Categorical Data Analysis: Chi-Squared Tests

13.1 Finding Chi-Square Critical Values

In this chapter, we will be working with random variables that have a Chi-Squared distribution. As a result, we will have a need to find Chi-Squared critical values. As we have seen with the other distributions we have worked with, we will find these values on a table (this time one of Chi-Squared values). For every problem we encounter, there will be two quantities required in order to determine the unique Chi-Squared value needed. Those quantities are the **area to the right of the critical value** (for us this will be the significance level) and the **degrees of freedom**.

In the following two examples, we will demonstrate how to use the Chi-Squared table:

Example 173.5 Find $\chi^2_{\alpha,df} = \chi^2_{0.10,6}$

Example 173.6 Find $\chi^2_{\alpha,df} = \chi^2_{0.005,3}$

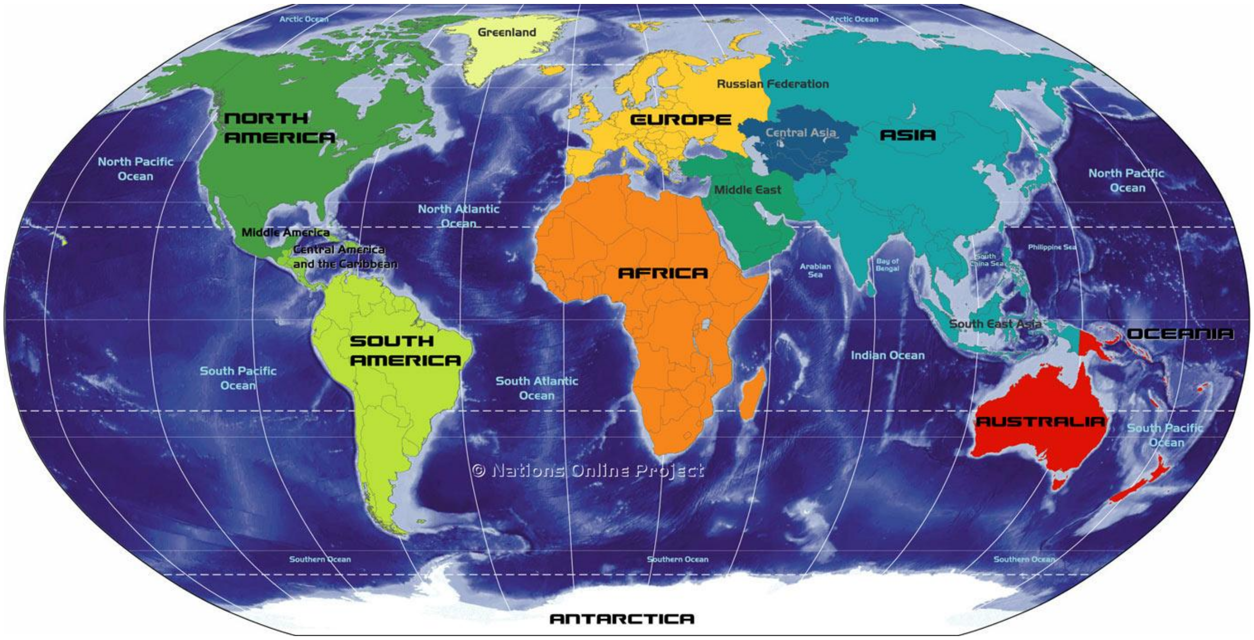
13.2 Checking the Assumptions for a Chi-Square Goodness-of-Fit Test

Categorical Data Analysis and the Multinomial Experiment

Properties of the Multinomial Experiment

1. The experiment consists of n identical trials
2. There are k possible outcomes for each trial. These outcomes are sometimes called classes, categories, or cells.
3. The probabilities of the k outcomes, denoted by $p_1, p_2, \ldots, p_k$, remain the same from trial to trial and they sum to one.
4. The trials are independent.
5. The random variables of interest are the cell counts, $n_1, n_2, \ldots, n_k$, which are the number of observations that fall in each of the k categories.

Example 174: If I ask 36 randomly selected students if they are American (N & S), European, Asian, African, or other. Is this experiment Multinomial?



Example 175: Suppose three candidates are running for office, and 150 voters are asked their preferences. Is this experiment Multinomial?

Solution: Yes, this experiment meets all of the 5 criteria above for a multinomial experiment.

Example 175.5 Consider the data below: 276 people were asked to pick the city they would most like to travel to between Rome, Paris, and London. Is this experiment multinomial in nature?

Which of the 3 cities below would you most like to visit?		
Rome	Paris	London
92	108	76

The table above is called a one-way table because each cell count corresponds to a single category.

The researchers conducting this survey may want to detect a difference between traveler's preferences regarding these three cities. To test this in a formal way, we will set up the following pair of competing hypotheses:

$$H_0 : p_R = p_P = p_L$$
$$H_A : \textit{At least} \text{ one proportion exceeds } 1/3.$$

(Why 1/3? What would it be if we had 5 categories? Answer: 1/5)

We could specify any value we want to for these proportions in our null hypothesis above as long as their sum adds to 1.00. If we do that, we then rephrase our alternative hypothesis to say, "At least one of the probabilities differs from its hypothesized value." This kind of test is often referred to as a **Goodness-of-fit Test.** A goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution. How will we extract the information available in the one-way table in order to test our null hypothesis?

13.3 The Chi-Square Test Statistic

If the null hypothesis were true, we would expect that the number of people who would prefer to go to Rome would be $E_{Rome} = np_R = 276\left(\frac{1}{3}\right) = 92$.

The other two cities would also have 92 people saying they prefer to visit them (assuming the null is correct in saying the proportions of people wishing to travel to each of these three cities are all equal). Consider the following test statistic. It measures the distance between the actual observation and the hypothesized null value for that observation:

Chi-Squared Test Statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$$

O_i represents the observed frequency of an outcome. k represents the number of different categories or classes, and $E_i = np_{i0}$ represents the expected frequency of an outcome (this is the sample size multiplied by the hypothesized proportion for the ith category).

If the value of the above statistic is large with respect to our critical value, we will reject the null hypothesis. Our critical value will be found using the Chi-squared table. Goodness-of-fit hypothesis tests are always *right-tailed*.

The **criteria for rejection** will be as follows: Reject the null when $\chi^2 > \chi^2_\alpha$ where χ^2_α has k – 1 degrees of freedom (remember k is the number of categories we have in the multinomial experiment).

Like the other hypothesis test we have discussed thus far, the Chi-square goodness-of-fit test has certain **assumptions** or requirements:

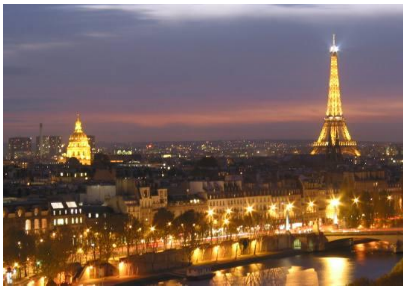
1. The data must have been randomly selected.
2. The sample data consist of frequency counts for each of the different categories.
3. For each category, the expected frequency is at least 5. (The expected frequency for a category is the frequency that would occur if the data actually have the distribution that is being claimed. There is no requirement that the *observed* frequency for each category must be at least 5.)

13.4 Testing Categorical Probabilities: One-Way Table

Now let's work out our travel preference example from start to finish:

Example 176: 276 people were asked to pick the city they would most like to travel to when given the three choices: Rome, Paris, and London. This experiment was multinomial in nature. Test the claim that the proportion of people who choose Rome equals the proportion who will choose London equals the proportion who will choose Paris.

Which of the 3 cities below would you most like to visit?		
Rome	Paris	London
92	108	76


Solution:

1. Claim: The proportion of the population that would select a trip to Rome equals the proportion that would select a trip to Paris which equals the proportion that would select London.
2. Hypotheses: $H_0 : p_R = p_P = p_L$
   $H_A : At\ least$ one proportion exceeds 1/3.
3. Get Data and determine your alpha level:

Which of the 3 cities below would you most like to visit?		
Rome	Paris	London
92	108	76

   Let's set alpha at 5%

4. Calculate the Test Stat: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(92-92)^2}{92} + \frac{(108-92)^2}{92} + \frac{(76-92)^2}{92} = 5.565$
5. Get Your Critical Value: $\chi^2_{\alpha,k-1} = \chi^2_{.05,2} = 5.991$
6. Form Your Initial Conclusion: Fail to reject the null
7. Final Conclusion: There is not sufficient evidence to warrant rejection of the claim that the proportions are all equal at the 5% significance level.

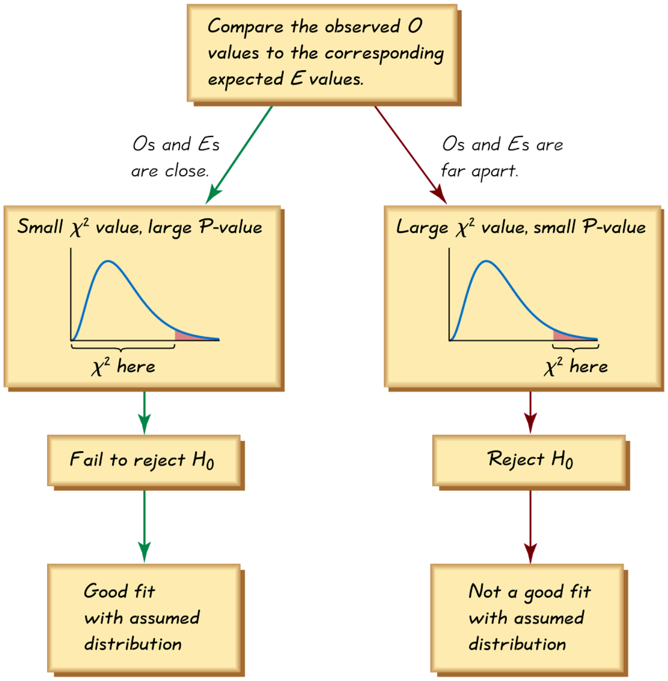
Finally, our test assumptions:

1. The experiment which produced our data was a multinomial experiment.
2. The sample size is such that the individual cell expectations are all greater than or equal to 5.

Let's analyze what just happened. We noticed there weren't huge differences between the expected number of responses for each city and the responses we observed during the survey. If you look at the formula for our test stat you will notice the following things:

- A close agreement between observed and expected values will lead to a small value of χ^2 and a large *P*-value.

- A large disagreement between observed and expected values will lead to a large value of χ^2 and a small *P*-value.



Example 177: Statistics can be used to detect fraud. There is a pattern that turns up when observing the leading digit in real data (as opposed to fake data like you might find in falsified financial records). That pattern is expressed by Benford's law. The table below lists the percentages for leading digits from Benford's Law that we would expect to observe. It also lists the number of leading digits actually observed on a batch of 784 checks that are believed to be fraudulent.

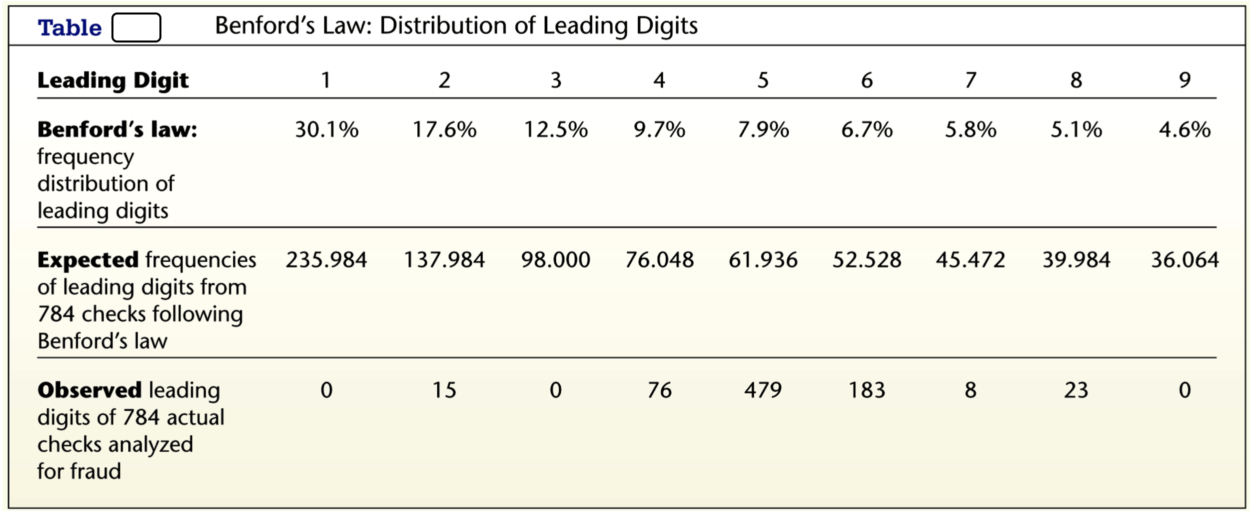


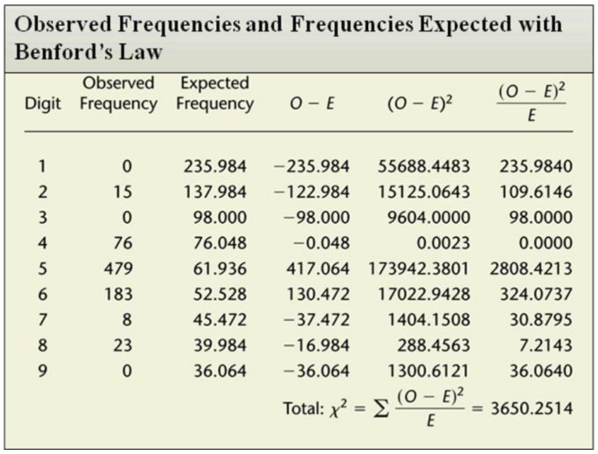
Table Benford's Law: Distribution of Leading Digits

Leading Digit	1	2	3	4	5	6	7	8	9
Benford's law: frequency distribution of leading digits	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
Expected frequencies of leading digits from 784 checks following Benford's law	235.984	137.984	98.000	76.048	61.936	52.528	45.472	39.984	36.064
Observed leading digits of 784 actual checks analyzed for fraud	0	15	0	76	479	183	8	23	0

Use this data to test the claim at the 1% significance level that there is a significant discrepancy between the leading digits expected from Benford's Law and the leading digits from the 784 checks.

Solution:

1. Claim: There is a significant discrepancy between the leading digits expected from Benford's Law and the leading digits from the 784 checks.
2. Hypotheses: $H_0: \rho_1 = 0.301, \rho_2 = 0.176, \ldots, \rho_9 = 0.046$
   H_A : At least one of the proportions is different from the claimed values.
3. Get Data and determine your alpha level: The data is above and the alpha level is 1%.
4. Calculate the Test Stat: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

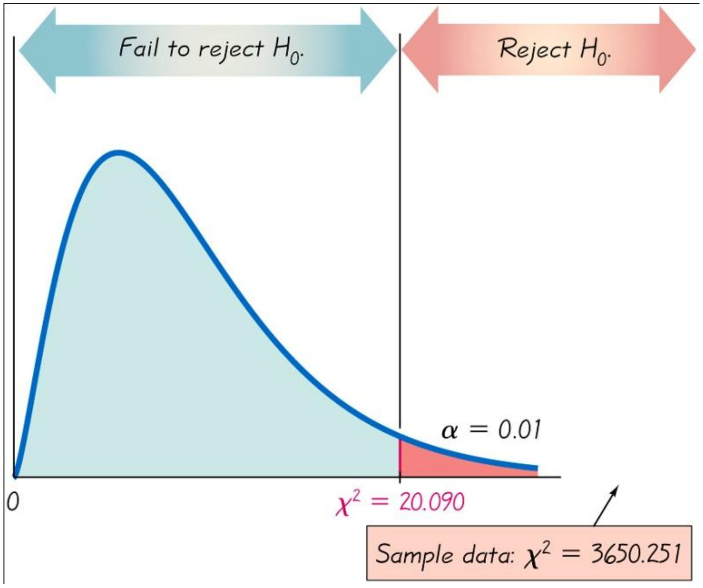


Observed Frequencies and Frequencies Expected with Benford's Law

Digit	Observed Frequency	Expected Frequency	$O - E$	$(O - E)^2$	$\frac{(O-E)^2}{E}$
1	0	235.984	−235.984	55688.4483	235.9840
2	15	137.984	−122.984	15125.0643	109.6146
3	0	98.000	−98.000	9604.0000	98.0000
4	76	76.048	−0.048	0.0023	0.0000
5	479	61.936	417.064	173942.3801	2808.4213
6	183	52.528	130.472	17022.9428	324.0737
7	8	45.472	−37.472	1404.1508	30.8795
8	23	39.984	−16.984	288.4563	7.2143
9	0	36.064	−36.064	1300.6121	36.0640

Total: $\chi^2 = \sum \frac{(O-E)^2}{E} = 3650.2514$

5. Get Your Critical Value: $\chi^2_{\alpha,k-1} = \chi^2_{.01,8} = 20.090$
6. Form Your Initial Conclusion: Reject the null



7. Final Conclusion: There is sufficient evidence to support the claim that there is a significant discrepancy at the 1% significance level.

Example 178: Use the information below to test the claim at the 1% significance level that the airing of a television series about marijuana possession has altered the public's opinions about marijuana possession.

Distribution of Opinions About Marijuana Possession **Before** Television Series has Aired

Legalization	Decriminalization	Existing Law	No Opinion
7%	18%	65%	10%

Distribution of Opinions About Marijuana Possession **After** Television Series has Aired

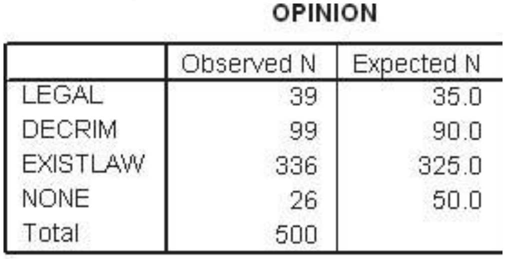
Legalization	Decriminalization	Existing Law	No Opinion
39	99	336	26

Solution:

1. Claim: The television series about marijuana possession has altered the public's opinions about marijuana possession.

2. Hypotheses: $H_0: p_1 = .07, p_2 = .18, p_3 = .65, p_4 = .10$
   $H_a:$ At least one of the proportions differs from its null hypothesis value.

3. Get Data and determine your alpha level: Alpha is set at 1%, and our observed and expected values are given below:



OPINION

	Observed N	Expected N
LEGAL	39	35.0
DECRIM	99	90.0
EXISTLAW	336	325.0
NONE	26	50.0
Total	500	

4. Calculate the Test Stat:

$$\chi^2 = \frac{(39-35)^2}{35} + \frac{(99-90)^2}{90} + \frac{(336-325)^2}{325} + \frac{(26-50)^2}{50} = \chi^2 = 13.249$$

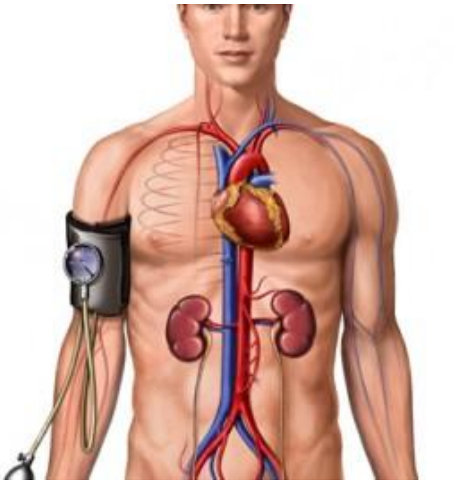
5. Rejection region: $\chi^2 > \chi^2_{\alpha=.01, df=3} = 11.3449$
6. Form Your Initial Conclusion: Reject the null
7. Final Conclusion: There is sufficient evidence to support the claim that the public's opinion has been altered at the 1% significance level.

13.5 Finding Expected Cell Counts for a Two-Way Table

In the section above, we considered data classified by a single criterion. We now consider multinomial experiments in which the data are classified according to two criteria (two qualitative factors).

Consider the following example from *JAMA* (April 18th, 2001), the data is from a study of alcohol consumption's effect on congestive heart failure. The patients were classified under two criteria: 1) the average number of alcoholic drinks consumed per week and 2) whether or not they had congestive heart failure. The results of the study involving 1,913 patients are summarized below in a two-way table called a ***contingency table.***

		Alcohol Consumption		
		Abstainers	Less than 7 drinks	7 or more drinks
Congestive Heart Failure	Yes	146	106	29
	No	750	590	292
	Totals	**896**	**696**	**321**



Before attempting to analyze this data, let's look at the table above in symbolic fashion:

		Alcohol Consumption			
		Abstainers	Less than 7 drinks	7 or more drinks	Totals
Congestive Heart Failure	Yes	n_{11}	n_{12}	n_{13}	r_1
	No	n_{21}	n_{22}	n_{23}	r_2
	Totals	c_1	c_2	c_3	n

In the above table, the n_{ij} values are the observed cell counts. The c_j are the column totals. The r_i are the row totals, and *n* is the grand total (total number of observations).

Now consider the same table except this time with cell probabilities:

		Alcohol Consumption			
		Abstainers	Less than 7 drinks	7 or more drinks	Totals
Congestive Heart Failure	Yes	p_{11}	p_{12}	p_{13}	P r_1
	No	p_{21}	p_{22}	p_{23}	P r_2
	Totals	P c_1	P c_2	P c_3	1

The row and column totals here are referred to as **marginal probabilities**. For example**,** the probability a subject is an abstainer is given by: $p_{11} + p_{21} = p_{c1}$.

The experiment above is a multinomial experiment with a total of 1,913 trials, (2)(3) = 6 possible outcomes, and probabilities for each cell as shown in the table above.

Suppose we want to know whether our two classifications (alcohol consumption and congestive heart failure) are dependent. That is, if we know a person's drinking habits, does that give us any information regarding the likelihood that person will have congestive heart failure?

Our hypotheses would be:

H_0 : The two classifications are independent
H_A : The two classifications are dependent

From probability, we know if events A and B are independent then $P(AB) = P(A)P(B)$. Similarly, the probability that a subject is classified in any particular cell should be equal to the product of the marginal probabilities for that cell if the two classifications are independent.

If we hypothesize independence, we are saying that the following is true:

		Alcohol Consumption			
		Abstainers	Less than 7 drinks	7 or more drinks	Totals
Congestive Heart Failure	Yes	$p_{11} = \text{P}r_1(\text{P}c_1)$	$p_{12} = \text{P}r_1(\text{P}c_2)$	$p_{13} = \text{P}r_1(\text{P}c_3)$	$\text{P}r_1$
	No	$p_{21} = \text{P}r_2(\text{P}c_1)$	$p_{22} = \text{P}r_2(\text{P}c_2)$	$p_{23} = \text{P}r_2(\text{P}c_3)$	$\text{P}r_2$
	Totals	$\text{P}c_1$	$\text{P}c_2$	$\text{P}c_3$	1

To create a test statistic designed for use in a test of the hypothesis of independence, we will use the same logic as in the one-way contingency table scenario. We will compare the expected cell counts to our observed cell counts.

Recall the expected value for the first cell should be: $E_{11} = np_{11} = (under H_0) np_{r1}p_{c1}$

We can estimate this by: $\hat{E}_{11} = n\left(\frac{r_1}{n}\right)\left(\frac{c_1}{n}\right) = \frac{r_1c_1}{n}$

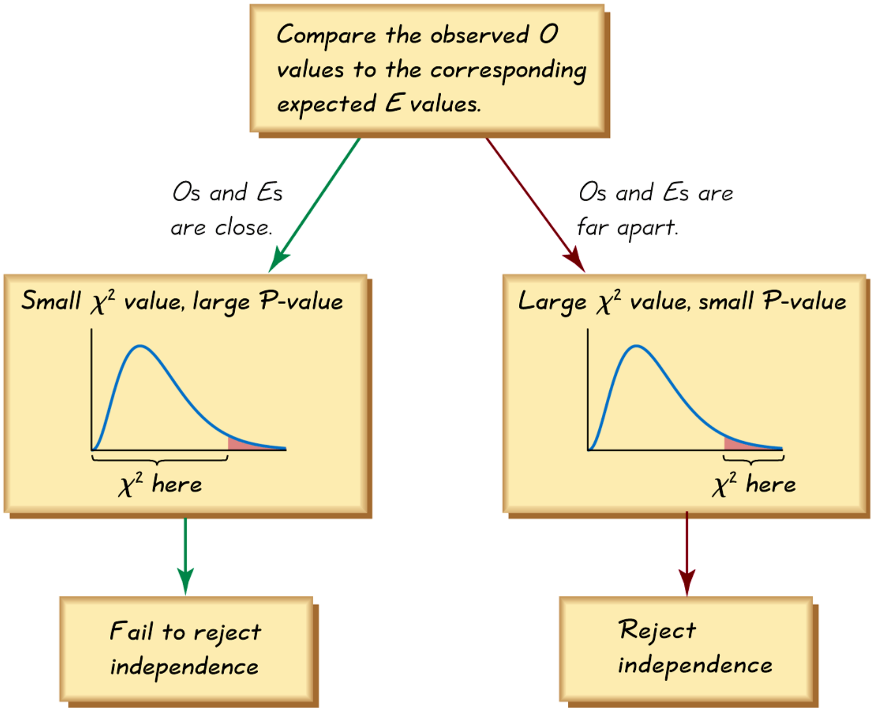
Similarly, all **<u>expected cell counts</u>** can be found by:

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

Then our **test statistic** can be found by using:

$$\chi^2 = \sum \frac{\left(n_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}} = \sum \frac{\left(O_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$$

And similar to the one-way table problems, we will **reject the null when our test statistic is larger than our critical value** $\chi^2 > \chi^2_\alpha$ where χ^2_α has (r – 1)(c – 1) degrees of freedom.



Example 178.5 Use the data below to find the expected value for the cell count in the second row and first column (i.e. - find E_{21})

		Alcohol Consumption		
		Abstainers	Less than 7 drinks	7 or more drinks
Congestive Heart Failure	Yes	146	106	29
	No	750	590	292
	Totals	**896**	**696**	**321**

Finally, our **test assumptions** are:

1. The experiment which produced our data was a multinomial experiment.
2. The sample size is such that the individual cell expectations are all greater than or equal to 5.

Now we are ready to work our example:

13.6 Testing Categorical Probabilities: Two-Way (Contingency) Table

Example 179: At the 1% significance level, test the claim that drinking habits and congestive heart failure are independent.

		Alcohol Consumption		
		Abstainers	Less than 7 drinks	7 or more drinks
Congestive Heart Failure	Yes	146	106	29
	No	750	590	292
	Totals	**896**	**696**	**321**

Solution:

1. Claim: Drinking habits and congestive heart failure are independent.
2. Hypotheses: H_0 : The two classifications are independent
   H_A : The two classifications are dependent
3. Calculate **Expected Cell Values** for each cell:

		Alcohol Consumption		
		Abstainers	Less than 7 drinks	7 or more drinks
Congestive Heart Failure	Yes	146 (131.6)	106 (102.2)	29 (47.2)
	No	750 (764.4)	590 (593.8)	292 (273.9)
	Totals	**896**	**696**	**321**

4. Calculate the Test Stat: $\chi^2 = \sum \frac{\left(n_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}} = \sum \frac{\left(O_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}} =$

$$\frac{(146-131.6)^2}{131.6} + \frac{(106-102.2)^2}{102.2} + \frac{(29-47.2)^2}{47.2} + \frac{(750-764.4)^2}{764.4} + \frac{(590-593.8)^2}{593.8} + \frac{(292-273.9)^2}{273.9} \approx 10.197$$

5. Get Your Critical Value: $\chi^2_{.01,(2-1)(3-1)} = 9.210$
6. Form Your Initial Conclusion: Reject the Null
7. Final Conclusion: At the 1% significance level, the sample data warrant rejection of the claim that alcohol consumption and congestive heart failure are independent.

Example 180: Is the color of the motorcycle helmet used by riders somehow related to the risk of crash related injuries? Use the data below to test the claim at the 5% level of significance that the colors of motorcycle helmets are independent of crash injuries.

	Black	White	Yellow/Orange	Totals
Not Injured	491	377	31	899
Injured or Killed	213	112	8	333
Totals	704	489	39	1232

Example 181: Use a 5% significance level and the provided computer output to test the pair of hypotheses below:

Religious Affiliation

Marital Status	A	B	C	D	None	Totals
Divorced	39	19	12	28	18	116
Married, never divorced	172	61	44	70	37	384
Totals	211	80	56	98	55	500

H_0: Marital status and religious affiliation are independent
H_a: Marital status and religious affiliation are dependent

```
                     The FREQ Procedure

                Table of MARITAL by RELIGION

MARITAL      RELIGION

Frequency|
Expected |A        |B        |C        |D        |NONE     |  Total
DIVORCED |      39 |      19 |      12 |      28 |      18 |    116
         |  48.952 |   18.56 |  12.992 |  22.736 |   12.76 |
NEVER    |     172 |      61 |      44 |      70 |      37 |    384
         |  162.05 |   61.44 |  43.008 |  75.264 |   42.24 |
Total         211        80        56        98        55      500

          Statistics for Table of MARITAL by RELIGION

    Statistic                        DF       Value      Prob
    ------------------------------------------------------
    Chi-Square                        4      7.1355    0.1289
    Likelihood Ratio Chi-Square       4      6.9854    0.1367
    Mantel-Haenszel Chi-Square        1      6.4943    0.0108
    Phi Coefficient                          0.1195
    Contingency Coefficient                  0.1186
    Cramer's V                               0.1195

                   Fisher's Exact Test
              ----------------------------------
              Table Probability (P)   6.936E-06
              Pr <= P                    0.1251

                    Sample Size = 500
```