Unit 30
Chi-Square Test Concerning Independence

Objectives:

- To perform the chi-square hypothesis test concerning independence in a contingency table
- To describe the relationship between two categorical variables in a contingency table

While the chi-square goodness-of-fit test focuses on categories defined by one qualitative variable, the chi-test concerning independence in a contingency table focuses on the relationship between two qualitative variables. For instance, consider Table 30-1, which displays the raw frequencies when a simple random sample of voters in a city are classified according to "Sex" and "Opinion on a Smoking Bill." If these two variables were independent, then we would expect the relative frequencies corresponding to the three opinion categories not to change for the two sexes; alternatively, we could say that we would expect the relative frequencies corresponding to the two sexes not to change across the three opinion categories.

Table 30-1
Contingency Table for
"Sex" and "Opinion on Bill"

	Opinion on Smoking Bill			
Sex	*Against*	*Neutral*	*Favor*	
Male	104	80	152	336
Female	88	48	168	304
	192	128	320	640

The chi-square test concerning independence can be used to see if there is any evidence of a relationship between the two variables in a contingency table. The null hypothesis states that the two variables are independent, and the alternative hypothesis states that there is a relationship between the two variables. Once the observed frequencies (O) and the expected frequencies (E) have been obtained, the formula for the test statistic is exactly the same as in the chi-square goodness-of-fit test. Recall that the degrees of freedom for the chi-square goodness-of-fit test statistic is one less than the number of categories for the quantitative variable. There are two quantitative variables involved in the calculation of the chi-square test statistic concerning independence, and the degrees of freedom for this test statistic is the product of one less than the number of categories for one quantitative variable and one less than the number of categories for the other quantitative variable. If we let r represent the number of categories for the row variable in the contingency table, and let c represent the number of categories for the column variable, then the degrees of freedom for the chi-square test statistic concerning independence is $(r-1)(c-1)$. We may write the chi-square test statistic concerning independence as

$$\chi^2_{(r-1)(c-1)} = \sum \frac{(O-E)^2}{E} \quad .$$

With a chi-square goodness-of-fit test statistic, we have already seen how to obtain the expected frequencies (E) by multiplying the sample size by each of the hypothesized proportions. With a chi-square test statistic concerning independence, the expected frequencies (E) can be obtained from

$$E = \frac{(row\ total)(column\ total)}{n}$$

Go ahead and multiply the first row total in Table 30-1 (336) by the first column total (192), and divide this product by the sample size (n = 640) to obtain the expected frequency (100.8) for the corresponding cell; then check that this expected frequency (100.8) is displayed in the appropriate place in Table 30-2. You can then verify that all the other expected frequencies in Table 30-2 are correct. Remember that an expected frequency is

to be interpreted as the average frequency we would expect to occur when repeatedly taking random samples, if the null hypothesis is true. Consequently, expected frequencies are not always integers (as is the case in Table 30-2), even though it is only possible to observe integer frequencies.

Table 30-2
Expected Frequencies for Table 30-2

	Opinion on Smoking Bill			
Sex	*Against*	*Neutral*	*Favor*	
Male	100.8	67.2	168.0	336
Female	91.2	60.8	152.0	304
	192	128	320	640

Obtaining all the expected frequencies in a contingency table can be a bit time consuming, but then also having to substitute these expected frequencies into the formula for the chi-square test statistic can become very tedious. Of course, access to the appropriate statistical software or programmable calculator allows one to avoid much of this calculation.

To illustrate the chi-square test concerning independence, let us choose a 0.05 significance level to see if there is any evidence of a relationship between sex of the voter and opinion on the smoking bill, using the data of Table 30-1. Note that there are two methods by which the data of Table 30-1 could have been obtained. One might first select one simple random sample of voters, and then classify each of them by sex and opinion on the bill. Alternatively, one might first select two independent simple random samples of voters, one sample of males and one sample of females; then, each selected voter is classified according to opinion on the bill. In either case, the data will be appropriate for a chi-square test concerning independence.

The same four steps we have used in the past to perform hypothesis tests are used in a chi-square test concerning independence. We can complete the first step in our hypothesis test as follows:

H_0: Sex of the voter and opinion on the smoking bill are independent
vs. ($\alpha = 0.05$)
H_1: There is a relationship between sex of the voter and opinion on the smoking bill

Since this hypothesis test concerns two qualitative variables in a contingency table, there is no abbreviated way to write the hypotheses as there has been for the previous hypothesis tests we have discussed. Consequently, we have written the hypotheses as complete sentences. However, there are some equivalent ways that we could state the hypotheses. First of all, saying that two variables are independent is the same thing as saying the two variables are not related or not associated, and saying that there is a relationship between two variables is the same thing as saying that there is a dependency or association between the two variables.

Also, recall that if two qualitative variables in a contingency table are independent, then the distribution of column categories is the same for each row category, and the distribution of row categories is the same for each column category. With this in mind, we could choose to write the hypotheses as follows:

H_0: The distribution of opinions on the smoking bill is the same for males and females
vs. ($\alpha = 0.05$)
H_1: The distribution of opinions on the smoking bill is not the same for males and females

We could alternatively choose to write the hypotheses as follows:

H_0: The distribution of sexes is the same for each of the opinions on the smoking bill
vs. ($\alpha = 0.05$)
H_1: The distribution of sexes is not the same for each of the opinions on the smoking bill

In practice, the way that hypotheses are stated in a chi-square test concerning independence often depends on the personal preference of the researcher(s) and/or on how the data were collected.

The second step is to collect data and calculate the value of the test statistic. Tables 30-1 and 30-2 provide us with the observed and expected frequencies. The sample size necessary for the chi-square test concerning independence, like with the chi-square goodness-of-fit test, is sufficiently large only if each expected frequency (E) is greater than or equal to 5. If one or more of the expected frequencies is less than 5, categories

can be combined in a way to insure that $E \geq 5$ for each cell of the contingency table. Note that each of the expected frequencies in Table 30-2 is considerably greater than 5.

Having verified that the sample size is sufficiently large for the chi-square test concerning independence, we are now ready to calculate the test statistic. Since our contingency table has $r = 2$ rows and $c = 3$ columns, the degrees of freedom for the test statistic is $(r-1)(c-1) = (2-1)(3-1) = 2$. Substituting the observed and expected frequencies into the test statistic formula, we have that

$$\chi^2_2 = \frac{(104-100.8)^2}{100.8} + \frac{(80-67.2)^2}{67.2} + \frac{(152-168.0)^2}{168.0} + \frac{(88-91.2)^2}{91.2} + \frac{(48-60.8)^2}{60.8} + \frac{(168-152.0)^2}{152.0} = 8.555.$$

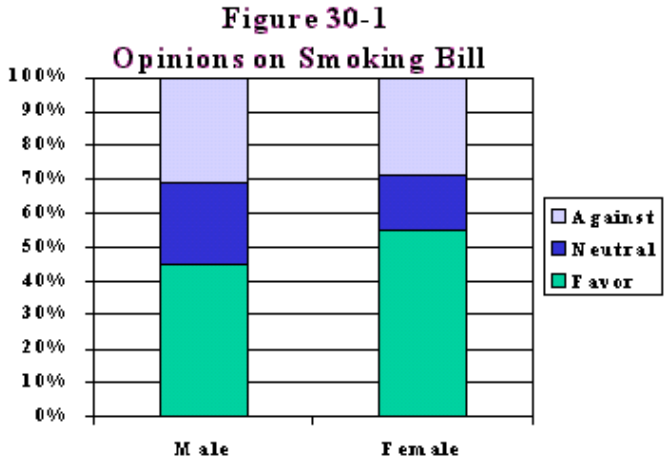
You can verify that the observed and expected frequencies have been correctly entered into the formula and then do this calculation to verify the final answer. Of course, you may choose to do this calculation using appropriate statistical software or a programmable calculator.

The third step is to define the rejection region, decide whether or not to reject the null hypothesis, and obtain the *p*-value of the test. The shaded area in the figure at the top of the first page of Table A.5 graphically illustrates our rejection region, which is defined by the χ^2-score above which lies 0.05 of the area under the density curve for a χ^2 distribution with $df = 2$. From Table A.5, we find that $\chi^2_{2;\,0.05} = 5.991$, and we can then define the rejection region algebraically as follows:

$$\chi^2_2 \geq 5.991 .$$

Since our test statistic, calculated in the second step, was found to be $\chi^2_2 = 8.555$, which is in the rejection region, our decision is to reject the null hypothesis which states that sex of the voter and opinion on the smoking bill are independent; in other words, our data provides sufficient evidence of a relationship between these two variables.

The shaded area in the figure at the top of the first page of Table A.5 graphically illustrates the *p*-value, which is the area above our observed test statistic value $\chi^2_2 = 8.555$ under the density curve for the χ^2 distribution with $df = 2$. This shaded area would represent the probability of obtaining a test statistic value χ^2_2 which represents greater differences between the observed and expected frequencies than the observed test statistic value $\chi^2_2 = 8.555$. By looking at the entries in Table A.5 corresponding to $df = 2$, we find that the observed test statistic $\chi^2_2 = 8.555$ is between $\chi^2_{2;\,0.025} = 7.378$ and $\chi^2_{2;\,0.01} = 9.210$. This tells us that the *p*-value is between 0.01 and 0.025, which we can designate by writing $0.01 < p\text{-value} < 0.025$. The fact that $0.01 < p\text{-value} < 0.025$ confirms to us that H_0 is rejected with $\alpha = 0.05$. However, this also tells us that H_0 would not be rejected with $\alpha = 0.01$ but would of course be rejected with $\alpha = 0.10$.



To complete the fourth step of the hypothesis test, we summarize the results as follows:

> Since $\chi^2_2 = 8.555$ and $\chi^2_{2;\,0.05} = 5.991$, we have sufficient evidence to reject H_0. We conclude that there is a relationship between sex of the voter and opinion on the smoking bill ($0.01 < p\text{-value} < 0.025$). Since H_0 is rejected, we need to describe the relationship.

When we do not reject the null hypothesis in a chi-square test concerning independence, no further analysis is called for, since we are concluding that the two variables are independent. However, rejecting the null hypothesis in a chi-square test concerning independence prompts us to investigate the relationship which we are concluding exists. A stacked bar chart to display the data used in the chi-square test concerning independence will be helpful. Figure 30-1 is a stacked bar chart with sex labeled on the horizontal axis and

stacks defined by opinion on the smoking bill. Verify that from this stacked bar chart, we might describe the relationship we concluded exists as follows:

> It appears that the percentage of voters favoring the bill is higher among females than among males, and that the percentage of voters neutral toward the bill is higher among males than among females.

A two-sided *z* test to compare two proportions is actually just a different version of the chi-square test concerning independence. To see this, return to hypothesis test corresponding to Figures 25-1, 25-2, and 25-3, where the *z* test to compare two proportions is used with $\alpha = 0.10$ to see if there is any evidence of a difference between the channel 4 and channel 8 viewing areas in the proportion of residents who viewed the 11:00 newscast telecast by both channels. In a simple random sample of 175 residents in the channel 4 viewing area, 49 viewed the newscast; in a simple random sample of 225 residents in the channel 8 viewing area, 81 viewed the newscast. We can organize this data into the contingency table displayed as Table 30-3.

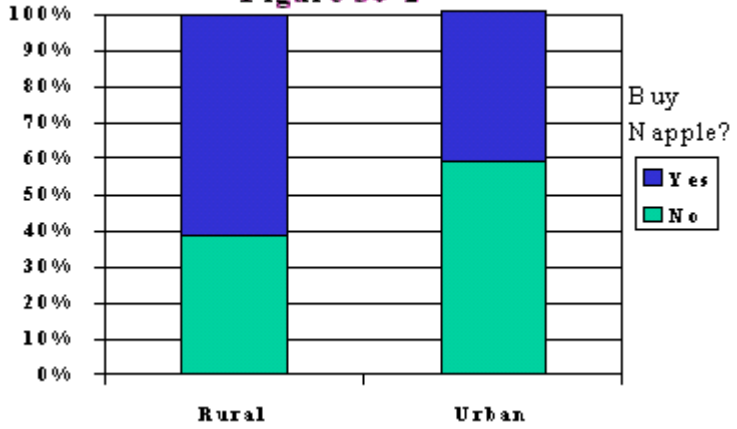
Table 30-3
Contingency Table for "Viewing Area" and "Viewed the Newscast"

	Viewed the Newscast		
Viewing Area	*Yes*	*No*	
Channel 4	49	126	175
Channel 8	81	144	225
	130	270	400

Looking for evidence of a difference between the channel 4 and channel 8 viewing areas in the proportion of residents who viewed the 11:00 newscast is exactly the same as looking for evidence of a relationship between viewing area and watching the newscast. It should come as no surprise then that applying a chi-square test concerning independence in a contingency table with two rows and two columns will produce exactly the same results as the two sample *z* test to compare proportions. As we stated previously, each of the χ^2-scores with 1 degree of freedom in Table A.5 is the square of a corresponding *z*-score. For instance, you can check that $[z_{0.025}]^2 = 1.960^2 = 3.841$, and that $\chi^2_{1;\,0.05} = 3.841$.

For the data organized into Table 30-3, we previously found that $z = -1.695$ in the two-sided *z* test to look for evidence of a difference between the channel 4 and channel 8 viewing areas in the proportion of residents who viewed the 11:00 newscast. We shall leave it to you to use the data displayed in Table 30-3 to calculate the chi-square test statistic concerning the independence of the two qualitative variables viewing area and watching the newscast. You should find this test statistic to be $\chi^2_1 = 2.872$. To see how the two-sided *z* test statistic and the chi-square test statistic concerning independence are related, you can check that $z^2 = (-1.695)^2 = 2.873 \approx 2.872 = \chi^2_1$; if each statistic were calculated to more decimal places of accuracy, then one would find that $z^2 = \chi^2_1$ exactly.

Figure 30-2



Self-Test Problem 30-1. A 0.05 significance level is chosen for a hypothesis test to see if there is any evidence of a relationship between location of residence and whether or not consumers buy the soft drink Napple. The results of the market survey are displayed in the contingency table on the right.

	Buy Napple	
Residence	*Yes*	*No*
Rural	127	81
Urban	248	344

(a) Choose all correct ways of completing the following sentence: Looking for evidence of a relationship between location of residence and buying Napple is the same as looking for evidence
   (i) of a difference between location of residence and buying Napple.
   (ii) of a relationship between buying Napple and not buying Napple.
   (iii) of a difference in the proportion of residents who buy Napple between the rural area and the urban area.
   (iv) of a relationship between locations.
   (v) of a difference difference in the proportion of rural residents between those who buy Napple and those who do not buy Napple.
   (vi) of a difference between the proportion of residents who buy Napple and the proportion of residence who do not buy Napple.

(b) Explain how the data for this hypothesis test is appropriate for a chi-square test concerning independence.

(c) Complete the four steps of the hypothesis test by completing the table titled *Hypothesis Test for Self-Test Problem 30-1*. You should find that $\chi^2_1 = 22.704$.

(d) If necessary, describe the relationship; if this is not necessary, explain why not.

(e) Verify that the sample size is sufficiently large for the chi-square test concerning independence to be appropriate.

(f) Considering the results of the hypothesis test, decide which of the Type I or Type II errors is possible, and describe this error.

(g) Decide whether H_0 would have been rejected or would not have been rejected with each of the following significance levels: (i) $\alpha = 0.01$, (ii) $\alpha = 0.10$.

(h) Construct an appropriate graphical display for the data used in this hypothesis test.

Hypothesis Test for Self Test Problem 30-1

Step 1 H_0:

H_1: $\alpha =$

Step 2

Step 3

Step 4

Answers to Self-Test Problems

30-1 (a) (iii) and (v)

(b) The data consists of a random sample of observations of two qualitative variables "location of residence" and "whether or not consumers buy the soft drink Napple," and the purpose of a chi-square test concerning independence is to decide whether or not there is a relationship between two qualitative variables.

(c) Step 1: H_0: Location of residence and buying Napple are independent

vs. ($\alpha = 0.05$)

H_1: At least one hypothesized proportion is not correct

OR

H_0: The proportion of consumers buying Napple is the same for rural and urban locations

vs. ($\alpha = 0.05$)

H_1: The proportion of consumers buying Napple is not the same for rural and urban locations

OR

H_0: The distribution of rural and urban locations is the same for consumers buying and not buying Napple

vs. ($\alpha = 0.05$)

H_1: The distribution of rural and urban locations is not the same for consumers buying and not buying Napple

Step 2: $\chi^2_1 = 22.704$

Step 3: The rejection is $\chi^2_1 \geq 3.841$. H_0 is rejected; p-value < 0.001.

Step 4: Since $\chi^2_1 = 22.704$ and $\chi^2_{1;\,0.05} = 3.841$, we have sufficient evidence to reject H_0. We conclude that there is a relationship between location and buying Napple (p-value < 0.001). Since H_0 is rejected, we need to describe the relationship.

(d) It appears that the proportion of residents buying Napple is higher in the rural area than in the urban area.

(e) Since the expected frequencies are all greater than 5, the sample size is sufficiently large for the chi-square test concerning independence to be appropriate.

(f) Since H_0 is rejected, the Type I error is possible, which is concluding that a relationship exists when in reality location of residence and buying Napple are independent.

(g) H_0 would have been rejected with both $\alpha = 0.01$ and $\alpha = 0.10$.

(h) See Figure 30-2.

Summary

With a sufficiently large sample size, the chi-square test concerning independence in a contingency table with r rows and c columns can be performed by using the chi-square test statistic

$$\chi^2_{(r-1)(c-1)} = \sum \frac{(O-E)^2}{E} \quad ,$$

where O observed frequencies and E represents frequencies that are expected if H_0 were actually true. The null hypothesis, which states that two qualitative variables in the contingency table are independent, is rejected when the chi-square test statistic is larger than a value determined from the *chi-square distribution with* $(r-1)(c-1)$ *degrees of freedom*. An easy way to check that the sample size is sufficiently large is to verify that all $E \geq 5$.

When we do not reject the null hypothesis in a chi-square test concerning independence, no further analysis is called for, since we are concluding that the two qualitative variables are independent. When we reject the null hypothesis in a chi-square test concerning independence, we need to describe the relationship between the two qualitative variables; a stacked bar chart can be helpful in describing the relationship. Concluding that the two qualitative variables are not independent is the same thing as concluding that there is an association or dependency between the two variables or that the distribution of distribution of categories for one qualitative variable is the same for each category of the other qualitative variable.