Lecture 20: Chi Square

Note the first type of chi square in this lecture -Goodness of Fit- is not currently on the take home test. See “contingency” or “independence” test below. I write this on the test but you’d be surprised how many miss it.

Introduction

Up until now, we done statistical test using means, but the assumptions for means have eliminated certain types of variables. For example, since the mean is not an appropriate measure of central tendency for nominal data, we have not been able to use these sorts of variables. Chi-square tests allow us to do so.

There are two types of Chi-square tests:

Goodness of Fit tests look at one variable only, and
Contingency Table Tests (or Tests of Independence) allow us to examine two variables at a time.

Goodness of Fit – note this is not currently on the take home test. See “contingency” or “independence” test below

Some Common Sense Assumptions Chi-Square Goodness of Fit test:

1. Discrete data only. Chi-square is best for Nominal Data. **(Hint for exam: no student project should ever violate this nor have to assume it. Your data set will have this sort of variable.)**
2. We examine only one variable at a time. **(Hint for exam: no student project should ever violate this nor have to assume it. Your data set will one variable.)**
3. The categories of the variable are mutually exclusive. **(Hint for exam: no student project should ever violate this nor have to assume it. Your data set will have this sort of variable.)**
4. The data comes from a random sample. **(Hint for exam: all student projects violate this assumption.)**
5. Frequency counts can be obtained for each of the possible categories of the variable. **(Hint for exam: no student project should ever violate this nor have to assume it. Your data set will this sort of variable.)**
6. The expected frequency count for each category of the variable is *at least 5.* (This suggests that large n’s are best**). (Hint for exam: if you violate this SPSS will give you and “error note” along with your output.)**
7. With Chi-Square here is not an assumption of a “normal distribution” of any kind.

Theory Behind Goodness of Fit

A goodness of fit test looks to see whether a single variable follows or “fits” some hypothesized probability distribution. For example, maybe we would expect that in a population of surfers, there would be an equal number of longboarders and shortboarders. Maybe we would expect it not to be equal! Goodness of Fit allows us to test either theory. The point here is that these probability distributions can follow any pattern that *theory* suggests.

We can also look at “market share.” Maybe we could look at type of plate lunch ordered and we would expect people to order an equal number of each. Maybe a beer marketing firm says that 50% of beer drinkers on the island drink Bud products, 25% drink Coors, and another 25% drink Miller. We could see if that is true. The point here is that these probability distributions can follow any pattern that *theory* suggests.

We could look at “market share” of felons in prison for drug dealing. Pretend want to measure policing of drug crimes and we want to know what sorts of drug dealers are being caught? Pretend the mayor has asked the cops to concentrate on arresting methamphetamine dealers. So we could test the theory that the police arresting an equal number of Ice (or methamphetamine), marijuana, and crack dealers. If so, then about 33% of the arrests should be for crack, 33% for Ice, and 33% for marijuana. If we can prove this theory wrong, the police are NOT arresting all types of drug dealers equally.

By the way, we could also change the percentages in each category and test them. We won’t do that below, but pretend the mayor wanted at least 50% of all drug arrests for Ice and the remaining split evenly: 25% for marijuana and 25% for crack. *The point is you can test any proportions or percentages that theory suggests, it just makes for slightly easier math to in step 6 to make all categories equal.*

You could investigate whether an equal percentage of the population supports a particular public policy. So for example, in recent years many state legislatures have looked at abolishing the death penalty. Hawaii does not have the death penalty, but getting rid of the death penalty is attractive to some states for a variety of reasons such as it costs 10 times more to execute someone as to put them in prison “for life,” the death penalty does not lower murder rates, we accidently kill innocent offenders often, etc. In my criminal justice courses I go into greater detail on these matters. But let’s look at support for the death penalty.

Example of Goodness of Fit Test

Let's use the death penalty example. Pretend we have a question "Do you support or oppose the death penalty?"
1 = support the death penalty
2 = oppose the death penalty

DP	**# of persons**
Support	50
Oppose	25

In plain English we will test the theory that the population distribution is uniform – an equal percentage of the population supports as opposes the death penalty. Or we will try to prove the population distribution is not uniform – <u>an unequal</u> percentage of the population supports as opposes the death penalty

1. **State null and alternative hypothesis**.

H_o: The population distribution is uniform – an equal percentage of the population ***an equal percentage of the population supports as opposes the death penalty***
H_1: The population distribution is not uniform – ***– <u>an unequal</u> percentage of the population supports as opposes the death penalty***

2. State level of significance or a "alpha."
For this example we'll use alpha =.05

3. **Determine the test distribution to use –** Chi Square tests use X^2 distribution.

4. **Define the rejection regions. And draw a picture!**
   df= k-1 (where k= # of categories in variable). In this case df = 2-1 = 1. Using Appendix 6 (or the chi square table linked in our course schedule) the critical value df=1area in tail - .05 = 3.84

5. **State the decision rule.**

Reject the null if the TR >3.84, otherwise FTR.

6. **Perform necessary calculations on data and compute TR value.**
   Recall in the take home tests you do NOT have to do the math by hand. In this step write, "SPSS says TR =______" and report the correct SPSS TR. Below I show you how the math is done by hand for illustrative purposes.

$$TR = \sum \frac{(O - E)^2}{E}$$

O = # of observed cases E = # of expected cases

note how the TR formula above looks ***<u>a bit</u>*** like a variance formula

$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$ = This (population) variance formulas sort of asks "on average how much to all of these x's differ from the mean?"

Well the TR formula above sort of asks "on average how much do our observations differ from what our theory expected?" If they differ a lot from the theory, the we will reject the theory. If they do not differ a lot from the theory, then the theory is probably correct, or at least we do not have sufficient evidence to reject the theory.
Large differences between expected and observed values will tend to have larger numerators in the fraction and thus larger TR values and thus "tend towards significance."
What happens when the number on the top of the fraction gets larger relative the number on the bottom of the fraction? The number represented by the fraction "gets bigger:"
1/10 2/10 3/10 5/10 10/10 20/10 note how the numbers get bigger as the top number of the fraction grows larger.

Test Ratio Computation					
DP	**Observed**	**Expected**	**O - E**	**$(O - E)^2$**	**$(O - E)^2/E$**
Support	50	37.5	12.5	156.25	4.167
Oppose	25	37.5	-12.5	156.25	4.1667
Σ (sum)	**80**	**75**			**8.333**

TR = 8.33

7. **Compare TR value with the decision rule and make a statistical decision. (Write out decision in English! -- my addition)**

In this case TR is more than 3.84 and therefore we reject the null and conclude the population distribution is not uniform – ***an unequal* percentage of the population supports as opposes the death penalty**

What is the wording when we fail to reject the null hypothesis?

When we fail to reject the null hypothesis we say "Insufficient evidence to reject theory that __________ " [insert Ho in plain English.]

Using SPSS

Below is the output for these data using SPSS. See lecture 20c_SPSS.pdf for how to have SPSS create this output.

Do you support the Death Penalty

	Observed N	Expected N	Residual
support	50	37.5	12.5
opppose	25	37.5	-12.5
Total	75		

Test Statistics

	Do you support the Death Penalty
Chi-Square	8.333[a]
df	1
Asymp. Sig.	.004

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 37.5.

Contingency Table or Test for Independence

A contingency table test allows us to test whether 2 variables are independent of each other. That is to say, does one variable affect the outcome of another? If they are truly independent then one variable DOES NOT affect the outcome of the other.

Some Common Sense Assumptions for Contingency Table Test:

1. Discrete data only. Chi-square is best for Nominal Data. **(Hint for exam: no student project should ever violate this nor have to assume it. Your data set will have this sort of variable.)**
2. We examine only two variables at a time. **(Hint for exam: no student project should ever violate this nor have to assume it. Your data set will two variables.)**
3. The data comes from a random sample. **(Hint for exam: all student projects violate this assumption.)**
4. The categories for each of the two variables are mutually exclusive. **Hint for exam: no student project should ever violate this nor have to assume it. Your data set will have this sort of variable.)**
5. Frequency counts can be obtained for each of the possible categories of the two variables. **(Hint for exam: no student project should ever violate this nor have to assume it. Your data set will have two such variables.)**
6. The expected frequency count for each cell of the contingency table is *at least 5.* (This suggests that large n's are best). **(Hint for exam: It is likely you will violate this assumption. If you violate it, SPSS will give you and "error note" along with your output.)**
7. With Chi-Square here is not an assumption of a "normal distribution" of any kind.

Example of Contingency Table Test

Public opinion surveys in the US tend to show that there is a relationship between gender and support for the death penalty. In general, men are stronger supporters of the death penalty than women.

So we have two variables: gender (1=male 2= female) and a question "Do you support or oppose the death penalty?"
(1 = support the death penalty 2 = oppose the death penalty).

So, in plain English, we will do a Contingency Table Chi Square (or a Chi-Square Test of Independence) to test the theory that whether or not someone supports the death penalty is ***independent of or NOT related*** *to whether they identify as male or female. Or conversely, attempt do the chi-square test to prove that whether or not someone supports the death penalty is* ***dependent upon or related to*** *whether they identify as male or female.*

Say the two variables "cross-tabulate" like this:

	Male	Female
Support	26	3
Oppose	24	22

1. **State null and alternative hypothesis**.

H_o: *whether or not someone supports the death penalty is **independent of or NOT related to** whether they identify as male or female*

H_1: *whether or not someone supports the death penalty is **dependent upon or related to** whether they identify as male or female.*

2. **State level of significance or a "alpha."**

For this example we'll use alpha =.05

3. **Determine the test distribution to use –** Chi Square tests use X^2 distribution.

4. **Define the rejection regions. And draw a picture!**

df= (r-1) (c -1) (where r= # of rows in table and c= # of columns in the table). In this case df = (2-1) (2-1) = 1. Using Appendix 6 (or the chi square table linked in our course schedule) the critical value df=1 area in tail - .05 = 3.84

5. **State the decision rule.**

Reject the null if the TR > 3.84, otherwise FTR.

6. **Perform necessary calculations on data and compute TR value.**

   Recall in the take home tests you do NOT have to do the math by hand. In this step write, "SPSS says TR =______" and report the correct SPSS TR. Below I show you how the math is done by hand for illustrative purposes.

   $$TR = \sum \frac{(O - E)^2}{E}$$

   O = # of observed cases   E = # of expected cases

Again, note how the TR formula above looks *a bit* like a variance formula

$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$ = This (population) variance formulas sort of asks "on average how much to all of these x's differ from the mean?"

Well the TR formula above essentially asks "on average how much do our observations differ from what our theory expected?" If they differ a lot from the theory, the we will reject the theory. If they do not differ a lot from the theory, then the theory is probably correct, or at least we do not have sufficient evidence to reject the theory.
Large differences between expected and observed values will tend to have larger numerators in the fraction and thus larger TR values and thus "tend towards significance."
What happens when the number on the top of the fraction gets larger relative the number on the bottom of the fraction? The number represented by the fraction "gets bigger:"
1/10 2/10 3/10 5/10 10/10 20/10 note how the numbers get bigger as the top number of the fraction grows larger.

E = # of expected cases [E=(row total) (column total)/grand total]

_	Male	Female	total
Support (o)	26	3	29
(E)	19.3	9.7	
Oppose (0)	24	22	46
(E)	30.7	15.3	
column totals	50	25	75

row-col (cell)	Observed	Expected	O -E	(O - E)2	(O - E)2/E
1 1	26	19.3	6.7	44.44	2.299
1 2	3	9.7	-6.7	44.44	4.598
2 1	24	30.7	-6.7	44.44	1.449
2 2	22	15.3	6.7	44.44	2.899
				$\sum=$	**11.244**

TR= 11.244

7. **Compare TR value with the decision rule and make a statistical decision. (Write out decision in English! -- my addition)**

Since TR is greater than 3.84 we reject null and conclude alternative. In plain English we say *whether or not someone supports the death penalty is* ***dependent upon or related to*** *whether they identify as male or female.*

Again, TR value is essentially a ratio that computes a variance or "an average of the sum of the squared deviations from the expected value" where large differences in the numerator tend to make the fraction (or ratio) large. Large differences thus tend towards significance.

What is the wording when we fail to reject the null hypothesis?

When we fail to reject the null hypothesis we say "Insufficient evidence to reject theory that __________ " [insert Ho in plain English.] or

Insufficient evidence to reject theory that *whether or not someone supports the death penalty is* ***independent of or NOT related to*** *whether they identify as male or female*

Using the SPSS

Below is the output from SPSS for these data. To see how to make SPSS create this output see lecture 20c_SPSS.pdf. For the TR value you look at "Pearson Chi-Square" Row under "Value" and the p value is under the "Asymp. Sig. (2-sided)

1=male 2=female * Do you support the Death Penalty Crosstabulation

			Do you support the Death Penalty		
			support	oppose	Total
1=male 2=female	male	Count	26	3	29
		Expected Count	19.3	9.7	29.0
	female	Count	24	22	46
		Expected Count	30.7	15.3	46.0
Total		Count	50	25	75
		Expected Count	50.0	25.0	75.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	11.244[a]	1	.001		
Continuity Correction[b]	9.621	1	.002		
Likelihood Ratio	12.504	1	.000		
Fisher's Exact Test				.001	.001
Linear-by-Linear Association	11.094	1	.001		
N of Valid Cases	75				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 9.67.

b. Computed only for a 2x2 table

Table: Chi-Square Probabilities

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it fror

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169