11.1: Chi-Square Test for Independence

Remember, qualitative data is where you collect data on individuals that are categories or names. Then you would count how many of the individuals had particular qualities. An example is that there is a theory that there is a relationship between breastfeeding and autism. To determine if there is a relationship, researchers could collect the time period that a mother breastfed her child and if that child was diagnosed with autism. Then you would have a table containing this information. Now you want to know if each cell is independent of each other cell. Remember, independence says that one event does not affect another event. Here it means that having autism is independent of being breastfed. What you really want is to see if they are not independent. In other words, does one affect the other? If you were to do a hypothesis test, this is your alternative hypothesis and the null hypothesis is that they are independent. There is a hypothesis test for this and it is called the **Chi-Square Test for Independence**. Technically it should be called the Chi-Square Test for Dependence, but for historical reasons it is known as the test for independence. Just as with previous hypothesis tests, all the steps are the same except for the assumptions and the test statistic.

Hypothesis Test for Chi-Square Test

1. State the null and alternative hypotheses and the level of significance
   H_o: the two variables are independent (this means that the one variable is not affected by the other)
   H_A: the two variables are dependent (this means that the one variable is affected by the other)
   Also, state your α level here.
2. State and check the assumptions for the hypothesis test
   a. A random sample is taken.
   b. Expected frequencies for each cell are greater than or equal to 5 (The expected frequencies, E, will be calculated later, and this assumption means $E \geq 5$).
3. Find the test statistic and p-value
   Finding the test statistic involves several steps. First the data is collected and counted, and then it is organized into a table (in a table each entry is called a cell). These values are known as the observed frequencies, which the symbol for an observed frequency is *O*. Each table is made up of rows and columns. Then each row is totaled to give a row total and each column is totaled to give a column total.
   The null hypothesis is that the variables are independent. Using the multiplication rule for independent events you can calculate the probability of being one value of the first variable, *A*, and one value of the second variable, *B* (the probability of a particular cell $P(A \text{ and } B)$). Remember in a hypothesis test, you assume that H_o is true, the two variables are assumed to be independent.

$$\begin{aligned} P(A \text{ and } B) &= P(A) \cdot P(B) \text{ if } A \text{ and } Bare independent \\ &= \frac{\text{number of ways } A \text{ can happen}}{\text{total number of individuals}} \cdot \frac{\text{number of ways } B \text{ can happen}}{\text{total number of individuals}} \\ &= \frac{\text{row total}}{n} * \frac{\text{column total}}{n} \end{aligned}$$

Now you want to find out how many individuals you expect to be in a certain cell. To find the expected frequencies, you just need to multiply the probability of that cell times the total number of individuals. Do not round the expected frequencies.

$$\begin{aligned} \text{Expected frequency } (\text{cell } A \text{ and } B) &= E(A \text{ and } B) \\ &= n\left(\frac{\text{row total}}{n} \cdot \frac{\text{column total}}{n}\right) \\ &= \frac{\text{row total -column total}}{n} \end{aligned}$$

If the variables are independent the expected frequencies and the observed frequencies should be the same. The test statistic here will involve looking at the difference between the expected frequency and the observed frequency for each cell. Then you want to find the "total difference" of all of these differences. The larger the total, the smaller the chances that you could find that test statistic given that the assumption of independence is true. That means that the assumption of independence is not true. How do you find the test statistic? First find the differences between the observed and expected frequencies. Because some of these differences will be positive and some will be negative, you need to square these differences. These squares could be large just

because the frequencies are large, you need to divide by the expected frequencies to scale them. Then finally add up all of these fractional values. This is the test statistic.

Test Statistic:

The symbol for Chi-Square is χ^2

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

where *O* is the observed frequency and *E* is the expected frequency

Distribution of Chi-Square

χ^2 has different curves depending on the degrees of freedom. It is skewed to the right for small degrees of freedom and gets more symmetric as the degrees of freedom increases (see *Figure* 11.1.1). Since the test statistic involves squaring the differences, the test statistics are all positive. A chi-squared test for independence is always right tailed.

Figure 11.1.1: Chi-Square Distribution

p-value:

Using the TI-83/84: χ cdf (lower limit, 1E99, df)

Using R: $1 -$ pchisq (x^2, df)

Where the degrees of freedom is $df = (\#\text{ of rows } - 1) * (\#\text{ of columns } - 1)$

4. Conclusion

This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value < α, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o.

5. Interpretation

This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Example 11.1.1 hypothesis test with chi-square test using formula

Is there a relationship between autism and breastfeeding? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what time period they breastfed their children. The data is in table #11.1.1 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to show that that breastfeeding and autism are independent? Test at the1% level.

Table 11.1.1: Autism Versus Breastfeeding

Autis261m	Breast Feeding Timelines				Row Total
	None	Less than 2 months	2 to 6 months	More than 6 months	
Yes	241	198	164	215	818
No	20	25	27	44	116
Column Total	261	223	191	259	934

Solution

1. State the null and alternative hypotheses and the level of significance

H_o: Breastfeeding and autism are independent

H_A: Breastfeeding and autism are dependent

α = 0.01

2. State and check the assumptions for the hypothesis test

a. A random sample of breastfeeding time frames and autism incidence was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. All expected frequencies are more than 5.

3. Find the test statistic and p-value

Test statistic:

First find the expected frequencies for each cell

$$E(\text{ Autism and no breastfeeding }) = \frac{818^{*}261}{934} \approx 228.585$$

$$E(\text{ Autism and } < 2 \text{ months }) = \frac{818^{*}223}{934} \approx 195.304$$

$$E(\text{ Autism and 2 to 6 months }) = \frac{818^{*}191}{934} \approx 167.278$$

$$E(\text{ Autism and more than 6 months }) = \frac{818 * 259}{934} \approx 226.833$$

Others are done similarly. It is easier to do the calculations for the test statistic with a table, the others are in table #11.1.2 along with the calculation for the test statistic. (Note: the column of *O-E* should add to 0 or close to 0.)

Table 11.1.2: Calculations for Chi-Square Test Statistic

O	*E*	*O-E*	$(O-E)^2$	$(O-E)^2/E$
241	228.585	12.415	154.132225	0.674288448
198	195.304	2.696	7.268416	0.03721591
164	167.278	-3.278	10.745284	0.064236086
215	226.833	-11.833	140.019889	0.617281828
20	32.4154	-12.4154	154.1421572	4.755213792
25	27.6959	-2.6959	7.26787681	0.262417066
27	23.7216	3.2784	10.74790656	0.453085229
44	32.167	11.833	140.019889	4.352904809
Total		0.0001		11.2166432 = χ^2

The test statistic formula is $\chi^2 = \sum \frac{(O-E)^2}{E}$, which is the total of the last column in Example 11.1.2

p-value:

$df = (2-1)^{*}(4-1) = 3$

Using TI-83/84: χcdf$(11.2166432, 1\text{E}99, 3) \approx 0.01061$

Using R: $1 - \text{pchisq}(11.2166432, 3) \approx 0.01061566$

4. Conclusion

Fail to reject H_o since the p-value is more than 0.01.

5. Interpretation

There is not enough evidence to show that breastfeeding and autism are dependent. This means that you cannot say that the whether a child is breastfed or not will indicate if that the child will be diagnosed with autism.

Example 11.1.2 hypothesis test with chi-square test using technology

Is there a relationship between autism and breastfeeding? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what time period they breastfed their children. The data is in Example 11.1.1 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to show that that breastfeeding and autism are independent? Test at the1% level.

Solution

1. State the null and alternative hypotheses and the level of significance

H_o: Breastfeeding and autism are independent

H_A: Breastfeeding and autism are dependent

α = 0.01

2. State and check the assumptions for the hypothesis test

a. A random sample of breastfeeding time frames and autism incidence was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. All expected frequencies are more than 5.

3. Find the test statistic and p-value

Test statistic:

To use the TI-83/84 calculator to compute the test statistic, you must first put the data into the calculator. However, this process is different than for other hypothesis tests. You need to put the data in as a matrix instead of in the list. Go into the MATRX menu then move over to EDIT and choose 1:[A]. This will allow you to type the table into the calculator. *Figure* 11.1.2 shows what you will see on your calculator when you choose 1:[A] from the EDIT menu.

Figure 11.1.2: Matrix Edit Menu on TI-83/84

The table has 2 rows and 4 columns (don't include the row total column and the column total row in your count). You need to tell the calculator that you have a 2 by 4. The 1 X1 (you might have another size in your matrix, but it doesn't matter because you will change it) on the calculator is the size of the matrix. So type 2 ENTER and 4 ENTER and the calculator will make a matrix of the correct size. See *Figure* 11.1.3

Figure 11.1.3: Matrix Setup for Table

Now type the table in by pressing ENTER after each cell value. *Figure* 11.1.4 contains the complete table typed in. Once you have the data in, press QUIT.

Figure 11.1.4: Data Typed into Matrix

To run the test on the calculator, go into STAT, then move over to TEST and choose χ^2-Test from the list. The setup for the test is in *Figure* 11.1.5

Figure 11.1.5: Setup for Chi-Square Test on TI-83/84

Once you press ENTER on Calculate you will see the results in *Figure* 11.1.6

Figure 11.1.6: Results for Chi-Square Test on TI-83/84

The test statistic is $\chi^2 \approx 11.2167$ and the p-value is $p \approx 0.01061$. Notice that the calculator calculates the expected values for you and places them in matrix B. To eview the expected values, go into MATRX and choose 2:[B]. *Figure* 11.1.7 shows the output. Press the right arrows to see the entire matrix.

Figure 11.1.7: Expected Frequency for Chi-Square Test on TI-83/84

To compute the test statistic and p-value with R,
row1 = c(data from row 1 separated by commas)
row2 = c(data from row 2 separated by commas)
keep going until you have all of your rows typed in.
data.table = rbind(row1, row2, …) – makes the data into a table. You can call it what ever you want. It does not have to be

data.table.
data.table – use if you want to look at the table
chisq.test(data.table) – calculates the chi-squared test for independence
chisq.test(data.table)$expected – let's you see the expected values

For this example, the commands would be
row1 = c(241, 198, 164, 215)
row2 = c(20, 25, 27, 44)
data.table = rbind(row1, row2)
data.table

Output:
[,1] [,2] [,3] [,4]
row1 241 198 164 215
row2 20 25 27 44

chisq.test(data.table)

Output:
Pearson's Chi-squared test

data: data.table
X-squared = 11.217, df = 3, p-value = 0.01061

chisq.test(data.table)$expected

Output: [,1] [,2] [,3] [,4]
row1 228.58458 195.30407 167.27837 226.83298
row2 32.41542 27.69593 23.72163 32.16702

The test statistic is $\chi^2 \approx 11.217$ and the p-value is $p \approx 0.01061$.

4. Conclusion

Fail to reject H_o since the p-value is more than 0.01.

5. Interpretation

There is not enough evidence to show that breastfeeding and autism are dependent. This means that you cannot say that the whether a child is breastfed or not will indicate if that the child will be diagnosed with autism.

Example 11.1.3 hypothesis test with chi-square test using formula

The World Health Organization (WHO) keeps track of how many incidents of leprosy there are in the world. Using the WHO regions and the World Banks income groups, one can ask if an income level and a WHO region are dependent on each other in terms of predicting where the disease is. Data on leprosy cases in different countries was collected for the year 2011 and a summary is presented in *Table* 11.1.3 ("Leprosy: Number of," 2013). Is there evidence to show that income level and WHO region are independent when dealing with the disease of leprosy? Test at the 5% level.

Table 11.1.3: Number of Leprosy Cases

WHO Region	World Bank Income Group				Row Total
	High Income	Upper Middle Income	Lower Middle Income	Low Income	
Americas	174	36028	615	0	36817
Eastern Mediterranean	54	6	1883	604	2547
Europe	10	0	0	0	10
Western Pacific	26	216	3689	1155	5086

Africa	0	39	1986	15928	17953
South-East Asia	0	0	149896	10236	160132
Column Total	264	36289	158069	27923	222545

Solution

1. State the null and alternative hypotheses and the level of significance

H_o: WHO region and Income Level when dealing with the disease of leprosy are independent

H_A: WHO region and Income Level when dealing with the disease of leprosy are dependent

α = 0.05

2. State and check the assumptions for the hypothesis test

a. A random sample of incidence of leprosy was taken from different countries and the income level and WHO region was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. There are actually 4 expected frequencies that are less than 5, and the results of the test may not be valid. If you look at the expected frequencies you will notice that they are all in Europe. This is because Europe didn't have many cases in 2011.

3. Find the test statistic and p-value

Test statistic:

First find the expected frequencies for each cell.

$$E(\text{ Americas and High Income }) = \frac{36817 * 264}{222545} \approx 43.675$$

$$E(\text{ Americas and Upper Middle Income }) = \frac{36817 * 36289}{222545} \approx 6003.514$$

$$E(\text{ Americas and Lower Middle Income}) = \frac{36817 * 158069}{222545} \approx 26150.335$$

$$E(\text{ Americas and Lower Income }) = \frac{36817 * 27923}{222545} \approx 4619.475$$

Others are done similarly. It is easier to do the calculations for the test statistic with a table, and the others are in Example 11.1.4along with the calculation for the test statistic.

Table 11.1.4: Calculations for Chi-Square Test Statistic

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
174	43.675	130.325	16984.564	388.8838719
54	3.021	50.979	2598.813	860.1218328
10	0.012	9.988	99.763	8409.746711
26	6.033	19.967	398.665	66.07628214
0	21.297	-21.297	453.572	21.29722977
0	189.961	-189.961	36085.143	189.9608978
36028	6003.514	30024.486	901469735.315	150157.0038
6	415.323	-409.323	167545.414	403.4097962
0	1.631	-1.631	2.659	1.6306365
216	829.342	-613.342	376188.071	453.5983897
39	2927.482	-2888.482	8343326.585	2850.001268

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
0	26111.708	-26111.708	681821316.065	26111.70841
615	26150.335	-25535.335	652053349.724	24934.7988
1883	1809.080	73.290	5464.144	3.020398811
0	7.103	-7.103	50.450	7.1027882
3689	3612.478	76.522	5855.604	1.620938405
1986	12751.636	-10765.636	115898911.071	9088.944681
149896	113738.368	36157.632	1307374351.380	11494.57632
0	4619.475	-4619.475	21339550.402	4619.475122
604	319.575	284.425	80897.421	253.1404187
0	1.255	-1.255	1.574	1.25471253
1155	638.147	516.853	267137.238	418.6140882
15928	2252.585	13675.415	187016964.340	83023.25138
10236	20091.963	-9855.963	97140000.472	4834.769106
Total		0.000		328594.008 = χ^2

The test statistic formula is $\chi^2 = \sum \frac{(O-E)^2}{E}$, which is the total of the last column in Example 11.1.2

p-value:

$df = (6-1)*(4-1) = 15$

Using the TI-83/84: χcdf$(328594.008, 1E99, 15) \approx 0$

Using R: $1 -$ pchisq $(328594.008, 15) \approx 0$

4. Conclusion

Reject H_o since the p-value is less than 0.05.

5. Interpretation

There is enough evidence to show that WHO region and income level are dependent when dealing with the disease of leprosy. WHO can decide how to focus their efforts based on region and income level. Do remember though that the results may not be valid due to the expected frequencies not all be more than 5.

Example 11.1.4 hypothesis test with chi-square test using technology

The World Health Organization (WHO) keeps track of how many incidents of leprosy there are in the world. Using the WHO regions and the World Banks income groups, one can ask if an income level and a WHO region are dependent on each other in terms of predicting where the disease is. Data on leprosy cases in different countries was collected for the year 2011 and a summary is presented in *Table* 11.1.3 ("Leprosy: Number of," 2013). Is there evidence to show that income level and WHO region are independent when dealing with the disease of leprosy? Test at the 5% level.

Solution

1. State the null and alternative hypotheses and the level of significance
H_o: WHO region and Income Level when dealing with the disease of leprosy are independent

H_A: WHO region and Income Level when dealing with the disease of leprosy are dependent

α = 0.05

2. State and check the assumptions for the hypothesis test

a. A random sample of incidence of leprosy was taken from different countries and the income level and WHO region was taken.
b. Expected frequencies for each cell are greater than or equal to 5 (ie. $E \geq 5$). See step 3. There are actually 4 expected frequencies that are less than 5, and the results of the test may not be valid. If you look at the expected frequencies you will notice that they are all in Europe. This is because Europe didn't have many cases in 2011.

3. Find the test statistic and p-value
Test statistic:
Using the TI-83/84. See Example 11.1.2 for the process of doing the test on the calculator. Remember, you need to put the data in as a matrix instead of in the list.

Figure 11.1.8: Setup for Matrix on TI-83/84

Figure 11.1.9: Results for Chi-Square Test on TI-83/84

$$\chi^2 \approx 328594.0079$$

Figure 11.1.10: Expected Frequency for Chi-Square Test on TI-83/84

Press the right arrow to look at the other expected frequencies.

p-value:

$p - \text{value} \approx 0$

Using R:

```
row1=c(174, 36028, 615, 0)
row2=c(54, 6, 1883, 604)
row3=c(10, 0, 0, 0)
row4=c(26, 216, 3689, 1155)
row5=c(0, 39, 1986, 15928)
row6=c(0, 0, 149896, 10236)
chisq.test(data.table)
```

Pearson's Chi-squared test

data: data.table
X-squared = 328590, df = 15, p-value < 2.2e-16

Warning message:
In chisq.test(data.table) : Chi-squared approximation may be incorrect

chisq.test(data.table)$expected

	[,1]	[,2]	[,3]	[,4]
row1	43.67515783	6003.514404	$2.615034e+04$	4619.475122
row2	3.02144735	415.323117	$1.809080e+03$	319.575281
row3	0.01186277	1.630637	$7.102788e+00$	1.254713
row4	6.03340448	829.341724	$3.612478e+03$	638.146793
row5	21.29722977	2927.481709	$1.275164e+04$	2252.585405
row6	189.96089780	26111.708410	$1.137384e+05$	20091.962686

Warning message:
In chisq.test(data.table) : Chi-squared approximation may be incorrect

$\chi^2 = 328590$ and p-value = 2.2×10^{-16}

4. Conclusion

Reject H_o since the p-value is less than 0.05.

5. Interpretation

There is enough evidence to show that WHO region and income level are dependent when dealing with the disease of leprosy. WHO can decide how to focus their efforts based on region and income level. Do remember though that the results may not be valid due to the expected frequencies not all be more than 5.

Homework

Exercise 11.1.1

In each problem show all steps of the hypothesis test. If some of the assumptions are not met, note that the results of the test may not be correct and then continue the process of the hypothesis test.

1. The number of people who survived the Titanic based on class and sex is in Example 11.1.5 ("Encyclopedia Titanica," 2013). Is there enough evidence to show that the class and the sex of a person who survived the Titanic are independent? Test at the 5% level.

Table 11.1.5: Surviving the Titanic

Class	Sex		Total
	Female	Male	
1st	134	59	193
2nd	94	25	119
3rd	80	58	138
Total	308	142	450

2. Researchers watched groups of dolphins off the coast of Ireland in 1998 to determine what activities the dolphins partake in at certain times of the day ("Activities of dolphin," 2013). The numbers in Example 11.1.6 represent the number of groups of dolphins that were partaking in an activity at certain times of days. Is there enough evidence to show that the activity and the time period are independent for dolphins? Test at the 1% level.

Table 11.1.6: Dolphin Activity

Activity	Period				Row Total
	Morning	Noon	Afternoon	Evening	
Travel	6	6	14	13	39
Feed	28	4	0	56	88
Social	38	5	9	10	62
Column Total	72	15	23	79	189

3. Is there a relationship between autism and what an infant is fed? To determine if there is, a researcher asked mothers of autistic and non-autistic children to say what they fed their infant. The data is in Example 11.1.7 (Schultz, Klonoff-Cohen, Wingard, Askhoomoff, Macera, Ji & Bacher, 2006). Do the data provide enough evidence to show that that what an infant is fed and autism are independent? Test at the 1% level.

Table 11.1.7: Autism Versus Breastfeeding

Autism	Feeding			Row Total
	Breast feeding	Formula with DHA/ARA	Formula without DRA/ARA	
Yes	12	39	65	116
No	6	22	10	38

Column Total	18	61	75	164

4. A person's educational attainment and age group was collected by the U.S. Census Bureau in 1984 to see if age group and educational attainment are related. The counts in thousands are in Example 11.1.8("Education by age," 2013). Do the data show that educational attainment and age are independent? Test at the 5% level.

Table 11.1.8: Educational Attainment and Age Group

Education	Age Group					Row Total
	25-34	35-44	45-54	55-64	>64	
Did not complete HS	5416	5030	5777	7606	13746	37575
Completed HS	16431	1855	9435	8795	7558	44074
College 1-3 years	8555	5576	3124	2524	2503	22282
College 4 or more years	9771	7596	3904	3109	2483	26863
Column Total	40173	20057	22240	22034	26290	130794

5. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important good grades were to them (1 very important and 4 least important). The data is in Example 11.1.9 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of grades are independent? Test at the 5% level.

Table 11.1.9: Personal Goal and Importance of Grades

Goal	Grades Importance Rating				Row Total
	1	2	3	4	
Grades	70	66	55	56	247
Popular	14	33	45	49	141
Sports	10	24	33	23	90
Column Total	94	123	133	128	478

6. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important being good at sports were to them (1 very important and 4 least important). The data is in Example 11.1.10("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of sports are independent? Test at the 5% level.

Table 11.1.10: Personal Goal and Importance of Sports

Goal	Sports Importance Rating				Row Total
	1	2	3	4	
Grades	83	81	55	28	247
Popular	32	49	43	17	141
Sports	50	24	14	2	90
Column Total	165	154	112	47	478

7. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important having good looks were to them (1 very important and 4 least important). The data is in Example 11.1.11("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of looks are independent? Test at the 5% level.

Table 11.1.11: Personal Goal and Importance of Looks

Goal	Looks Importance Rating				Row Total
	1	2	3	4	
Grades	80	66	66	35	247
Popular	81	30	18	12	141
Sports	24	30	17	19	90
Column Total	185	126	101	66	478

8. Students at multiple grade schools were asked what their personal goal (get good grades, be popular, be good at sports) was and how important having money were to them (1 very important and 4 least important). The data is in Example 11.1.12 ("Popular kids datafile," 2013). Do the data provide enough evidence to show that goal attainment and importance of money are independent? Test at the 5% level.

Table 11.1.12: Personal Goal and Importance of Money

Goal	Money Importance Rating				Row Total
	1	2	3	4	
Grades	14	34	71	128	247
Popular	14	29	35	63	141
Sports	6	12	26	46	90
Column Total	34	75	132	237	478

Answer

For all hypothesis tests, just the conclusion is given. See solutions for the entire answer.

1. Reject Ho
3. Reject Ho
5. Reject Ho
7. Reject Ho