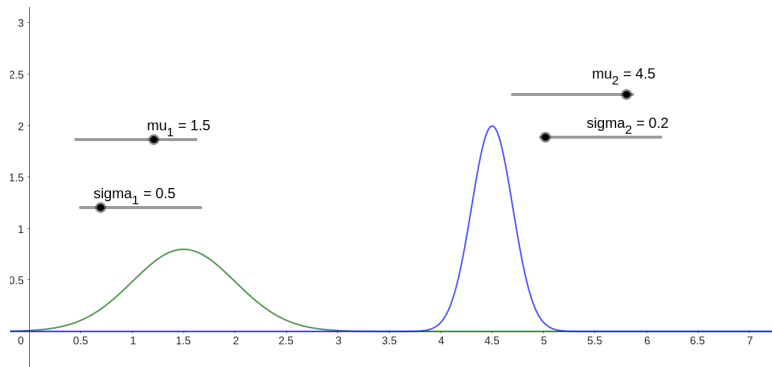


Figure 1: Two sample test - For example, we would like to test whether $\mu_1 = \mu_2$, where μ_1 and μ_2 are unknown to us.

What's Next!

- For more than two populations, for example for p populations we will assume all the population variances are equal, for example they are equal to σ^2 , and do the following test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

H_a : at least one of the μ_i 's is different or H_0 is not true

- Here we will use the ANOVA technique developed by Ronald Fisher (1890 - 1962).
- In this case we will form F statistic and the *under the Null this statistic will follow F distribution* with some numerator and denominator degrees of freedom.
- The slides of this chapter will closely follow chapter 9.6 and chapter 11 [DeGroot and Schervish \(2012\)](#) (probably one of the best textbooks in Statistics). You can find similar treatments in [Rice \(2007\)](#). We will solve some examples from [Anderson et al. \(2020\)](#).

What's Next!



Figure 2: Young Ronald Fisher (Source - Wiki). For his work in statistics, he has been described as "a genius who almost single-handedly created the foundations for modern statistical science."

- A quote from Ronald Fisher (1890 - 1962) - *"In Scientific subject the natural remedy for dogmatism has been found in research."*

What's Next!

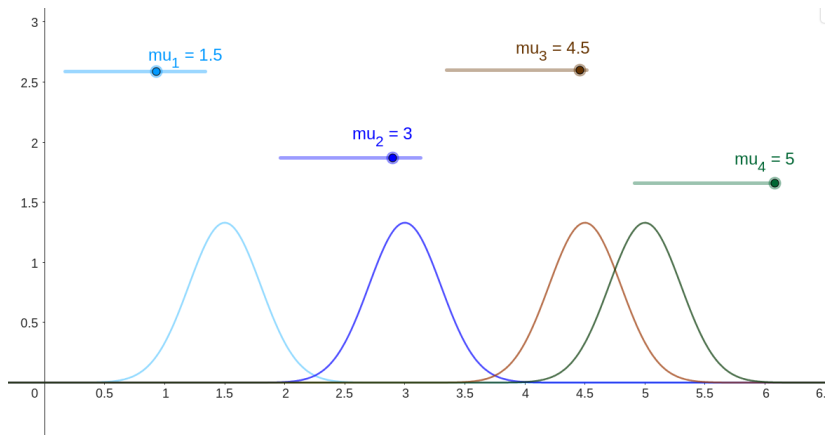


Figure 3: For more than two samples, we would like to test whether $\mu_1 = \mu_2 = \mu_3 = \mu_4$, where μ_1, μ_2, μ_3 and μ_4 are unknown to us.

► So let's get started 🚶 🚶 🚶 ...

1. Comparing Two Means

- σ_1^2 and σ_2^2 known
- σ_1^2 and σ_2^2 unknown

2. Comparing Several Means (ANOVA)

- 1. The One-Way Layout
- 2. Two - Way Layout
- Two - Way ANOVA without replication

Comparing Two Means

Comparing Two Means

σ_1^2 and σ_2^2 known

Two-Sample Test

Known Variances

- ▶ The *two-sample test* is very similar to the *one-sample test* that we have seen in Chapter 2.
- ▶ The idea is we will form a statistic, then under the Null with Normality assumption we expect that this statistic will follow normal distribution or t -distribution (with certain Df) in finite samples.
- ▶ Of course under no distributional assumption, our only hope is we have large samples, then the statistic will follow normal distribution, but here we are interested to use the result in finite samples or we would like to use *exact distribution*.
- ▶ Let's write the problem. Suppose we have two random samples,
 - ▶ $(Y_{11}, Y_{12}, \dots, Y_{1n_1})$ form a random sample of n_1 observations from a normal distribution for which both the mean μ_1 and the variance σ_1^2 are unknown, so in this case for all $i = 1, 2, \dots, n_1$, $Y_{1i} \sim \mathcal{N}(\mu_1, \sigma_1^2)$.
 - ▶ Then we also have $(Y_{21}, Y_{22}, \dots, Y_{2n_2})$ form a random sample of n_2 observations from a normal distribution for which both the mean μ_2 and the variance σ_2^2 are unknown, so in this case for all $i = 1, 2, \dots, n_2$, $Y_{2i} \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
- ▶ You can think about following table as a random sample,

Two-Sample Test

Known Variances

1 st Sample	2 nd Sample
Y_{11}	Y_{21}
Y_{12}	Y_{22}
\vdots	\vdots
\vdots	Y_{2n_2}
Y_{1n_1}	\vdots

- ▶ Since this is a random sample, you can think all random variables in the first column are independent and follows the same distribution $\mathcal{N}(\mu_1, \sigma_1^2)$.
- ▶ Similarly all random variables in the second column are independent and follows the same distribution $\mathcal{N}(\mu_2, \sigma_2^2)$.
- ▶ And second column is completely independent of the first column.

Two-Sample Test

Known Variances

- Using sample we can calculate sample means

$$\bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i} \quad \text{and} \quad \bar{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{2i}$$

- Under Normality assumption, it is possible to show,

$$\bar{Y}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2/n_1) \quad \text{and} \quad \bar{Y}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2/n_2)$$

- Now we can assume σ_1^2 and σ_2^2 are known, then we can show that

$$\bar{Y}_1 - \bar{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- Or equivalently,

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

Two-Sample Test

Known Variances

- ▶ We can use this distributional result both for testing and constructing confidence intervals. We will focus on testing now. For testing the general structure of the z-statistic will be

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- ▶ Again we can repeat the same old story as Chapter 2,
- ▶ We can do a two - tail test

$$H_0 : \mu_1 - \mu_2 = 0 \quad (1)$$

$$H_a : \mu_1 - \mu_2 \neq 0 \quad (2)$$

- ▶ Or a lower - tail test

$$H_0 : \mu_1 - \mu_2 \geq 0 \quad (3)$$

$$H_a : \mu_1 - \mu_2 < 0 \quad (4)$$

- ▶ Or an upper tail test,

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad (5)$$

$$H_a : \mu_1 - \mu_2 > 0 \quad (6)$$

Two-Sample Test

Known Variances

- In all cases the test statistic will be same and that is

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Only the rejection region (hence the critical values) will be different, we already know this!
- When you calculate this for a sample you should write z_{calc} , then you need to find critical values and do the test (or follow the p-value approach!)

Two-Sample Test

Known Variances

- If we assume the population variances are equal, meaning $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the standard errors will slightly change, this will be

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- Now the distributional result will be adjusted to,

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim \mathcal{N}(0, 1)$$

- and the z-statistic for the test in (1), (5) and (3) will be

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- The rest of the procedures are exactly same as before.

Comparing Two Means

σ_1^2 and σ_2^2 unknown

Two-Sample Test

t - test

- ▶ When we don't know the population variances (which is a realistic case), we don't know standard error, then we cannot use the z-statistic anymore, what's the solution?... use some kind of t-statistic, where we will estimate the standard error.
- ▶ Recall the idea of the t-statistic is just to replace the standard error with the estimate (or estimator in repeated sampling) of the standard error.
- ▶ Here we *need some assumptions*, if we *assume equal variances* this means $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then we can propose following estimator,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 \quad \text{and} \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2$$

- ▶ This is sometimes called the *pooled sample variance*, where we calculated the sample variances separately and then pooled them together.
- ▶ Now we can propose the following t-statistic

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (7)$$

Two-Sample Test

t - test

- ▶ it is possible to show,

$$t \sim t_{n_1+n_2-2}$$

- ▶ This means the random variable in (7) is distributed as t distribution with $n_1 + n_2 - 2$ degrees of freedom.
- ▶ **Warning :** I know that in Chapter 2, I wrote T_n with upper case T , this was to clearly mention that it's a random variable and we know its distribution for any n , but actually the convention is to write with lower case t , this is why I am writing t here, but you should always keep in mind that it's a random variable and its distribution is $t_{n_1+n_2-2}$
- ▶ When you calculate, you can definitely write t_{calc} , otherwise there will be other t values which are critical values and you might get yourself confused.
- ▶ This *equal variance case for the two-sample t-test* can be easily extended to the situation when we know the ratio of the two variances, i.e., we know k , where $k = \frac{\sigma_1^2}{\sigma_2^2}$, we will not cover it here, you can see it in [DeGroot and Schervish \(2012\)](#) if you are interested!

Two-Sample Test

t - test

- If we assume two populations have unknown variances and we don't know their ratio, then you may think you will use following standard error,

$$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

- and maybe you can think about the following *t*-statistic,


$$t^* = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (8)$$

- However in this case, in finite sample, this statistic does not follow any known distribution, so we cannot use anything to get critical values. This is a very hard problem and still an open problem, it is famously known as *Behrens-Fisher problem*.
- In this case the work around is to use *some approximation*. The most popular method is use t^* and take critical values from *t* distribution with the Df given by the *Welch-Satterthwaite* approximation given below

$$Df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}$$

Two-Sample Test

t - test

- ▶ It has been shown that the distribution of the t^* can be closely approximated by the t distribution with Df by the above formula.
- ▶ But don't worry you don't have to memorize this. In  the function `t.test()` function implements this test for unequal variances, so we will see how to do this in coming sections.

1. Comparing Two Means

- σ_1^2 and σ_2^2 known
- σ_1^2 and σ_2^2 unknown

2. Comparing Several Means (ANOVA)

- 1. The One-Way Layout
- 2. Two - Way Layout
- Two - Way ANOVA without replication

Comparing Several Means (ANOVA)

Comparing Several Means

ANOVA

- So far, we studied methods for comparing the means of two normal distributions. In this section, we shall consider the method which will help us to compare the means for two or more normal distributions.

Comparing Several Means (ANOVA)

1. The One-Way Layout

The One-Way Layout of ANOVA

- ▶ Here is the problem setup, suppose now we have random samples from p different normal distributions and each of these distributions has the same variance σ^2 , and the means of the p distributions are to be compared now.
- ▶ We will assume for $i = 1, \dots, p$, the random variables Y_{i1}, \dots, Y_{in_i} , form a random sample of n_i observations from the Normal distribution with mean μ_i and variance σ^2 , and the values of μ_1, \dots, μ_p and σ^2 are unknown.

1 st Sample	2 nd Sample	3 rd Sample	...	p^{th} Sample
Y_{11}	Y_{21}	Y_{31}	...	Y_{p1}
Y_{12}	Y_{22}	Y_{32}	...	Y_{p2}
\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	Y_{2n_2}	\vdots	\vdots	Y_{pn_p}
Y_{1n_1}		Y_{3n_3}		

- ▶ The sample sizes n_1, \dots, n_p are not necessarily the same. We shall let $n = \sum_{i=1}^p n_i$ denote the total number of observations in the p samples, and we shall assume that all n observations are independent.
- ▶ It follows from the assumptions we have just made that for $j = 1, \dots, n_i$ and $i = 1, \dots, p$, we have

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2) \text{ for all } j = 1, 2, \dots, n_i$$

The One-Way Layout of ANOVA

- ▶ This means $\mathbb{E}(Y_{ij}) = \mu_i$ and $\mathbb{V}\text{ar}(Y_{ij}) = \sigma^2$.
- ▶ For $i = 1, \dots, p$, we shall let \bar{Y}_i denote the sample mean of the n_i observations in the i th sample. Thus,

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

- ▶ Before we develop an appropriate test procedure, we shall carry out some preparatory algebraic manipulations. First, define the *overall average of all n observations*,

$$\bar{\bar{Y}} = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^p n_i \bar{Y}_i,$$

- ▶ We shall *partition the total sum of squares* into two smaller sums of squares, each of which will be associated with a certain type of variation among the n observations, first note,

$$SS_{\text{Total}} = \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{\bar{Y}})^2 \quad (9)$$

- ▶ Note that SS_{Total}/n would be the estimator of σ^2 if we believed that all of the observations came from a single normal distribution rather than from p different normal distributions. This means that we can interpret SS_{Total} as an overall measure of variation between the n observations.

The One-Way Layout of ANOVA

- ▶ Here are the two terms which we can use to partition SS_{Total} into two smaller sums of squares,

- ▶ First SS_{Between} , this represents how much is the *the variation between the p different samples*,

$$SS_{\text{Between}} = \sum_{i=1}^p n_i (\bar{Y}_i - \bar{\bar{Y}})^2.$$

- ▶ Second SS_{Within} , this represents how much is the *the variation between the observations within each of the samples*,

$$SS_{\text{Within}} = \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2,$$

- ▶ And it is possible to show that,

$$SS_{\text{Total}} = SS_{\text{Within}} + SS_{\text{Between}}$$

or in other words,

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{\bar{Y}})^2 = \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^p n_i (\bar{Y}_i - \bar{\bar{Y}})^2$$

- ▶ The test of the hypotheses that we shall develop will be based on *the ratio of these two measures of variation*. For this reason, we have the name *analysis of variance*.

The One-Way Layout of ANOVA

Important Remarks Regarding the three SS terms:

- ▶ SS_{Total} can be regarded as the *total variation of the observations around overall mean*.
- ▶ SS_{Within} can be regarded as the *total variation of the observations around their particular sample means*, or the total residual variation within the samples. So if this high then this means the different sample means are not doing a good job and residuals are high. Maybe we should consider the same population mean and calculate the overall mean.
- ▶ SS_{Between} can be regarded as the *total variation of the sample means around the overall mean*, or the variation between the sample means. If this is high then this means there is a lot of variation between the sample means, and we should consider different population means.
- ▶ This partitioning is summarized in a table, which is called the *ANOVA table for the one-way layout* and is presented below.

Source of variation	Degrees of freedom	Sum of squares	Mean square
Between samples	$p - 1$	SS_{Between}	$SS_{\text{Between}} / (p - 1)$
Within Samples	$n - p$	SS_{Within}	$SS_{\text{Within}} / (n - p)$
Total	$n - 1$	SS_{Total}	

- ▶ The numbers in the "Mean square" column are just the sums of squares divided by the degrees of freedom. They are used for testing the hypotheses.

The One-Way Layout of ANOVA

- Now with the Mean Squares we can construct following test statistic, which is a F-statistic,

$$F = \frac{SS_{\text{Between}} / (p - 1)}{SS_{\text{Within}} / (n - p)} \quad (10)$$

- And under the Null (with normality assumption and equal variance), it is possible to show that this statistic follows F distribution with $p - 1$ and $n - p$ degrees of freedom. We can write it as,

$$F \sim F_{p-1, n-p} \quad (11)$$

- So to do the test we can find a critical value and if F is greater than the critical value we reject the null hypothesis.
- What's the intuition of this test? - we already explained this, note that when the null hypothesis H_0 is NOT true (i.e., population means are different), then on average the numerator of the F statistic in (10) will be larger than the denominator.

The One-Way Layout of ANOVA

- ▶ We will see how to use ANOVA technique using a data set from a randomized experiment. In particular *randomized experiment with one factor* (Hence the name *one-way ANOVA*) (side note, under the assumptions that we stated we can also use this technique for observational studies!)
- ▶ Question: What is a randomized study or *randomized experiment*? What's the difference between a randomized experiment and an observational study?
- ▶ In a randomized experiment, the researcher randomly assigns some subjects to different levels of *treatments*. And the goal for a randomized experiment is to study the effect of a *treatment* (cause) on a *response* (effect) variable.
- ▶ Note that with this we can actually answer the casual effect of the treatment on the response variable.
- ▶ The crucial difference between an observational study and a randomized experiment is that in a randomized experiment, the researcher controls everything, and since the subjects are randomly assigned to different levels of the treatment, the researcher can make causal claims.
- ▶ However in an observational study, the researcher does not have control who gets the treatment, and just observes the subjects and collects the data. So in this case we cannot make causal claims unless we have some additional information / strong assumptions.
- ▶ Because usually the treatment variable is categorical, often these are also called *factor* variable. And the response variable is usually a quantitative.

The One-Way Layout of ANOVA

- ▶ Here is an example - suppose a tutoring service plans to offer two new programs for their students, service A and service B. To check on the effectiveness of these services 21 students were randomly chosen, then out of this 21
 - ▶ 7 students were randomly assigned to *service A*,
 - ▶ 7 students were randomly assigned to *service B*, and
 - ▶ the remaining 7 did not take the service.

Their scores on the examination, expressed out of 100, are given in the following table.

Service A	Service B	No Service
61	91	99
65	88	83
71	74	75
52	56	95
67	76	62
58	81	69
78	66	80

- ▶ We want to test the null hypothesis that the *three population mean* of scores are same? Question what does this imply if the population means are same?
- ▶ Now let's generate the ANOVA table in excel, you should get following ANOVA table

The One-Way Layout of ANOVA

Service A	Service B	No Service				
61	91	99				
65	88	83				
71	74	75				
52	56	95				
67	76	62				
58	81	69				
78	66	80				
Anova: Single Factor						
SUMMARY						
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
Service A	7	452	64.57143	73.61905		
Service B	7	532	76	149.6667		
No Service	7	563	80.42857	177.2857		
ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	937.2381	2	468.619	3.509629	0.05164	3.554557
Within Groups	2403.4286	18	133.5238			
Total	3340.6667	20				

- Here $SS_{Between} = 937.2381$, $SS_{Within} = 2403.4286$ and $SS_{Total} = 3340.6667$. The degrees of freedom are $p - 1 = 3 - 1 = 2$ and $n - p = 21 - 3 = 18$, and the mean squares are $MS_{Between} = 468.6190$ and $MS_{Within} = 133.5238$, the F statistic is $F = 3.509629$ and the p -value is 0.05, so this means at 5%, we do not have enough evidence to reject the null hypothesis that the three population means are different.
- Note, we calculated using Excel, but you should be able to manually calculate this,

The One-Way Layout of ANOVA

- First calculate the sample means for each of the three samples, you will get $\bar{Y}_1 = 64.57$, $\bar{Y}_2 = 76$ and $\bar{Y}_3 = 80.42$. Then you should calculate the overall mean $\bar{\bar{Y}} = 73.667$. With this you can easily calculate the SS_{Total} and SS_{Between} .
- There is an interesting relation between ANOVA and MLR, infact ANOVA is a special case of MLR. To see, you can create a column where you put all Y values and then run a dummy variable regression (create two dummies for the three groups). You should get the same results.

Comparing Several Means (ANOVA)

2. Two - Way Layout

Comparing Several Means (ANOVA)

Two - Way ANOVA without replication

The TwoWay Layout of ANOVA

- ▶ In the one-way layout, we analyzed p samples, which differed in one way (i.e., one factor or one treatment), and we were interested in comparing the means of the p different populations.
- ▶ But in sometimes there might be more than one factor or treatment that might affect the response variable. In this case we need to use the two-way layout of ANOVA.
- ▶ Following is the generic two-way layout, note we have two factors A and B , sometimes one of the factor is also called a *block* (specially in a factorial experiment, more on this later!) .

Factor A (Block)	Factor B (Groups)			
	1	2	...	J
1	Y_{11}	Y_{12}	...	Y_{1J}
2	Y_{21}	Y_{22}		Y_{2J}
\vdots				
I	Y_{I1}	Y_{I2}		Y_{IJ}

- ▶ Because now we have two factors the notation becomes a little messy, but trust me things are not that complicated! In the table in general cells are represented with Y_{ij} , the first index is for block and the second index is for factor B .
- ▶ We shall assume that there are I possible different values, or different levels, of factor A , so $i = 1, \dots, I$. And there are J possible different values, or different levels, of factor B . This means $j = 1, \dots, J$.
- ▶ The total sample size will be $n = I \times J$.

The Two-Way Layout of ANOVA

- ▶ First note we can get three means here
 - ▶ $\bar{Y}_{i.}$ - the mean of the i th block (summing over the J factor levels)
 - ▶ $\bar{Y}_{.j}$ - the mean of the j th factor (summing over the I blocks) and
 - ▶ $\bar{\bar{Y}}$ - the overall mean.
- ▶ The SS terms are

$$SS_{\text{Total}} = \sum_{i=1}^I \sum_{j=1}^J \left(Y_{ij} - \bar{\bar{Y}} \right)^2$$

$$SS_{\text{Between A}} = \sum_{i=1}^I J \left(\bar{Y}_{i.} - \bar{\bar{Y}} \right)^2$$

$$SS_{\text{Between B}} = \sum_{j=1}^J I \left(\bar{Y}_{.j} - \bar{\bar{Y}} \right)^2$$

$$SS_{\text{Error}} = \sum_{i=1}^I \sum_{j=1}^J \left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{\bar{Y}} \right)^2$$

It can be shown that

$$SS_{\text{Total}} = SS_{\text{Between A}} + SS_{\text{Between B}} + SS_{\text{Error}}$$

- ▶ To give a little bit more explanation, note that
 - ▶ SS_{Total} - total variation of all the observations around the overall mean $\bar{\bar{Y}}$,

The Two-Way Layout of ANOVA

- ▶ $SS_{\text{Between A}}$ - variation of the block means around the overall mean $\bar{\bar{Y}}$,
- ▶ $SS_{\text{Between B}}$ - variation of the factor means around the overall mean $\bar{\bar{Y}}$,
- ▶ SS_{Error} - variation of the observations around their particular block means, factor means and overall mean .

- ▶ And it is possible to show that,

$$SS_{\text{Total}} = SS_{\text{Between A}} + SS_{\text{Between B}} + SS_{\text{Error}}$$

or in other words,

$$\sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{\bar{Y}})^2 = \sum_{i=1}^I J (\bar{Y}_{i.} - \bar{\bar{Y}})^2 + \sum_{j=1}^J I (\bar{Y}_{.j} - \bar{\bar{Y}})^2 + \sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{\bar{Y}})^2$$

- ▶ Here we can perform two kinds of testing

- ▶ Test - 1: First we can do a test for the *factor A*, i.e., we can test if the population means for the I blocks are same.
- ▶ Test - 2: Second we can do a test for the *factor B*, i.e., we can test if the population means for the J groups are same.

- ▶ To do Test - 1, we can construct the following F statistic,

$$F_{\text{Factor A}} = \frac{SS_{\text{Between A}} / (I - 1)}{SS_{\text{Error}} / (I - 1)(J - 1)}$$

- ▶ Under the null hypothesis that the population means for the I blocks are same, this statistic follows F distribution with $I - 1$ and $(I - 1)(J - 1)$ degrees of freedom. We can write it as,

The Two-Way Layout of ANOVA

- To do Test - 2, we can construct the following F statistic,

$$F_{\text{Factor B}} = \frac{SS_{\text{Between B}} / (J - 1)}{SS_{\text{Error}} / (I - 1)(J - 1)}$$

- Again under the null hypothesis that the population means for the J groups are same, this statistic follows F distribution with $J - 1$ and $(I - 1)(J - 1)$ degrees of freedom. We can write it as,
- Here is the ANOVA table, that we can write,

Source of Variation	Sum Of Squares	Degrees of Freedom	Mean Squares	F Ratio
Between groups	$SS_{\text{Between B}}$	$J - 1$	$MS_{\text{Between B}} = \frac{SS_{\text{Between B}}}{J - 1}$	$\frac{MS_{\text{Between B}}}{MS_{\text{Error}}}$
Between blocks	$SS_{\text{Between A}}$	$I - 1$	$MS_{\text{Between A}} = \frac{SS_{\text{Between A}}}{I - 1}$	$\frac{MS_{\text{Between A}}}{MS_{\text{Error}}}$
Within groups	SS_{Error}	$(I - 1)(J - 1)$	$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{(I - 1)(J - 1)}$	

The Two-Way Layout of ANOVA

- ▶ Now let's see an example, we will extend the last example, suppose now the tutoring company introduced 4 services, service A, service B, service C and service D (you can also think D is No service!). And they want to test whether the 4 services are same.
- ▶ One way to design a randomized study is to call some students in a place, randomly allocating the different services to them collect their score (or some performance measures).
- ▶ For example suppose we can randomly allocate 20 students to 4 services (so 5 students in each services) and collect their score after taking some kind of performance measures. This design is called *completely randomized design* and since we have one factor we can use *one-way ANOVA* technique.
- ▶ This method seems okay but there might some issues, an obvious issue is students have different level of skills, which might have some *extraneous impact* in our conclusion. So in this case perhaps it's better to control the student variations and use following *randomized block design*.
- ▶ In this design we will only call 5 students. First we will pick 1st student, and then we will randomly allocate one of the 4 services to the student and record the score, then we will pick the 2nd student and again randomly allocate one of the 4 services to the student and record the score and so on, until we finish all 5 students and then record scores for at least one of the 4 services. Then again we start from the 1st student and do the whole process again, we continue like this until we finish all 5 students and each received all 4 services. This design is called *randomized block design*.

The Two-Way Layout of ANOVA

- ▶ The randomized aspect of the randomized block design is the random order in which the treatments (services) are assigned to the students. If every student were to test the three services in the same order, any observed difference in services might be due to in which order we provided different services rather than to true differences in the services.
- ▶ Also if we allocate the same services to all 5 students in a same time, probably we will only see the difference between the students, and not the difference between the services.
- ▶ Here the students are another factor, but for a randomized block design they are also called *blocks*, and again we are randomizing the techniques within each block.
- ▶ Let's do it using a data.

The Two-Way Layout of ANOVA

Students	Services			
	A	B	C	D
1	8	12	7	13
2	9	9	8	12
3	12	10	9	10
4	11	10	10	12
5	9	8	10	14

- ▶ This is an example of a randomized block design, where the blocks are the students, and the treatments are the services
 - ▶ a. Prepare the two-way analysis of variance table.
 - ▶ b. Test the null hypothesis that the population mean scores are same for all services.
 - ▶ b. Test the null hypothesis that the population mean scores are same for all students. Usually for randomized block design with a blocking variable we don't do this, but in general for many factors and when we have repeated measures we can also do this, in that case maybe we can categorize students in 5 categories.
- ▶ Let's do the complete calculation in Excel

References

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. J. and Ohlmann, J. W. (2020), *Statistics for Business & Economics*, 14th edn, Cengage, Boston, MA.

DeGroot, M. H. and Schervish, M. J. (2012), *Probability and Statistics*, 4th edn, Addison-Wesley, Boston.

Rice, J. A. (2007), *Mathematical Statistics and Data Analysis*, Duxbury advanced series, 3rd edn, Thomson/Brooks/Cole, Belmont, CA.