

CHAPTER SIX

CHI SQUARE TESTING

Categorical Analysis

Most hypothesis tests involve one variable being categorical in nature so we can analyze subgroups (e.g., scores for men vs. scores for women, with GENDER being the categorical variable). In these parametric tests, the other variable must be a scale variable (i.e., interval or ratio) since the tests involve computing a mean. Yet there are many occasions in which we have only categorical variables, and the technique is quite simple. Its applications range from games of chance to analyzing surveys and polls.

Goodness of Fit

Let's start with the simplest application in which we have a categorical variable in which we expect all values to have equal occurrences. When we flip a coin, we expect to get 50% heads and 50% tails. When we roll a die, we expect to roll a one 1/6 of the time, and likewise for a two through six. With categorical data, there is no averaging; we merely count the frequencies and compute the percentages (i.e., relative frequencies). So if you flipped 12 heads and 8 tails, you would have 60% heads and 40% tails; the question is whether that is significantly different from the 50/50 you expected to get. In evaluating if it is a significant difference and if the coin flips do not fit with the distribution of a perfectly fair coin, we are essentially looking at two things – the percent breakdown of the sample and the size of the sample. So getting 75% heads is not significant if it is based on only 4 coin flips, and yet 55% heads would be significant if it were based on 10,000 flips. We see newspapers report poll results and tell us how a candidate has a lead, but that lead may be insignificant if the sample is not large enough.

The proper way to word our test is to state whether the results fit the expected distribution. So if we flipped a coin 100 times and got 59 heads and 41 tails, does that outcome fit with the distribution of a fair coin?

Ho: The coin flips fit the distribution of a fair coin.

Ha: The coin flips do not fit the distribution of a fair coin.

[illegible]

The math behind this analysis is actually quite simple. The statistics is called CHI SQUARE and the computation involves computing the differences between the observed and expected values, squaring those differences, and dividing by the expected value. We had 59 heads but out of 100 coin flips we would have expected 50. The difference between the observed 59 and the expected 50 is 9, and 9 squared is 81, and if we divide that 81 by the expected value of 50 we get 1.62, which is the value shown in the Chi Square column for Heads. We do the same for Tails and also get 1.62. Adding them up we get 3.24, our Chi Square test statistic. Statistics tables tell us the magic number above which the Chi Square statistic is considered to be significant. If our observed values exactly matched the expected values (i.e., we flipped 50 heads and 50 tails), the computed Chi Square statistic would be zero, which is as small as it gets. So the larger this value, the more that the results veer from the expected results. In this case we have a p-value of .0719. This means that if you took a perfectly fair coin and flipped it 100 times, there is a 7.19% probability of getting a 59/41 split. Since we set a significance level of 5%, that means that if these results could happen by chance more than 5% of the time, we don't draw conclusions about the coin being unfair (i.e., it could have come from a fair coin). We would not reject the null hypothesis in this case.

Had there been 60 heads and 40 tails, the p-value would drop to .0455, and there is less than a 5% chance of getting such a result by random chance. We would then reject the null hypothesis and conclude that the

Chi-Square Goodness of Fit Test (Assuming Equal Expected)			
Items	Observed	Expected	Chi Square
Heads	60	50.00	2.00
Tails	40	50.00	2.00
		-	-
		-	-
		-	-
		-	-
		-	-
		-	-
		-	-
		-	-

Data	
Level of Significance	0.05
Degrees of Freedom	1

Results	
Critical Value	3.8415
Chi-Square Test Statistic	4.00
p-Value	0.0455
Reject the null hypothesis	

results are not from the distribution of a fair coin (i.e., the coin is not fair).

Be careful not to jump to conclusions merely on the percentage of heads or tails. As stated earlier, it can be deceiving and depends on the sample size too.

Here you see 75% heads, and it is not significant since it is based on only 4 coin flips.

Chi-Square Goodness of Fit Test (Assuming Equal Expected)					
Items	Observed	Expected	Chi Square		
Hheads	3	2.00	0.50		
Tails	1	2.00	0.50		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		

Data	
Level of Significance	0.05
Degrees of Freedom	1

Results	
Critical Value	3.8415
Chi-Square Test Statistic	1.00
p-Value	0.3173
Do not reject the null hypothesis	

And here you see 51% heads which is significant because it is based on 10,000 flips. The closer the results are to equality, the larger the sample you need to reject the null, and vice versa. So for a small sample to have significant results, the deviation from equality must be large.

Chi-Square Goodness of Fit Test (Assuming Equal Expected)			
Items	Observed	Expected	Chi Square
Heads	5,100	5,000.00	2.00
Tails	4,900	5,000.00	2.00
	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-

Data	
Level of Significance	0.05
Degrees of Freedom	1

Results	
Critical Value	3.8415
Chi-Square Test Statistic	4.00
p-Value	0.0455
Reject the null hypothesis	

With only 10 coin flips,
we got significant results
because of the 90% heads
in the sample.

Chi-Square Goodness of Fit Test (Assuming Equal Expected)					
Items	Observed	Expected	Chi Square		
Heads	9	5.00	3.20		
Tails	1	5.00	3.20		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		
	-	-	-		

Data

Level of Significance	0.05
Degrees of Freedom	1

Results

Critical Value	3.8415
Chi-Square Test Statistic	6.40
p-Value	0.0114
Reject the null hypothesis	

Here you can see how this test can be applied to a 6-sided die to determine if the rolls follow a uniform distribution in which all six sides have an equal likelihood of occurrence.

Chi-Square Goodness of Fit Test (Assuming Equal Expected)			
Items	Observed	Expected	Chi Square
Roll a 1	13	20.00	2.45
Roll a 2	33	20.00	8.45
Roll a 3	14	20.00	1.80
Roll a 4	7	20.00	8.45
Roll a 5	36	20.00	12.80
Roll a 6	17	20.00	0.45
	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-
	-	-	-

Data	
Level of Significance	0.01
Degrees of Freedom	5

Results	
Critical Value	15.0863
Chi-Square Test Statistic	34.40
p-Value	0.0000
Reject the null hypothesis	

Now what if we don't expect an even spread of the values? Maybe we expect one value to occur more or less than the others (which is typically the case). We can still use the Chi Square Goodness of Fit test, but we just need to load the expected values manually. Let's start with another game of chance – Roulette. On a roulette wheel there are 38 numbers (1 – 36, 0 and 00). Numbers 1 through 36 are evenly split between red and black, while the 0 and 00 are both green. So the wheel has 18 red, 18 black and 2 green. When people bet on red or black, the payoff is even money (which is fine if the green numbers weren't there, but they are what gives the house its edge). 95% of the time, red or black wins and the house pays off, but 5% of the time, green wins and the house pretty much cleans the table. The casinos have computers that monitor every game in the house, and they essentially conduct Chi Square testing to determine if a pattern doesn't fit.

K
chi S
E
N
D
A
L
L
,

free
% m
the
nty
gre
ca
no
atin
5
us
in
iat

This Goodness of Fit technique is the method used to test if a sample comes from a Normal distribution; essentially we expect to see a certain percentage within 1 standard deviation, within 2 standard deviations, and within 3 standard deviations, and the frequencies are compared to those expected results.

Crosstabulations

We've seen how Chi Square testing can be used to test for the goodness of fit of a single categorical variable. Its other common use is to test for the relationship of two categorical variables.

With so many employee surveys and customer surveys conducted regularly, the temptation is to analyze them by merely looking at questions individually, but that tells us less than you might think. On employee surveys that use Likert scales, the questions on satisfaction with pay and benefits typically have the lowest scores in terms of the percent of employees who are satisfied or agree that their pay is fair. Does this mean that if an employer gives pay raises, the employee satisfaction will increase? Actually, no! But analyzing one dimensionally will often lead to such false conclusions. On employee surveys there is often a question regarding overall satisfaction, and if we analyze the relationship of the responses to that question against other questions, we can determine if one's satisfaction is independent of pay, benefits, working conditions, their boss, etc. And this can be accomplished easily with the Chi Square Test of Independence, which uses the same mathematics as the Chi Square Goodness of Fit Test but in a crosstabulation.

To illustrate, let's look at two examples likely to get very different results. If men and women were asked whether they enjoyed shopping for shoes, we could test whether one's enjoyment of shoe shopping was independent of his/her gender. Suppose 10 of the 100 men surveyed indicated that they liked shoe shopping, while 50 of the 200 women surveyed felt likewise.

	Likes Shoe Shopping	Does NOT Like Shoe Shopping	TOTAL
Men	10	90	100
Women	50	150	200
TOTAL	60	240	300

In total, 60 of the 300 people (20%) surveyed answered YES. So if one's response had nothing to do with gender and the two variables were independent, you would expect 20% of the men and 20% of the women to answer YES. So we would have expected 20 men to answer YES and 80 to answer NO, and 40 women to answer YES and 160 to answer NO. These expected values would be compared to the observed values in the table in computing the Chi Square statistic.

Here you can see how the Chi Square Test of Independence appears, and the table of Expected Frequencies reflects the computations just stated.⁴ The hypotheses in this case are:

Ho: One's enjoyment of shoe shopping is independent of one's gender.

Ha: One's enjoyment of shoe shopping is dependent on one's gender.

Adding up the Chi Square values for the 4 cells results in 9.375 which is relatively high for a crosstabulation with only two rows and two columns. The resulting p-value of .0022 is less than .05, and you would reject the null hypothesis and conclude a dependent relationship. In this case, conclude that women enjoy shoe shopping more than men. .

Chi-Square Test of Independence						
Observed Frequencies						
		Likes to Shop for Shoes				Total
Gender		YES	NO			
MEN		10	90			100
WOMEN		50	150			200
						0
						0
						0
Total		60	240	0	0	300
Expected Frequencies						
		Likes to Shop for Shoes				Total
Gender		YES	NO	0	0	
MEN		20.00	80.00	0.00	0.00	100
WOMEN		40.00	160.00	0.00	0.00	200
0		0.00	0.00	0.00	0.00	0
0		0.00	0.00	0.00	0.00	0
0		0.00	0.00	0.00	0.00	0
Total		60	240	0	0	300
Data						
Level of Significance	0.05					
Number of Rows	2					
Number of Columns	2					
Degrees of Freedom	1					
Results						
Critical Value	3.84146					
Chi-Square Test Statistic	9.375					
p-Value	0.0022					
Reject the null hypothesis						

Now suppose we asked the same 300 people whether they enjoyed watching football, and 80% of the men said YES while 70% of the women said YES.

Ho: One's enjoyment of watching football is independent of one's gender.

Ha: One's enjoyment of watching football is dependent on one's gender.

Chi-Square Test of Independence						
Observed Frequencies						
		Likes to Watch Football				Total
Gender		YES	NO			
MEN		80	20			100
WOMEN		140	60			200
						0
						0
						0
Total		220	80	0	0	300
Expected Frequencies						
		Likes to Watch Football				Total
Gender		YES	NO	0	0	
MEN		73.33	26.67	0.00	0.00	100
WOMEN		146.67	53.33	0.00	0.00	200
0		0.00	0.00	0.00	0.00	0
0		0.00	0.00	0.00	0.00	0
0		0.00	0.00	0.00	0.00	0
Total		220	80	0	0	300
Data						
Level of Significance	0.05					
Number of Rows	2					
Number of Columns	2					
Degrees of Freedom	1					
Results						
Critical Value	3.84146					
Chi-Square Test Statistic	3.40909					
p-Value	0.0648					
Do not reject the null hypothesis						

In this case the p-value is .0648 which is greater than .05, so the null hypothesis is not rejected. Thus we have insufficient evidence to conclude that one's enjoyment of football depends on one's gender. This doesn't mean that men and women enjoy it equally; it merely means that if you asked someone whether they enjoyed football, regardless of the response, you would not feel confident in guessing if that respondent were a man or woman.

Now onto the employee surveys. What often happens is that almost everyone responds negatively to the pay question, but the overall satisfaction question may or may not get high marks.

Chi-Square Test of Independence						
Observed Frequencies						
	Overall Satisfaction					
Satisfied with Pay	YES	NO				Total
YES	40	10				50
NO	720	230				950
						0
						0
						0
Total	760	240	0	0	0	1000
Expected Frequencies						
	Overall Satisfaction					
Satisfied with Pay	YES	NO	0	0	0	Total
YES	38.00	12.00	0.00	0.00	0.00	50
NO	722.00	228.00	0.00	0.00	0.00	950
0	0.00	0.00	0.00	0.00	0.00	0
0	0.00	0.00	0.00	0.00	0.00	0
0	0.00	0.00	0.00	0.00	0.00	0
Total	760	240	0	0	0	1000
Data						
Level of Significance	0.05					
Number of Rows	2					
Number of Columns	2					
Degrees of Freedom	1					
Results						
Critical Value	3.84146					
Chi-Square Test Statistic	0.46168					
p-Value	0.4968					
Do not reject the null hypothesis						

Here we see only 50 of the 1000 employees claim to be satisfied with their pay (5%) while 760 of the 1000 (76%) are satisfied with their jobs overall. This is very common, but can you really attribute pay as a contributing factor to dissatisfaction? If the two variables were truly independent, you would expect 76% of those satisfied with their pay to be satisfied with their jobs (38 out of 50) and 76% of those not satisfied with their pay to be satisfied with their jobs (722 of the 950). These expected numbers are fairly close to the observed numbers, and so the p-value of .4968 was much greater than .05, resulting in our not rejecting the null hypothesis. Thus we have insufficient evidence to conclude one's satisfaction with their job depends on their pay.

Chi-Square Test of Independence						
Observed Frequencies						
	Overall Satisfaction					
Treated with Respect	YES	NO				Total
YES	750	50				800
NO	10	190				200
						0
						0
						0
Total	760	240	0	0	0	1000
Expected Frequencies						
	Overall Satisfaction					
Treated with Respect	YES	NO	0	0	0	Total
YES	608.00	192.00	0.00	0.00	0.00	800
NO	152.00	48.00	0.00	0.00	0.00	200
0	0.00	0.00	0.00	0.00	0.00	0
0	0.00	0.00	0.00	0.00	0.00	0
0	0.00	0.00	0.00	0.00	0.00	0
Total	760	240	0	0	0	1000
Data						
Level of Significance	0.05					
Number of Rows	2					
Number of Columns	2					
Degrees of Freedom	1					
Results						
Critical Value	3.84146					
Chi-Square Test Statistic	690.927					
p-Value	0.0000					
Reject the null hypothesis						

Had we looked at questions regarding whether an employee is treated with respect or whether they feel valued or whether they like their boss, you would probably find a dependent relationship exist; those who feel they are treated with respect overwhelming are satisfied with their jobs and vice versa. To the right is an example of what you might see.

With a p-value of .0000, we would reject the null hypothesis and conclude that one's overall satisfaction depends on whether one is treated with respect. This technique can be applied to every pair of variables or at least a key indicator vs. each of the demographic variables and survey questions. Once again we see how a simple tool can be used to turn simple nominal data into useful information.

CHAPTER SIX KNOWLEDGE ASSESSMENT

CHI SQUARE TESTING

Discussion Questions

DISCUSSION QUESTION 1

PUTTING YOU ON: Many alleged studies are misleading in that the conclusions are often based on undisclosed premises. Parenting books state that an advantage to breastfeeding is that it makes women healthier; what is not stated is that women who breastfeed tend to follow stricter diets, and it is the diets that actually makes them healthier. If you looked at statistics for the number of tires replaced in a given year, would you be surprised to find out that Toyota Camry tops the list for wearing out tires? It shouldn't since it is the most popular car in America. And in 2000, Sen. Daschle (D-South Dakota) pushed his gun control agenda against Pres. Bush when he stated "Do you realize that the majority of vendors at gun shows don't do background checks on their customers;" while true, what he failed to mention is that the majority of vendors at gun shows don't sell guns (they sell food, t-shirts, souvenirs, etc.), but not guns. Nonetheless, this quasi-truth aroused the voters. Name another conclusion that, while true, is based on misconception. There are multitudes of them in regards to health and politics, so keep your eyes open. If you cannot find one, then make one up (use your creativity and silliness if you wish).

DISCUSSION QUESTION 2

TRUTH IN ADVERTISING: Infomercials often peddle products under the guise of "studies show...". While some of the products are surprisingly good (e.g., FoodSaver, Foreman Grill, Ronco Rotisserie), many of the products are not. Why should you be initially suspicious of any products advertised on these shows, despite the "existence" of data? What factors would help to convince you that the claims might be true?

Practice Problems: Real Estate

Solutions are provided to practice problems so you can check your work.

Use the Real_Estate.xls file which consists of 100 homes purchased in 2007 and appraised in 2008. It includes variables regarding the number of bedrooms, number of bathrooms, whether the house has a pool or garage, the age, size and price of the home, what the house is constructed from, how far it is to the city center, and the appraisals from two agents.

PRACTICE PROBLEM 1:

It has been noted that swimming pools do nothing to increase value of a home, and yet, understandably, many homes have them. Whether a homeowner puts in a pool depends largely on the size of their backyard, whether the subdivision has a community pool, and the general climate in their location. Using the sample of 100 homes (in the data file), conduct a Chi Square Goodness of Fit test to determine if there is an equal split in those homeowners who have or do not have a swimming pool installed. Use a .05 significance level.

CHAPTER SIX KNOWLEDGE ASSESSMENT

CHI SQUARE TESTING

PRACTICE PROBLEM 2:

Brick is better for maintaining the value of a home, no doubt, but it costs more up front and can prove a barrier to prospective buyers. If the historical data for your city showed 40% of the homes sold in the USA in the last 25 years were made of wood, 30% of brick and 30% of stucco, how does the sample of homes in your city compare (i.e., the 100 homes in the data file)? Using the sample of 100 homes (in the data file), conduct a Chi Square Goodness of Fit test to determine if the homes purchased recently fit the distribution of homes in the USA in terms of construction material. Use a .05 significance level.

PRACTICE PROBLEM 3:

A swimming pool is considered a luxury, especially since it doesn't add value to a home in most cases. And brick homes cost more to build than those of wood or stucco. So if you follow the money, might you see that those with brick homes are more inclined to spend the money on a swimming pool? Or perhaps those who saved money by buying a wood home used their savings to get a pool? Using the sample of 100 homes (in the data file), conduct a Chi Square Test of Independence to determine if owning a swimming pool is independent of the construction of one's home. Use a .05 significance level.

Assigned Problems: Student Data

Use the Student_Data.xls file which consists of 200 MBA at Whatsamattu U. It includes variables regarding their age, gender, major, GPA, Bachelors GPA, course load, English speaking status, family, weekly hours spent studying.

ASSIGNED PROBLEM 1:

Historically, the MBA program at Whatsamattu U. has about 40% of their students choose a Leadership major, 30% choose a Finance major, 20% choose a Marketing major, and 10% choose no major. Does the most recent class of 200 MBA students fit that same pattern or has there been a shift in the choice of majors. Using the sample of 200 students (in the data file), conduct a Chi Square Goodness of Fit test to determine if the current distribution fits the historical pattern. Use a .05 significance level.

ASSIGNED PROBLEM 2:

While job opportunities for men and women are considerably more balanced than they were 40 years ago, the career aspirations may still differ. Is there a difference in majors chosen by men and women? Using the sample of 200 MBA students (in the data file), conduct a Chi Square Test of Independence to determine if one's choice of major is independent of their gender. Use a .05 significance level.