CSSS/SOC/STAT 321
Case-Based Social Statistics I

Analyzing Tabular Data

Christopher Adolph

Department of Political Science

and

Center for Statistics and the Social Sciences

University of Washington, Seattle

Inference for a Sample Mean

Last time:

Inference from the Sample Mean to the Population Mean

Inference of the Difference of Population Means

(which we saw was also inference for a 2×2 table)

Both used the t-test

What if we wanted to make inferences about associations
in a larger $R \times C$ table?

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

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

Recall these data from earlier in the quarter

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

Recall that to see associations, we converted to "column percentages."
Most useful presentation of a cross-tab

Inference for Tabular Data

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

Powerful technique for detecting 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 indexed by $i \in \{1, \ldots, I\}$

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 indexed by $i \in \{1, \ldots, I\}$
- the columns of the table are indexed by $j \in \{1, \ldots J\}$

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 indexed 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}

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 indexed 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 $\pi_{i\cdot}$

Independence

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

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

Independence

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

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

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

Independence

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

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

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

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

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

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	0.00	0.01	0.00	0.00	0.00	0.01
	Sum	0.15	0.50	0.08	0.17	0.09	1.00

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 $\hat{n}_{ij} = N\pi_i\pi_j$

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

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 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

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

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

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

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

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

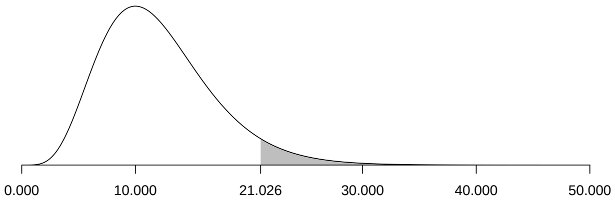
We can calculate this probability of seeing a particular X^2 by summing the area to the right of that value in the χ^2 distribution with $(I-1)(K-1)$ degrees of freedom

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, would need methods beyond CSSS 321

The χ^2 distribution with 12 df

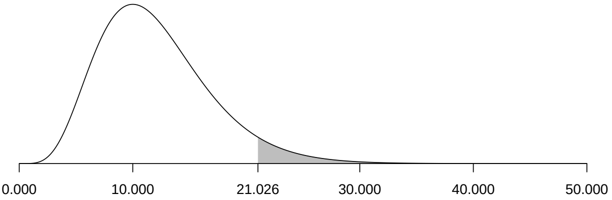


Only 5% of this distribution has a value higher than 20.026

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

(Why are we only testing for extreme values in one tail of χ^2?)

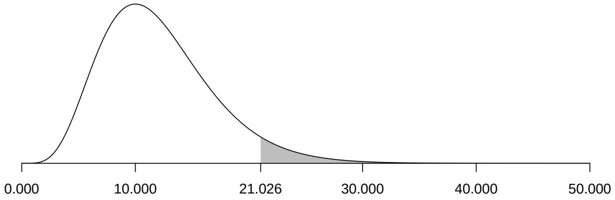
The χ^2 distribution with 12 df



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

The impact of tuition hikes on first-in-family college attendees

For the next example, we will revisit an example from a class survey of University of Washington undergraduates (January 2012 convenience sample):

Does a student's parents and/or grandparents college attendance predict that student's self-reported ability to cope with tuition hikes?

We expect a positive association.

Possible mechanisms: older generations' college could produce wealth, income, knowledge about college aid/admission/preparation, or a pro-education ethic

The impact of tuition hikes on first-in-family college attendees

PG Whether any of a student's parents and/or grandparents attended college. Ordered, from oldest family history of college to newest, in three categories:

1. at least one parent and at least one grandparent attended
2. at least one parent but no grandparents attended
3. no parents and no grandparents attended.

Tuition Self-reported ability to cope with recent UW tuition hikes. Ordered in four categories from greatest to least ability to cope:

1. No material effect
2. Difficult but manageable
3. Taking out more loans
4. Time off or transfer

2012 Class survey of UW students: Raw counts

		Family college attendance history			
		At least one parent and one grand-parent	At least one parent and no grand-parents	No parents and no grand-parents	Total
Tuition Hike	No material effect	228	119	59	406
	Difficult but manageable	248	185	146	579
	Took out more loans	89	75	70	234
	Taking time off or transferring	2	10	9	21
	Total	567	389	284	1240

2012 Class survey of UW students: Column percentages

		Family college attendance history			
		At least one parent and one grand-parent	At least one parent and no grand-parents	No parents and no grand-parents	Mean
Tuition Hike	No material effect	0.402	0.306	0.208	0.305
	Difficult but manageable	0.437	0.476	0.514	0.476
	Took out more loans	0.157	0.193	0.246	0.199
	Taking time off or transferring	0.004	0.026	0.032	0.020
	Total	1.000	1.000	1.000	1.000

$N = 1,240$. Pearson $X^2 = 44.63$ with 6 df. $p < 0.0000000554$.

(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 the Gamma statistic

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 positive related)

2012 Class survey of UW students: Column percentages

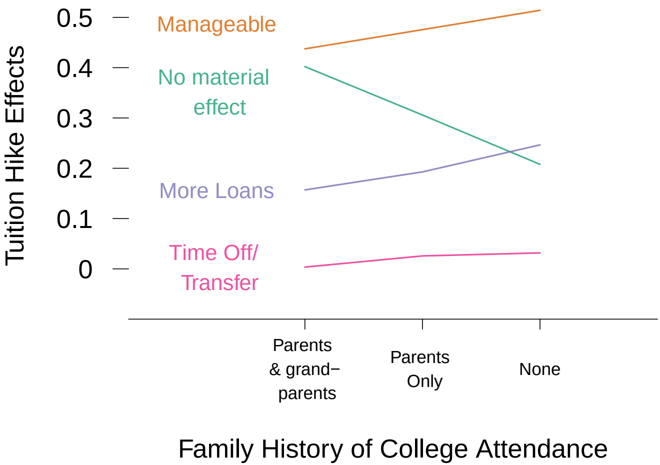
		Family college attendance history			
		At least one parent and one grand-parent	At least one parent and no grand-parents	No parents and no grand-parents	Mean
Tuition Hike	No material effect	0.402	0.306	0.208	0.305
	Difficult but manageable	0.437	0.476	0.514	0.476
	Took out more loans	0.157	0.193	0.246	0.199
	Taking time off or transferring	0.004	0.026	0.032	0.020
	Total	1.000	1.000	1.000	1.000

$N = 1,240$. Pearson $X^2 = 44.63$ with 6 df. $p < 0.001$. Gamma $= 0.244$.

Knowing a student's family college attendance history reduces error in predicting effects of tuition hikes by 24.4%.

Note that just as on the midterm, we aren't sure if this relationship is causal, or just the result of confounders

A graph is still a useful summary



Proportion of students self-reporting difficult with tuition hikes by family history of college attendance. Data taken from 221 class suvery (convenience sample of fellow University of Washington students). $N = 1,240$. Pearson $X^2 = 44.63$ with 6 df. $p < 0.001$. Gamma $= 0.244$.

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 often 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 between two discrete variables regardless of the number of categories in each

Still missing:

1. Methods for continuous variables
2. Controls for confounders