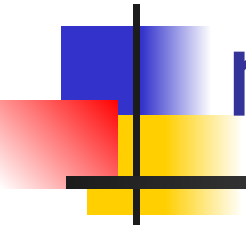
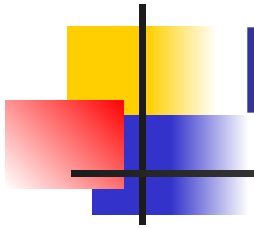
Review of ANOVA and linear regression





Review of simple ANOVA

ANOVA

for comparing means between more than 2 groups

Why ANOVA?

- ~~In real life things do not typically result in two~~ groups being compared
  - Test lines on I-64 in Frankfort
- Two-sample t-tests are problematic
  - Increasing the risk of a Type I error
  - At .05 level of significance, with 100 comparisons, 5 will show a difference when none exists (experimentwise error)
  - So the more t-tests you run, the greater the risk of a type I error (rejecting the null when there is no difference)
- ANOVA allows us to see if there are differences between means with an **OMNIBUS** test

When ANOVA?

- Data must be experimental
- If you do not have access to statistical software, an ANOVA can be computed by hand
- With many experimental designs, the sample sizes must be equal for the various factor level combinations
- A regression analysis will accomplish the same goal as an ANOVA.
- ANOVA formulas change from one experimental design to another

Variance – why do scores vary?

- A representation of the spread of scores
- What contributes to differences in scores?
  - Individual differences
  - Which group you are in

Variance to compare Means

- We are applying the variance concept to means
  - How do means of different groups compare to the overall mean
- Do the means vary so greatly from each other that they exceed individual differences within the groups?

Between/Within Groups

- Variance can be separated into two major components
  - **Within groups** – variability or differences in particular groups (individual differences)
  - **Between groups** - differences depending what group one is in or what treatment is received

Formulas: page 550

Bottom Line

- We are examining the ratio of differences (variances) from **treatment** to variances from **individual differences**
- If the ratio is large there is a significant impact from treatment.
- We know if a ratio is "large enough" by calculating the ratio of the MST to MSE and conducting an F test.

Fundamental Concepts

- You are able to compare MULTIPLE means
- Between-group variance reflects differences in the way the groups were treated
- Within-group variance reflects individual differences
- Null hypothesis: no difference in means
- Alternative hypothesis: difference in

Sum of Squares

- We are comparing "variance estimates"
  - Variance = SS/df
- The charge is to partition the variance into between and within group variance
- Critical factors:
  - BETWEEN GROUP VARIANCE
  - WITHIN GROUP VARIANCE
- How does the between group variance compare with the within group variance?

Designed Experiments of Interest

- One-factor completely randomized designs (Formulas: p. 558)

  Total SS = Treatment SS + Error SS

  SS(Total) = SST + SSE

- Randomized Block Designs (Formulas: p. 575)

  Total SS = Treatment SS + Block SS + Error SS

  SS(Total) = SST + SSB + SSE

- Two-Factor Factorial Experiments (Formulas: p. 593)

  Total SS = Main effect SS Factor A + Main effect SS Factor B + AB Interaction SS + Error SS

  SS(Total) = SS(A) + SS (B) + SS (AB) + SSE

Word check

- When I talk about between groups variability, what am I talking about?
- What does SS between represent?
- What does MS (either within or between) represent?
- What does the F ratio represent?

Multiple Comparisons (do the pairs of numbers capture 0)

THESE ARE CONFIDENCE INTERVALS

We can tell if there are differences but now we must determine which is better

- See MINITAB (Tukey family error rate)

Tukey's pairwise comparisons

Intervals for (column level mean) - (row level mean)

	1	2	3
2	-3.854 1.320		
3	-4.467 0.467	-3.320 1.854	
4	-6.854 -1.680	-5.702 -0.298	-4.854 0.320

Hypotheses of One-Way ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \square = \mu_c$
  - All population means are equal
  - i.e., no treatment effect (no variation in means among groups)

- H_1 : Not all of the population means are the same
  - At least one population mean is different
  - i.e., there is a treatment effect
  - Does not mean that all population means are different (some pairs may be the same)

The F-distribution

- A ratio of variances follows an F-distribution:

$$\frac{\sigma^2_{between}}{\sigma^2_{within}} \sim F_{n,m}$$

- The F-test tests the hypothesis that two variances are equal.
- F will be close to 1 if sample variances are equal.

$$H_0 : \sigma^2_{between} = \sigma^2_{within}$$
$$H_a : \sigma^2_{between} \neq \sigma^2_{within}$$

How to calculate ANOVA's by hand…

Treatment 1	Treatment 2	Treatment 3	Treatment 4
y_{11}	y_{21}	y_{31}	y_{41}
y_{12}	y_{22}	y_{32}	y_{42}
y_{13}	y_{23}	y_{33}	y_{43}
y_{14}	y_{24}	y_{34}	y_{44}
y_{15}	y_{25}	y_{35}	y_{45}
y_{16}	y_{26}	y_{36}	y_{46}
y_{17}	y_{27}	y_{37}	y_{47}
y_{18}	y_{28}	y_{38}	y_{48}
y_{19}	y_{29}	y_{39}	y_{49}
y_{110}	y_{210}	y_{310}	y_{410}

n=10 obs./group

k=4 groups

$$\bar{y}_{1\bullet} = \frac{\sum_{j=1}^{10} y_{1j}}{10} \qquad \bar{y}_{2\bullet} = \frac{\sum_{j=1}^{10} y_{2j}}{10} \qquad \bar{y}_{3\bullet} = \frac{\sum_{j=1}^{10} y_{3j}}{10} \qquad \bar{y}_{4\bullet} = \frac{\sum_{j=1}^{10} y_{4j}}{10}$$

The group means

$$\frac{\sum_{j=1}^{10}(y_{1j} - \bar{y}_{1\bullet})^2}{10-1} \qquad \frac{\sum_{j=1}^{10}(y_{2j} - \bar{y}_{2\bullet})^2}{10-1} \qquad \frac{\sum_{j=1}^{10}(y_{3j} - \bar{y}_{3\bullet})^2}{10-1} \qquad \frac{\sum_{j=1}^{10}(y_{4j} - \bar{y}_{4\bullet})^2}{10-1}$$

The (within) group variances

Sum of Squares Within (SSW), or Sum of Squares Error (SSE)

$$\frac{\sum_{j=1}^{10}(y_{1j}-\bar{y}_{1\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{2j}-\bar{y}_{2\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{3j}-\bar{y}_{3\bullet})^2}{10-1} \quad \frac{\sum_{j=1}^{10}(y_{4j}-\bar{y}_{4\bullet})^2}{10-1}$$

The (within) group variances

$$\sum_{j=1}^{10}(y_{1j}-\bar{y}_{1\bullet})^2+\sum_{j=1}^{10}(y_{2j}-\bar{y}_{2\bullet})^2+\sum_{j=3}^{10}(y_{3j}-\bar{y}_{3\bullet})^2+\sum_{j=1}^{10}(y_{4j}-\bar{y}_{4\bullet})^2$$

$$=\sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij}-\bar{y}_{i\bullet})^2$$

Sum of Squares Within (SSW) (or SSE, for chance error)

Sum of Squares Between (SSB), or Sum of Squares Regression (SSR)

Overall mean of all 40 observations ("grand mean")

$$\bar{\bar{y}}_{\bullet\bullet} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{10} y_{ij}}{40}$$

$$10x\sum_{i=1}^{4}(\bar{y}_{i\bullet} - \bar{\bar{y}}_{\bullet\bullet})^2$$

Sum of Squares Between (SSB). Variability of the group means compared to the grand mean (the variability due to the treatment).

Total Sum of Squares (SST)

$$\sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij}-\bar{\bar{y}}_{\bullet\bullet})^2$$

Total sum of squares(TSS). Squared difference of every observation from the overall mean. (numerator of variance of Y!)

Partitioning of Variance

$$\sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij}-\bar{y}_{i\bullet})^2 \;+\; 10x\sum_{i=1}^{4}(\bar{y}_{i\bullet}-\bar{\bar{y}}_{\bullet\bullet})^2 \;=\; \sum_{i=1}^{4}\sum_{j=1}^{10}(y_{ij}-\bar{\bar{y}}_{\bullet\bullet})^2$$

SSW + SSB = TSS

ANOVA Table

Source of variation	d.f.	Sum of squares	Mean Sum of Squares	F-statistic	p-value
Between (k groups)	**k-1**	**SSB** (sum of squared deviations of group means from grand mean)	SSB/k-1	$\frac{SSB/k-1}{SSW/nk-k}$	**Go to $F_{k-1,nk-k}$ chart**
Within (n individuals per group)	**nk-k**	**SSW** (sum of squared deviations of observations from their group mean)	**s^2=SSW/nk-k**		
Total variation	**nk-1**	**TSS** (sum of squared deviations of observations from grand mean)		TSS=SSB + SSW	

Example

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

Example

Step 1) calculate the sum of squares between groups:

Mean for group 1 = 62.0

Mean for group 2 = 59.7

Mean for group 3 = 56.3

Mean for group 4 = 61.4

Grand mean= 59.85

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

$SSB = [(62-59.85)^2 + (59.7-59.85)^2 + (56.3-59.85)^2 + (61.4-59.85)^2] \; xn \; per \; group = 19.65 x10 = \mathbf{196.5}$

Example

Step 2) calculate the sum of squares within groups:

$(60-62)^2+(67-62)^2+(42-62)^2+(67-62)^2+(56-62)^2+(62-62)^2+(64-62)^2+(59-62)^2+(72-62)^2+(71-62)^2+(50-59.7)^2+(52-59.7)^2+(43-59.7)^2+67-59.7)^2+(67-59.7)^2+(69-59.7)^2$…+….(sum of 40 squared deviations) = **2060.6**

Treatment 1	Treatment 2	Treatment 3	Treatment 4
60 inches	50	48	47
67	52	49	67
42	43	50	54
67	67	55	67
56	67	56	68
62	59	61	65
64	67	61	65
59	64	60	56
72	63	59	60
71	65	64	65

Step 3) Fill in the ANOVA table

Source of variation	d.f.	Sum of squares	Mean Sum of Squares	F-statistic	p-value
Between	3	**196.5**	65.5	1.14	**.344**
Within	36	**2060.6**	57.2		
Total	39	2257.1			

Step 3) Fill in the ANOVA table

Source of variation	d.f.	Sum of squares	Mean Sum of Squares	F-statistic	p-value
Between	3	**196.5**	65.5	1.14	**.344**
Within	36	**2060.6**	57.2		
Total	39	2257.1			

INTERPRETATION of ANOVA:

How much of the variance in height is explained by treatment group?

R^2="Coefficient of Determination" = SSB/TSS = 196.5/2275.1=9%

Coefficient of Determination

$$R^2 = \frac{SSB}{SSB + SSE} = \frac{SSB}{SST}$$

The amount of variation in the outcome variable (dependent variable) that is explained by the predictor (independent variable).

ANOVA example

Table 6. Mean micronutrient intake from the school lunch by school

		S1[a], *n*=25	**S2[b], *n*=25**	**S3[c], *n*=25**	***P*-value[d]**
Calcium (mg)	Mean	117.8	158.7	206.5	**0.000**
	SD[e]	62.4	70.5	86.2	
Iron (mg)	Mean	2.0	2.0	2.0	0.854
	SD	0.6	0.6	0.6	
Folate (µg)	Mean	26.6	38.7	42.6	**0.000**
	SD	13.1	14.5	15.1	
Zinc (mg)	Mean	1.9	1.5	1.3	0.055
	SD	1.0	1.2	0.4	

[a] School 1 (most deprived; 40% subsidized lunches).
[b] School 2 (medium deprived; <10% subsidized).
[c] School 3 (least deprived; no subsidization, private school).
[d] ANOVA; significant differences are highlighted in bold (*P*<0.05).

FROM: Gould R, Russell J, Barker ME. School lunch menus and 11 to 12 year old children's food choice in three secondary schools in England-are the nutritional standards being met? *Appetite*. 2006 Jan;46(1):86-92.

Answer

Step 1) calculate the sum of squares between groups:

Mean for School 1 = 117.8

Mean for School 2 = 158.7

Mean for School 3 = 206.5

Grand mean: 161

SSB = $[(117.8-161)^2 + (158.7-161)^2 + (206.5-161)^2]$ *x25 per group*= 98,113

Answer

Step 2) calculate the sum of squares within groups:

S.D. for S1 = 62.4

S.D. for S2 = 70.5

S.D. for S3 = 86.2

Therefore, sum of squares within is:

$(24)[\ 62.4^2 + 70.5^2 + 86.2^2] = 391{,}066$

Answer

Step 3) **Fill in your ANOVA table**

Source of variation	d.f.	Sum of squares	Mean Sum of Squares	F-statistic	p-value
Between	2	98,113	49056	9	<.05
Within	72	391,066	5431		
Total	74	489,179			

****R^2=98113/489179=20%**

School explains 20% of the variance in lunchtime calcium intake in these kids.

Beyond one-way ANOVA

Often, you may want to test more than 1 treatment. ANOVA can accommodate more than 1 treatment or factor, so long as they are independent. Again, the variation partitions beautifully!

$$TSS = SSB1 + SSB2 + SSW$$