STA 321 Spring 2015

Lecture 18 *Tuesday, April 14*

- **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**

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**

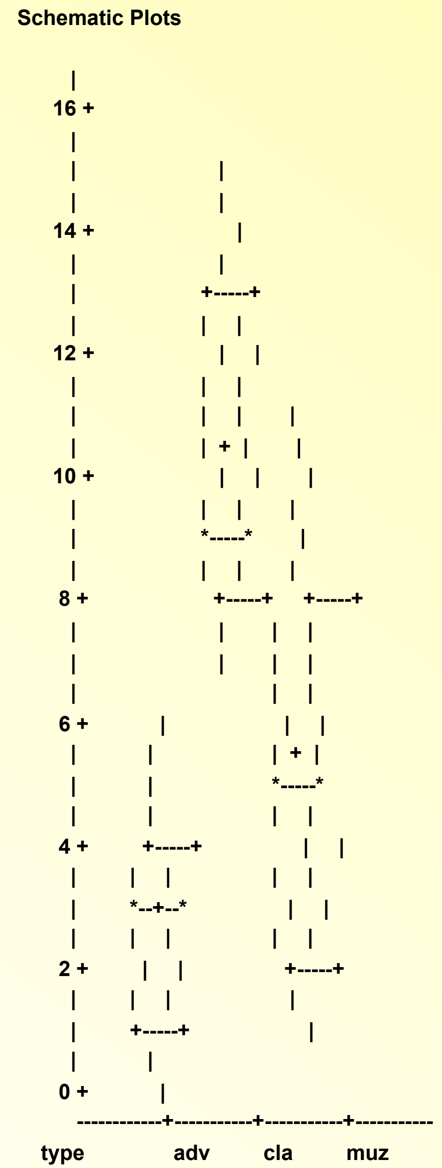
N	5
Mean	**2.8**
Std Deviation	2.38746728
Variance	5.7
Skewness	0.2057528
Median	**3.000000**

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**



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

                 Sum of     Expected      Std Dev        Mean
inst      N       Scores     Under H0     Under H0        Score
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
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	
No	86	41	36	18	
Total					

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					
No					
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)

Chi-Squared Test of Independence

- Assumptions
  - Two categorical variables
  - Random sampling (perhaps stratified with respect to the categories of one variable)
  - Expected cell count at least 5 in all cells
- Hypotheses
  - Null hypothesis: Statistical independence of the two variables
  - Alternative hypothesis: Statistical dependence

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)

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	78.6	52.4	41.9	26.2	199
No	71.4	47.6	38.1	23.8	181
Total	150	100	80	50	380

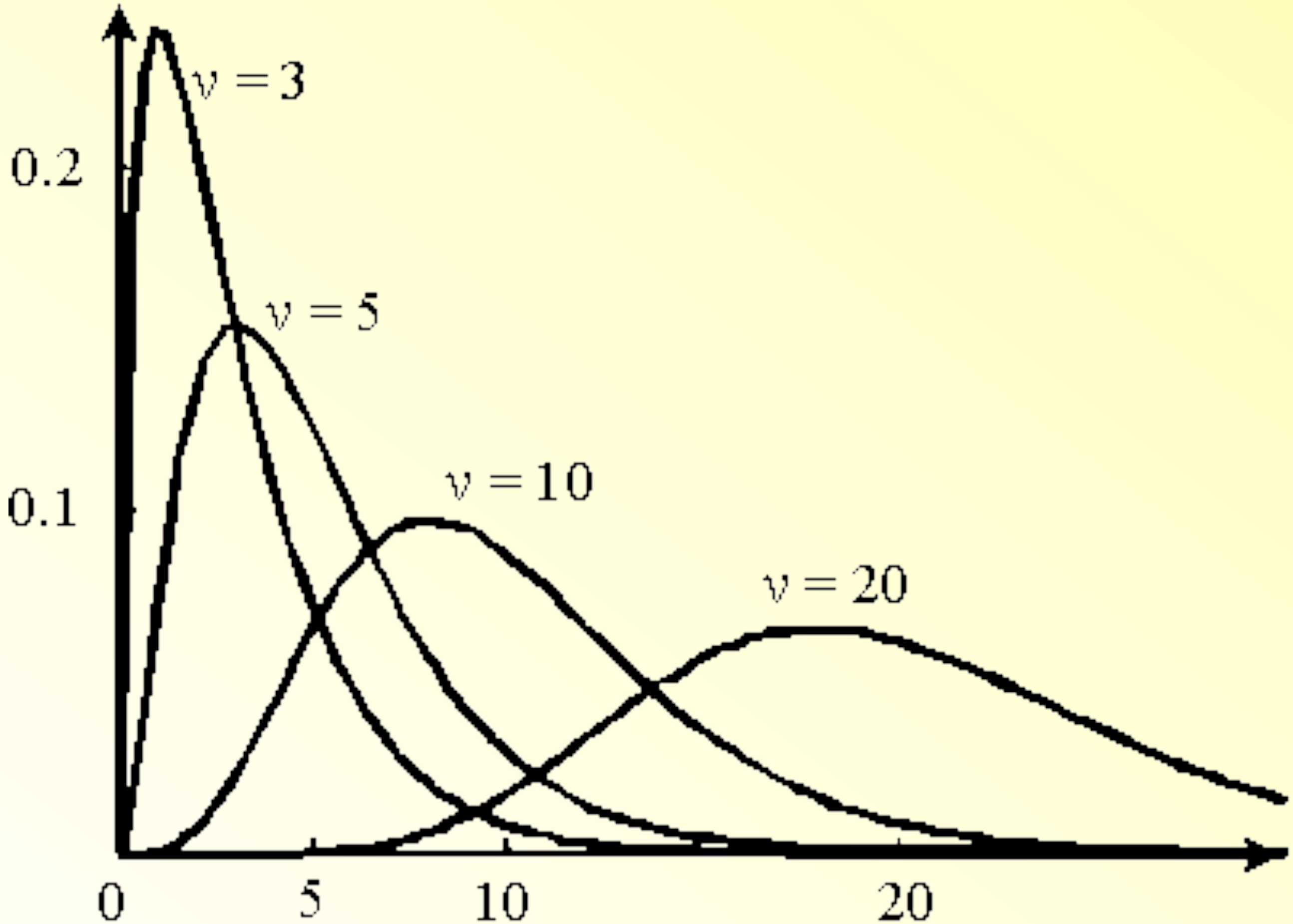
Expected

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 http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

Chi square distribution



Chi-Squared Test of Independence

- Test Statistic

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

$$\text{where } f_e = \frac{(\text{Row total}) \cdot (\text{Column total})}{\text{Total sample size}}$$

- In our example:

Chi-Squared Test of Independence

- P-Value
  - P = right-hand tail probability above the observed chi-squared value for chi-squared distribution with *df=(r-1)(c-1)*
  - For the chi-squared test, always use the right-hand tail probability!
- Report P-value, reject null hypothesis at alpha-level if P is less than alpha
- In our example, P-value =

Degrees of Freedom

- *(r-1)x(c-1)*
- Given the row marginals and the column marginals, this is the number of frequencies that we need to determine all the other cell frequencies
- In our example, *(r-1)x(c-1)=(2-1)x(4-1)=3*
- This is the number of cell frequencies that are free to vary, because once we know row and column totals, they determine the remaining ones

(the remaining ones are not free to vary)

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

Chi-Squared Test, Properties

- The chi-squared test treats the classifications as nominal
- Any reordering of rows or columns of the table leaves the value of the chi-squared test unchanged
- If either of the classifications is in fact ordinal, this information is not used
- If the response variable is in fact ordinal, one should use the Kruskal-Wallis test instead

Chi-Squared Test, Misuse

- The chi-squared test should not be used when any of the expected frequencies is less than five
- For smaller sample sizes, there is a procedure that can be used
  - generalized version of Fisher's exact test
  - SAS: PROC FREQ, option EXACT
- Also, the test statistic must be calculated using the observed/expected frequencies, and not using percentages!
- This test can not be used when the samples are dependent.
- For example, when each row or each column has observations on the same subjects, the samples are dependent (McNemar's test can be used then)

Special Case: Chi-Squared Test, 2x2 Table

- For the 2x2 table with large enough sample sizes, we can use
  - Either the test for a difference of proportions (using normal scores)
  - Or the chi-squared test for association
  - Fortunately, the two tests are equivalent

Chi-Squared Test, 2x2 Table: Example

- 340 commercial motor vehicle drivers who had accidents in Kentucky from 1998 to 2002
- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)

		Accident Fatal		
		Yes	No	
Seat Belt	Yes	30	212	242
	No	33	52	85
		63	264	327

Chi-Squared Test, 2x2 Table: Example

- Testing whether the two variables “Seat Belt” and “Fatal” are associated or independent is equivalent to
- testing whether the fatality rate is the same for the two groups “Seat Belt: Yes” and “Seat Belt: No”
- The row variable is explanatory, the column variable is response
- Calculate the p-value for both tests

Chi-Squared Test, 2x2 Table

- For the 2x2-table, the chi-squared statistic is exactly the square of the z statistic
- Also, squaring z-scores for certain tail probabilities yields chi-squared scores with *df=1* for the same tail probabilities
- Squared normal = chi-squared with *df=1*

- *In short: For the special case of a 2x2 table, the **chi-squared test for independence is equivalent to the test for equal proportions** of two independent samples*

Summary: Investigating Association Between Two Variables

		Response Variable		
		Unordered Categorical (Nominal)	Ordinal	Quantitative
Explanatory Variable	Nominal with 2 Levels	Comparing Proportions of 2 samples	Nonparametric Wilcoxon-Mann-Whitney Test	Comparing Means of 2 samples, t-test for independent samples
	Unordered Categorical (Nominal) More than 2 Levels	Analyzing Association, Chi Squared Tests	Nonparametric Kruskal-Wallis Test	ANOVA
	Ordinal		*Spearman Rank Correlation*	
	Quantitative	*Logistic Regression*		*Regression*