STA 321 Spring 2016

Lecture 17 *Tuesday, April 5*

Comparing Several Independent Samples of… …
Quantitative Data

- **12.1 Analysis of Variance**
- **12.2 Multiple Comparison of Means**

…Ordinal Data

- **12.8 Kruskal-Wallis Test**

…Nominal Data

- 8.2 Chi-Squared Test of Independence

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:

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

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

Source	SS	DF	MS	F
Treatment	BSS	g-1	BSS/(g-1)	MSH/MSE
Error	SSE	N-g	WSS/(N-g)	
Total	SS	N-1		

ANOVA 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 (experimentwise) 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 for Multiple Comparisons

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 (*post-hoc)* comparisons to find out which pairs are significantly different
  - The alpha-level needs to be adjusted: Divide the ***experimentwise alpha*** by the number of comparisons to obtain the ***individual alphas***

ANOVA Assumptions

- Moderate departures from normality and equal standard deviations can be tolerated
- Caution if
  - Samples are not random
  - Population distributions are highly skewed ***and*** the sample size/number of samples is small
  - There are large differences among the standard deviations (largest sample standard deviation several times as large as the smallest one) ***and*** the sample sizes are unequal

Multiple Choice

- Select the correct response(s).
- ANOVA provides relatively more evidence that the null hypothesis of equal population means is false

a) The smaller the "between variation" and the larger the "within variation"
b) The smaller the "between variation" and the smaller the "within variation"
c) The larger the "between variation" and the smaller the "within variation"
d) The larger the "between variation" and the larger the "within variation"

More on ANOVA

STA 671: Regression and Correlation

- Simple linear regression, elementary matrix algebra and its application to simple linear regression; general linear model, multiple regression, ***analysis of variance tables, testing of subhypotheses***, nonlinear regression, step-wise regression; partial and multiple correlation. Emphasis upon use of computer library routines; other special topics according to the interests of the class.

STA 672: Design and Analysis of Experiments

- ***Review of one-way analysis of variance; planned and unplanned individual comparisons, including contrasts and orthogonal polynomials; factorial experiments; completely randomized, randomized block, Latin square, and split-plot designs: relative efficiency, expected mean squares;*** multiple regression analysis for balanced and unbalanced experiments, analysis of covariance.

Comparing Several Independent Samples

- **Quantitative Data**
  - ❑ **Analysis of Variance**
- **Ordinal Data**
  - ❑ **Kruskal-Wallis Test**
    - Should also be used for quantitative data that does not satisfy the ANOVA assumptions
- **Nominal Data**
  - ❑ **Chi-Squared Test for Contingency Tables**

Example: on Hold

- An airline analyzed whether telephone callers to their call center would remain on hold longer if they heard
  - (A) Advertisements about the airline,
  - (B) Muzak, or
  - (C) Classical music.

	Advertise-ments	Muzak	Classical
Holding Times (min)	0,1,3,4,6	1,2,5,8,11	7,8,9,13,15

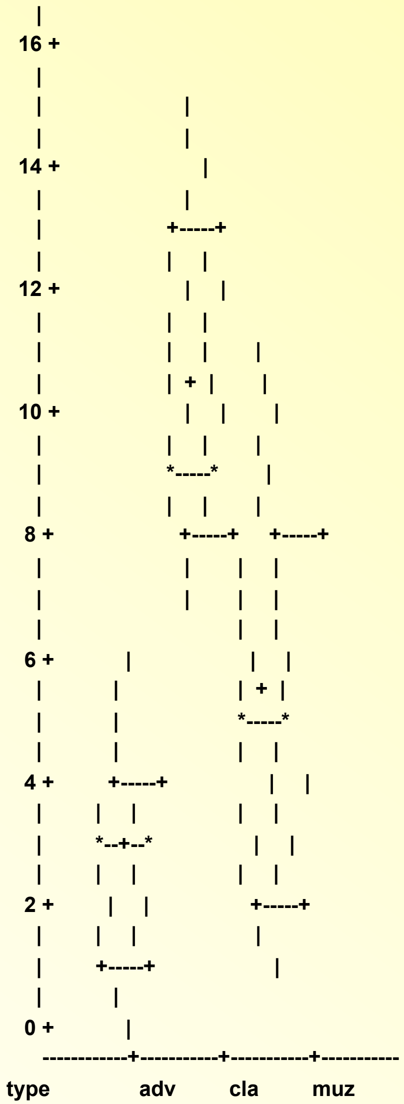
Example (on Hold, contd.)

type=**adv**
N 5
Mean 2.8
Std Deviation 2.38746728
Variance 5.7
Skewness 0.2057528
Median 3.000000

type=**cla**
N 5
Mean 10.4
Std Deviation 3.43511281
Variance 11.8
Skewness 0.60689296
Median 9.00000

type=**muz**
N 5
Mean 5.4
Std Deviation 4.15932687
Variance 17.3
Skewness 0.39746316
Median 5.000000

Schematic Plots



Example (on Hold, contd.): ANOVA Table, post-hoc analysis

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	149.2000000	74.6000000	6.43	0.0126
Error	12	139.2000000	11.6000000		
Corrected Total	14	288.4000000			

Bonferroni (Dunn) t Tests for time

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	12
Error Mean Square	11.6
Critical Value of t	2.77947
Minimum Significant Difference	5.9872

Means with the same letter are not significantly different.

Bon Grouping		Mean	N	type
	A	10.400	5	cla
	A			
B	A	5.400	5	muz
B				
B		2.800	5	adv

Recall: ANOVA Assumptions

- Moderate departures from normality and equal standard deviations can be tolerated
- Caution if
  - Samples are not random
  - Population distributions are highly skewed ***and*** the sample size/number of samples is small
  - There are large differences among the standard deviations (largest sample standard deviation several times as large as the smallest one) ***and*** the sample sizes are unequal

Not Sure if the Assumptions are Met? Kruskal-Wallis Test!

(A nonparametric test)

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable time

Classified by Variable type

type	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
adv	5	22.50	40.0	8.150372	4.50
cla	5	60.50	40.0	8.150372	12.10
muz	5	37.00	40.0	8.150372	7.40

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	7.3814
DF	2
Pr > Chi-Square	0.0250

Comparing Ordinal Samples Kruskal-Wallis Test

- The nonparametric Kruskal-Wallis test can be used to ***compare independent samples*** if
  – ***The data is quantitative, but the assumptions for an ANOVA may not be met.***
  – ***The data is ordinal.***
  – ***ANOVA can never be used for ordinal data.***
- Example for comparing ordinal data: Which instructor gives better grades in parallel classes?

Instructor	1	2	3
Grades	A C B E A	B B A C D	D C B A C

Example: Grade Comparison

Instructor	1	2	3
Grades	A C B E A	B B A C D	D C B A C

Wilcoxon Scores (Rank Sums) for Variable grade
Classified by Variable inst

inst	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	5	43.00	40.0	7.935754	8.60
2	5	40.50	40.0	7.935754	8.10
3	5	36.50	40.0	7.935754	7.30

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square 0.2276
DF 2
Pr > Chi-Square 0.8924

Comparing Nominal Samples
Chi-Squared Test of Independence

- Example: Family Structure and Sexual Activity
- Sociologists think that family structure may have an influence on sexual activity of teenagers
- 380 randomly selected females between 15 and 19 years of ages are asked to disclose
  - Family structure at age 14
  - Whether or not she has had sexual intercourse
- Response variable is binary (nominal)

Example: Family Structure, Sexual Activity

Sexual activity	Both parents	Single Parent	Parent and Stepparent	Nonparental Guardian
Yes	64	59	44	32
No	86	41	36	18

First step: Descriptive Statistics

Calculate a table with conditional proportions per column.

In this example, the different columns represent different categories of the explanatory variable.

The rows represent different categories of the response variable

Sexual activity	Both parents	Single Parent	Parent and Stepparent	Nonparental Guardian	Total
Yes					
No					
Total	100%	100%	100%	100%	100%

Comparing Nominal Samples Chi-Squared Test of Independence

- Null hypothesis: The two variables are statistically independent
- Alternative hypothesis: The variables are statistically dependent
- Even for independent variables, we do not expect the sample conditional distribution to be exactly the same
- Reason: Sampling variability

Observed and Expected Frequencies

- The chi-squared test compares the observed frequencies in the cells of the contingency table with the values that we would expect under the null hypothesis
- Notation:
  - f_o = observed frequency in a cell
  - f_e = expected frequency in a cell assuming that the variables are independent

Observed and Expected Frequencies

Sexual activity	Both parents	Single Parent	Parent and Stepparent	Nonparental Guardian	Total
Yes	64	59	44	32	199
No	86	41	36	18	181
Total	150	100	80	50	380

Observed

- The expected frequency f_e in a cell equals the product of row and column totals for that cell, divided by the total sample size

Sexual activity	Both parents	Single Parent	Parent and Stepparent	Nonparental Guardian	Total
Yes	78.55263				
No	71.44737				
Total					

Expected

Chi-Squared Test Statistic

- Karl Pearson (1900)
- Sum of the squared differences between observed and expected cell frequencies, each divided by the expected frequency

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Chi-Squared Test Statistic

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- When the null hypothesis of independence is true, then the observed frequencies are close to the expected frequencies, so the chi-squared statistic takes a relatively small value
- A large value of the chi-squared statistic is evidence *against* the null hypothesis
- In order to quantify the evidence and calculate a P-value, we need the sampling distribution of the statistic
- Chi-Squared Distribution (another online tool)