

## ANalysis Of VAriance

Dr. Tom Ilvento  
FREC 408

## ANOVA

- ANOVA provides a strategy to compare two or more population means associated with various treatments
- It is used when we have
  - A dependent variable (AKA Response Variable)
  - One or more categorical variables (ordinal) or continuous variables that are thought of as independent variables that influence the dependent variables
  - E.g. levels of fertilizer; group membership; different treatments

## ANOVA

- ANOVA is used heavily in experimental designs in the biological sciences
  - Treatment versus control groups
  - Levels of treatment of a drug
  - Levels of applications of fertilizers or pesticides
- It is possible to have continuous and categorical independent variables
- Its origins were in agricultural studies

## Elements of a Designed Experiment

- **Response Variable:** the variable of interest to be measured in the experiment. (Def11.3 p650)
  - Also known as the **dependent variable**.
- **Factors:** variables which are thought to influence the response variable (Def11.5 p650)
  - Quantitative
  - Qualitative
- **Factor Levels:** the levels of the factor that are experimentally manipulated (Def11.6 p650)
  - In a single factor experiment, the factor levels are the **treatments**

## Elements of a Designed Experiment

- **Treatments:** when two or more factors are utilized, the treatments are the combinations of factor levels used in the experiment. (Def11.7 p650)
  - Factor 1: fertilizer (low; medium; high)
  - Factor 2: water (low; high)
    - Treatment 1: low fertilizer, low water
    - Treatment 2: low fertilizer, high water
    - Treatment 3: medium fertilizer, low water
    - And so forth.....

## Elements of a Designed Experiment

- **Experimental Unit:** the objects on which the response variable and factors are observed (Def11.4 p650)
  - People
  - Plants
  - Animals
  - Schools

## Designed versus observational

- Designed Experiment**
  - The specification of the treatments
  - And the way experimental units are assigned to treatments is under the control of the researcher
- Observational Study**
  - The researcher observes the treatments and the response on a sample of experimental units

## Completely Randomized Design

- The treatments are **randomly assigned** to the experimental units
- Or independent random samples** of experimental units are selected from target populations for each treatment (Def11.8 p655)
  - The book refers to both designed and observational studies as being randomized designs
  - Most think of ANOVA for designed experiments

## What is ANOVA? HINT!!!!!!

It is all about the variance!

*It is all about the variance!*

It is all about the variance!

*It is all about the variance!*

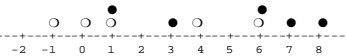
It's the variance!

It is all about the variance!

## Book Example of contrasts of two samples, p653

Obs	Sample 1	Sample 2
1	6	8
2	-1	1
3	0	3
4	4	7
5	1	6
Sum	10.0	25.0
Mean	2.0	5.0
Var	8.5	8.5
Std Dev	2.9	2.9

The difference between sample means is relatively small when compared to the variability within the sample observations



## Difference of Means Test

t-Test: Two-Sample Assuming Equal Variances

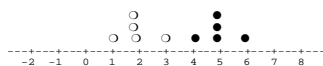
	Sample 2	Sample 1
Mean	5.00	2.00
Variance	8.50	8.50
Observations	5	5
Pooled Variance	8.5	
Hypothesized Mean	0	
df	8	
t Stat	1.627	
P(T<=t) one-tail	0.071	
t Critical one-tail	1.860	
P(T<=t) two-tail	0.142	
t Critical two-tail	2.306	

Based on a  $t^* = 1.627$  and an  $\alpha = .05$ , I would fail to reject  $H_0: \mu_2 - \mu_1 = 0$

## Book Example of contrasts of two samples, p653

Obs	Sample 1	Sample 2
1	2	5
2	3	5
3	2	5
4	2	4
5	1	6
Sum	10.0	25.0
Mean	2.0	5.0
Var	0.5	0.5
Std Dev	0.7	0.7

The difference between sample means is relatively large when compared to the variability within the sample observations



## Difference of Means Test

t-Test: Two-Sample Assuming Equal Variances		
	Sample 2	Sample 1
Mean	5.00	2.00
Variance	0.50	0.50
Observations	5	5
Pooled Variance	0.5	
Hypothesized Mean Difference	0	
df	8	
t Stat	6.708	
P(T<=t) one-tail	0.000	
t Critical one-tail	1.860	
P(T<=t) two-tail	0.000	
t Critical two-tail	2.306	

Based on a  
 $t^* = 6.708$  and  
 an  $\alpha = .05$ ,  
 I would reject  
 $H_0: \mu_2 - \mu_1 = 0$

## Book Example of contrasts of two samples, p653

- One way to determine whether a difference exists between the population means is to examine the difference between the sample means and compare it to a measure of variability within the samples. P653
- The difference of the two means is only part of the story.
- The other part is the variability and separation of the two samples
- ANOVA uses this strategy to compare two or more means

## ANOVA

- We will **decompose** the variance of our dependent variable
  - Part **due to the treatments** or independent variables – part that is explained
  - Part that is unexplained or **random "error"**
- I will adjust these variances from different sources by dividing by degrees of freedom to get an average deviation

## ANOVA

- We will decompose
- **Total Sum of Squares =**
  - **Sum of Squares for Treatment +**
  - **Sum of Squares for Error**

$$SS(Total) = \sum_{i=1}^n (y_i - \bar{Y})^2$$

## Let's Look at the data comparing males and female cholesterol level

Cholesterol Level	Females	Males
Mean	200.32	196.09
Standard Error	0.88	0.97
Median	201	196
Mode	194	196
Standard Deviation	10.72	12.37
Sample Variance	114.94	153.07
Kurtosis	-0.49	0.02
Skewness	-0.11	0.09
Range	47	61
Minimum	176	166
Maximum	223	227
Sum	29647	32158
Count	148	164
Confidence Level(95.0%)	1.74	1.91

## How might we approach this data?

- A large sample difference of means test
- ANOVA - randomized design for an observational study
  - **Response Variable:** cholesterol level
  - **Factor:** gender
    - Females
    - Males
  - **Experimental Units:** people

## Difference of Means Test

t-Test: Two-Sample Assuming Unequal Variances		
	Females	Males
Mean	200.32	196.09
Variance	114.94	153.07
Observations	148	164
Hypothesized Mean Difference	0	
df	310	
t Stat	3.24	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.65	
P(T<=t) two-tail	0.00	
t Critical two-tail	1.97	

## Difference of means

- Null hypothesis     ■  $H_0: (\mu_f - \mu_m) = 0$
- Alternative         ■  $H_a: (\mu_f - \mu_m) \neq 0$  two-tailed test
- Assumptions        ■ Large sample
- Test Statistic       ■  $z^* = 3.24$
- Rejection Region   ■  $Z_{.05/2} = \pm 1.96$
- Conclusion          ■  $z^* > z$
- $3.24 > 1.96$
- Reject  $H_0$

## ANOVA Results

Anova: Single Factor		Cholesterol Levels				
<b>SUMMARY</b>						
<b>Groups</b>	<b>Count</b>	<b>Sum</b>	<b>Average</b>	<b>Variance</b>		
Females	148	29647	200.32	114.94		
Males	164	32158	196.09	153.07		
<b>ANOVA</b>						
<b>Source of Variation</b>	<b>SS</b>	<b>df</b>	<b>MS</b>	<b>F</b>	<b>P-value</b>	<b>F crit</b>
Between Groups	1393.425	1	1393.425	10.322	0.001	3.872
Within Groups	41846.879	310	134.990			
<b>Total</b>	<b>43240.304</b>	<b>311</b>				

## ANOVA Output Shows

- Mean, variance, and sample size for each group for each group
- ANOVA Table with Sums of Squares breakdown
  - SS Total            43,240.304
  - SS Treatment    1,393.425    Between Groups
  - SS Error           41,846.879    Within Groups
- F-test

## How does ANOVA work?

- We compute the variability of the treatment means from the **Grand Mean** (the mean from the whole sample, i.e., all groups)
  - **Sum of Squares for Treatments (SST)**
  - **Which measures between-sample variation**
- And the variability within the treatment levels
  - **Sum of Squares for Error (SSE)**
  - **Which measures within-sample variation**

## How does ANOVA work?

- We adjust SST and SSE to reflect a mean sum of squares – **divide by degrees of freedom**
- Then we compare to see if the Mean Sum of Squares for Treatments (**MST**) is larger relative to the Mean Sum of Squares for Error (**MSE**)
- We do this by taking a ratio of the two sources of sums of squares

$$\frac{MST}{MSE}$$

## How does ANOVA work?

- We ask, "Is there more variability across the means for the factor levels (or treatments), than within the treatments?"
- If there is more variability across factor levels (treatments),
  - with respect to a probability framework
  - using an **F-distribution** with specified degrees of freedom
- **We will conclude that the factors influence the response variable**

## Sum of Squares

- Let **k** = # treatments
- And the Grand Mean is  $\bar{Y}$
- Each group mean is  $y_i$  as  $i$  goes from 1 to  $k$
- **Degrees of freedom**
  - Total Sum of Squares (**n-1** d.f.)
  - Sum of Squares for Treatments (**k-1** d.f.)
  - Sum of Squares Error (**n-k** d.f.)

## Sum of Squares

- **SS<sub>Total</sub> = SS<sub>Treatment</sub> + SS<sub>Error</sub>**
- Degrees of freedom
  - **n-1 = (k-1) + (n-k)**
  - e.g.  $n = 100$   $k=3$
  - $100-1 = (3-1) + (100-3)$
  - $99 = 2 + 97$

## Total Sum of Squares

- **Total Sum of Squares (n-1 d.f.)**

$$SS(Total) = \sum_{i=1}^n (y_i - \bar{Y})^2$$

- If I divided by the degrees of freedom I would have the sample variance

## Sum of Squares for Treatments

- **Sum of Squares for Treatments**
  - (**k-1** d.f.)

$$SST = \sum_{i=1}^k n_i (\bar{y}_i - \bar{Y})^2$$

- If I divide by the degrees of freedom I have the Mean Square for Treatment **MST**
- **SST/(k-1) = MST**

## Sum of Squared Error (SSE) with (n-k df) =

- Adding the squared deviations of each group around the group mean

$$SSE = \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2 + \dots + \sum_{j=1}^{n_k} (y_{kj} - \bar{y}_k)^2$$

- Another way to calculate this is:

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2$$

**Notice this is just like our pooled estimate for a difference of means problem**

## Sum of Squares Error (n-k d.f.)

- If I divide by the degrees of freedom I have the Mean Square for Error, referred to as: **MSE**
- **$SSE/(n-k) = MSE$**

## The computations for the cholesterol example

- $k = 2$
- $n_1 = 148$      $n_2 = 164$
- $\bar{y}_1 = 200.318$      $\bar{y}_2 = 196.085$
- $s_1^2 = 114.94$      $s_2^2 = 153.07$
- $\bar{Y} = 198.093$
- Total Sum of Squares = 43,240.304

$$SST = \sum_{i=1}^k n_i (\bar{y}_i - \bar{Y})^2$$

- $SST = 148(200.318 - 198.093)^2 + 164(196.085 - 198.093)^2$
- $SST = 732.441 + 660.984$
- $SST = 1,393.425$
- **$MST = 1,393.425/(2-1) = 1,393.425$**

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2$$

- $SSE = (148-1)114.94 + (164-1)153.07$
- $SSE = 16,896.18 + 24,950.41$
- $SSE = 41,846.59$
- **$MSE = 41,896.59/310 = 134.989$**

## ANOVA Results

Anova: Single Factor		Cholesterol Levels					
SUMMARY							
Groups	Count	Sum	Average	Variance			
Females	148	29647	200.32	114.94			
Males	164	32158	196.09	153.07			
ANOVA							
Source of Variation	SS	df	MS	F	P-value	F crit	
Between Groups	1393.425	1	1393.425	10.322	0.001	3.872	
Within Groups	41846.879	310	134.990				
Total	43240.304	311					

## The computations

The F statistic (F\*) =

$$F^* = \frac{MST}{MSE}$$

Page 657

$$F^* = \frac{1393.425}{134.989} = 10.323$$

## The F Statistic

- An  $F^*$  is the ratio of two variances
  - Between treatments
  - Within treatments
- An  $F^* = 1$  mean equality of variances

## F Statistic

- We compare our  $F^*$  to a  $F$  at a specified  $\alpha$ , with  $v_1$  and  $v_2$  degrees of freedom
  - $v_1$  represents the degrees of freedom for the numerator ( $k-1$ )
  - $v_2$  represents the degrees of freedom for the denominator ( $n-k$ )

## F Table

- Look at the F-Table starting on page 741
- For each level of  $\alpha$  (.10, .05, .025, .01) there is a different table with
  - Degrees of freedom for the numerator and denominator
  - The numerator degrees of freedom have less options ( $k-1$ )

## ANOVA Hypothesis F Test

- $H_0: \mu_1 = \mu_2 = \dots = \mu_k$  where  $k$  represents the # treatment levels
- $H_a$ : they differ (with 3 or more levels, at least 2 means differ)
- We conduct a F-test comparing sources of variability
- Assumptions:
  - Samples are selected randomly and independently
  - The populations for each treatment level are distributed normally
  - The population variances are equal

## ANOVA Hypothesis Test

- Null hypothesis  $H_0: \mu_1 = \mu_2$
- Alternative  $H_a$ : means differ
- Assumptions
  - Equal variances, normal distribution
- Test Statistic  $F^* = 10.322$
- Rejection Region  $F_{.05; 1, 120 \text{ d.f.}} = 3.92$
- Conclusion
  - $F^* > F$
  - $10.322 > 3.92$
  - Reject  $H_0$

## ANOVA Results

Anova: Single Factor		Cholesterol Levels					
<b>SUMMARY</b>							
Groups	Count	Sum	Average	Variance			
Females	148	29647	200.32	114.94			
Males	164	32158	196.09	153.07			
<b>ANOVA</b>							
Source of Variation	SS	df	MS	F	P-value	F crit	
Between Groups	1393.425	1	1393.425	10.322	0.001	3.872	
Within Groups	41846.879	310	134.990				
Total	43240.304	311					

## Compare the $t^*$ to $F^*$

- $t^*$  from difference of means test = 3.24
- $F^* = 10.322$
- In case of comparing just two means, if you square  $t^*$ , it roughly equals  $F^*$
- The two tests (t and F) result in the same conclusion
- The advantage of ANOVA is that you can compare and make inferences on **more than two means**

## Coupon Example

- A greeting company wanted to use a coupon offer to increase sales
- They developed four different coupon designs, and used each design with a number of customers
- They took a sample of 8 customers for each design and noted their purchase amount as a result of the coupon
- Did the coupons have different effects on sales?

## The data

Customer	Design1	Design 2	Design 3	Design 4
1	\$4.10	\$6.90	\$4.60	\$12.50
2	\$5.90	\$9.10	\$11.40	\$7.50
3	\$10.45	\$13.00	\$6.15	\$6.25
4	\$11.55	\$7.90	\$7.85	\$8.75
5	\$5.25	\$9.10	\$4.30	\$11.15
6	\$7.75	\$13.40	\$8.70	\$10.25
7	\$4.78	\$7.60	\$10.20	\$6.40
8	\$6.22	\$5.00	\$10.80	\$9.20
Mean	\$7.00	\$9.00	\$8.00	\$9.00
Variance	\$7.34	\$8.42	\$7.63	\$5.02
GRAND MEAN		\$8.25		

## The calculations

- $SST = 8(7.00-8.25)^2 + 8(9.00-8.25)^2 + 8(8.00-8.25)^2 + 8(9.00-8.25)^2$ 
  - $SST = 22.00$
  - $MST = 22.00/(4-1) = 7.33$
- $SSE = (8-1)7.34 + (8-1)8.42 + (8-1)7.63 + (8-1)5.02$ 
  - $SSE = 198.87$
  - $MSE = 198.87/(32-4) = 7.10$
- $F^* = 7.33/7.10 = 1.03$

**But you don't have to do the calculations, use Excel, Tools, Data Analysis, ANOVA: Single Factor**

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Design1	8	56	7	7.340971		
Design 2	8	72	9	8.422857		
Design 3	8	64	8	7.632143		
Design 4	8	72	9	5.015714		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	Fcrit
Between Groups	22	3	7.333333	1.032439	0.3933	2.9467
Within Groups	198.88	28	7.102921			
Total	220.88	31				

## ANOVA Hypothesis Test

- Null hypothesis ■  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
- Alternative ■  $H_a$ : At least two means differ
- Assumptions ■ Equal variances, normal distribution
- Test Statistic ■  $F^* = 1.032$
- Rejection Region ■  $F_{.05, 3, 28 \text{ d.f.}} = 2.95$
- Conclusion ■  $F^* < F$
- $1.032 < 2.95$
- Cannot Reject  $H_0$

## Using Excel to do ANOVA

- Arrange data in columns – each factor level is a column
- It is a good idea to label the columns
- Use: Tools, Data Analysis, ANOVA: Single Factor
- Identify
  - the input range of columns
  - Alpha for the test
  - Whether labels are present
  - Output range

## Using Excel to Do ANOVA

- I would also suggest using Data Descriptive on each column and on the total sample