STA 570 Spring 2011

Lecture 23 *Tuesday, April 19*

- **12.1 Analysis of Variance**

Summary: Investigating Association Between Two Variables

		Response Variable		
		Unordered Categorical (Nominal)	Ordinal	Quantitative
Explanatory Variable	Nominal with 2 Levels	Comparing Proportions of 2 Independent Samples	Wilcoxon-Mann-Whitney Rank Sum Test	Comparing Means of 2 Independent Samples
	Unordered Categorical (Nominal) More than 2 Levels	?	?	Analysis of Variance
	Ordinal		?	
	Quantitative	?		Regression

Analysis of Variance

- Earlier
  - Comparing Two Samples (Binary, Ordinal, Quantitative)
  - Example: Compare mean annual income of men and women
- Now
  - Several samples, Quantitative variable
  - Comparing Several Means
  - Example: Compare mean statistics exam score for students from several different departments/colleges

Analysis of Variance

- Short: ANOVA
- Developed by Sir R.A. Fisher in the 1920s for agricultural data
- Uses the *F*-distribution (named after Fisher)
- Goal: Detect evidence of differences between population means

ANOVA Notation

- Number of groups: g
- Means of the response variables for the g populations: $\mu_1, \mu_2, \ldots, \mu_g$
- Null hypothesis: All the means are equal, that is, $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$
- Group sample sizes: $n_1, n_2, \ldots, n_g$

  ANOVA works best when the sample sizes are equal (balanced design)
- Total sample size: $N = n_1 + n_2 + \ldots + n_g$
- Sample Means: $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_g$
- Sample standard deviations: $s_1, s_2, \ldots, s_g$

ANOVA Assumptions

- Independent random samples are selected from the *g* populations ***(very important)***
- Response is truly quantitative ***(very important)***
- The standard deviations of the population distributions for the *g* groups are equal, denoted by *sigma* *(if this is not the case, there is a statistical solution: variance stabilizing transformations)*
- The population distributions of the response variable are normal for each of the *g* groups ***(important for small sample sizes and small number of samples)***

Example

- Scores on the first quiz (maximum 10 points) in a beginning French course for ninth-grade students
- Three groups of students:
  - Group A: Never studied foreign language before, but have good English skills
  - Group B: Never studied foreign language before; have poor English skills
  - Group C: Studied other foreign language

Variability Between and Within Groups

- ANOVA compares two types of variability
  - Variability of the sample observations about their separate means (*within-group variability*)
  - Variability of the sample means from the different groups about the overall mean (*between-group variability*)

Variability Between and Within Groups

- ***The greater the variability between sample means and***

  The smaller the variability within each group of observations,

  The stronger the evidence that the null hypothesis of equal means is false
- The test statistic is the ratio of two variance estimates:
  - Between-groups estimate divided by
  - Within-groups estimate

Within-Groups Estimate of Variance Technical Details

- For each of the *g* groups, a variance estimate can be calculated
- Construct a ***pooled variance estimate*** by adding up weighted group variance estimates
- The weights are the degrees of freedom for each group: $n_i - 1$
- Divide by the total degrees of freedom: $N - g$

Within-Groups Estimate of Variance Technical Details

- This estimate is a weighted average of the separate sample variances, with greater weight given to larger samples
- It is unbiased and efficient
- Mathematical Formula:

$$\hat{\sigma}^2 = \frac{WSS}{N-g} = \frac{SSE}{N-g} = MSE$$

$$= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_g-1)s_g^2}{N-g}$$

Between-Groups Estimate of Variance

- This estimate is based on the variability between each sample mean and the overall mean (from all samples together)
- Under the null hypothesis, it is unbiased
- Mathematical Formula:

$$\frac{BSS}{g-1} = MSH$$

$$= \frac{n_1(\bar{Y}_{1.} - \bar{Y}_{..})^2 + n_2(\bar{Y}_{2.} - \bar{Y}_{..})^2 + \cdots + n_g(\bar{Y}_{g.} - \bar{Y}_{..})^2}{g-1}$$

$$\text{where } \bar{Y}_{..} = \frac{1}{N}\sum Y_{ij}$$

F Test Statistic

The test statistic for the null hypothesis

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$

is the ratio of the two variance estimates:

$$F_{OBS} = \frac{\text{Between-Groups Estimate}}{\text{Within-Groups Estimate}}$$

$$= \frac{BSS/(g-1)}{WSS/(N-g)} = \frac{MSH}{MSE}$$

Sampling distribution: F distribution with degrees of freedom $df_1 = g - 1$ and $df_2 = N - g$

P-value: Right-tail probability

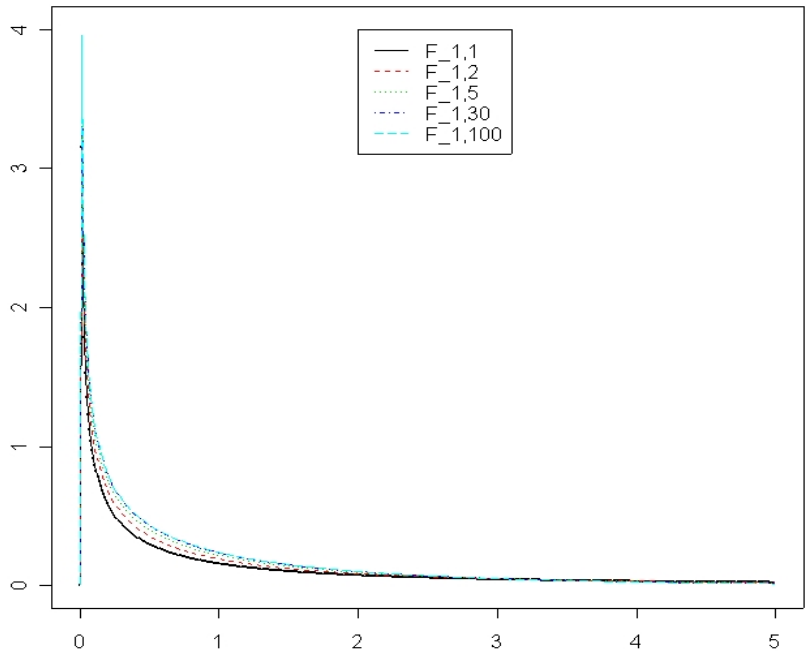
F Test Statistic

- It is called the ***analysis of variance F statistic*** or ***ANOVA F statistic***
- If the null hypothesis is true, its sampling distribution is the *F* distribution with degrees of freedom $df_1 = g-1$ and $df_2 = N-g$
- The P-value is the right-tail probability that the *F* test statistic takes a value at least as large as the observed *F* value
- The larger the *F* test statistic, the smaller the P-value

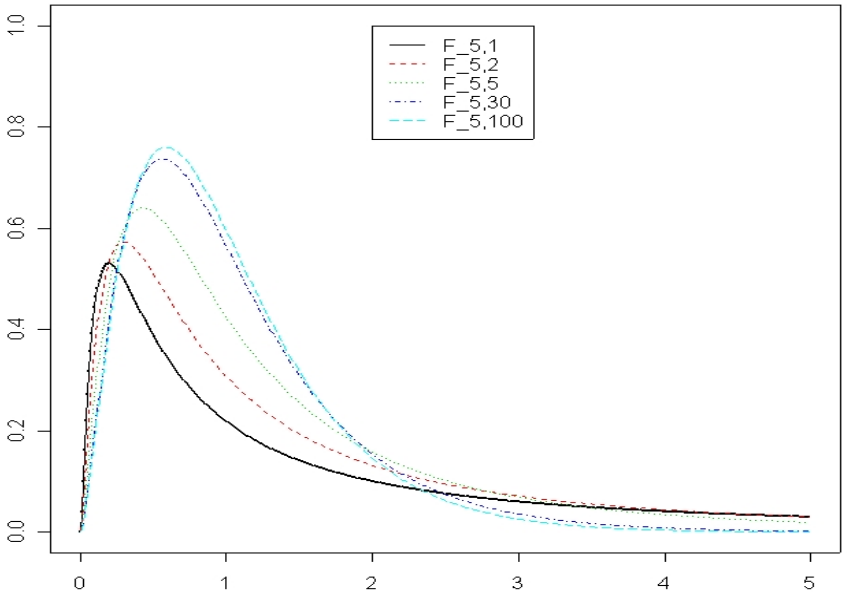
F Test Statistic

- If the null hypothesis is true, numerator and denominator are both unbiased estimates of the same quantity
- We expect the ratio to be around 1
- If the null hypothesis is not true, the numerator will be larger, and the test statistic takes larger values
- We reject the null hypothesis when the values are too large
- What is too large? Sampling distribution

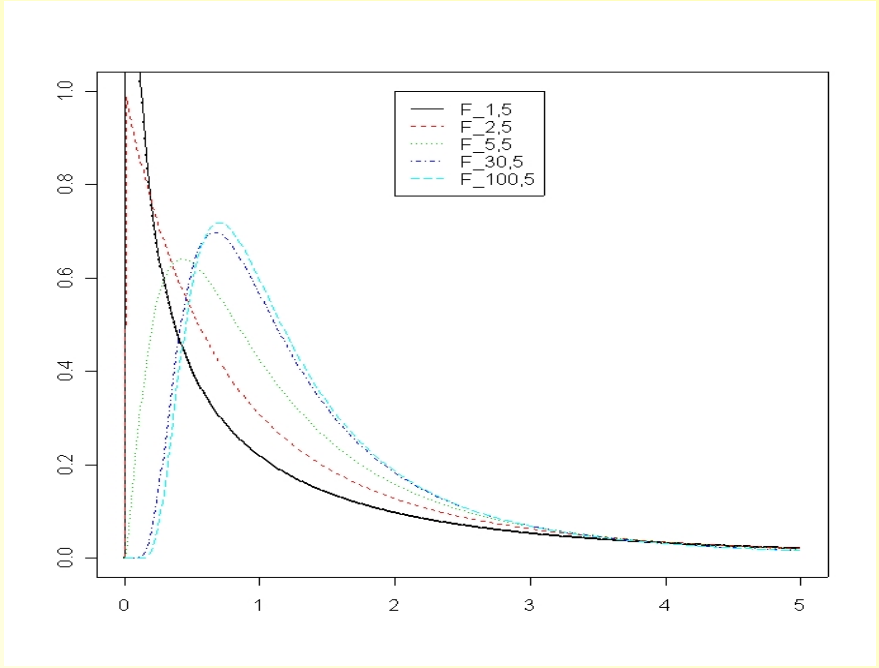
F Distributions (1,x)



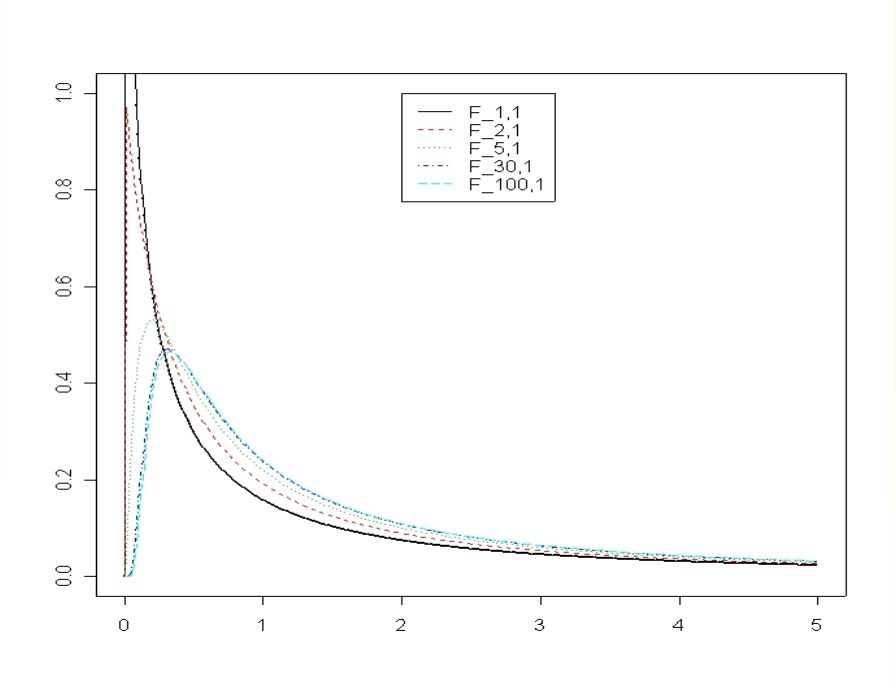
F Distributions (5,x)



F Distributions (x,5)



F Distributions (x,1)



Example (contd.)

- The quiz scores in the beginning French course are given in the table
- Calculate the F statistic for the quiz score example
- What is the P-value?
- F distribution online tool

Group A	Group B	Group C
4	1	9
6	5	10
8		5

ANOVA Table (SAS Output)

- Statistical software displays the results of ANOVA *F* tests in a table called ***ANOVA table***

The GLM Procedure

Class Level Information

Class	Levels	Values
group	3	A B C

Number of observations 8

Dependent Variable: score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	30.00000000	15.00000000	2.50	0.1768
Error	5	30.00000000	6.00000000		
Corrected Total	7	60.00000000			

Level of group	N	score Mean	score Std Dev
A	3	6.00000000	2.00000000
B	2	3.00000000	2.82842712
C	3	8.00000000	2.64575131

Another Example

Type 1	Type 2	Type 3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

There are three types of fertilizers and these numbers are plants in a particular pot.

Example cont.

- **Step 1:** Calculate the mean within each group:

$$\overline{Y}_1 = \frac{1}{6}\sum Y_{1i} = \frac{6+8+4+5+3+4}{6} = 5$$

$$\overline{Y}_2 = \frac{1}{6}\sum Y_{2i} = \frac{8+12+9+11+6+8}{6} = 9$$

$$\overline{Y}_3 = \frac{1}{6}\sum Y_{3i} = \frac{13+9+11+8+7+12}{6} = 10$$

Example cont.

- **Step 2:** Calculate the overall mean:

$$\overline{Y} = \frac{\sum_i \overline{Y}_i}{a} = \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3}{a} = \frac{5+9+10}{3} = 8$$

where *a* is the number of groups.

Step 3: Calculate the "between-group" sum of squares (BSS):

$$S_B = n(\overline{Y}_1 - \overline{Y})^2 + n(\overline{Y}_2 - \overline{Y})^2 + n(\overline{Y}_3 - \overline{Y})^2$$

$$= 6(5-8)^2 + 6(9-8)^2 + 6(10-8)^2 = 84$$

where *n* is the number of data values per group.

Example cont.

- The between-group degrees of freedom is one less than the number of groups

  $df_1 = 3 - 1 = 2$

so the between-group mean square value is

$MSH = 84 / 2 = 42$

Example cont.

- **Step 4:** Calculate the "within-group" sum of squares. Begin by centering the data in each group

Type 1	Type 2	Type 3
6 − 5 = 1	8 − 9 = -1	13 − 10 = 3
8 − 5 = 3	12 − 9 = 3	9 − 10 = -1
4 − 5 = -1	9 − 9 = 0	11 − 10 = 1
5 − 5 = 0	11 − 9 = 2	8 − 10 = -2
3 − 5 = -2	6 − 9 = -3	7 − 10 = -3
4 − 5 = -1	8 − 9 = -1	12 − 10 = 2

Example cont

The within-group sum of squares is the sum of squares of all 18 values in this table

$S_W = 1 + 9 + 1 + 0 + 4 + 1 + 1 + 9 + 0 + 4 + 9 + 1 + 9 + 1 + 1 + 4 + 9 + 4 = 68$

The within-group degrees of freedom is

$$df_2 = a(n - 1) = 3(6 - 1) = 15$$

Thus the within-group mean square value is

$MSE = WSS/df_2 = 68/15 = 4.5$

Example cont.

- **Step 5:** The F-ratio is
  MSH/MSE = 42/4.5 = 9.3
  In this case, $F_{crit}(2,15) = 3.68$ at $\alpha = 0.05$.
  Since $F = 9.3 > 3.68$, the results are significant at the 5% significance level.
  One would reject the null hypothesis, concluding that there is strong evidence that the expected values in the three groups differ. The p-value for this test is 0.002.

Another Example

- Three materials for making artificial teeth are compared with regard to hardness.
- The materials are Endura, Duradent, and Duracross.
- Six pairs of teeth are tested for each material.
- The response variable is the Vickers microhardness of the occlusal surfaces, measured with a load of 50 g and a loading time of 30 sec.

Example: Hardness of Artificial Teeth (contd.)

- Data table, with sample means and standard deviations

	Endura	Duradent	Duracross
Hardness	27.1 27.6 28 28.5 27.3 26.7	23.9 24.5 23.9 24.4 22.9 24.5	44.9 37.9 40.4 38.5 40.4 35.7
Sample Mean	*27.53*	*24.02*	*39.63*
Sample Standard Deviation	*0.65*	*0.61*	*3.12*

ANOVA Table (from SAS)

The SAS System 19:52 Monday, March 31, 2008
The GLM Procedure

Class Level Information

Class	Levels	Values
MATERIAL	3	Duracros Duradent Endura

Number of Observations Read 18
Number of Observations Used 18

Dependent Variable: HARDNESS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**2**	**805.3144444**	**402.6572222**	**114.71**	**<.0001**
Error	**15**	**52.6550000**	**3.5103333**		
Corrected Total	**17**	**857.9694444**			

Interpretation

- The null hypothesis for the ANOVA is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$$

- The P-value from the ANOVA table is <0.0001
- At 5% level, there is sufficient evidence against the null hypothesis
- So we can conclude that not all the population means are equal

Interpretation, contd.

- However, the conclusion of the test does not specify which means are different or how different they are
- More detailed inference is necessary to determine the nature of the differences

Multiple Comparisons of Means

- Confidence intervals are usually more informative than test results
- In practice, we would be interested in estimates of the population means and confidence intervals for their differences
- Compare groups A vs. B, A vs. C, B vs. C
- We can also perform pairwise t-tests
- “post-hoc (*after this*) analysis”

Multiple Comparisons of Means

- When we have many groups, the number of pairwise comparisons [(g)(g-1)/2] can be very large
- g=3: 3 comparisons
- g=4: (4)(3)/2=6 comparisons
- g=5: (5)(4)/2=10 comparisons
- g=10: (10)(9)/2=45 comparisons
- g=20: (20)(19)/2=190 comparisons

Dangers of Forming Many Confidence Intervals

- When g=20, we compare 190 pairs of means
- Suppose we form a 95% confidence interval for the difference between each pair
- Interpretation of confidence interval: In the long run, about 95% of them contain the true difference in means
- So, about 5% of them are not expected to contain the true difference
- 5% of 190 is (190)(0.05)=9.5

Dangers of Forming Many Confidence Intervals

- Suppose that in fact all the population means are equal
- With 20 groups, we expect that, *just by chance*, about 10 confidence intervals for pairwise differences will not contain 0
- The chance of at least one incorrect pairwise inference increases with the number of groups

Multiple Comparison Error Rate

- The probability that at least one interval is in error, not containing the true difference in means, is called the ***multiple comparison error rate*** or ***experimentwise error rate***
- The multiple comparison error rate is considerably larger than the error probability for one particular interval

Simultaneous Confidence Intervals

- Control the probability that *all* intervals contain the true differences
- "We are 95% confident that *all* intervals simultaneously contain the correct difference of means"
- A multiple comparison procedure that yields a set of simultaneous confidence intervals is the Bonferroni procedure

Bonferroni Procedure

- Assume we have $g=4$ groups, therefore 6 pairwise comparisons
- Suppose we want a multiple comparison error rate of 0.10
- That is, the probability that at least one interval is in error, is less than 10%

Bonferroni Procedure

- Bonferroni procedure: divide alpha=0.10 by the number of comparisons=6
- Result: 0.0167
- Use this number (0.0167) as the new error probability for *individual* confidence intervals
- That is, construct pairwise 98.33 confidence intervals
- They are wider than pairwise 90% confidence intervals
- *That is the price that we pay for making multiple comparisons*

Artificial Teeth Example (contd.)

Task:

Construct ***simultaneous*** 95% confidence intervals for the differences in hardness for each pair of materials.

Interpret the results and provide a diagram that indicates which types of material, if any, are judged to be different in mean hardness.

Example

- 3 groups
- 3 pairwise comparisons
  (Duracross-Duradent, Duracross-Endura, Endura-Duradent)
- If alpha=0.05 for the multiple comparison error rate, then the individual error rate is 0.05/3=0.0167=1.67%
- So, we construct 100%-1.67%=***98.33%*** confidence intervals ***for each pair***
- We will get a ***95%*** "***simultaneous (experimentwise)*** confidence level"

Multiple Comparisons Using SAS

```
data teeth;
input hardness material$;
cards;
27.1 Endura
27.6 Endura
28       Endura
28.5 Endura
27.3 Endura
26.7 Endura
23.9 Duradent
24.5 Duradent
23.9 Duradent
24.4 Duradent
22.9 Duradent
24.5 Duradent
44.9 Duracross
37.9 Duracross
40.4 Duracross
38.5 Duracross
40.4 Duracross
35.7 Duracross
;
```

```
proc glm data=teeth;
class material;
model hardness=material;
means material/bon
alpha=0.05;
run;
```

SAS Output

Bonferroni (Dunn) t Tests for ***hardness***

NOTE: This test controls the Type I ***experimentwise*** error rate

Alpha	***0.05***
Error Degrees of Freedom	15
Error Mean Square	3.510333
Critical Value of t	2.69374
Minimum Significant Difference	***2.9139***

Means with the same letter are not significantly different.

Bon Grouping		Mean	N	type
A	***39.633***	6	Duracros
B	***27.533***	6	Endura
C	***24.017***	6	Duradent

Interpretation

- ANOVA *F* test:
  - The population means are not all the same
- Pairwise comparisons:
  - Duracross is significantly harder than Endura and than Duradent
  - Endura is significantly harder than Duradent

Summary

- Use ANOVA to check whether population means for *g* groups are identical
- Quantitative response,

  qualitative explanatory variable (group)
- If (and ONLY IF) there is enough evidence that the population means are not all identical, perform pairwise comparisons to find out which pairs are significantly different

QUIZ

- For which of the following is not a required condition for ANOVA?

a. The populations are normally distributed.

b. The population variances are equal.

c. The samples are independent.

d. All of these choices are required conditions for ANOVA.