ANOVA II

Tom Ilvento
FREC 408

Review of ANOVA

- ANOVA – Analysis of Variance – is a technique which allows us to compare two or more group means
  - We will only focus on a few ideas used in ANOVA
  - Completely Randomized Design with a single Factor or two factors
  - Randomized Block Design

Review of ANOVA

- ANOVA has its own set of terms
  - Response Variable
  - Factors and Factor levels
  - Treatments
  - Experimental Units

Review of ANOVA

- ANOVA is based on focusing on the variability of a dependent variable and partitioning it into:
  - A part explained by independent variables, which are typically group membership
  - A part that is random error due to fluctuation within each of the groups

Review of ANOVA

- We then adjust each of these parts by dividing by degrees of freedom, calling them "Mean Squares"
- Next we compare the Mean Squares to see which is largest
  - The ratio of the Mean Squares is distribution as a **F random variable** with v_1 and v_2 degrees of freedom

Sum of Squares

- **Let** k **= # treatments**
- **And the Grand Mean is** $\overline{Y}$
- **Each group mean is** $\overline{y}_i$ 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

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

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

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

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

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

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

Another Formula for SSE

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

Sum of Squares

- **$SS_{Total} = SS_{Treatment} + SS_{Error}$**
- **Degrees of freedom**
  - **n-1 = (k-1) + (n-k)**
- **For example: n = 100 k=3**
  - **100-1 = (3-1) + (100-3)**
  - **99 = 2 + 97**

I want to introduce a concept called R-Square

- **R-square** (**R^2**) is a measure of association
- Measures of Association reflect the relationship between two or more variables
- **R^2** is a member of the class of measures of association called **PRE** measures – **Proportion in the Reduction in Error**
- It is based on fitting a model to the data based on information from an independent variable (or set of variables) and comparing our model to a baseline model

What is R-Square?

- **The baseline model is the Grand Mean**

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

- **Our model is one that is based on knowledge of the Factors/treatments**

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

What is R-Square?

- With R^2 we ask, "how much better do I understand the Response variable (dependent) by knowing something about the Factors/Treatments (independent variables)
- R^2 varies from 0 to 1
  - 0 means I explain nothing of the dependent variable
  - 1 mean I explain it perfectly
- **R^2 is a linear measure of association**
- **R^2 = SST/SS(Total)**
- **Or,**

R^2 = 1 - SSE/SS(Total)

ANOVA Table Terms

Book	Excel	SAS
SST – Sum of Squares Treatment	Between Groups	Model
SSE – Sum of Squares Error	Within Groups	Error
SS(Total) – Sum of Squares Total	Total	Corrected Total
Factors	Groups	Class

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 = $\mathbf{8(7.00-8.25)^2 + 8(9.00-8.25)^2 + 8(8.00-8.25)^2 + 8(9.00-8.25)^2}$
  - **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	F crit
Between Groups	22	3	7.333333	1.032439	0.3933	2.9467
Within Groups	198.88	28	7.102921			
Total	220.88	31				

What is R^2 for this model?

- **22/220.88 = .0996**
- **R^2 = 10%**
- **Not very large!!!**

ANOVA Hypothesis Test

- Null hypothesis
- Alternative
- Assumptions
- Test Statistic
- Rejection Region
- Conclusion

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_a: At least two means differ
- Equal variances, normal distribution
- $F^* = 1.032$
- $F_{.05,\ 3,\ 28\ d.f.} = 2.95$
- $F^* < F$
- $1.032 < 2.95$
- 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

Example Problem

- Experiment to determine the differences in mean increases in plant growth from 5 different inoculins
- The experiment involved 20 cuttings of a shrub (all of equal weight), with four cuttings assigned to the five different inoculins
- The data represent the increase in weight in grams
- We will use α =.05

The data

A	B	C	D	E
15	21	22	10	6
18	13	19	14	11
9	20	24	21	15
16	17	21	13	8

Descriptive Statistics

	A	*B*	*C*	*D*	*E*	*TOTAL*
Mean	14.50	17.75	21.50	14.50	10.00	15.65
Standard Error	1.94	1.80	1.04	2.33	1.96	1.15
Median	15.50	18.50	21.50	13.50	9.50	15.50
Mode	#N/A	#N/A	#N/A	#N/A	#N/A	21.00
Standard Deviation	3.87	3.59	2.08	4.65	3.92	5.13
Sample Variance	15.00	12.92	4.33	21.67	15.33	26.34
Kurtosis	2.36	-0.58	0.39	2.12	-0.77	-0.93
Skewness	-1.38	-0.89	0.00	1.19	0.60	-0.23
Range	9.00	8.00	5.00	11.00	9.00	18.00
Minimum	9.00	13.00	19.00	10.00	6.00	6.00
Maximum	18.00	21.00	24.00	21.00	15.00	24.00
Sum	58.00	71.00	86.00	58.00	40.00	313.00
Count	4.00	4.00	4.00	4.00	4.00	20.00

ANOVA Results $R^2 = .585$
$R^2 = 292.80/500.55$

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
A	4	58	14.50	15.00		
B	4	71	17.75	12.92		
C	4	86	21.50	4.33		
D	4	58	14.50	21.67		
E	4	40	10.00	15.33		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	292.80	4	73.20	5.29	0.01	3.06
Within Groups	207.75	15	13.85			
Total	500.55	19				

ANOVA Hypothesis Test for the Factor

- **Null hypothesis**
- **Alternative**
- **Assumptions**
- **Test Statistic**
- **Rejection Region**
- **Conclusion**

- **H_0: $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$**
- **H_a: At least two means differ**
- **Equal variances, normal distribution**
- **$F^* = 5.29$**
- **$F_{.05,\ 4,\ 15\ d.f.} = 3.06$**
- **$F^* > F$**
- **$5.29 > 3.06$**
- **Reject H_0: $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$**

There are differences across inoculins

A few points

- We have five levels of the factors
- That means we could have made 10 different comparisons of means
  - A to B; A to C; A to D; A to E; B to C; B to D; B to E; C to D; C to E; D to E
  - ANOVA just tests that at least two of the means are different
- I will address multiple comparisons later in the lecture

When we deal with more than one Factor

- Called a **Factorial Design**
- **A Complete Factorial Design** is an experiment that includes all possible treatments. (Def11.10 p674)

Hair color and pain

- **Studies at the University of Melbourne indicate that there may be a difference between the pain thresholds of blondes and brunettes.**
- **Men and women of various ages were divided into four categories according to hair color**
  - **Light blonde**
  - **Dark blonde**
  - **Light brunette**
  - **Dark brunette**

Hair color and pain

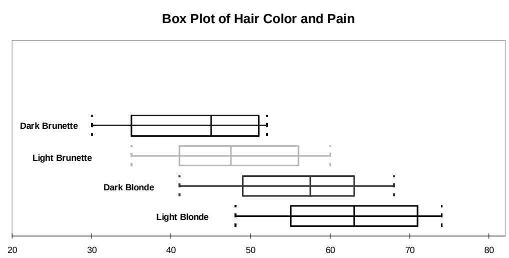
- **Pain was measured by a pain threshold test – the higher the score, the higher the pain tolerance**
- **I fudged the data from the book by adding a gender component to the data to show a two factor experiment**
- **The output is from Excel**

The data

	Females	Males
Mean	57.3	46.85
Variance	86.64	117.50
Standard Deviation	9.31	10.84

	Light Blonde	Dark Blonde	Light Brunette	Dark Brunette
Males	62	63	42	32
	60	57	50	39
	64	52	41	51
	55	41	37	30
	48	43	35	35
Females	65	68	45	46
	55	49	58	51
	72	59	56	47
	71	64	60	44
	74	58	52	52
Mean	62.60	55.40	47.60	42.70
Variance	70.27	80.71	78.93	67.12
Standard Deviation	8.38	8.98	8.88	8.19

Box Plot of Data



One-Way ANOVA

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Light Blonde	10	626	62.600	70.267
Dark Blonde	10	554	55.400	80.711
Light Brunette	10	476	47.600	78.933
Dark Brunette	10	427	42.700	67.122

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2297.475	3	765.825	10.313	0.000	2.866
Within Groups	2673.300	36	74.258			
Total	4970.775	39				
R Squared	0.462	More than 46% of the variability in pain is explained by hair color				

Results from One-Way ANOVA

- **R-Square is**
  - **SST/SS(Total) or, Between Groups/Total**
  - **2297.475/4970.775 = .462**
  - **46.2% of the variability in pain threshold is "explained" by knowing hair color**
  - **F* = 10.313 and the p-value < .001**
    - We can reject the Null Hypothesis
    - We have evidence that some of the means differ from each other
    - We can conclude that hair color is related to pain thresholds

ANOVA Hypothesis Test for the Hair Color Experiment

- **Null hypothesis**
- **Alternative**
- **Assumptions**
- **Test Statistic**
- **Rejection Region**
- **Conclusion**

- **H_0: $\mu_1=\mu_2=\mu_3=\mu_4$**
- **H_a: At least two means differ**
- **Equal variances, normal distribution**
- **F* = 10.313**
- **$F_{.05,\ 3,\ 36\ d.f.} = 2.866$**
- **F* > F**
- **10.313 > 2.866**
- **Reject H_0: $\mu_1=\mu_2=\mu_3=\mu_4$**

There are differences across hair color

Two-Factor Approach

- **Now I will take into account Gender as a second factor**
- **I will use Excel**
  - **Tools**
  - **Data Analysis**
  - **ANOVA: Two-Factor With Replication**
    - Include the row labels
    - Note there are 5 rows per sample
    - Set alpha and where the output should go

Anova: Two-Factor With Replication

SUMMARY	Light Blonde	Dark Blonde	Light Brunette	Dark Brunette	Total
Males					
Count	5	5	5	5	20
Sum	289	256	205	187	937
Average	57.800	51.200	41.000	37.400	46.850
Variance	41.200	86.200	33.500	69.300	117.503
Females					
Count	5	5	5	5	20
Sum	337	298	271	240	1146
Average	67.400	59.600	54.200	48.000	57.300
Variance	59.300	51.300	35.200	11.500	86.642
Total					
Count	10	10	10	10	
Sum	626	554	476	427	
Average	62.600	55.400	47.600	42.700	
Variance	70.267	80.711	78.933	67.122	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	1092.025	1	1092.025	22.545	0.000	4.149
Columns	2297.475	3	765.825	15.811	0.000	2.901
Interaction	31.275	3	10.425	0.215	0.885	2.901
Within	1550.000	32	48.438			
Total	4970.775	39				

R Squared	**0.688**	Nearly 69% of the variability in pain threshold is exaplined by hair color and gender

ANOVA: Two Factor with Replication

- **Shows the means for hair color and gender**
- **In the ANOVA Table**
  - **Sample represents gender, with 2-1 =1 degree of freedom**
  - **Columns represents the hair color with 4-1=3 degrees of freedom**
  - **The model always tests for an Interaction between the two factors – something different happening with the presence of both factors**

ANOVA: Two Factor with Replication

- **Notice that the SS and the MS for hair color is the same in the One-Way and Two-Way models**
- **But now in the Two-Way model, gender accounts for some of the variability in pain**
  - **Thus, the SSE is reduced**
  - **R Squares goes up to .688**

ANOVA: Two Factor with Replication

- **Hypothesis Test**
  - **For Gender, look at the Sample row: F* = 22.545, p-value < .001**
  - **For Hair Color, look at the Columns Row: F* = 15.811 p-value < .001**
  - **For the Interaction: F* = .215 p-value = .885**

Randomized Block Design (Not in Book)

- In a Randomized Block Design, we use matched sets of experimental units to each treatment
- The matched sets of experimental units are called **BLOCKS**
- The strategy is to reduce the MSE by accounting for another source of variability in the data (other than the factors)

Definition

- **Randomized Block Designs consist of two steps**
  - Matched sets of experimental units, called Blocks, are formed, each block consisting of k experimental units (where k is the number of treatments).
  - The blocks should consist of experimental units that are as similar as possible.
  - One experimental unit from each block is randomly assigned to each treatment, resulting in n=bk responses.

Randomized Block Design

- The blocks can be experimental units matched on a set of characteristics – age, strength, weight, race, soil type
- **Or the same units**
  - The book gives a good example relating to golfers hitting four different club brands
  - Rather than 40 different golfers randomly assigned to four different brands
  - A Randomized Block Design would have 10 golfers each hitting a ball using each brand of club
  - This allows us to account for the effect of the blocks

Randomized Block Design

- The Blocks account for some of the variation in the response variable
  - **Sum of Squares for Blocks (SSB)**

$$SSB = \sum_{i=1}^{b} k(y_{B_i} - \overline{Y})^2$$

- And the SSE is a residual, derived as:
  - **SSE = SS(Total) - SST - SSB**

Randomized Block Design

- **Degrees of Freedom**
  - **SS(Total)** **n-1**
  - **SST** **k-1**
  - **SSB** **b-1**
  - **SSE** **n-b-k+1**
  - **or** **(b-1)(k-1)**

Note: n=bk

Housing Appraisal Example

- A bank out-sources home appraisals to three companies
- They want to conduct a formal test to see if the companies tend to differ on average – some being higher or lower
- One approach would be to randomly assign different homes to different companies

Housing Appraisal Example

- However, even a random approach could lead to one company receiving larger or higher quality homes relative to the others
- The Blocks in this case would need to match similar homes by size, quality, and location
- Or, they could assign the same house to all three companies, and try to account for that variation in our model

The Data

- Three companies (Factor Levels)
- Five houses (Blocks)
- Response variable is the appraised value in $1,000s

	Appriasal Companies			**Block**
Property	**Allen**	**Heist**	**AL**	**Mean**
1	78	82	79	79.67
2	102	102	99	101.00
3	68	74	70	70.67
4	83	88	86	85.67
5	95	99	92	95.33
Factor Mean	85.20	89.00	85.20	86.47
Grand Mean	86.47			

Single Factor Approach

$R^2 = 48.13/1829 = .026!!$

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Allen	5	426	85.2	182.7		
Heist	5	445	89	136		
AL	5	426	85.2	126.7		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	48.133	2	24.067	0.162	0.852	3.885
Within Groups	1781.600	12	148.467			
Total	1829.733	14				

This test indicates there are no differences between the companies, F* = .162

Block Approach using Excel

- Use Tools, Data Analysis
- ANOVA: **Two-Factor Without Replication**
  - The Rows reflect the Blocks
  - The Columns the Factor Levels

Excel Output

Anova: Two-Factor Without Replication				
SUMMARY	*Count*	*Sum*	*Average*	*Variance*
1	3	239	79.67	4.33
2	3	303	101.00	3.00
3	3	212	70.67	9.33
4	3	257	85.67	6.33
5	3	286	95.33	12.33
Allen	5	426	85.20	182.70
Heist	5	445	89.00	136.00
Al	5	426	85.20	126.70

Excel Output

- Rows reflect Blocks SSB
- Columns reflect Treatments SST
- R-square is 1 – (22.53/1829.73) = **.9877**

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1759.07	4	439.77	156.13	0.00	3.84
Columns	48.13	2	24.07	8.54	0.01	4.46
Error	22.53	8	2.82			
Total	1829.73	14				

Most of the variability in the data is due to the Blocks

Notice

- **The Sum of Squares due to the Factors (I.e., the companies was the same in both approaches**
  - **48.13 for SST**
  - **24.07 for the MST**
- **Accounting for the Blocks (properties) reduces the SSE**

ANOVA Hypothesis Test for the Factor

- **Null hypothesis**
- **Alternative**
- **Assumptions**
- **Test Statistic**
- **Rejection Region**
- **Conclusion**

- **H_0: $\mu_1 = \mu_2 = \mu_3$**
- **H_a: At least two means differ**
- **Equal variances, normal distribution**
- **F* = 8.54**
- **$F_{.05, 2, 8 \text{ d.f.}} = 4.46$**
- **F* > F**
- **8.54 > 4.46**
- **Reject H_0: $\mu_1 = \mu_2 = \mu_3$**

There are differences across companies

ANOVA Hypothesis Test for the Blocks

- **Null hypothesis** — $H_0: \mu_{b1} = \mu_{b2} = \mu_{b3} = \mu_{b4} = \mu_{b5}$
- **Alternative** — H_a: **At least two means differ**
- **Assumptions** — **Equal variances, normal distribution**
- **Test Statistic** — **$F^* = 156.13$**
- **Rejection Region** — **$F_{.05,\ 4,\ 8\ d.f.} = 3.84$**
- **Conclusion**
  - **$F^* > F$**
  - **$156.13 > 3.84$**
  - **Reject $H_0: \mu_{b1} = \mu_{b2} = \mu_{b3} = \mu_{b4} = \mu_{b5}$**

It was useful to account (control) for the Blocks

SAS Output for Appraisal Data

```
The ANOVA Procedure
       Class Level Information
   Class          Levels    Values
   BLOCK               5     1 2 3 4 5
   COMPANY             3     1 2 3
      Number of observations    15

Dependent Variable: VALUE
                              Sum of
Source              DF       Squares       Mean Square     F Value     Pr >
F
Model                6     1807.200000      301.200000      106.93
<.0001
Error                8       22.533333        2.816667
Corrected Total     14     1829.733333

                R-Square      Coeff Var       Root MSE     VALUE Mean
                0.987685       1.940971       1.678293       86.46667

Source              DF       Anova SS      Mean Square     F Value     Pr > F
BLOCK               4     1759.066667      439.766667      156.13     <.0001
COMPANY             2       48.133333       24.066667        8.54     0.0103
```

SAS gives additional information – difference of means tests

```
Bonferroni (Dunn) t Tests for VALUE
NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha 0.05
Error Degrees of Freedom              8
Error Mean Square              2.816667
Critical Value of t             3.01576
Minimum Significant Difference   3.2011

Means with the same letter are not significantly different.
Bon Grouping   Mean      N    COMPANY
     A        89.000     5    2

     B        85.200     5    1
     B
     B        85.200     5    3
```

Difference of Means Tests with multiple comparisons

- **Covered on pages 665-71 in the book**
- **Here is what I want you to be aware of:**
  - When conducting multiple comparisons of means (more than 2) the level of alpha will not be constant across the levels - it changes
  - If we expect 5% error on a single test, or 5 out of 100 times we will be wrong in our conclusions
  - Then when we conduct multiple tests from the same sample we will end up with a higher level of alpha across all the tests

Difference of Means Tests with multiple comparisons

- To keep a constant level of alpha for all of the test, we need to adjust the rejection region
- This is called the experiment-wise error rate (EER)
- **The methods used are:**
  - **Tukey**
  - **Bonferroni**
  - **Scheffe**

SAS gives additional information – difference of means tests

```
Bonferroni (Dunn) t Tests for VALUE
NOTE: This test controls the Type I experimentwise error rate,
but it generally has a higher Type II error rate than REGWQ.

Alpha 0.05
Error Degrees of Freedom              8
Error Mean Square              2.816667
Critical Value of t             3.01576
Minimum Significant Difference   3.2011

Means with the same letter are not significantly different.
Bon Grouping   Mean      N    COMPANY
     A        89.000     5    2

     B        85.200     5    1
     B
     B        85.200     5    3
```