Course Outline through Week 11

Between now and the 3rd exam we will focus on measuring the association between two variables, X & Y

1. When X is discrete and Y is continuous, we will use "analysis of variance" techniques (Today, Before Break)
2. When X and Y are both discrete, we will use cross-tabular and χ^2 analyses (Today, After Break)
3. When X and Y are both continuous, we will use correlation & regression analyses (Next Week)

Introduction to ANOVA

ANalysis **O**f **VA**riance (**ANOVA**) techniques compare the mean of continuous variable Y across *J* populations

Do people's political views (Y) vary by the type of community they live in (X … e.g., rural vs. urban vs. suburban)?

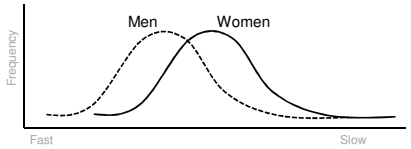
Do people from different religious traditions (X) vary with respect to how much money they donate to political parties (Y)?

Such questions amount to comparing the mean of some continuous variable Y across the *J* categories of some discrete variable X

Introduction to ANOVA

We saw something like this before…

How does the mean of continuous variable "race times" (Y) vary by levels of the categorical variable sex (X) ?

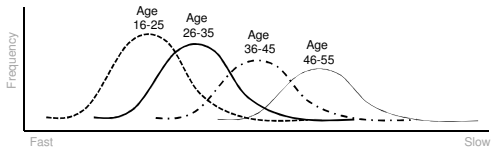


Using what we already know, we could test hypotheses or construct CI's for the difference in means of race times between men and women

Introduction to ANOVA

Now we want to allow for more categories of X

For example: How does the mean of continuous variable "race times" (Y) vary by levels of the categorical variable age (X) ?



Introduction to ANOVA

ANOVA amounts to a test of the hypothesis that all of the *J* population means are equal:

H_0: $\mu_1=\mu_2=\mu_3=...=\mu_J$

H_1: Not all of the means are equal

If we reject H_0, we do not know exactly which of the *J* means differs from the others, whether there are *J* different means, or what … we just know that not all of the *J* means are equal (Follow-up analyses may be necessary)

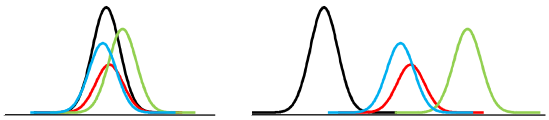
Introduction to ANOVA

The test statistic that we use to perform the ANOVA hypothesis test is called the *F* statistic

Conceptually, the more the *J* sample means differ, the larger the *F* statistic

If *F* is larger than we would expect by chance … if the observed value of the *F* test statistic exceeds a critical value determined in advance … then we reject H_0

Introduction to ANOVA



Small *F* Value Larger *F* Value

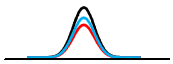
The Logic of F Tests

The *F* statistic (defined later) is essentially a ratio:

$$F = \frac{\textit{Variation Between the j Sample Means}}{\textit{Variation Within Each of the j Groups}}$$

Imagine that we have *J*=3 groups & we compute 95% confidence intervals for the mean of Y within groups:

Group #1 95% C.I. = 10 ± 0.5
Group #2 95% C.I. = 10 ± 0.5
Group #3 95% C.I. = 10 ± 0.5



Here, the *J* sample means do not vary between groups, but there is variability within groups. *F* would be tiny.

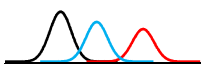
The Logic of F Tests

The *F* statistic (defined later) is essentially a ratio:

$$F = \frac{\textit{Variation Between the j Sample Means}}{\textit{Variation Within Each of the j Groups}}$$

Now imagine that we have *J*=3 groups with these 95% confidence intervals:

Group #1 95% C.I. = 5 ± 0.1
Group #2 95% C.I. = 10 ± 0.1
Group #3 95% C.I. = 15 ± 0.1



Here, the *J* sample means vary a lot between groups (and somewhat within groups). *F* would be large.

The Logic of F Tests

Y varies from observation to observation

The *F* test amounts to asking how much of the variability in Y occurs between the *J* groups as opposed to within each of the *J* groups

Does which group you belong to—as indexed by the categorical variable X— "explain" or "account for" variability in Y?

The Logic of F Tests

Another way to think about it: If μ is the overall mean of Y and μ_j is the mean for category J, then the "effect" of being in category J of the categorical variable X is a_j:

$$a_j = \mu_j - \mu$$

Rearranging this, the mean for category j equals

$$\mu_j = \mu + a_j$$

If the J group means are equal, then $\mu_j = \mu$ and $a_j = 0$ … implying that an individual's value of Y does not depend on the category of X to which they belong

The Logic of F Tests

For individual i we can express the value of Y as:

$$Y_{ij} = \mu + a_j + e_{ij}$$

where Y_{ij} is the value of Y for individual i in group j, μ is the overall mean of Y, a_j is the effect of being in category j of variable X, and e_{ij} is a random error term

ANOVA is an effort to determine how much of the variance in Y_{ij} is attributable to group membership (the a_j) and how much is due to other things (e_{ij})

How do we know how much of the variation in Y_{ij} is due to these two source of variability?

The Logic of F Tests

The core idea of ANOVA is that we can separate the variable in Y into two components

1. Variability between groups
2. Variability within groups

The *F*-statistic is based on a comparison of between-group variability to within-group variability

How do we measure between-group variability?

How do we measure within-group variability?

The Logic of F Tests

Return to the formula for sample variance:

$$s_Y^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1}$$

If we then subscript Y with "i" to represent each individual and "j" to represent the category of X to which they belong, then we can re-write the numerator of the formula for sample variance as

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2$$

The Logic of F Tests

The quantity

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2$$

is called the "total sum of squares," or SS_{TOTAL}

This quantity can be decomposed into two parts:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2$$

The Logic of F Tests

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_j)^2 + \sum_{j=1}^{J} n_j(\bar{Y}_j-\bar{Y})^2$$

SS_{TOTAL} = SS_{WITHIN} + $SS_{BETWEEN}$

SS_{TOTAL} = Sum of squared deviations of each individual's value of Y from the overall sample mean of Y

SS_{WITHIN} = Sum of the squared deviations of each individual's value of Y from their group's mean of Y

$SS_{BETWEEN}$ = Sum of the squared deviations of each group's mean value of Y from the overall sample mean of Y

The F Statistic

The F statistic is defined as

$$F_{J-1,N-J} = \frac{SS_{BETWEEN}/J-1}{SS_{WITHIN}/N-J} = \frac{MS_{BETWEEN}}{MS_{WITHIN}}$$

where

$$MS_{BETWEEN} = \frac{SS_{BETWEEN}}{J-1} \qquad MS_{WITHIN} = \frac{SS_{WITHIN}}{N-J}$$

$MS_{BETWEEN}$ is an estimate of the amount of variance in Y attributable to the category of X to which cases belong

MS_{WITHIN} is an estimate of the amount of variance in Y attributable to everything else

The F Statistic

The F statistic is defined as

$$F_{J-1,N-J} = \frac{SS_{BETWEEN}/J-1}{SS_{WITHIN}/N-J} = \frac{MS_{BETWEEN}}{MS_{WITHIN}}$$

where

$$MS_{BETWEEN} = \frac{SS_{BETWEEN}}{J-1} \qquad MS_{WITHIN} = \frac{SS_{WITHIN}}{N-J}$$

If the mean of Y does not vary across groups, then the ratio F will be small

Hypothesis test: Is the value of F larger than we would expect by chance if $\mu_1=\mu_2=...=\mu_j$?

The F Statistic

The F statistic is defined as

$$F_{J-1,N-J} = \frac{SS_{BETWEEN}/J-1}{SS_{WITHIN}/N-J} = \frac{MS_{BETWEEN}}{MS_{WITHIN}}$$

where

$$MS_{BETWEEN} = \frac{SS_{BETWEEN}}{J-1} \qquad MS_{WITHIN} = \frac{SS_{WITHIN}}{N-J}$$

$MS_{BETWEEN}$ has j — 1 degrees of freedom

MS_{WITHIN} has N — j degrees of freedom

F thus has two degrees of freedom (df): The "numerator" df_{NUM} (ν_1=j—1) and the "denominator" df_{DENOM} (ν_2=N—j)

ANOVA Example

Say that Y is people's political views (where 1=Extremely Liberal and 7=Extremely Conservative) and X represents the type of community that people live in; here, J=5

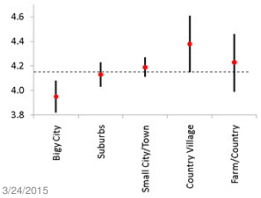
N for the full sample equals 2,779 with an overall sample mean of Y equal to 4.15

Big City	Mean=3.95	n=522	(95% C.I. = 3.82 to 4.08)
Suburbs	Mean=4.13	n=724	(95% C.I. = 4.03 to 4.23)
Small City/Town	Mean=4.19	n=1,093	(95% C.I. = 4.11 to 4.27)
Country Village	Mean=4.38	n=122	(95% C.I. = 4.15 to 4.61)
Farm/Country	Mean=4.30	n=318	(95% C.I. = 3.99 to 4.46)

ANOVA Example

Say that Y is people's political views (where 1=Extremely Liberal and 7=Extremely Conservative) and X represents the type of community that people live in; here, J=5

N for the full sample equals 2,779 with an overall sample mean of Y equal to 4.15



ANOVA Example

Hypothesis Testing in 6 Steps … Just Like Before

1. State the null (H_0) and alternative (H_1) hypotheses
2. Check that the sample data conform to basic assumptions; if they do not, then do not go any further
3. Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis
4. Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level
5. Calculate the test statistic … F
6. Compare the test statistic to the critical value

ANOVA Example

State the null (H_0) and alternative (H_1) hypotheses

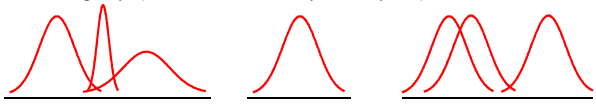
H_0: $\mu_{Big\ City} = \mu_{Suburbs} = \mu_{Small\ City} = \mu_{Village} = \mu_{Farm}$

H_1: Not all of the population group means are equal

ANOVA Example

Check that the sample data conform to basic assumptions; if they do not, then do not go any further

1. The *j* samples are independent random samples
2. Within each population group, Y is normally distributed
3. The standard deviation of Y is equal across the *j* population groups (the "homoskedasticity" assumption)



Assumptions Are Not Met | **H_0: $\mu_1=\mu_2=\mu_3$** | **H_1: Not all Means Are Equal**

ANOVA Example

Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis

Let's choose α=0.05

Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level

As shown in an F table, the critical value of F depends on α, df_{NUM} (or ν_1), and df_{DENOM} (or ν_2)

Worksheet

With α=0.05, what is the critical value of *F* when…

df_{NUM}=3, and df_{DENOM}=120 ?
df_{NUM}=10, and df_{DENOM}=20 ?
df_{NUM}=20, and df_{DENOM}=30 ?

With α=0.01, what is the critical value of *F* when…

df_{NUM}=1, and df_{DENOM}=1,000 ?
df_{NUM}=4, and df_{DENOM}=20 ?

ANOVA Example

In our example, N=2,779 and *j*=5
So, df_{NUM}= *j*—1 = 4 and df_{DENOM}= N — *j* = 2,774
With α=0.05, the critical value of *F* is 2.37

We can thus re-write the hypotheses:

H_0: $\mu_{Big\ City} = \mu_{Suburbs} = \mu_{Small\ City} = \mu_{Village} = \mu_{Farm}$
H_1: Not all of the population group means are equal

H_0: $F \leq 2.37$
H_1: $F > 2.37$

ANOVA Example

Calculate the test statistic ... *F*

This is not something we do by hand for large samples

STATA Output for our example:

ANOVA

Think of self as liberal or conservative

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	35.229	4	8.807	4.807	.001
Within Groups	5082.393	2774	1.832		
Total	5117.622	2778			

ANOVA Example

Compare the test statistic to the critical value

1. If the test statistic is larger than the critical value, then reject H_0 (with probability of α of doing so even though H_0 should not be rejected)
2. If the test statistic is less than or equal to the critical value, then do no reject H_0 (with probability of β of doing so even though H_0 should be rejected)

We determined that our critical value of *F* is 2.37

We observed an *F* statistic of 4.807

H_0: $F \leq 2.37$ H_1: $F > 2.38$

Conclusion: Reject H_0

ANOVA vs Difference in Means When J=2

We have now seen two techniques for comparing the mean of the continuous variable Y across two population groups

First, we can conduct a test of the hypothesis that $\mu_{j=1}=\mu_{j=2}$ using a *t* test

Second, we can conduct an ANOVA with *J*=2

We will get the same result. This is because

$$t_{df_{DENOM}} = \sqrt{F_{1,df_{DENOM}}}$$

Worksheet

There are three delivery companies: A, B, & C

I had all three mail me 5 packages. Below are the number of days it took for me to get the packages

A	2	2	3	4	6
B	1	2	2	5	5
C	2	2	3	3	4

Hint/Help:

$SS_{Between} = 0.9333$

$SS_{Within} = 28$

Test the hypothesis that the mean number of days that each company takes to deliver packages is equal; use α=0.05 *(For this example, relax/ignore the basic assumptions that must be met in order to perform ANOVA)*

Worksheet

There are three delivery companies: A, B, & C

I had all three mail me 3 packages. Below are the number of days it took for me to get the packages

A	1	1	2
B	2	2	3
C	3	3	4

Hint/Help:

$SS_{Between} = 6$

$SS_{Within} = 2$

Test the hypothesis that the mean number of days that each company takes to deliver packages is equal; use α=0.05 *(For this example, relax/ignore the basic assumptions that must be met in order to perform ANOVA)*

Want More?

David Lane's Book

http://onlinestatbook.com/2/analysis_of_variance/ANOVA.html

Chapters 13 and 14 of Richard Lowry's Book

http://vassarstats.net/textbook/

Gerard Dallal's Book

http://www.jerrydallal.com/LHSP/anova1.htm

http://www.jerrydallal.com/LHSP/aov1out.htm

BREAK

Course Outline through Week 11

Between now and the 3rd exam we will focus on measuring the association between two variables, X & Y

1. When X is discrete and Y is continuous, we will use "analysis of variance" techniques (Today, Before Break)
2. When X and Y are both discrete, we will use cross-tabular and χ^2 analyses (Today, After Break)
3. When X and Y are both continuous, we will use correlation & regression analyses (Next Week)

Crosstabulations of Discrete Variables

How Important Are "Loose Morals and Drunkenness" in Explaining Poverty

Source: GSS

		Very Important	*Somewhat Important*	*Not Important*	Row Total
Political Party	*Democrat*	164	161	131	456
	Independent	179	139	100	418
	Republican	173	154	103	430
	Column Total	516	454	334	N=1,304

The Logic of χ^2 Tests

Even if there is no association between two discrete variables X and Y in the population, we may observe an association between X and Y in sample data because of random error or sampling variability

How can we tell whether the association observed between X and Y in sample data is strong enough to rule out the hypothesis that in the population X and Y are statistically independent (or not associated with one another)?

The Logic of χ^2 Tests

We begin with the null hypothesis that there is no association between X and Y in the population … that is, we assume "statistical independence"

We then compute the cell frequencies that we would expect to observe under the null hypothesis and compare them to the actually observed cell frequencies

The χ^2 test statistic quantifies the degree to which the observed frequencies differ from the frequencies that we would expect to observe under the null hypothesis

The Logic of χ^2 Tests

Imagine (1) that 10% of people are left-handed, (2) that 50% of people are male and 50% are female, and (3) that there is no relationship between gender and whether someone is left-handed.

What would you expect in the cells of the table below if you sampled 1,000 people?

	Right	Left	*Total*
Male			
Female			
Total			1,000

χ^2 Tests

Hypothesis Testing in 6 Steps … Just Like Before

1. State the null (H_0) and alternative (H_1) hypotheses
2. Check that the sample data conform to basic assumptions; if they do not, then do not go any further
3. Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis
4. Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level
5. Calculate the test statistic … χ^2
6. Compare the test statistic to the critical value

χ^2 Tests

State the null (H_0) and alternative (H_1) hypotheses

We begin with the assumption that there is no association between X and Y

H_0: X and Y are statistically independent

H_1: X and Y are not statistically independent

χ^2 Tests

Check that the sample data conform to basic assumptions; if they do not, then do not go any further

X and Y must be collected from a random sample of individuals from the population

Standard χ^2 testing procedures should be used with extreme caution — or not at all — if any cell frequency is less than 5

χ^2 Tests

Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis

Let's use α=0.05 in our example

χ^2 Tests

Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level

Critical values of χ^2 corresponding to common levels of α and to specific numbers of degrees of freedom (*df*) are given in a χ^2 table

- For χ^2, *df*=(*R*—1)(*C*—1) where *R*=number of rows in the crosstable and *C*=number of columns in the crosstable

In our example, α=0.05 and df=(3—1)(3—1)=4, so the critical value of χ^2 equals 9.49

χ^2 Tests

What is the critical value of χ^2 when α=0.05 and…

- …the table has 2 rows and 3 columns?
- …the table has 5 rows and 3 columns?

What is the critical value of χ^2 when α=0.01 and…

- …the table has 3 rows and 4 columns?
- …the table has 4 rows and 4 columns?

χ^2 Tests

Calculate the test statistic ... χ^2

The χ^2 test statistic is based on a comparison of the observed cell frequencies to the cell frequencies that we expect under the null hypothesis

For the cell in row i and column j, the "expected" frequency under the null hypothesis is

$$\hat{f}_{ij} = \frac{(f_{i\bullet})(f_{\bullet j})}{N}$$

where $f_{i\bullet}$ is the number of cases in row i, $f_{\bullet j}$ is the number of cases in column j, and N is the sample size

χ^2 Tests

Calculate the test statistic ... χ^2

For the cell in row *i*=1 and column *j*=1 (the top left cell):

$$\hat{f}_{1,1} = \frac{(f_{1\bullet})(f_{\bullet 1})}{N} = \frac{(456)(516)}{1{,}304} = 180$$

For the cell in row *i*=1

and column *j*=2:

$$\hat{f}_{1,2} = \frac{(456)(454)}{1{,}304} = 159$$

And so forth…

How Important Are "Loose Morals and Drunkenness" in Explaining Poverty

Political Party	Very Important	Somewhat Important	Not Important	Row Total
Democrat	164 180	161 159	131 117	456
Independent	179 165	139 146	100 107	418
Republican	173 170	154 150	103 110	430
Column Total	516	454	334	N=1,304

χ^2 Tests

Calculate the test statistic ... χ^2

The χ^2 test statistic equals

$$\chi^2 = \text{Sum of } \frac{(\text{Expected} - \text{Observed})^2}{\text{Expected}} = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(\hat{f}_{ij} - f_{ij})^2}{\hat{f}_{ij}}$$

where

R is the number of rows in the crosstable,

C is the number of columns in the crosstable,

f_{ij} is the observed frequency in the cell in row i and column j, &

$\hat{f}_{ij}$ is the expected frequency in the cell in row i and column j

χ^2 Tests

Calculate the test statistic … χ^2

$$\chi^2 = \frac{(180-164)^2}{180} + \frac{(159-161)^2}{159} + \ldots + \frac{(110-103)^2}{110} = 5.767$$

In this example, χ^2=5.767

How Important Are "Loose Morals and Drunkenness" in Explaining Poverty

		Very Important	*Somewhat Important*	*Not Important*	Row Total
Political Party	*Democrat*	164 180	161 159	131 117	456
	Independent	179 165	139 146	100 107	418
	Republican	173 170	154 150	103 110	430
	Column Total	516	454	334	N=1,304

χ^2 Tests

Compare the test statistic to the critical value

1. If the test statistic is as large or larger than the critical value, then reject H_0 (with probability of α of doing so even though H_0 should not actually be rejected)
2. If the test statistic is less than the critical value, then do no reject H_0 (with probability of β of doing so even though H_0 should be rejected)

We can restate the hypotheses as

H_0: X & Y are independent ➔ Fail to Reject if $\chi^2 \leq 9.49$
H_1: X & Y not independent ➔ Reject if $\chi^2 > 9.49$

Since χ^2=5.767, we fail to reject H_0

χ^2 Example

Is there an association between people's childhood family income and their income as adults?

Source: GSS

Family Income in Childhood

		Below Average	*Average*	*Above Average*	Row Total
Family Income as an Adult	*Below Average*	4,085	4,027	1,054	9,166
	Average	4,796	9,853	2,121	16,770
	Above Average	1,494	2,886	2,075	6,455
	Column Total	10,375	16,766	5,250	N=32,391

χ^2 Example

Is there an association between people's childhood family income and their income as adults?

Source: GSS

		Family Income in Childhood			
		Below Average	*Average*	*Above Average*	Row Total
Family Income as an Adult	*Below Average*	39%	24%	20%	28%
	Average	46%	59%	40%	52%
	Above Average	14%	17%	40%	20%
	Column Total	100%	100%	100%	100%

χ^2 Example

State the null (H_0) and alternative (H_1) hypotheses

Again, we begin with the assumption that there is no association between X and Y

H_0: Family income in childhood and family income in adulthood are statistically independent

H_1: Family income in childhood and family income in adulthood are not statistically independent

χ^2 Example

Check that the sample data conform to basic assumptions; if they do not, then do not go any further

X and Y must be collected from a random sample of individuals from the population (OK)

Standard χ^2 testing procedures should be used with extreme caution — or not at all — if any cell frequency is less than 5 (OK)

χ^2 Example

Choose an α probability level … that is, a probability associated with incorrectly rejecting the null hypothesis

Let's use α=0.01 in our example

χ^2 Example

Determine the "critical value" … that is, how large the test statistic must be in order to reject the null hypothesis at the given α level

In our example, α=0.01 and df=(3—1)(3—1)=4

According to a χ^2 table, the critical value of χ^2 thus equals 13.28

χ^2 Example

Calculate the test statistic … χ^2

$$\hat{f}_{i,j} = \frac{(f_{i\bullet})(f_{\bullet j})}{N}$$

		Family Income in Childhood			
		Below Average	*Average*	*Above Average*	Row Total
Family Income as an Adult	*Below Average*	4,085 2,935.92	4,027 4,744.44	1,054 1,485.64	9,166
	Average	4,796 5,371.52	9,853 8,680.37	2,121 2,718.12	16,770
	Above Average	1,494 2,067.57	2,886 3,341.19	2,075 1,046.24	6,455
	Column Total	10,375	16,766	5,250	N=32,391

χ^2 Example

Calculate the test statistic ... χ^2

$$\chi^2 = \text{Sum of } \frac{(\text{Expected} - \text{Observed})^2}{\text{Expected}} = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{\left(\hat{f}_{ij} - f_{ij}\right)^2}{\hat{f}_{ij}}$$

In this example, χ^2=2,267.59

χ^2 Example

Compare the test statistic to the critical value

1. If the test statistic is as large or larger than the critical value, then reject H_0 (with probability of α of doing so even though H_0 should not actually be rejected)
2. If the test statistic is less than the critical value, then do no reject H_0 (with probability of β of doing so even though H_0 should be rejected)

We can restate the hypotheses as

H_0: X & Y are independent ➔ Fail to Reject if $\chi^2 \leq 13.28$

H_1: X & Y not independent ➔ Reject if $\chi^2 > 13.28$

Since χ^2=2,267.59, we reject H_0

Worksheet

Is there an association between the kind of truck that people drive and their favorite kind of hat?

		Kind of Pickup Truck Driven			
		Ford	*Chevy*	*Dodge*	Row Total
Favorite Type of Hat	*Cowboy Hat*	10	5	5	20
	Baseball Cap	5	10	5	20
	Other	5	5	10	20
	Column Total	20	20	20	N=60

Worksheet

Are political views related to whether people view the bible as the literal word of God?

		Political Views			
		Liberal	*Moderate*	*C'servative*	Row Total
Bible is…	*Word of God*	1,500	2,888	3,234	7,622
	Something Else	4,670	5,961	5,079	15,710
	Column Total	6,170	8,849	8,313	23,332

χ^2 vs Tests for Two Proportions

Does the proportion living past 60 vary by family SES?
Different methods yield the same results…

Using what we learned earlier about hypothesis tests for differences in proportions: Z=-2.453; p-value=0.014
Using a χ^2 test, we find: χ^2=6.027; p-value=0.014

		Lived Past Age 60?	
		Yes=1	*No=0*
Family SES	*Low=0*	603	330
	High=1	456	319

Factors Driving Significance

Two things affect our chances of observing a statistically significant association

1. The strength of the association in the population
2. The size of the sample … consider the following two cross-tabs/results:

		Voted in 1996?			Voted in 1996?		
		Yes	No	Row Total	Yes	No	Row Total
Political Party Affiliation	Democrat	**6**	**4**	**10**	**300**	**200**	**500**
	Republican	**8**	**2**	**10**	**400**	**100**	**500**
	Column Total	**14**	**6**	**20**	**700**	**300**	**1,000**
		χ^2 = 0.95			χ^2 = 47.62		

Want More?

David Lane's Book
- http://onlinestatbook.com/2/chi_square/Chi_Square.html

Chapter 8 of Richard Lowry's Book
- http://vassarstats.net/textbook/

Gerard Dallal's Book
- http://www.jerrydallal.com/LHSP/ctab.htm

Stat Trek
- http://stattrek.com/chi-square-test/homogeneity.aspx?tutorial=ap

Association Between Discrete Variables

We use χ^2 to assess whether there is any statistically significant association between two categorical variables, X and Y, in the population

We have said nothing about measuring the direction or strength of that association

If we conclude that categorical variables X and Y are associated, how can we quantify the direction and strength of that association?

Association Between Discrete Variables

Measures of Association
- Statistics that show the direction and/or magnitude of a relationship between pairs of variables

When X and Y are Both Ordinal:
- Gamma

Others (Not discussed today)

When X and Y are Both Dichotomous:
- Gamma
- Relative Risk (RR)
- Odds Ratio (OR)

Others (Not discussed today)

Association Between Ordinal Variables

		Family Income in Childhood			
		Below Average	*Average*	*Above Average*	Row Total
Family Income as an Adult	*Above Average*	1,494	2,886	2,075	6,455
	Average	4,796	9,853	2,121	16,770
	Below Average	4,085	4,027	1,054	9,166
	Column Total	10,375	16,766	5,250	N=32,391

Source: GSS 1972-2008

Gamma

χ^2 = 2,267.59; rejected H_0 … that is, we rejected the null hypothesis that there is no association in the population

But how strong is the association? And in which direction?

Gamma is a measure of association for use when both variables are ordinal and have 2 or more categories

		Family Income in Childhood			
		Below Average	*Average*	*Above Average*	Row Total
Family Income as an Adult	*Above Average*	1,494	2,886	2,075	6,455
	Average	4,796	9,853	2,121	16,770
	Below Average	4,085	4,027	1,054	9,166
	Column Total	10,375	16,766	5,250	32,391

Source: GSS, 1972-2008

Gamma

Here is this analysis done separately for 1972 and 2008. Are the two variables associated in both years? Has the strength of the association changed over time?

1972

χ^2=130.79 (p<0.01)

		Family Income in Childhood			
		Below Average	*Average*	*Above Average*	**Row Total**
Family Income as an Adult	*Above Average*	40	72	71	183
	Average	173	591	135	899
	Below Average	181	241	65	495
	Column Total	402	904	271	1,577

Source: GSS

2008

χ^2=181.19 (p<0.01)

		Family Income in Childhood			
		Below Average	*Average*	*Above Average*	**Row Total**
Family Income as an Adult	*Above Average*	87	124	159	370
	Average	263	472	165	900
	Below Average	331	274	100	705
	Column Total	681	870	424	1,975

Source: GSS

Gamma

The possible values of gamma range from -1 to +1

Gamma = -1 = a perfect negative association

Gamma = +1 = a perfect positive association

Gamma = 0 = no association

Gamma = +0.986

	x=1	*x=2*	*x=3*
y=3	1	1	100
y=2	1	100	1
y=1	100	1	1

Gamma = -0.986

	x=1	*x=2*	*x=3*
y=3	100	1	1
y=2	1	100	1
y=1	1	1	100

Gamma = 0.000

	x=1	*x=2*	*x=3*
y=3	1	1	1
y=2	1	1	1
y=1	1	1	1

Gamma

Calculating gamma is semi-complicated; see the "bonus slides" at the end for instructions and an example

Focus for now on how it is interpreted…

Is the association positive or negative?

Is it weak or strong?

Gamma

Here is this analysis done separately for 1972 and 2008. Are the two variables associated in both years? Has the strength of the association changed over time?

1972

Gamma= 0.312

Source: GSS

Family Income as an Adult	Family Income in Childhood: *Below Average*	*Average*	*Above Average*	**Row Total**
Above Average	40	72	71	183
Average	173	591	135	899
Below Average	181	241	65	495
Column Total	402	904	271	1,577

2008

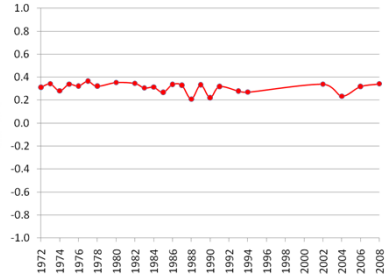
Gamma= 0.342

Source: GSS

Family Income as an Adult	Family Income in Childhood: *Below Average*	*Average*	*Above Average*	**Row Total**
Above Average	87	124	159	370
Average	263	472	165	900
Below Average	331	274	100	705
Column Total	681	870	424	1,975

Gamma

And separately by year…



Association in 2x2 Tables

Cross-tabulations of categorical variables that each have 2 categories—that is, that are dichotomous—are a special case of cross-tabulations of ordinal variables

We will explore three measures of association that pertain to 2x2 tables

- Gamma
- Relative Risk
- Odds Ratio

Association in 2x2 Tables

Each measure relies on a particular naming convention for the four cells in the cross-tabulation

		X	
		X=1	*X=2*
Y	*Y=2*	Cell a	Cell b
	Y=1	Cell c	Cell d

Relative Risk

Gamma expresses the association between two dichotomous variables without regard to whether one variable affects change in the other … they are thus referred to as "symmetric" measures

Relative Risk (and later **Odds Ratios**) are "asymmetric" measures, in that they express the change in the chances of being in a particular category of the "dependent" (Y) variable that result from changing categories of the "independent" (X) variable

Relative Risk

When X=1, the "Risk" that Y=2 equals a/(a+c) … which is the same as P(Y=2|X=1)

When X=2, the "Risk" that Y=2 equals b/(b+d) … which is the same as P(Y=2|X=2)

The **relative risk** (RR) is the ratio of the two risks:

$$RR = \frac{b/(b+d)}{a/(a+c)}$$

		X	
		X=1	*X=2*
Y	*Y=2*	a	b
	Y=1	c	d

Relative Risk

The risk of being in excellent health (Y=2) if you do not smoke (X=1) is a/(a+c) = 1,955/(3,838+1,955)=0.337

The risk of being in excellent health (Y=2) if you do smoke (X=2) is b/(b+d) = 826/(2,356+826)=0.260

The Relative Risk (RR) is the ratio of the two risks:

$$RR = \frac{0.260}{0.337} = 0.77$$

		Current Smoker? *Source: GSS*	
		X=1: No	*X=2: Yes*
Health	*Y=2: Excellent*	(a) 1,955	(b) 826
	Y=1: Not Excellent	(c) 3,838	(d) 2,356

Relative Risk

Relative Risk can range from zero to infinity

RR 0.00 to 1.00 means that the risk that Y=2 is reduced when you move from X=1 to X=2

RR = 1.00 means that the risk that Y=2 is unchanged when you move from X=1 to X=2

RR > 1.00 means that the risk that Y=2 is increased when you move from X=1 to X=2

Relative Risk

Interpret the value of RR with reference to the number 1

RR = 0.77: The risk that Y=2 is reduced by 23% when you move from X=1 to X=2

RR = 1.20: The risk that Y=2 is increased by 20% when you move from X=1 to X=2

RR = 2.50: The risk that Y=2 is increased by 150% (or is 2.5 times greater) when you move from X=1 to X=2

Odds Ratio

When X=1, the "Odds" that Y=2 equals a/c

When X=2, the "Odds" that Y=2 equals b/d

The "Odds Ratio" (OR) is the ratio of the two odds:

$$OR = \frac{b/d}{a/c} = \frac{bc}{ad}$$

		X	
		X=1	*X=2*
Y	*Y=2*	a	b
	Y=1	c	d

Odds Ratio

The odds of being in excellent health (Y=2) if you do not smoke (X=1) is a/c = 1,955/3,838=0.509

The odds of being in excellent health (Y=2) if you do smoke (X=2) is b/d = 826/2,356=0.351

The Odds Ratio (OR) is the ratio of the two odds:

$$OR = \frac{0.351}{0.509} = 0.69$$

Health

	Current Smoker? Source: GSS	
	X=1: No	*X=2: Yes*
Y=2: Excellent	(a) 1,955	(b) 826
Y=1: Not Excellent	(c) 3,838	(d) 2,356

Odds Ratio

Odds Ratios can range from zero to infinity

OR 0.00 to 1.00 means that the odds that Y=2 is reduced when you move from X=1 to X=2

OR = 1.00 means that the odds that Y=2 is unchanged when you move from X=1 to X=2

OR > 1.00 means that the odds that Y=2 is increased when you move from X=1 to X=2

Odds Ratio

Interpret the value of OR with reference to the number 1

OR = 0.77: The odds that Y=2 is reduced by 23% when you move from X=1 to X=2

OR = 1.20: The odds that Y=2 is increased by 20% when you move from X=1 to X=2

OR = 2.50: The odds that Y=2 is increased by 150% (or is 2.5 times greater) when you move from X=1 to X=2

Worksheet

How is race (white vs. black) associated with political party preference? (χ^2=2,214.2 w/ 1 df; Reject H_0 w/ α=0.01). Compute **and interpret** RR and OR

		Race		
		X=1: White	*X=2: Black*	Row Total
Party	*Y=2: Democrat*	15,594	4,401	19,995
	Y=1: Republican	13,288	505	13,793
	Column Total	28,882	4,906	N=33,788

Source: GSS

Worksheet

Is there are a relationship between X and Y in the population from which the sample data were collected?

Use χ^2 (w/ α=0.05) to test the hypothesis of no association.

Use RR, and OR to describe the association

		X		
		X=1	*X=2*	Row Total
Y	*Y=2*	10	5	15
	Y=1	5	10	15
	Col. Total	15	15	30

Practical vs. Statistical Significance

Even when we can reject the hypothesis of no association, we should always investigate the direction and magnitude of the association

		Ever Smokes Pot?		
		Yes	**No**	Row Total
Right or Left Handed	**Right**	100,000	200,000	300,000
	Left	10,000	20,550	30,550
	Column Total	110,000	220,550	330,550

χ^2 = 4.5 w/ df=1
Reject H_0 at α=0.05

Odds Ratio = 1.03

Right handed people are 3% more likely to have ever smoked pot … is this a practically meaningful result?

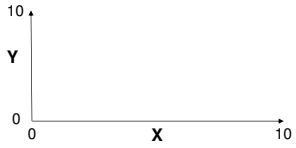
Bonus:
The Next Slides Show
How to Compute Gamma
(If you are interested…)

Bonus: Computing Gamma

In order to compute Gamma, the crosstable must be arranged such that

(1) the column variable is arranged from lowest to highest going from left to right

(2) the row variable is arranged from lowest to highest going from the bottom to the top



Family Income in Childhood

Source: GSS, 1972-2008

Family Income as an Adult	*Below Average*	*Average*	*Above Average*	**Row Total**
Above Average	1,494	2,886	2,075	6,455
Average	4,796	9,853	2,121	16,770
Below Average	4,085	4,027	1,054	9,166
Column Total	10,375	16,766	5,250	32,391

Bonus: Computing Gamma

The formula for Gamma requires evaluating all pairs of observations in a cross-tabulation, counting the total number that are untied concordant pairs (n_c) and the total number that are untied discordant pairs (n_d)

An untied pair is "one in which both cases have different values on two variables"

Family Income in Childhood

Source: GSS, 1972-2008

Family Income as an Adult	*Below Average*	*Average*	*Above Average*	**Row Total**
Above Average	1,494	2,886	2,075	6,455
Average	4,796	9,853	2,121	16,770
Below Average	4,085	4,027	1,054	9,166
Column Total	10,375	16,766	5,250	32,391

Bonus: Computing Gamma

Concordant Pairs

"One observation has a higher rank on both variables than does the other member of the pair"

Example: On both variables, the 2,075 cases in the top right cell have higher rank than the 4,796+9,853+4,085+4,027=22,761 cases in the four cells in the bottom left

Thus the observations in the top right cell contribute (2,075)(22,761)=47,229,075 concordant pairs

Family Income in Childhood — Source: GSS, 1972-2008

Family Income as an Adult	Below Average	Average	Above Average	Row Total
Above Average	1,494	2,886	2,075	6,455
Average	4,796	9,853	2,121	16,770
Below Average	4,085	4,027	1,054	9,166
Column Total	10,375	16,766	5,250	32,391

Bonus: Computing Gamma

Cells contribute concordant pairs if there are other cells with lower rank on both variables

For the top-right: (2,075)(22,761)=47,229,075

For the top-middle: (2,886)(4,796+4,085)=25,630,566

For the middle-right: (2,121)(4,085+4,027)=17,205,552

For the middle-middle: (9,853)(4,085)=40,249,505

Summing, there are n_c= 47,229,075+ 25,630,566 + 17,205,552 + 40,249,505 = 130,314,698 concordant pairs

Family Income in Childhood — Source: GSS, 1972-2008

Family Income as an Adult	Below Average	Average	Above Average	Row Total
Above Average	1,494	2,886	2,075	6,455
Average	4,796	9,853	2,121	16,770
Below Average	4,085	4,027	1,054	9,166
Column Total	10,375	16,766	5,250	32,391

Bonus: Computing Gamma

Discordant Pairs

"One member of a pair of observations ranks higher than the other member on one variable but ranks lower on the other variable"

Example: The 1,494 cases in the top left cell have higher rank than the 9,853+2,121+4,027+1,054=17,055 cases in the four cells in the bottom right on one variable,

but lower rank on the other

Thus the observations in the top left cell contribute (1,494)(17,055)=25,480,170 discordant pairs

Family Income in Childhood — Source: GSS, 1972-2008

Family Income as an Adult	Below Average	Average	Above Average	Row Total
Above Average	1,494	2,886	2,075	6,455
Average	4,796	9,853	2,121	16,770
Below Average	4,085	4,027	1,054	9,166
Column Total	10,375	16,766	5,250	32,391

Bonus: Computing Gamma

Cells contribute discordant pairs if there are other cells with lower rank on one variable, higher rank on another

- For the top-left: (1,494)(17,055)=25,480,170
- For the top-middle: (2,886)(2,121+1,054)=9,163,050
- For the middle-left: (4,796)(4,027+1,054)=24,368,476
- For the middle-middle: (9,853)(1,054)=10,385,062

Summing, there are n_d=

25,480,170+ 9,163,050 + 24,368,476 + 10,385,062 = 69,396,758 discordant pairs

Family Income in Childhood (Source: GSS, 1972-2008)

Family Income as an Adult	Below Average	Average	Above Average	Row Total
Above Average	1,494	2,886	2,075	6,455
Average	4,796	9,853	2,121	16,770
Below Average	4,085	4,027	1,054	9,166
Column Total	10,375	16,766	5,250	32,391

Bonus: Computing Gamma

Gamma can be expressed as:

$$G = \frac{n_c - n_d}{n_c + n_d}$$

So in our example:

$$G = \frac{130{,}314{,}698 - 69{,}396{,}758}{130{,}314{,}698 + 69{,}396{,}758}$$

$$= 0.305$$

Family Income in Childhood (Source: GSS, 1972-2008)

Family Income as an Adult	Below Average	Average	Above Average	Row Total
Above Average	1,494	2,886	2,075	6,455
Average	4,796	9,853	2,121	16,770
Below Average	4,085	4,027	1,054	9,166
Column Total	10,375	16,766	5,250	32,391

Bonus: Computing Gamma Example

Is there an association between X and Y?

χ^2=11.011 w/ 4 df (reject H_0 w/ α=0.05)

Gamma = ____

		X: *Low*	X: *Medium*	X: *High*	Row Total
Y	*High*	12	3	2	17
	Medium	7	10	4	21
	Low	6	8	9	23
	Column Total	25	21	15	N=61

Bonus: Computing Gamma Example

Is there an association between X and Y?

χ^2=11.011 w/ 4 df (reject H_0 w/ α=0.05)

n_c = (2)(7+10+6+8)+(3)(7+6)+(4)(6+8)+(10)(6) = 217

n_d = (12)(10+4+8+9)+(3)(4+9)+(7)(8+9)+(10)(9) = 620

$$G = \frac{n_c - n_d}{n_c + n_d} = \frac{217 - 620}{217 + 620}$$

Gamma = -0.481

		X			
		Low	*Medium*	*High*	Row Total
Y	*High*	12	3	2	17
	Medium	7	10	4	21
	Low	6	8	9	23
	Column Total	25	21	15	N=61