Lecture 13: Contingency Tables

Jacob M. Montgomery

Quantitative Political Methodology

Contingency tables

Roadmap

- **Before**: Comparing two independent samples
- **Today** we will learn how to see if two variables are dependent.
  - How to display the data informatively
  - Chi-squared test of independence
  - Standardized residuals

Comparing populations with categorical outcomes

We have two categorical variables and we want to see if there is some relation.

Comparing populations with categorical outcomes

We have two categorical variables and we want to see if there is some relation. If we have three samples, our data might look like this.

Variable 1 (Outcome or response)	Variable 2 (Explanatory or grouping)
1	1
2	0
3	1
5	2
3	2
2	0
4	0
$\vdots$	$\vdots$

Cross-tabs: The basics

Assume we have two variables that are nominal.

- Gender and eye color
- Party-ID and racial/ethnicity

Cross-tabs: The basics

Assume we have two variables that are nominal.

- Gender and eye color
- Party-ID and racial/ethnicity
- WARNING: The calculations are different when variables are ordinal . . . especially if BOTH are ordinal.

Cross-tabs: The basics

Assume we have two variables that are nominal.

- Gender and eye color
- Party-ID and racial/ethnicity
- WARNING: The calculations are different when variables are ordinal ... especially if BOTH are ordinal.

We will use a contingency table, which is usually (at least by me) referred to as a cross-tabulation.

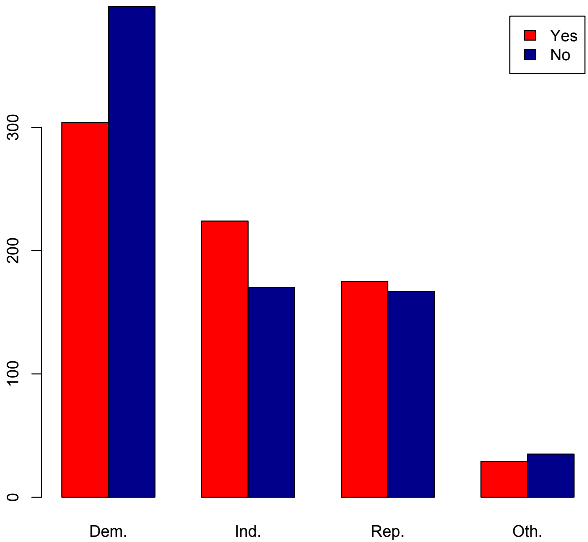
“Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if

"Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if . . . the family has a very low income and cannot afford any more children?"

"Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if . . . the family has a very low income and cannot afford any more children?"

	Yes	No	Total
Democrats	304	398	702
Independents	224	170	394
Republicans	175	167	342
Other	29	35	64
Total	732	770	1502

Plot of data: 1972 GSS



How to display a conditional distribution

What we want to examine the the **conditional distribution** of the outcome variable.

How to display a conditional distribution

What we want to examine the the **conditional distribution** of the outcome variable. What is the distribution of the outcome variable conditioned on the independent/grouping variable?

	Yes	No
Democrats	0.43	0.57
Independents	0.57	0.43
Republicans	0.51	0.49
Other	0.45	0.55

How to display a conditional distribution

What we want to examine the the **conditional distribution** of the outcome variable. What is the distribution of the outcome variable conditioned on the independent/grouping variable?

	Yes	No
Democrats	0.43	0.57
Independents	0.57	0.43
Republicans	0.51	0.49
Other	0.45	0.55

Note that the rows total to 100% because the rows indicate the independent variable.

How to display a conditional distribution

What we want to examine the the **conditional distribution** of the outcome variable. What is the distribution of the outcome variable conditioned on the independent/grouping variable?

	Yes	No
Democrats	0.43	0.57
Independents	0.57	0.43
Republicans	0.51	0.49
Other	0.45	0.55

Note that the rows total to 100% because the rows indicate the independent variable. We might have the columns add up to 100% if that was our explanatory variable.

Chi-square test of independence

***Statistical independence**: Two variables are statistically independent if the **population** conditional distributions of a variable are identical across categories.*

Chi-square test of independence

***Statistical independence**: Two variables are statistically independent if the **population** conditional distributions of a variable are identical across categories.*

	1st child boy	1st child girl	No children
Next is a boy	50	50	50
Next is a girl	50	50	50

Chi-square test of independence

Statistical independence: *Two variables are statistically independent if the* ***population*** *conditional distributions of a variable are identical across categories.*

	1st child boy	1st child girl	No children
Next is a boy	50	50	50
Next is a girl	50	50	50

What if it's a little bit off? This is, after all, a sample.

Chi-square test of independence

***Statistical independence**: Two variables are statistically independent if the **population** conditional distributions of a variable are identical across categories.*

	1st child boy	1st child girl	No children
Next is a boy	50	50	50
Next is a girl	50	50	50

What if it's a little bit off? This is, after all, a sample.

	1st child boy	1st child girl	No children
Next is a boy	49	51	50
Next is a girl	51	49	50

Chi-square test: The intuition

H_0: The variables are statistically independent.

H_a: The variables are statistically dependent.

Chi-square test: The intuition

H_0: The variables are statistically independent.

H_a: The variables are statistically dependent.

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution.

Chi-square test: The intuition

H_0: The variables are statistically independent.

H_a: The variables are statistically dependent.

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution.

$f_{observed} = f_o =$ observed frequency $=$ the raw count

Chi-square test: The intuition

H_0: The variables are statistically independent.

H_a: The variables are statistically dependent.

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution.

$f_{observed} = f_o =$ observed frequency $=$ the raw count (NOT THE %)

Chi-square test: The intuition

H_0: The variables are statistically independent.

H_a: The variables are statistically dependent.

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution.

$f_{observed} = f_o =$ observed frequency = the raw count (NOT THE %)

$f_{expected} = f_e =$ what we would expect for independent samples =

Chi-square test: The intuition

H_0: The variables are statistically independent.

H_a: The variables are statistically dependent.

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution.

$f_{observed} = f_o =$ observed frequency = the raw count (NOT THE %)

$f_{expected} = f_e =$ what we would expect for independent samples =

$$= \frac{\text{Row total}}{\text{Grand total}} \times \text{Column total}$$

Chi-square test: The intuition

H_0: The variables are statistically independent.

H_a: The variables are statistically dependent.

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution.

$f_{observed} = f_o =$ observed frequency $=$ the raw count (NOT THE %)

$f_{expected} = f_e =$ what we would expect for independent samples $=$

$$= \frac{\text{Row total}}{\text{Grand total}} \times \text{Column total}$$

If H_0 is true, then we would expect $f_{observed} = f_{expected}$

Chi-square statistic: The calculations

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Chi-square test: Example

"Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the family has a very low income and cannot afford any more children?"

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
Independents	f_o=224	f_o=170	394
Republicans	f_o=175	f_o=167	342
Other	f_o=29	f_o=35	64
Total	732	770	1502

Chi-square test: Example

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
	$\mathbf{f_e = 342.12}$	$f_e = 359.88$	
Independents	f_o=224	f_o=170	394
	$f_e = 192.12$	$f_e = 201.98$	
Republicans	f_o=175	f_o=167	342
	$f_e = 166.67$	$f_e = 175.33$	
Other	f_o=29	f_o=35	64
	$f_e = 31.19$	$f_e = 32.81$	
Total	732	770	1502

Chi-square test: Example

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
	$\mathbf{f_e = 342.12}$	$f_e = 359.88$	
Independents	f_o=224	f_o=170	394
	$f_e = 192.12$	$f_e = 201.98$	
Republicans	f_o=175	f_o=167	342
	$f_e = 166.67$	$f_e = 175.33$	
Other	f_o=29	f_o=35	64
	$f_e = 31.19$	$f_e = 32.81$	
Total	732	770	1502

Now we want to calculate the χ^2 statistic.

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

Now we want to calculate the χ^2 statistic.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(304 - 342.12)^2}{342.12}$$

Now we want to calculate the χ^2 statistic.

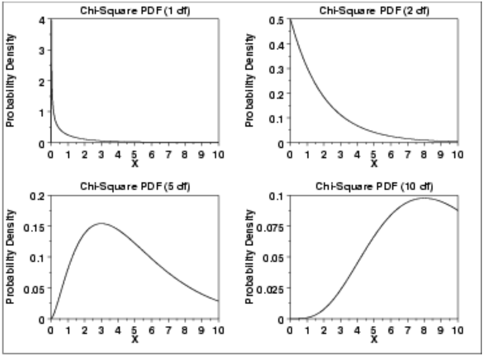
$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(304 - 342.12)^2}{342.12} + \frac{(398 - 359.88)^2}{359.88} + \ldots \approx 19.79$$

Calculating p-values for Chi-squared tests

Is the χ^2 statistic "big enough?"

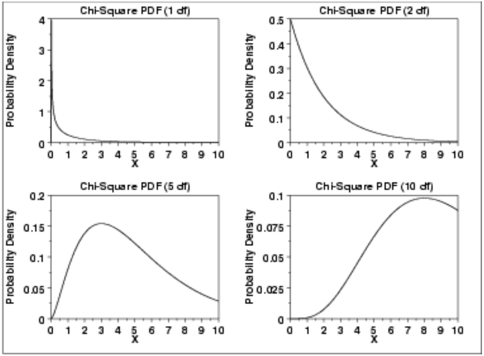
Calculating p-values for Chi-squared tests

Is the χ^2 statistic "big enough?"



Calculating p-values for Chi-squared tests

Is the χ^2 statistic "big enough?"



- We are going to need to calculate the degrees of freedom.
- This is skewed right and strictly positive.
- $\sum Z^2 \sim \chi^2$
- Always use the upper-tail (no $\times$ 2).

Calculating p-values for Chi-squared tests

- Frequency $\geq 5 \ \forall$ cells.

Calculating p-values for Chi-squared tests

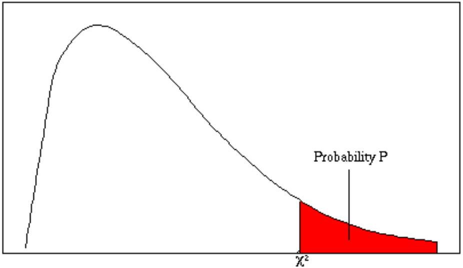
- Frequency $\geq 5 \; \forall$ cells.
- $df = (rows - 1)(columns - 1)$

Calculating p-values for Chi-squared tests

- Frequency $\geq 5 \; \forall$ cells.
- $df = (rows - 1)(columns - 1)$
- `pchisq(`χ^2`, df = (rows-1)(columns-1), lower.tail=FALSE)`

Calculating p-values for Chi-squared tests

- Frequency $\geq 5 \; \forall$ cells.
- $df = (rows - 1)(columns - 1)$
- `pchisq(`χ^2`, df = (rows-1)(columns-1), lower.tail=FALSE)`



Let's first use the table for our example. For what value or greater would this be significant with $df = 3$ and $\alpha = .025$?

Let's first use the table for our example. For what value or greater would this be significant with $df = 3$ and $\alpha = .025$?

p-value = `pchisq(19.78999, df=3, lower.tail=F) = 0.00019`

Standardized residuals

Now, we have evidence that the two variables are not independent. Where does the deviation from independence take place?

Standardized residuals

Now, we have evidence that the two variables are not independent. Where does the deviation from independence take place? Why did we reject the null? What does it mean?

Standardized residuals

Now, we have evidence that the two variables are not independent. Where does the deviation from independence take place? Why did we reject the null? What does it mean?

- We need to find the **adjusted residual**:

$$z = \frac{f_{observe} - f_{expected}}{se} = \frac{f_{observe} - f_{expected}}{\sqrt{f_e(1 - \text{row prop.})(1 - \text{column prop.})}}$$

- The denominator is the standard error of the quantity $f_o - f_e$ under the null hypothesis

Example: Calculating standardized residuals

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
	$f_e = 342.12$	$f_e = 359.88$	
Independents	f_o=224	f_o=170	394
	$f_e = 192.12$	$f_e = 201.98$	
Republicans	f_o=175	f_o=167	342
	$f_e = 166.67$	$f_e = 175.33$	
Other	f_o=29	f_o=35	64
	$f_e = 31.19$	$f_e = 32.81$	
Total	732	770	1502

Example: Calculating standardized residuals

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
	$f_e = 342.12$	$f_e = 359.88$	
Independents	f_o=224	f_o=170	394
	$f_e = 192.12$	$f_e = 201.98$	
Republicans	f_o=175	f_o=167	342
	$f_e = 166.67$	$f_e = 175.33$	
Other	f_o=29	f_o=35	64
	$f_e = 31.19$	$f_e = 32.81$	
Total	732	770	1502

$$z_{11} = \frac{304 - 342.12}{\sqrt{342.12(1 - \frac{702}{1502})(1 - \frac{732}{1502})}} \approx -2.395$$

2008 GSS

"Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the family has a very low income and cannot afford any more children?""

	Yes	No	Total
Democrats	222	225	447
Independents	201	277	478
Republicans	113	223	336
Other	18	12	30
Total	554	737	1291

Challenging jurors

- North Carolina Racial Justice Act of 2009
- Act specifically identified the kind of evidence that could be considered
- Defendant must prove that race was a significant factor in the imposition of the death penalty
- The evidence before you is exactly what was used to commute the sentence of Marcus Raymond Robinson
- Repealed by legislator in 2012