



Notes for Students- Lesson 12

ANOVA

Q1




	Snapzi	Irisa	LolaMoon
	15	39	65
	12	45	45
	14	48	32
	11	60	38



Which of these brands have significantly different prices?

- ☐ Snapzi and Irisa
- ☐ Snapzi and LolaMoon
- ☐ Irisa and LolaMoon
- ☐ There is not a significant difference between any of the brands



Q2

Lesson 11: Test significant difference between two independent samples

Four samples: A B C D

How many t-tests would we need to compare 4 samples?

Q3.

Compare three or more samples:

distance/variability between means
error

How can we compare three or more samples?

- Use the maximum distance between any two sample means
- Use the average deviation of each sample mean from the mean of all values in all samples
- Find the distance each sample mean is from each of the other sample means
- Find the average squared deviation of each sample mean from the total mean
- Find the average squared deviation of each value in each sample from the total mean

Q4.

Grand mean \bar{X}_G

A	B	C	D
\bar{X}_A	\bar{X}_B	\bar{X}_C	\bar{X}_D

Will the mean of sample means be the same as the mean of all values in each sample?

$$\frac{\bar{X}_A + \bar{X}_B + \bar{X}_C + \bar{X}_D}{4} = \text{mean of sample means}$$

- Always
- Sometimes
- Never

Q5.

What conclusions can we draw from the deviation of each sample mean from the mean of means?

- The greater the distance between sample means, the less likely population means will differ significantly.
- The smaller the distance between sample means, the less likely population means will differ significantly.
- The greater the distance between sample means, the more likely population means will differ significantly.
- The smaller the distance between sample means, the more likely population means will differ significantly.

Q6.

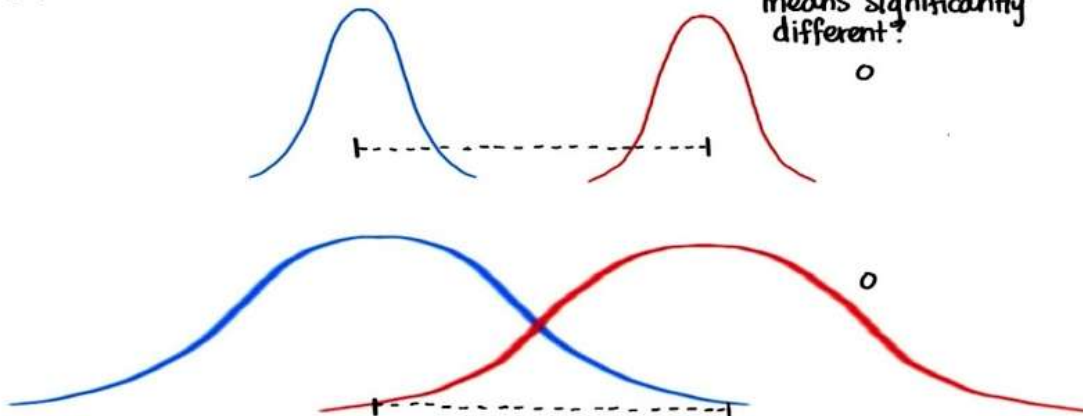
Between-group variability

The smaller the distance between sample means, the less likely population means will differ significantly.

The greater the distance between sample means, the more likely population means will differ significantly.

$$t = \frac{\text{difference}}{\text{error}}$$

In which situation are the means significantly different?



Q7.

Between-group variability

The smaller the distance between sample means, the less likely population means will differ significantly.

The greater the distance between sample means, the more likely population means will differ significantly.

$$t = \frac{\text{difference}}{\text{error}}$$

What does this say about the process of comparing three or more samples?

- The greater the variability of each individual sample, the less likely population means will differ significantly
- The smaller the variability of each individual sample, the less likely population means will differ significantly
- The greater the variability of each individual sample, the more likely population means will differ significantly
- The smaller the variability of each individual sample, the more likely population means will differ significantly

ANOVA

Between-group variability

The smaller the distance between sample means, the less likely population means will differ significantly.

The greater the distance between sample means, the more likely population means will differ significantly.

Analysis of Variance (ANOVA)

Within-group variability

The greater the variability of each individual sample, the less likely population means will differ significantly

The smaller the variability of each individual sample, the more likely population means will differ significantly

When we compare samples, we're simply extending the idea of the t test. We can compare samples to each other by seeing how far each sample mean is from the mean of means, or the grand mean, and this is between group variability. But we also want to look at the variability of each sample, and that's within group variability. Because this impacts whether or not the samples are significantly different. Since we're analyzing variabilities, this process is called analysis of variance, shortened to ANOVA. ANOVA can compare as many means as we want with just one test. We say one way ANOVA when we have one independent variable sometimes called a factor.

Q8. What are our null and alternate hypothesis ?

Analysis of Variance
(ANOVA)

- $H_0: \mu_1 = \mu_2 = \mu_3$
 $H_A: \mu_1 \neq \mu_2 \neq \mu_3$
- $H_0: \mu_1 \neq \mu_2 \neq \mu_3$
 $H_A: \mu_1 = \mu_2 \neq \mu_3$
- $H_0: \mu_1 = \mu_2 = \mu_3$
 $H_A: \text{At least one pair of samples is significantly different}$

Q9.

If the variants of an individual sample becomes bigger, all else held constant. Does this lean more in favor of the null or alternative hypotheses?

Analysis of Variance
(ANOVA)

- $H_0: \mu_1 = \mu_2 = \mu_3$
- $H_A: \text{At least one pair of samples is significantly different}$

As the between group variability increases meaning the sample means get further apart from each other all else held constant. Does this lean in favor of the null or alternative hypothesis?

Analysis of Variance (ANOVA)

- $H_0: \mu_1 = \mu_2 = \mu_3$
- H_A : At least one pair of samples is significantly different

F-Ratio

Q11.

Between-group variability

The smaller the distance between sample means, the less likely population means will differ significantly.
The greater the distance between sample means, the more likely population means will differ significantly.

Within-group variability

The greater the variability of each individual sample, the less likely population means will differ significantly.
The smaller the variability of each individual sample, the more likely population means will differ significantly.

Analysis of Variance (ANOVA)

$F = \text{ratio}$

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

H_A : At least one pair of samples is significantly different

- Between-group variability should be the numerator;
Within-group variability should be the denominator.
- Within-group variability should be the numerator;
Between-group variability should be the denominator.

Visualize Statistical Outcome

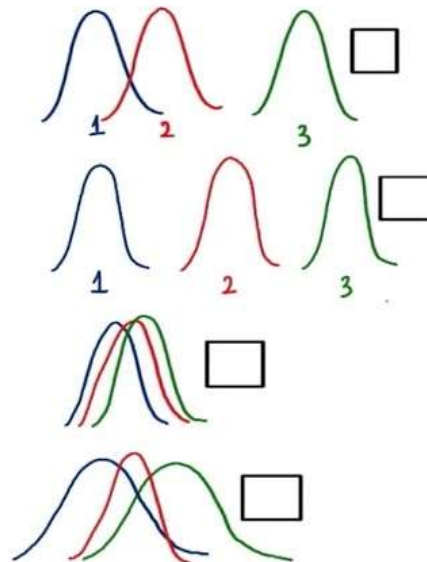
Q12.

Analysis of Variance
(ANOVA)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

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

H_A : At least one pair of samples is significantly different



A $\mu_1 = \mu_2 = \mu_3$

B $\mu_1 \neq \mu_2 \neq \mu_3$

C $\mu_1 \neq \mu_2$
 $\mu_2 = \mu_3$
 $\mu_1 \neq \mu_3$

D $\mu_1 = \mu_2$
 $\mu_1 \neq \mu_3$
 $\mu_2 \neq \mu_3$

Formalize Within-Group Variability

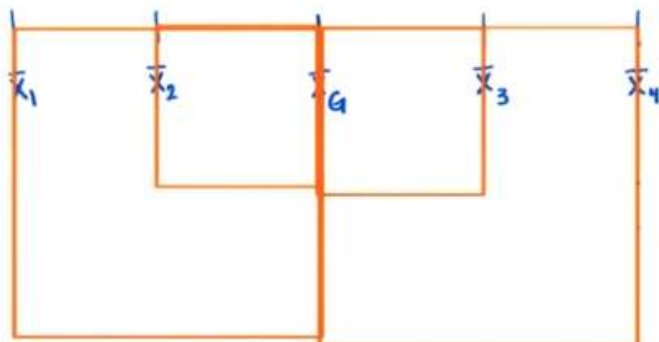
We need to formalize how we will precisely measure each type of variability.

Analysis of Variance
(ANOVA)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

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

H_A : At least one pair of samples is significantly different



Analysis of Variance (ANOVA)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{n \sum (\bar{x}_k - \bar{x}_a)^2 / (k-1)}{\sum (x_i - \bar{x}_k)^2 / df}$$

Q13.

Analysis of Variance (ANOVA)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{n \sum (\bar{x}_k - \bar{x}_a)^2 / (k-1)}{\sum (x_i - \bar{x}_k)^2 / df}$$

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

H_A : At least one pair of samples is significantly different

What will be df ?

$$\square n_1 + n_2 + n_3 - 1$$

$$\square n_1 + n_2 + n_3 - 3$$

$$\square n_1 + n_2 + n_3 - k \quad (\text{where } k \text{ is the number of samples})$$

$$\square N - k \quad (\text{where } N \text{ is the total number of values from all samples and } k \text{ is the number of samples})$$

Formula for F-Ratio

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{n \sum (\bar{x}_k - \bar{x}_a)^2 / (k-1)}{\sum (x_i - \bar{x}_k)^2 / (N-k)} = \frac{SS_{\text{between}} / df_{\text{between}}}{SS_{\text{within}} / df_{\text{within}}} = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

Q14.
**Analysis of Variance
(ANOVA)**

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{n \sum (\bar{x}_k - \bar{x}_G)^2 / (k-1)}{\sum (x_i - \bar{x}_k)^2 / (N-k)} = \frac{SS_{\text{between}} / df_{\text{between}}}{SS_{\text{within}} / df_{\text{within}}} = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

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

H_A : At least one pair of samples is significantly different

What is $df_{\text{between}} + df_{\text{within}}$?

- o $N+1$
- o N
- o $N+k$
- o $N-1$
- o $k+1$

Total Variation

We have total degrees of freedom and likewise we have a total variation. This is also the sum of the sum of squares for between and within group variability. This is the total sum of squares that we mentioned earlier. Each value, minus the grand mean squared and then sum them up. Basically what ANOVA does is partition the total variation into between group variation and within group variation. Difference is in the dependent variable or treatment is due to both between group differences and individual differences within each group. This means that only some of the variation can be explained by knowing which group is subject is in and the rest is unexplained variants.

$$SS_{\text{between}} + SS_{\text{within}} = SS_{\text{total}} = \sum (x_i - \bar{x}_G)^2$$

$$df_{\text{between}} + df_{\text{within}} = df_{\text{total}} = N-1$$

F – Distribution

If we could take all possible f-statistics, what would this distribution look like?

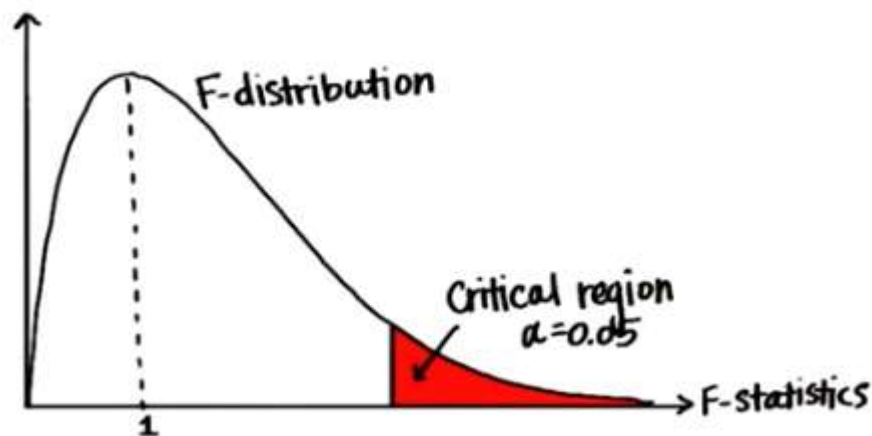
Let's start with

Q15

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{n \sum (\bar{x}_k - \bar{x}_G)^2 / (k-1)}{\sum (x_i - \bar{x}_k)^2 / (N-k)} = \frac{SS_{\text{between}} / df_{\text{between}}}{SS_{\text{within}} / df_{\text{within}}} = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

The F-statistic is _____ negative.

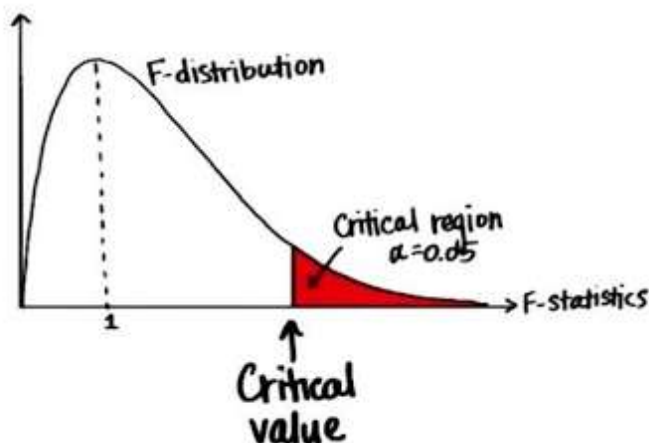
- always
- sometimes
- never



- Unlike the Z and T tests, the distribution of the F-statistics is not symmetrical. The F distribution is positively skewed.
- This distribution peaks at 1.
- Also note that when we're doing an F test, this will always be non-directional.
- With the F test, we only know if there's a significant difference.
- Our critical region will always be out here in the upper tail.
- Everything else is essentially the same. We choose an alpha level, usually 0.05. And then we find out if our F-statistic lies in the critical region or not.
- If it lies in the critical region, we know that at least two population means will be significantly different.

Q16.

We need critical value



What do you think the table we need is called?

- o A table
- o M table
- o F table
- o Y table

Q17. Let's revisit the clothing example and apply ANOVA

Snapzi	Irisa	LolaMoon	
15	39	65	
12	45	45	
14	48	32	
11	60	38	
$\bar{x}_S =$	$\bar{x}_I =$	$\bar{x}_L =$	$\bar{x}_G =$



Q18.

Snapzi	Irisa	LolaMoon	
15	39	65	
12	45	45	
14	48	32	
11	60	38	
$\bar{x}_S = 13$	$\bar{x}_I = 48$	$\bar{x}_L = 45$	$\bar{x}_G = 35.33$

$$SS_{\text{between}} = n \sum (\bar{x}_k - \bar{x}_G)^2 =$$

Q19.

Snapzi	$x_i - \bar{x}_s$	$(x_i - \bar{x}_s)^2$	Irisa	$x_i - \bar{x}_I$	$(x_i - \bar{x}_I)^2$	LolaMoon	$x_i - \bar{x}_L$	$(x_i - \bar{x}_L)^2$
15			39			65		
12			45			45		
14			48			32		
11			60			38		
$\bar{x}_s = 13$		SS_1	$\bar{x}_I = 48$		SS_2	$\bar{x}_L = 45$		SS_3

$$SS_{\text{between}} = 3010.67$$

$$SS_{\text{within}} = \sum (x_i - \bar{x}_k)^2 =$$

Q20.

Snapzi	Irisa	LolaMoon
15	39	65
12	45	45
14	48	32
11	60	38
$\bar{x}_s = 13$	$\bar{x}_I = 48$	$\bar{x}_L = 45$

$$SS_{\text{between}} = 3010.67 \quad df_{\text{between}} =$$

$$SS_{\text{within}} = 862 \quad df_{\text{within}} =$$

Q21.

Snapzi	Irisa	LolaMoon
15	39	65
12	45	45
14	48	32
11	60	38
$\bar{x}_s = 13$	$\bar{x}_I = 48$	$\bar{x}_L = 45$

$SS_{\text{between}} = 3010.67$	$df_{\text{between}} = 2$	$MS_{\text{between}} =$
$SS_{\text{within}} = 862$	$df_{\text{within}} = 9$	$MS_{\text{within}} =$

Q22.

$SS_{\text{between}} = 3010.67$	$df_{\text{between}} = 2$	$MS_{\text{between}} = 1505.33$	$F =$
$SS_{\text{within}} = 862$	$df_{\text{within}} = 9$	$MS_{\text{within}} = 95.78$	

Q23.

$SS_{\text{between}} = 3010.67$	$df_{\text{between}} = 2$	$MS_{\text{between}} = 1505.33$	$F = 15.72$
$SS_{\text{within}} = 862$	$df_{\text{within}} = 9$	$MS_{\text{within}} = 95.78$	

Find the F critical value using the F table. $\alpha = 0.05$

USE http://www.socr.ucla.edu/applets.dir/f_table.html

Q24.

$SS_{\text{between}} = 3010.67$	$df_{\text{between}} = 2$	$MS_{\text{between}} = 1505.33$	$F = 15.72$
$SS_{\text{within}} = 862$	$df_{\text{within}} = 9$	$MS_{\text{within}} = 95.78$	

F critical value at $\alpha = 0.05$ is 4.2565.

- $H_0: \mu_s = \mu_I = \mu_L$
- H_A : At least one brand has significantly different prices