POLS 205
Political Science as a Social Science

Analyzing Tables of Data

Christopher Adolph

University of Washington, Seattle

May 17, 2010

Inference for a Sample Mean

Last time:

Inference from the Sample Mean to the Population Mean

Inference of the Difference of Population Means

Both used the *t*-test

More from the SurveyUSA 500 adult Washington sample

	Young (<35)	Older (35+)	Sum
For Income Tax	117	216	333
Against Income Tax	43	124	167
Sum	160	340	500

Let's put our survey data in tabular form

Percentages may be easier to read

More from the SurveyUSA 500 adult Washington sample

	Young (<35)	Older (35+)	*Sum*
For Income Tax	73.1%	63.5%	66.6%
Against Income Tax	26.9	36.5	33.3
Sum	100.0	100.0	100.0

Three questions:

1. Is Young support for the Income Tax significantly different from 50%?
2. Is Older support significantly different from 50%?
3. Are Young adults more supportive than Older adults?

Are Young Adults supportive of the income tax?

We use a t-test:

$$t = \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})}$$

Are Young Adults supportive of the income tax?

We use a t-test:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\ &= \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \end{aligned}$$

Are Young Adults supportive of the income tax?

We use a *t*-test:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\
&= \frac{0.731 - 0.5}{0.443/\sqrt{160}}
\end{aligned}
$$

Are Young Adults supportive of the income tax?

We use a t-test:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\ &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{0.731 - 0.5}{0.443/\sqrt{160}} \\ &= 6.596 \end{aligned}$$

A t this big would appear by chance less than 1 time in 1,526,717,557 random samples, (1 in 1.5 billion), for a $p = 0.000000000655$

We can just write this as $p < 0.001$

Are Older Adults supportive of the income tax?

We use a *t*-test:

$$t = \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})}$$

Are Older Adults supportive of the income tax?

We use a t-test:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\ &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \end{aligned}$$

Are Older Adults supportive of the income tax?

We use a *t*-test:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\
&= \frac{0.635 - 0.5}{0.481/\sqrt{340}}
\end{aligned}
$$

Are Older Adults supportive of the income tax?

We use a *t*-test:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\ &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{0.635 - 0.5}{0.481/\sqrt{340}} \\ &= 5.175 \end{aligned}$$

A t this big would appear by chance less than 1 time in 2,557,545 random samples, (1 in 2.5 million), for a $p < 0.000000391$

We can just write this as $p < 0.001$

Are Younger Adults More Supportive of the Income Tax than Older?

We use a t-test for comparison of means, which has this form:

$$t = \frac{\bar{x} - \bar{y}}{\text{se}(\bar{x} - \bar{y})}$$

However, the se's and dfs get a bit hairy, so we have STATA do it:

$t = 2.190$, $\text{df} = 335.365$, $p = 0.029$.

Are Younger Adults More Supportive of the Income Tax than Older?

We use a t-test for comparison of means, which has this form:

$$t = \frac{\bar{x} - \bar{y}}{\text{se}(\bar{x} - \bar{y})}$$

However, the se's and dfs get a bit hairy, so we have STATA do it:

$t = 2.190$, df = 335.365, $p = 0.029$.

A difference this big would occur by chance one in 34 draws from a populaton where Young and Older adults have the same view on the income tax

More from the SurveyUSA 500 adult Washington sample

	Young (<35)	Older (35+)	Sum
For Income Tax	117	216	333
Against Income Tax	43	124	167
Sum	160	340	500

What if this table had more than two columns or rows?

How could we infer whether there is an *association* between the row variable and the column variable?

Outline

Contingency Tables for Two Discrete Variables

Simple Inference for Contingency Tables

Multidimensional Tables (if time)

Example: Education & Partisan Identification

We have two variables from the General Social Survey:

Education Highest degree attained: No degree, High School diploma, Associates Degree, Bachelors Degree, Graduate Degree

Party Identification Strong Democrat, Democrat, Leans Democratic, Independent, Leans Republican, Republican, Strong Republican, Other

We take these data from the 1990 and 2006 samples of the GSS

What is the level of measurement of these variables?

How can we ascertain the relationship between them?

Monotonicity

Monotonic relationships are those which either consistently move in the same direction, or at least "stay still":

- If adding years of education always increases the *expected* probability one is Republican, or at least never lowers it, then Republican ID is *monotonically increasing* in Education
- If adding years of education always decreases the *expected* probability one is Republican, or at least never raises it, then Republican ID is *monotonically decreasing* in Education
- If adding years of education at first raises the expected probability of Republican ID, but then lowers it (or vice versa), the relationship is *non-monotonic*

Constructing a contingency table

The simplest way to explore the relationship between two discrete variables is a *contingency table*:

1. We consider every possible combination of education and party ID
2. Total up all subjects with that combination
3. Enter the sum in a *cross-tabulation*, with one variable's categories as the columns, and the other variable's categories as the rows
4. Customarily, the "dependent variable" (to the extent we believe one variable depends on the other) is the row variable

2006 General Social Survey: Partisanship & Education

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Str. Dem.	97	347	54	110	91	699
	Dem.	115	384	52	116	69	736
	Lean Dem.	67	265	50	87	58	527
	Indep.	263	503	86	92	53	997
	Lean Rep.	39	168	28	60	32	327
	Rep.	56	307	64	158	52	637
	Str. Rep.	40	256	37	118	44	495
	Other	9	32	3	18	3	65
	Sum	686	2262	374	759	402	4483

The above is a *contingency table* or *cross-tabulation*.

Powerful way to get the data. Can be tweaked to be more informative.

2006 GSS: Collapse partisans, treat leaners as independent

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	212	731	106	226	160	1435
	Independent	369	936	164	239	143	1851
	Republican	96	563	101	276	96	1132
	Other	9	32	3	18	3	65
	Sum	686	2262	374	759	402	4483

The first thing we will do is collapse some similar categories

Create **Democrat** out of the old “Strong Democrat” and “Democrat”

Create **Indepedent** out of the old “Leans Democratic”, “Independent”, and “Leans Republican”

Create **Republican** out of the old “Strong Republican” and “Republican”

2006 GSS: Collapse partisans, treat leaners as independent

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	212	731	106	226	160	1435
	Independent	369	936	164	239	143	1851
	Republican	96	563	101	276	96	1132
	Other	9	32	3	18	3	65
	Sum	686	2262	374	759	402	4483

Consolidation of categories reduces noise in each cell, but at a price: we've lost some of the fine-grained nature of our data

Introduces a trade-off between borrowing strength by pooling cells and informative measuremnt

Tabular methods pose this dilemma when applied to detailed ordered variables

2006 GSS: Collapse partisans, treat leaners as independent

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	212	731	106	226	160	1435
	Independent	369	936	164	239	143	1851
	Republican	96	563	101	276	96	1132
	Other	9	32	3	18	3	65
	Sum	686	2262	374	759	402	4483

Collapsing Party ID has simplified our table, but it's still hard to see the relationship between the variables

What could we do?

2006 GSS: Collapse partisans, treat leaners as independent

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	212	731	106	226	160	1435
	Independent	369	936	164	239	143	1851
	Republican	96	563	101	276	96	1132
	Other	9	32	3	18	3	65
	Sum	686	2262	374	759	402	4483

Collapsing Party ID has simplified our table, but it's still hard to see the relationship between the variables

What could we do? Perhaps percentages would be easier?

Let's divide by $N = 4483$, the total number of observations

2006 GSS: Percent of *N*

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	4.7%	16.3%	2.4%	5.0%	3.6%	32.0%
	Independent	8.2	20.9	3.7	5.3	3.2	41.3
	Republican	2.1	12.6	2.3	6.2	2.1	25.3
	Other	0.2	0.7	0.1	0.4	0.1	1.4
	Sum	15.3	50.5	8.3	16.9	9.0	100.0

Does this help?

2006 GSS: Percent of *N*

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	4.7%	16.3%	2.4%	5.0%	3.6%	32.0%
	Independent	8.2	20.9	3.7	5.3	3.2	41.3
	Republican	2.1	12.6	2.3	6.2	2.1	25.3
	Other	0.2	0.7	0.1	0.4	0.1	1.4
	Sum	15.3	50.5	8.3	16.9	9.0	100.0

Does this help? Not really. It's still hard to see the effects of each variable *separately*

We see that the combination of Democrat and High School is common, and Republican and College is rare, but does that mean there is an association?

That is, does being College educated make one less likely to be Republican? Or is it just that there are more High School grads than College grads?

2006 GSS: Percent of *N*

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	4.7%	16.3%	2.4%	5.0%	3.6%	32.0%
	Independent	8.2	20.9	3.7	5.3	3.2	41.3
	Republican	2.1	12.6	2.3	6.2	2.1	25.3
	Other	0.2	0.7	0.1	0.4	0.1	1.4
	Sum	15.3	50.5	8.3	16.9	9.0	100.0

What can we do to zero in on the likelihood that one is Republican given that one has a College Degree?

That is, how do we estimate the conditional probability $\Pr(\text{Republican}|\text{College})$?

2006 GSS: Percent of *N*

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	4.7%	16.3%	2.4%	5.0%	3.6%	32.0%
	Independent	8.2	20.9	3.7	5.3	3.2	41.3
	Republican	2.1	12.6	2.3	6.2	2.1	25.3
	Other	0.2	0.7	0.1	0.4	0.1	1.4
	Sum	15.3	50.5	8.3	16.9	9.0	100.0

What can we do to zero in on the likelihood that one is Republican given that one has a College Degree?

That is, how do we estimate the conditional probability $\Pr(\text{Republican}|\text{College})$?

How about the percentage of College grads that vote Republican in the sample?

2006 GSS: Column percentages

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	30.9%	32.3%	28.3%	29.8%	39.8%	32.0%
	Independent	53.8	41.4	43.9	31.5	35.6	41.3
	Republican	14.0	24.9	27.0	36.4	23.9	25.3
	Other	1.3	1.4	0.8	2.4	0.7	1.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

How about the percentage of College grads that vote Republican in the sample?

That is, what if we divide each *column* by its sum, to see how people with a given level of the column variable Education get distributed on the row variable, Partisan ID?

This is called showing "column percentages". Most useful presentation of a cross-tab

2006 GSS: Column percentages

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	30.9%	32.3%	28.3%	29.8%	39.8%	32.0%
	Independent	53.8	41.4	43.9	31.5	35.6	41.3
	Republican	14.0	24.9	27.0	36.4	23.9	25.3
	Other	1.3	1.4	0.8	2.4	0.7	1.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

Notice that with column percentages, each column sums to 100%

The interesting comparisons appear when we look *across* each row

For each row, higher values show positive relationships between that column category and the current row.

Low values within the row show negative relationships between the column category and the current row.

2006 GSS: Column percentages

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	30.9%	32.3%	28.3%	29.8%	39.8%	32.0%
	Independent	53.8	41.4	43.9	31.5	35.6	41.3
	Republican	14.0	24.9	27.0	36.4	23.9	25.3
	Other	1.3	1.4	0.8	2.4	0.7	1.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

In this example, $\Pr(\text{Democrat})$ is higher for those without high school diplomas or with graduate degrees, but lower for those with college degrees

Republicans do best among College degree holders, and worse at the ends of the Education spectrum

That is, support for either party seems to be a *non-monotonic* function of Education

2006 GSS: Column percentages

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	30.9%	32.3%	28.3%	29.8%	39.8%	32.0%
	Independent	53.8	41.4	43.9	31.5	35.6	41.3
	Republican	14.0	24.9	27.0	36.4	23.9	25.3
	Other	1.3	1.4	0.8	2.4	0.7	1.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

Notice that comparisons *across* rows in the column percentage cross-tab mean something different from comparisons across rows

For instance, Democrats do almost as well as Republicans in the strongest Republican category, College.

Why? College grads are more likely to be Republicans than any other education group. *But* more people on average are Dems, so even in this relatively weak category, Dems are fairly strong

2006 GSS: Column percentages

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	30.9%	32.3%	28.3%	29.8%	39.8%	32.0%
	Independent	53.8	41.4	43.9	31.5	35.6	41.3
	Republican	14.0	24.9	27.0	36.4	23.9	25.3
	Other	1.3	1.4	0.8	2.4	0.7	1.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

What if you encounter a cross-tab "in the field"?

Check if it's in column percentages, then start looking for patterns in each row

Remember this mantra: **Sum Down, Compare Across**

2006 GSS: Row percentages

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	14.8%	50.9%	7.4%	15.7%	11.1%	100.0%
	Independent	19.9	50.6	8.9	12.9	7.7	100.0
	Republican	8.5	49.7	8.9	24.4	8.5	100.0
	Other	13.8	49.2	4.6	27.7	4.6	100.0
	Sum	15.3	50.5	8.3	16.9	9.0	100.0

Why don't we use row percentages?
Because they show the conditioning of the columns on the rows, and we normally put the "dependent variable" in the rows

Inference for Tabular Data

We've learned how to assess relationships between discrete variables using cross-tabs

Enormously powerful technique that can detect even complex non-monotonic relationships

What's missing?

1. Are we sure the population has the same relationship as this sample?

2. What about confounders? Might the relationship we see between two variables be a spurious effect of a third variable?

2006 GSS: Marginal sums only

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat						1435
	Independent						1851
	Republican						1132
	Other						65
	Sum	686	2262	374	759	402	4483

To tackle inference from a sample to a population, we need to focus first on the marginal counts of the cross-tab

2006 GSS: Marginal proportions

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat						0.32
	Independent						0.41
	Republican						0.25
	Other						0.01
	Sum	0.15	0.50	0.08	0.17	0.09	1.00

To convert the marginal counts to marginal probabilities, we divide through by $N = 4483$

Now we have the *distributions* of our two categorical variables

2006 GSS: Estimated probabilities

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat						$\Pr(\text{d})$
	Independent						$\Pr(\text{ind})$
	Republican						$\Pr(\text{rep})$
	Other						$\Pr(\text{oth})$
	Sum	$\Pr(\text{ND})$	$\Pr(\text{HS})$	$\Pr(\text{AS})$	$\Pr(\text{CO})$	$\Pr(\text{GR})$	$\sum \Pr(\cdot)$

To emphasize this, we can replace these specific probabilities with their formal names

Independence

If Education and Party ID vary *independently*, what is the expected probability of having a specific combination of values?

Our point is broader than the two variables in our example, so let's imagine

- the rows of the table are index by $i \in \{1, \ldots, I\}$
- the columns of the table are indexed by $j \in \{1, \ldots J\}$
- the count in cell i, j is n_{ij}
- the overall count is $N = \sum_i \sum_j n_{ij}$

Independence

Call the probability we are in the *ith* row $\pi_{i.}$

Call the probability that we are in the *jth* column $\pi_{.j}$

Call the probability we are in cell i, j as π_{ij}

Now, if the rows and columns are independent, π_{ij} has a simple form:

$$\pi_{ij} = \pi_{i.} \times \pi_{.j}$$

2006 GSS: Predicted cell probabilities under independence

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	0.05	0.16	0.03	0.05	0.03	0.32
	Independent	0.06	0.21	0.03	0.07	0.04	0.41
	Republican	0.04	0.13	0.02	0.04	0.02	0.25
	Other	0	0.01	0	0	0	0.01
	Sum	0.15	0.5	0.08	0.17	0.09	1

Assuming no dependence between the rows and cells, we obtain the above predicted probabilities

If Education and Party ID have nothing to do with each other, these are the sample estimates that a random person from the population falls in each cell

2006 GSS: Predicted cell counts under independence

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	219.6	724.1	119.7	243.0	128.7	1435.0
	Independent	283.2	934.0	154.4	313.4	166.0	1851.0
	Republican	173.2	571.2	94.4	191.7	101.5	1132.0
	Other	9.9	32.8	5.4	11.0	5.8	65.0
	Sum	686.0	2262.0	374.0	759.0	402.0	4483.0

To convert the predicted probabilities for each cell into predicted counts for the sample, we just multiply each probability by $N = 4483$

The above predictions are for the model assuming *independence*, or no relationship between education and party

2006 GSS: Error under Independence Model

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	−7.6	6.9	−13.7	−17.0	31.3	0.0
	Independent	85.8	2.0	9.6	−74.4	−23.0	0.0
	Republican	−77.2	−8.2	6.6	84.3	−5.5	0.0
	Other	−0.9	−0.8	−2.4	7.0	−2.8	0.0
	Sum	0.0	0.0	0.0	0.0	0.0	0.0

All models are simplifications, and thus predict real data with error

If we used independence to "predict" the sample, how many cases would we misclassify? That is, how much error is there?

Above are the *residuals*, or $n_{ij} - \hat{n}_{ij}$: the actual count in the cell minus the estimated count

The χ^2 test

If Education and Party ID are *not* related in the general population, then they should appear to be independent variables in our sample

If our table represents the cross-tabulation of two independent variables, then each cell should be approximately $\hat{n}_{ij} = N\pi_i\pi_j$

This *independence model* forms our null hypothesis; if we reject it, we find *some* relationship holds between our variables

As with estimating the mean of a population, we will construct a test statistic, and see if that statistic seems "too large" to have been likely to occur if the null hypothesis is true

The χ^2 test

We can construct a statistic, X^2, which is 0 when our sample is perfectly predicted by the independence model

The worse independence appears to predict our sample, the bigger X^2

Specifically, we calculate:

$$\text{Pearson } X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Notice the numerator is the *squared error* for the cell, which we divide by the independence model prediction

The χ^2 test

If the population really has independent education and party ID, then we will only see a large X^2 very rarely

To see how rarely a large X^2 occurs by chance, note that X^2, as the sum of a finite series of squared normal variables, follows the χ^2 distribution

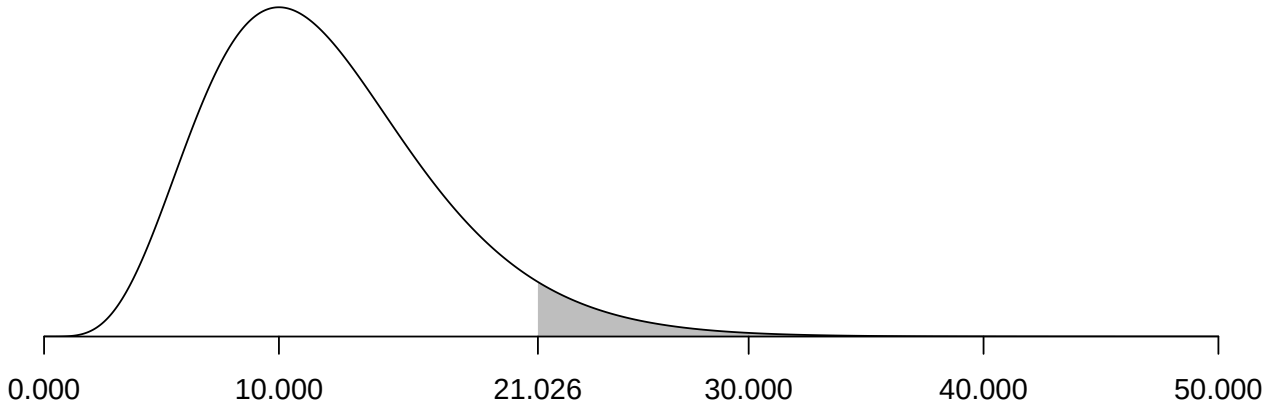
We can calculate this probability of seeing a particular X^2 by summing the area to the right of that value in the χ^2 distribution with $(I-1)(K-1)$ degrees of freedom

If this probability is very small, we consider that evidence against the chance that the variables are independent

Small p-values for the χ^2 suggest Education and Party ID depend on each other, but does not tell us the shape of this relationship, or the direction

To answer those questions, we'll need methods beyond POLS 205

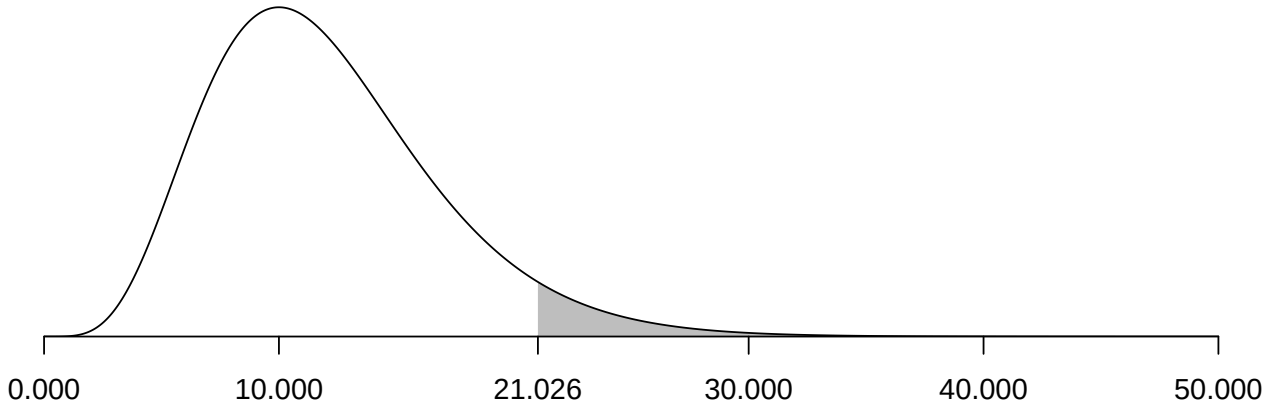
The χ^2 distribution with 12 df



Only 5% of this distribution has a value higher the 20.026

If we see a table with $X^2_{\text{df}=12} > 20.026$, we can conclude that table has an association between rows and columns that would occur by chance only 1 in 20 samples

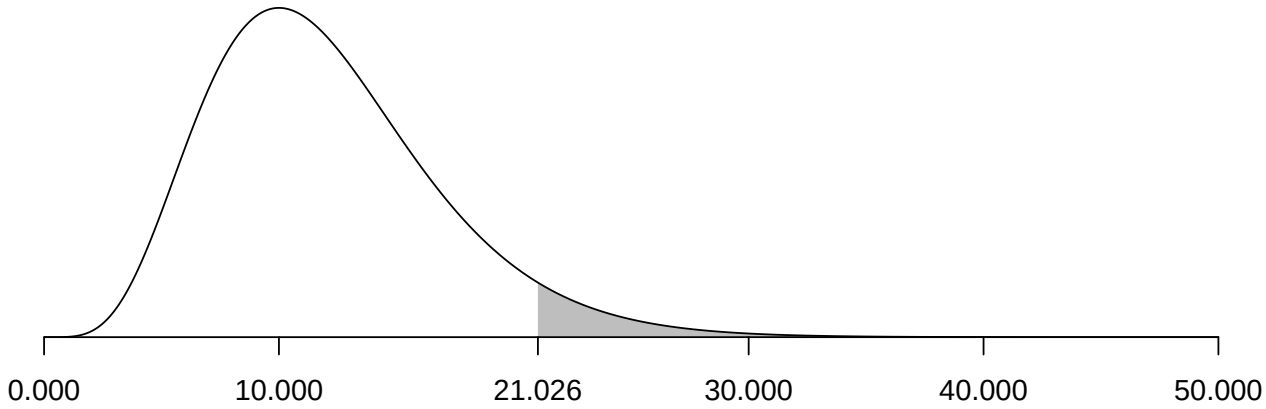
The χ^2 distribution with 12 df



This exercise is only valid if X^2 really follows this $\chi^2_{\text{df}=12}$distribution

That requires N be large, that all n_{ij} be above some threshold (e.g., 10 or so), and that each observation is an independently drawn random sample from the population

The χ^2 distribution with 12 df



If your test is "close" to the critical value, you should make sure the χ^2 approximation is appropriate

If your N or some n_{ij} are small, try one of the many available alternatives and corrections to χ^2 (e.g., Fisher's exact test, the Deviance, or X^2 with the Yates correction)

2006 GSS: Pearson Residuals

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	0.3	0.1	1.6	1.2	7.6	10.7
	Independent	26.0	0.0	0.6	17.7	3.2	47.4
	Republican	34.4	0.1	0.5	37.1	0.3	72.4
	Other	0.1	0.0	1.1	4.4	1.4	7.0
	Sum	60.7	0.2	3.7	60.4	12.5	137.5

The cell entries above are the Pearson residuals, $(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$

The sum of these, in the bottom right corner, is thus X^2

X^2 closer to 0 indicates a better fitting model; far from 0 a poor one.
If independence is a poor model, these variables are probably related

2006 GSS: Column percentages

		Highest Degree Attained					
		None	HS	Assoc	College	Grad	Sum
Party ID	Democrat	30.9%	32.3%	28.3%	29.8%	39.8%	32.0%
	Independent	53.8	41.4	43.9	31.5	35.6	41.3
	Republican	14.0	24.9	27.0	36.4	23.9	25.3
	Other	1.3	1.4	0.8	2.4	0.7	1.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

$N = 4483$. Pearson $X^2 = 137.5$ on 12 degrees of freedom, $p < 0.00000000000000022$.

If Education and Party ID are unrelated in the population, a X^2 this large would occur by chance in less than 1 in 4,500,000,000,000,000 large random samples.

Visualizing Tabular Data

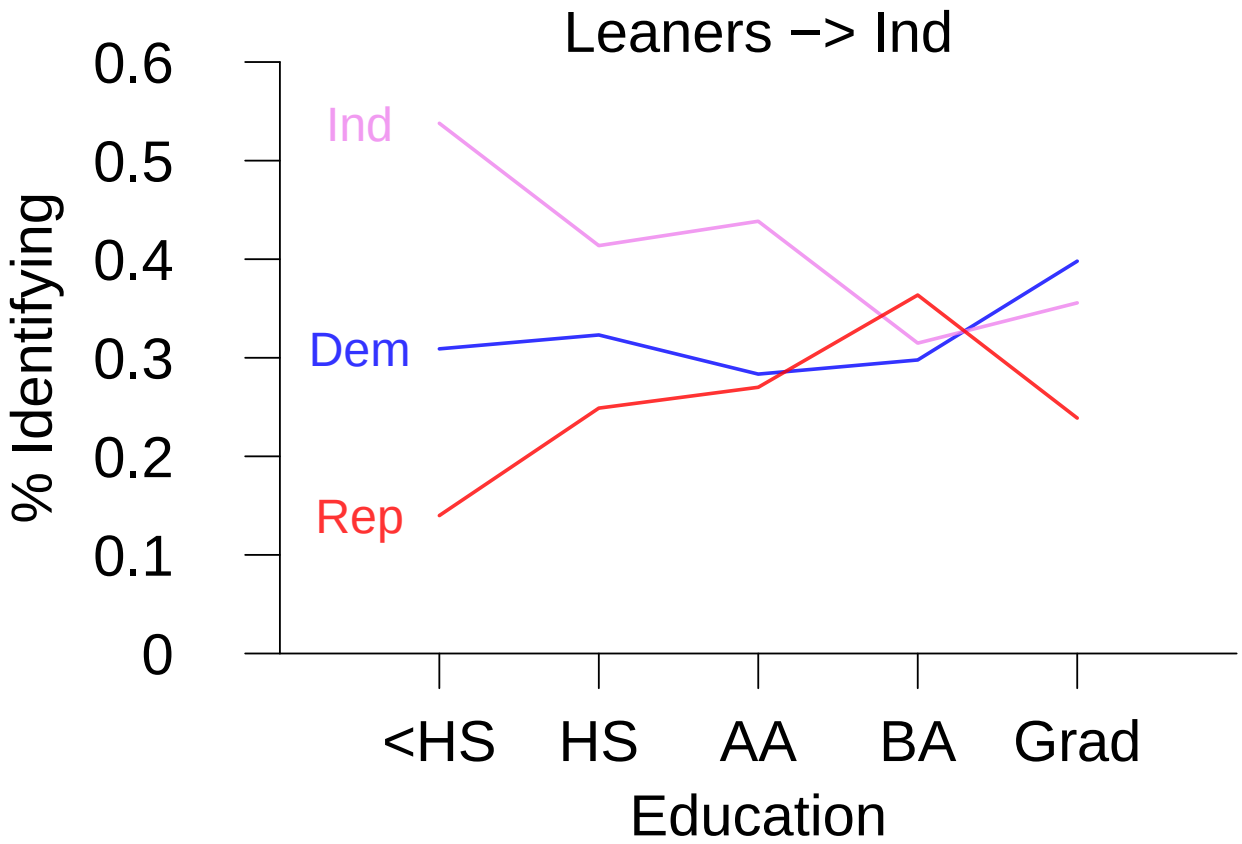
Just because our data come in a table doesn't mean we have to leave them there

A picture is often easier to sort out

But we need to plot the *right* numbers

What happens if we plot the *column percentages* from our tables?

The table as a graph



Exploring model sensitivity

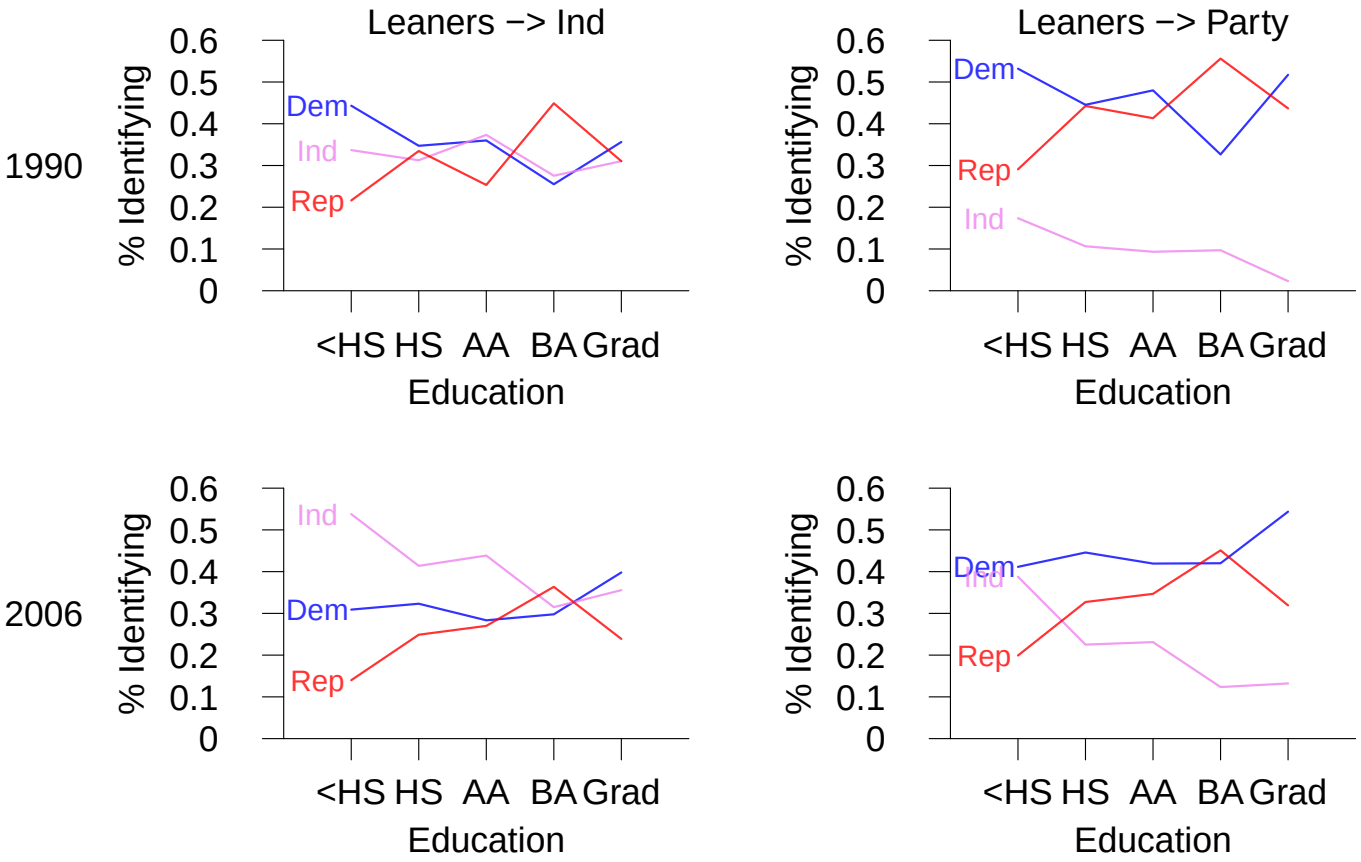
We made several assumptions in tabulating and analyzing our data

Categorizing Leaners We grouped leaners with other Independents. But many political scientists think they are actually intense partisans

Is 2006 special? We looked at just one year in American politics. Do our findings hold in other years? Is there interesting variation over time?

We could make more tables categorizing the leaners as partisans, or using data from, say 1990.

But who wants to pour over 4 cross-tabs?


1990
2006
Leaners −> Ind
Leaners −> Party
% Identifying
0.6
0.5
0.4
0.3
0.2
0.1
0
Dem
Ind
Rep
<HS HS AA BA Grad
Education

Monotonic Relationships

Let's examine the relationship between hours of TV watched & self-reported happiness.

The GSS includes:

TV Hours of TV watched per day. Reported by the hour; we will collapse to five categories: 0 hours, 1 to 2, 3 to 5, 6 to 10, and more than 10

Happiness "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?"

How might these variables be related?

2006 GSS: Raw counts

		Hours of TV watched per day					
		0	1 to 2	3 to 5	6 to 10	>10	Sum
Happiness	Very	513	4718	3387	600	101	93
	Pretty	702	7857	6731	1435	193	169
	Not Too	154	1278	1371	479	106	33
	Sum	1369	13853	11489	2514	400	296

2006 GSS: Column percentages

		Hours of TV watched per day					
		0	1 to 2	3 to 5	6 to 10	>10	Sum
Happiness	Very	37.5%	34.1%	29.5%	23.9%	25.3%	31.5%
	Pretty	51.3	56.7	58.6	57.1	48.3	57.1
	Not Too	11.2	9.2	11.9	19.1	26.5	11.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

$N = 29,625$. Pearson $X^2 = 397.8$ with 8 df. $p < 0.0000000000000022$.

(We can just write $p < 0.001$ to save space.)

What does this all mean, statistically *and* substantively?

Proportional Reduction in Error

Proportional Reduction in Error (PRE) statistics show how much of the variation in our dependent variable is explained by our independent variable

That is, if we know X, how much of the error in predicting Y can we eliminate?

χ^2 is not a PRE statistic

Instead, for monotonic relationships between (ordered) discrete variables, try Gamma

The Gamma Statistic

We will consider every possible "pair" of cases in our dataset, and classify into three groups:

Concordant pairs If case 1 is higher than case 2 on X, it is also higher on Y. The more concordant pairs, the more likely a positive, monotonic relationship

The Gamma Statistic

We will consider every possible "pair" of cases in our dataset, and classify into three groups:

Concordant pairs If case 1 is higher than case 2 on X, it is also higher on Y. The more concordant pairs, the more likely a positive, monotonic relationship

Discordant pairs If case 1 is higher than case 2 on X, it is *lower* on Y. The more discordant pairs, the more likely a negative, monotonic relationship

The Gamma Statistic

We will consider every possible "pair" of cases in our dataset, and classify into three groups:

Concordant pairs If case 1 is higher than case 2 on X, it is also higher on Y.
The more concordant pairs, the more likely a positive, monotonic relationship

Discordant pairs If case 1 is higher than case 2 on X, it is *lower* on Y.
The more discordant pairs, the more likely a negative, monotonic relationship

Tied pairs The cases share at least one value
The Gamma statistic ignores these pairs

The Gamma Statistic

Gamma has a simple form:

$$\text{Gamma} = \frac{\#\text{ of Concordant Pairs} - \#\text{ of Discordant Pairs}}{\#\text{ of Concordant Pairs} + \#\text{ of Discordant Pairs}}$$

Gamma has a possible range from:

- -1 (X completely explains Y, and is negatively related)
- 1 (X completely explains Y, and is negative related)

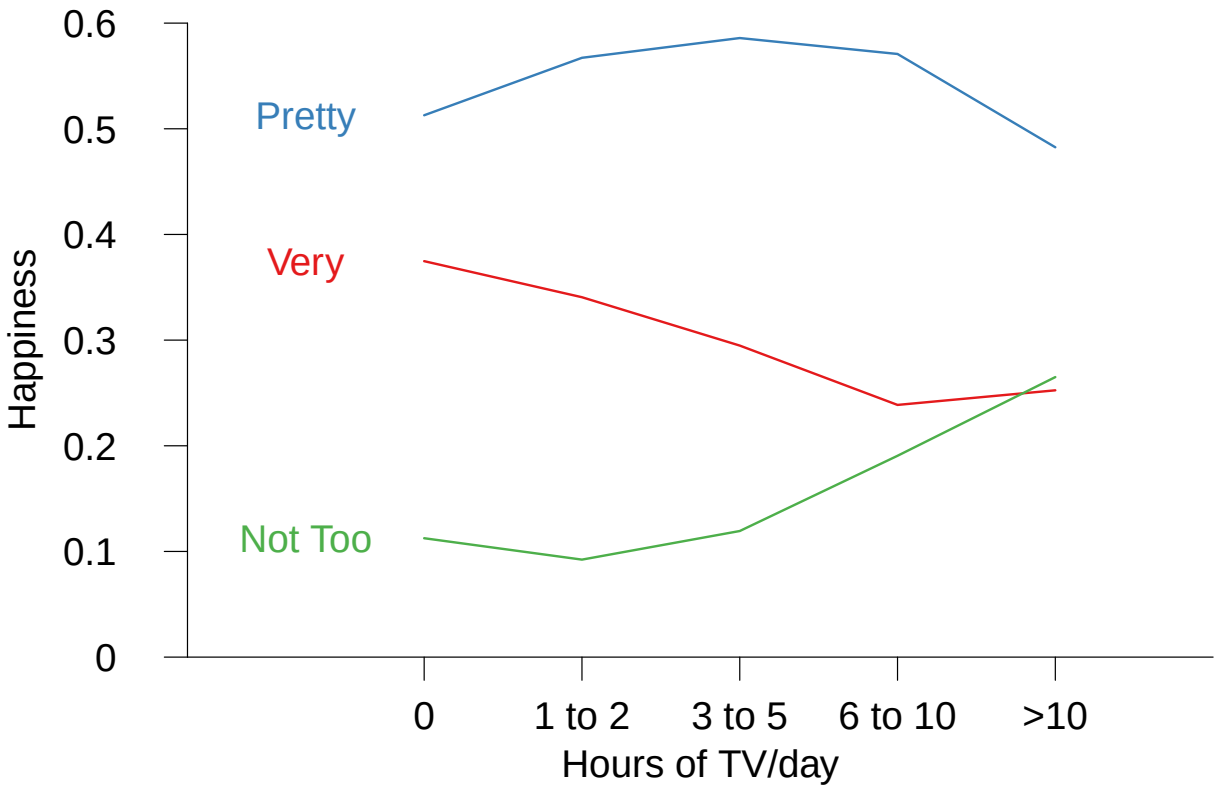
2006 GSS: Column percentages

		Hours of TV watched per day					
		0	1 to 2	3 to 5	6 to 10	>10	Sum
Happiness	Very	37.5%	34.1%	29.5%	23.9%	25.3%	31.5%
	Pretty	51.3	56.7	58.6	57.1	48.3	57.1
	Not Too	11.2	9.2	11.9	19.1	26.5	11.4
	Sum	100.0	100.0	100.0	100.0	100.0	100.0

$N = 29,625$. Pearson $X^2 = 397.8$ with 8 df. $p < 0.001$. Gamma $= 0.14$.

Knowing how much TV someone watches reduces error in predicting their happiness by 14%.

A graph is still the best summary



Final thoughts on 2-D cross-tabs

1. Inferential statistics like χ^2 and Gamma can help confirm your table isn't a mirage resulting from sampling error
2. Column percentages are essential for pining down the substance of the relationship
3. Graphics are best of all: easiest to read, and highlights the substantive size of the relationship

Contingency tables in the context of the course

Our study of associations between sampled variables began with comparison of means

That limited us to assessing the effect of a binary variable on one other variable

Crosstabulations allow us to infer relationships b/w two discrete variables regardless of the number of categories in each

Still missing:

1. Methods for continuous variables
2. Controls for confounders

Multidimensional Tables

If we want to consider possible confounders, we need more than two dimensions to our table

That is, we need one dimension for every independent variable, plus one for our dependent variable

This gets tricky fast: hard to visualize, or do our column percents trick

But important to consider: if we don't include confounders, we can make very incorrect inferences about relationships

Discrimination?

Suppose the (fictional) University of Tlon is sued for discriminatory hiring

Both sides stipulate that

- the best candidate can be determined uniquely
- should always be hired
- is equally likely to be male or female

The case turns on whether the University hired male and female candidates at the same rate

Discrimination?

Here is the data for the university's "eclectic" departments

Hiring data for Tlon University's "eclectic" departments

Departments	Men		Women	
	Hired	Applied	Hired	Applied

Discrimination?

Here is the data for the university's "eclectic" departments

Hiring data for Tlon University's "eclectic" departments

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5

Discrimination?

Here is the data for the university's "eclectic" departments

Hiring data for Tlon University's "eclectic" departments

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5
Navajo Cryptography	4	5	6	8

Discrimination?

Here is the data for the university's "eclectic" departments

Hiring data for Tlon University's "eclectic" departments

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5
Navajo Cryptography	4	5	6	8

The plaintiffs point out that in each dept, a greater % of men were hired:

Departments	Men		Women
Ancient Egyptian Algebra	25%	$>$	20%

Discrimination?

Here is the data for the university's "eclectic" departments

Hiring data for Tlon University's "eclectic" departments

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5
Navajo Cryptography	4	5	6	8

The plaintiffs point out that in each dept, a greater % of men were hired:

Departments	Men		Women
Ancient Egyptian Algebra	25%	$>$	20%
Navajo Cryptography	80%	$>$	75%

Discrimination?

"But wait!" says the defense. "Look at the *totals*"

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5
Navajo Cryptography	4	5	6	8

Discrimination?

"But wait!" says the defense. "Look at the *totals*"

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5
Navajo Cryptography	4	5	6	8
Total	6	13	7	13

"We actually hired *more* women at a higher rate than men!"

Discrimination?

"But wait!" says the defense. "Look at the *totals*"

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5
Navajo Cryptography	4	5	6	8
Total	6	13	7	13

"We actually hired *more* women at a higher rate than men!"

The plaintiffs in a lawsuit point out that in each dept, a greater % of men were hired:

Departments	Men		Women
Ancient Egyptian Algebra	25%	$>$	20%
Navajo Cryptography	80%	$>$	75%

Discrimination?

"But wait!" says the defense. "Look at the *totals*"

	Men		Women	
Departments	Hired	Applied	Hired	Applied
Ancient Egyptian Algebra	2	8	1	5
Navajo Cryptography	4	5	6	8
Total	6	13	7	13

"We actually hired *more* women at a higher rate than men!"

The plaintiffs in a lawsuit point out that in each dept, a greater % of men were hired:

Departments	Men		Women
Ancient Egyptian Algebra	25%	$>$	20%
Navajo Cryptography	80%	$>$	75%
Both departments	46%	$<$	54%

What's going on here?

Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

But these department level variables are confounders: correlated with the outcome *and* with our explanatory variable

Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

But these department level variables are confounders: correlated with the outcome *and* with our explanatory variable

Omitting them leads to this confusion, known as Simpson's Paradox (similar to the ecological fallacy)

Simpson's Paradox

The Departments are different. Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

But these department level variables are confounders: correlated with the outcome *and* with our explanatory variable

Omitting them leads to this confusion, known as Simpson's Paradox (similar to the ecological fallacy)

Presentation and Tabular Data

Devising good graphics takes creativity, attention to the reader, and hard work

Coming up with a good table requires the same things

Don't just slap down rows and columns in arbitrary order

Choose order, dimensions, scale, and nesting to highlight relationships

Avoid excessive digits; the first digit or two is always most important, and deserve exclusive attention

A multi-dimensional table: Titanic Example

A 4-D table groups all persons on the *Titanic* by gender, age, class, and survival

A multi-dimensional table: Titanic Example

A 4-D table groups all persons on the *Titanic* by gender, age, class, and survival

	Adult				Child			
	Male		Female		Male		Female	
	Died	Lived	Died	Lived	Died	Lived	Died	Lived
1st class	118	57	4	140	0	5	0	1
2nd	154	14	13	80	0	11	0	13
3rd	387	75	89	76	35	13	17	14
Crew	670	192	3	20	—	—	—	—

A multi-dimensional table: Titanic Example

A 4-D table groups all persons on the *Titanic* by gender, age, class, and survival

	Adult				Child			
	Male		Female		Male		Female	
	Died	Lived	Died	Lived	Died	Lived	Died	Lived
1st class	118	57	4	140	0	5	0	1
2nd	154	14	13	80	0	11	0	13
3rd	387	75	89	76	35	13	17	14
Crew	670	192	3	20	—	—	—	—

What do the "—"'s mean?

A multi-dimensional table: Titanic Example

A 4-D table groups all persons on the *Titanic* by gender, age, class, and survival

	Adult				Child			
	Male		Female		Male		Female	
	Died	Lived	Died	Lived	Died	Lived	Died	Lived
1st class	118	57	4	140	0	5	0	1
2nd	154	14	13	80	0	11	0	13
3rd	387	75	89	76	35	13	17	14
Crew	670	192	3	20	—	—	—	—

What do the "—"'s mean?

What patterns leap out?

A multi-dimensional table: Titanic Example

A 4-D table groups all persons on the *Titanic* by gender, age, class, and survival

	Adult				Child			
	Male		Female		Male		Female	
	Died	Lived	Died	Lived	Died	Lived	Died	Lived
1st class	118	57	4	140	0	5	0	1
2nd	154	14	13	80	0	11	0	13
3rd	387	75	89	76	35	13	17	14
Crew	670	192	3	20	—	—	—	—

What do the "—"'s mean?

What patterns leap out?

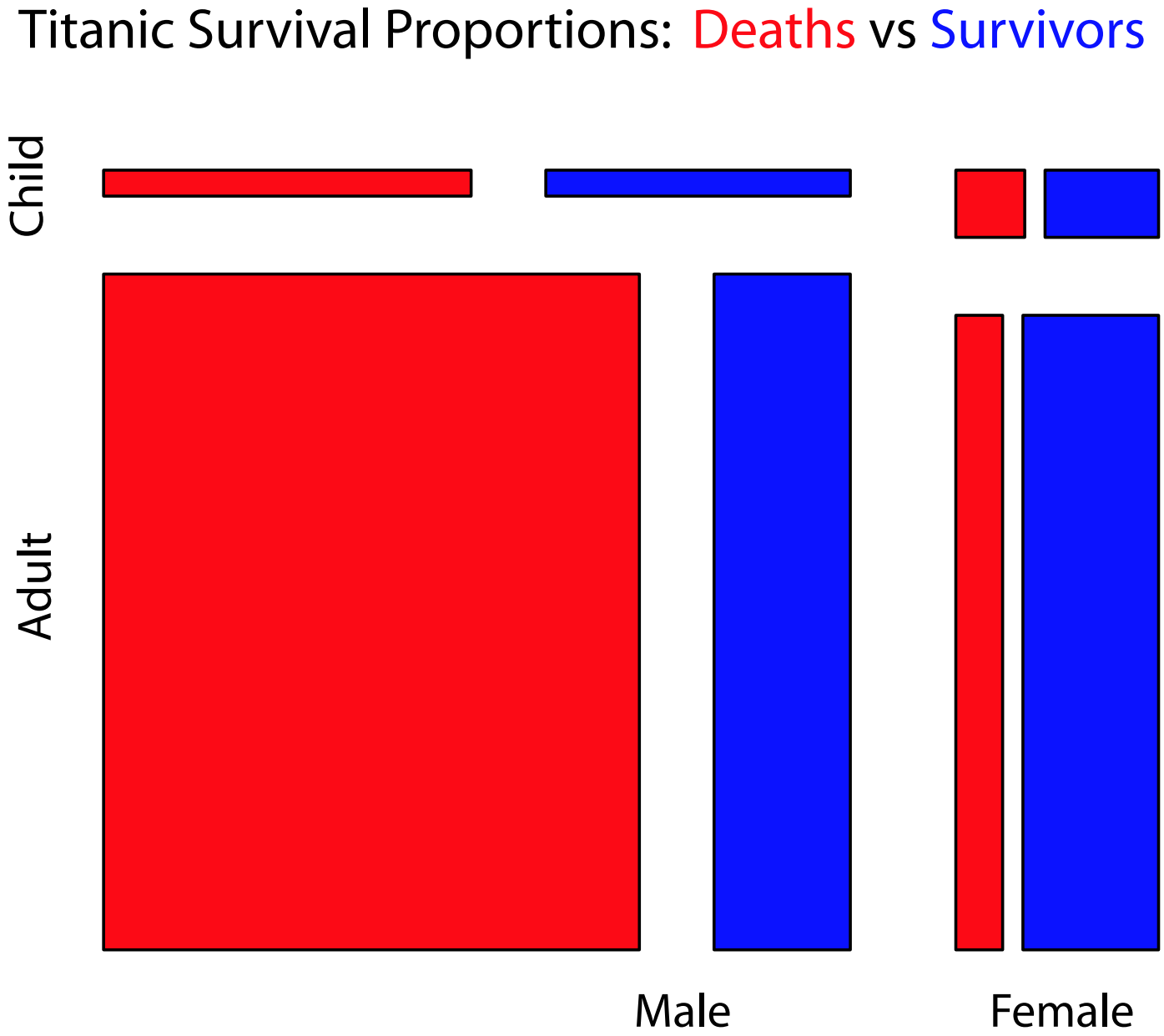
Could we make a graphic of all this?

Titanic Example: Mosaic plots

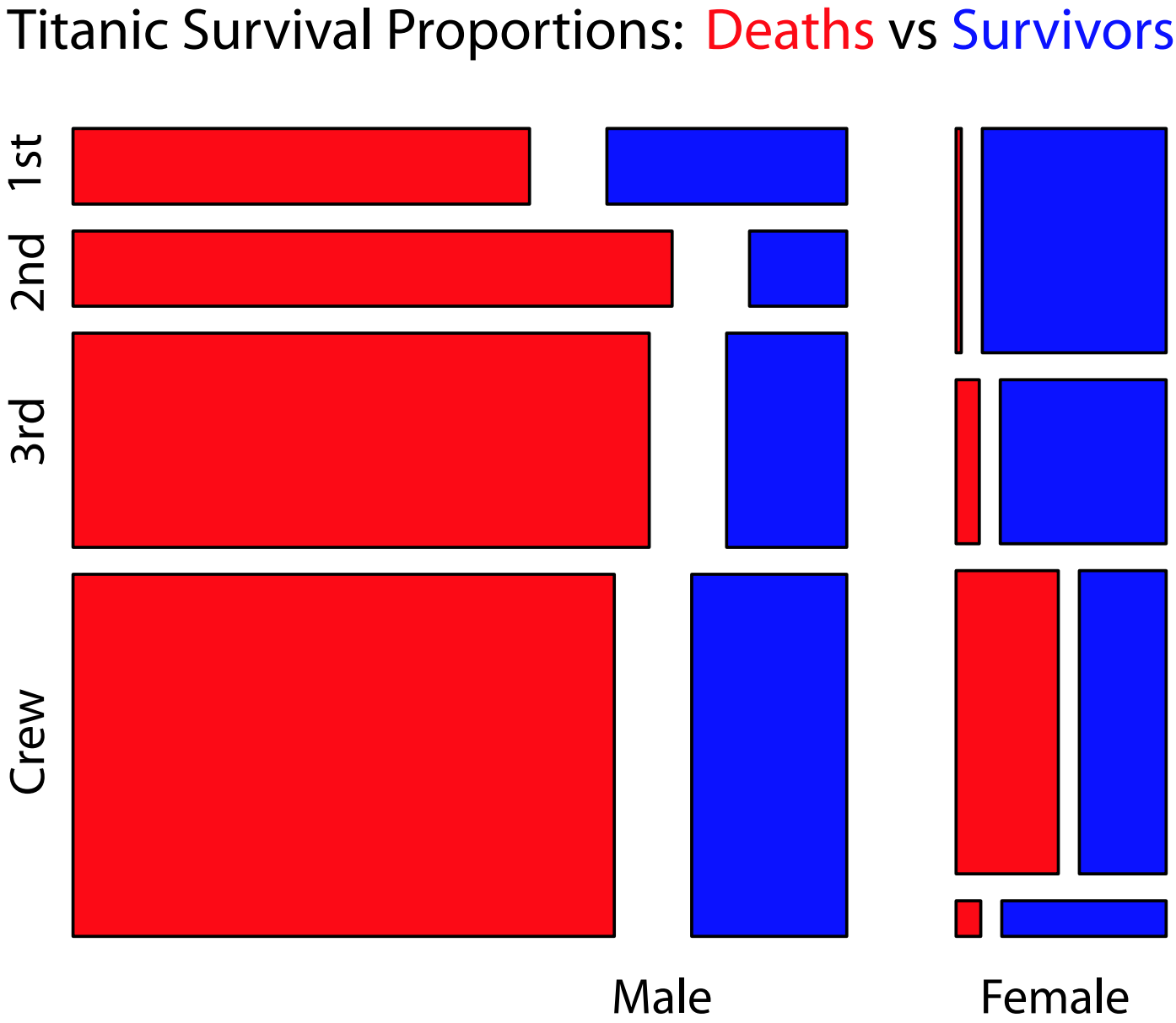
Mosaic plots:

- An exploratory tool for hihg-dimension contingency tables
- Draw rectangles for each cell,
  with height and width showing the proportions of observations falling in the cell
- Take some effort to understand at first . . .

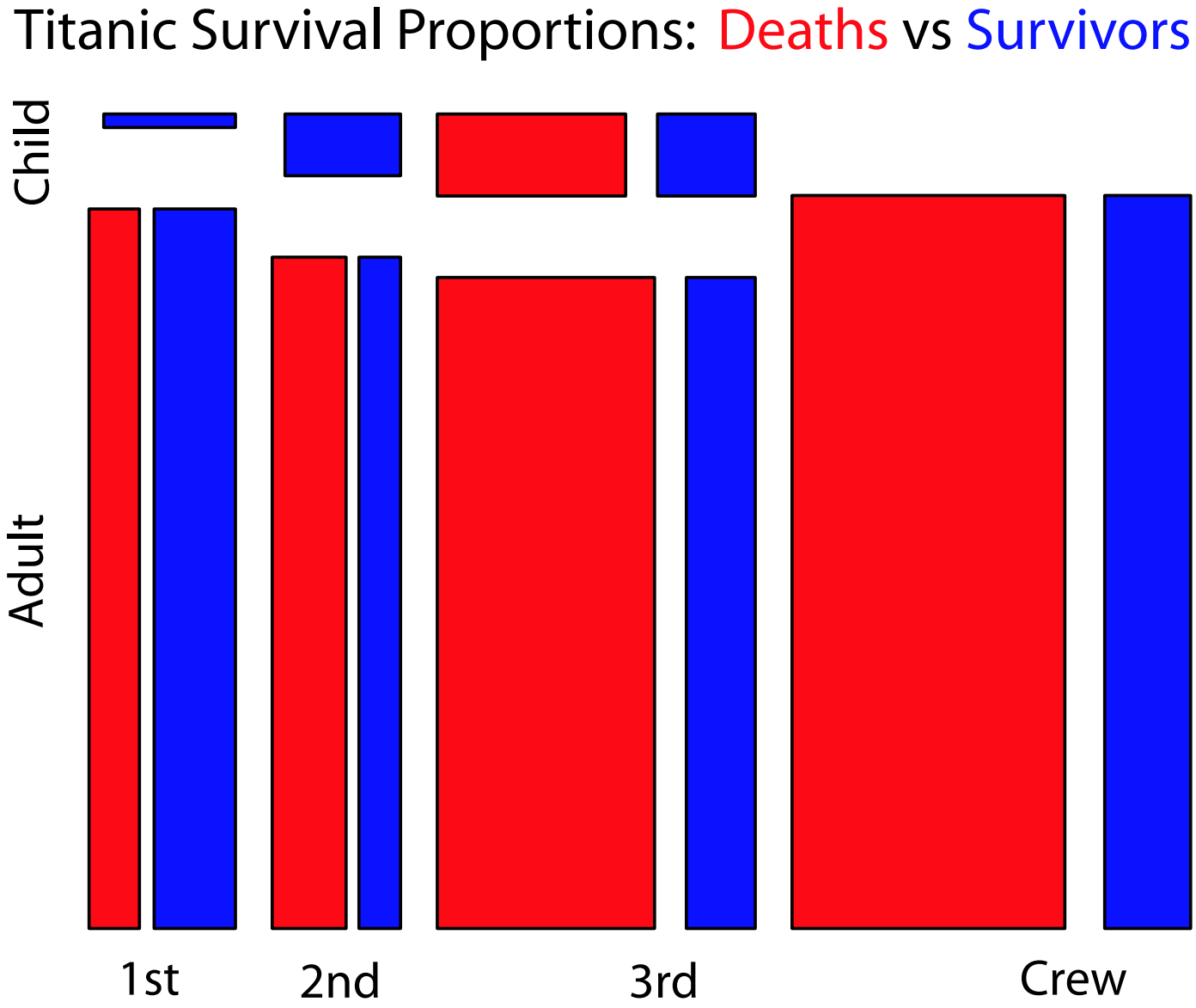
Mosaic: Age, Gender, and Survival



Mosaic: Class, Gender, and Survival



Mosaic: Age, Class, and Survival



Mosaic: Age, Class, Gender, and Survival