Inference for Difference Between 2 Population Proportions

<table>
<thead>
<tr>
<th>Chapter 10</th>
<th>Chapter 11</th>
</tr>
</thead>
</table>

We are interested in whether 2 pop. means are different.

**Example:**

Mean Income in Bradford = $38,010
Mean Income in Kane = $35,006

Difference = $3,004

Are the differences due to sampling error or are the significant differences that allow us to conclude that population parameters are different.

We are interested in whether 2 pop. proportions are different.

**Example:**

2013 Data Entry Error Rate = 0.142
2014 Data Entry Error Rate = 0.107

Sample proportions
1. Confidence Interval for difference between 2 pop. proportions
2. Hypothesis Test for difference between 2 pop. proportions
3. Hypothesis test to check Equality of 2 or more pop. proportions using chi-square test statistic
4. Test of Independence of 2 categorical variables using chi-square test statistic

Chapter 10

Sampling Distribution of \( \bar{X}_{1} - \bar{X}_{2} \)

- \( \bar{X}_{1} - \bar{X}_{2} \) is a random variable.
- Mean: \( M_{1} - M_{2} = E(\bar{X}_{1} - \bar{X}_{2}) = \text{sum of all possible } \bar{X}_{1} - \bar{X}_{2} \)/\( \text{count of all } \bar{X}_{1} - \bar{X}_{2} \)
- Standard Error of Sampling Distribution of \( \bar{X}_{1} - \bar{X}_{2} \): \( SD = \sigma_{\bar{X}_{1} - \bar{X}_{2}} = \sqrt{\frac{\sigma_{1}^{2} + \sigma_{2}^{2}}{n_{1} + n_{2}}} \)

Chapter 11

Sampling Distribution of \( \bar{P}_{1} - \bar{P}_{2} \)

- \( \bar{P}_{1} - \bar{P}_{2} \) is a random variable.
- Mean: \( P_{1} - P_{2} = E(\bar{P}_{1} - \bar{P}_{2}) = \text{sum of all possible } \bar{P}_{1} - \bar{P}_{2} \)/\( \text{count of all } \bar{P}_{1} - \bar{P}_{2} \)
- Standard Error of Sampling Distribution of \( \bar{P}_{1} - \bar{P}_{2} \): \( SD = \sigma_{\bar{P}_{1} - \bar{P}_{2}} = \sqrt{\frac{\hat{P}_{1}(1-\hat{P}_{1})}{n_{1}} + \frac{\hat{P}_{2}(1-\hat{P}_{2})}{n_{2}}} \)

In this chapter we wont know population proportions \( \hat{P}_{1} \) & \( \hat{P}_{2} \) so we will use \( \bar{P}_{1} \) \& \( \bar{P}_{2} \) in formula to estimate \( \sigma_{\bar{P}_{1} - \bar{P}_{2}} \)

\( \sigma_{\bar{X}_{1} - \bar{X}_{2}} \) is needed when \( \frac{n_{1} + n_{2}}{N} \geq 0.05 \)

\( \checkmark \) see Excel example to prove these formulas work.
Remember proportions from chapter 7:

\[ \bar{p} = \frac{X}{n} = \text{Random variable} \]

\( X = \text{number of elements in sample that possess the characteristic of interest} \)

\( x = \text{categorical variable} \)

\( n = \text{sample size} \)

**Sampling Distribution of \( \bar{p} \) can be approximated by a normal distribution whenever:**

\[ n \times p \geq 5 \]

\[ n \times (1-p) \geq 5 \]

**Expected value of \( \bar{p} \)**

\[ \{ \text{Expected value of \( \bar{p} \)} \} = E(\bar{p}) = p \]

\[ \text{Population proportion} \]

**Standard Error / Deviation of Sampling Distribution of \( \bar{p} \)**

\[ \sigma_{\bar{p}} = \sqrt{\frac{p \times (1-p)}{n}} \times \sqrt{\frac{N-n}{N-1}} \]

**Correction Factor needed when \( \frac{n}{N} > 0.05 \)**
Confidence Interval for Difference Between 2 Population Proportions $P_1 - P_2$

* From Sample Data, we calculate a point estimate for $P_1 - P_2$, $\bar{P}_1 - \bar{P}_2$, and then add a margin of error to both sides to get an interval (lower & upper value) that will contain the population proportion difference ($P_1 - P_2$) about 95 out of a 100 times.

* If we don't know $P_1$ or $P_2$, we use our sample proportions $\bar{P}_1$ and $\bar{P}_2$ in all of our calculations. (That's why we use word "Estimate" on SE calc.)

* We must take 2 Independent and random samples.

* We can use the Normal Distribution to estimate the sampling distribution of $\bar{P}_1 - \bar{P}_2$ if 4 tests are passed:
  1. $n_1 * \bar{P}_1 \geq 5$
  2. $n_1 * (1-\bar{P}_1) \geq 5$
  3. $n_2 * \bar{P}_2 \geq 5$
  4. $n_2 * (1-\bar{P}_2) \geq 5$

Margin of Error

$$\bar{P}_1 - \bar{P}_2 \pm Z_{\alpha/2} \sqrt{\frac{\bar{P}_1 * (1-\bar{P}_1)}{n_1} + \frac{\bar{P}_2 * (1-\bar{P}_2)}{n_2}}$$

Estimate of Standard Deviation

Standard Error For Sampling Distribution of $\bar{P}_1 - \bar{P}_2$
Variables Defined for Confidence Interval:

\( p_1 = \text{population 1 proportion} \)
\( p_2 = \text{population 2 proportion} \)
\( \bar{p}_1 = \text{Sample proportion from pop. 1} \)
\( \bar{p}_2 = \text{Sample proportion from pop. 2} \)
\( n_1 = \text{Sample size from population 1} \)
\( n_2 = \text{Sample size from population 2} \)
\( \alpha = \text{alpha = risk that the population proportion, } p_1 - p_2, \text{ is not in interval} \)
\( 1 - \alpha = \text{Confidence Level / Coefficient} = \text{How sure we are that } p_1 - p_2 \text{ is in interval} \)
\( \sqrt{\frac{N-n}{N-1}} = \text{Correction Factor for Standard Error calculation if } \frac{n}{N} > 0.05 \)
Hypothesis Test to check if there is no difference between 2 population proportions $P_1 - P_2$ (z method, hypothesized difference = 0)

* From Sample Data, we run our 5-step hypothesis test "No Difference" between 2 population proportions.

* In this section we run H.T. for $P_1 - P_2$ using the z method. Next section we will see how to check difference between 2 or more population proportions using chi-square method. Chi-square can only do 2-tail test.

* z method Allows us to run 3 types of Tests:

<table>
<thead>
<tr>
<th>1 tail, Lower</th>
<th>1 tail upper</th>
<th>2 tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: P_1 - P_2 \geq 0$</td>
<td>$H_0: P_1 - P_2 \leq 0$</td>
<td>$H_0: P_1 - P_2 = 0$</td>
</tr>
<tr>
<td>$H_a: P_1 - P_2 &lt; 0$</td>
<td>$H_a: P_1 - P_2 &gt; 0$</td>
<td>$H_a: P_1 - P_2 \neq 0$</td>
</tr>
</tbody>
</table>

* we must take 2 Independent & Random Samples

* We can use Normal Distribution to estimate the sampling distribution of $\bar{P}_1 - \bar{P}_2$ if 4 Tests are passed:

1. $n_1 \times \bar{P}_1 \geq 5$
2. $n_1 \times (1-\bar{P}_1) \geq 5$
3. $n_2 \times \bar{P}_2 \geq 5$
4. $n_2 \times (1-\bar{P}_2) \geq 5$
Formulas for Z method Hypothesis Test of $p_1 - p_2$

1. If $H_0$ is TRUE as an Equality:

\[ p_1 - p_2 = 0 \]

\[ \downarrow \]

\[ p_1 = p_2 \]

population proportion becomes: $p_1 = p_2 = p$

2. Standard Error of $\bar{p}_1 - \bar{p}_2$ when $p_1 = p_2 = p$:

\[
\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 \times (1-p_1)}{n_1} + \frac{p_2 \times (1-p_2)}{n_2}} = \sqrt{p \times (1-p) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}
\]

if $p_1 = p_2 = p \rightarrow$ then

Now we just need estimate of $p$

3. Pooled Estimator of $p$ when $p_1 = p_2 = p$

\[
\hat{p} = \frac{n_1 \times \bar{p}_1 + n_2 \times \bar{p}_2}{n_1 + n_2}
\]

if $p = p_1 = p_2$ we can just take weighted average

4. Z Test Statistic

\[
Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{p \times (1-p) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}
\]

\[\text{SE}\]
Variables Defined for Hypothesis Test of $P_1 - P_2$

- $P_1 = \text{population 1 proportion}$
- $P_2 = \text{population 2 proportion}$
- $\overline{P}_1 = p_{\text{bar}_1} = \text{sample proportion from Pop. 1}$
- $\overline{P}_2 = p_{\text{bar}_2} = \text{sample proportion from Pop. 2}$
- $P = \text{population proportion when } P_1 = P_2$
- $\overline{P} = \text{Estimate of } P \text{ when we assume } P_1 = P_2 = P$, called "Pooled Estimator of } P" or weighted average of $\overline{P}_1$ and $\overline{P}_2$.

- $n_1 = \text{sample size taken from population 1}$
- $n_2 = \text{sample size taken from population 2}$

- $\alpha = \text{risk that we reject } H_0 \text{ when it is true (Type I Error)}$
An HMO wanted to check the accuracy of patient contact information data entry. It took independent random samples of records and rechecked the accuracy of the data. No = no errors, data accurate. Yes = errors were found.

Is there a significant reduction in errors from 2013 to 2014? (p1 = proportion for population 1; p2 = proportion for population 2)

- \( p_{1\text{bar}} \) = sample proportion for simple random sample from population 1
- \( p_{2\text{bar}} \) = sample proportion for simple random sample from population 2

### Variables

<table>
<thead>
<tr>
<th>n</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>316</td>
<td>385</td>
</tr>
<tr>
<td>Count Yes</td>
<td>45</td>
<td>49</td>
</tr>
<tr>
<td>Count No</td>
<td>271</td>
<td>336</td>
</tr>
</tbody>
</table>

- Yes Sample proportion = \( p_{1\text{bar}} \) = \( \frac{45}{316} \) = 0.142465065
- Yes Sample proportion = \( p_{2\text{bar}} \) = \( \frac{49}{385} \) = 0.12739843

- No Sample proportion = \( 1-p_{1\text{bar}} \) = 0.857534935
- No Sample proportion = \( 1-p_{2\text{bar}} \) = 0.87260156

### Proportions 1 vs. 2

<table>
<thead>
<tr>
<th></th>
<th>Proportion 1</th>
<th>Proportion 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>( p_{1\text{bar}} )</td>
<td>( p_{2\text{bar}} )</td>
</tr>
<tr>
<td>2014</td>
<td>( 1-p_{1\text{bar}} )</td>
<td>( 1-p_{2\text{bar}} )</td>
</tr>
</tbody>
</table>

### Calculations for Confidence Interval

- Used For Confidence Interval
  - Alpha: 0.05
  - Alpha/2: 0.025
  - Confidence Coefficient Level: 0.95
  - Upper Z: 1.96
  - Margin of Error: 0.0245
  - Standard Error: 0.00326564
  - Standard Error: 0.00258682

### Calculations for Confidence Interval

- Calculation 1
  - Standard Error
  - Margin of Error
  - Standard Error

### Conclusion

- We are 95% sure that the population proportion for difference between 2013 errors and 2014 errors is between 0.0144 and 0.0846.

Because error is in our interval, it seems reasonable to assume that there is not a significant reduction in errors.

### Hypothesized Difference

<table>
<thead>
<tr>
<th>Hypothesized Difference</th>
<th>Calculations for Hypothesis Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{1\text{bar}} - p_{2\text{bar}} )</td>
<td>( &lt; ) 0</td>
</tr>
<tr>
<td>( p_{1\text{bar}} - p_{2\text{bar}} )</td>
<td>( &gt; ) 0</td>
</tr>
</tbody>
</table>

### Step 1 (Estimation of p₁ \& p₂)

- \( p_{1\text{bar}} \) = \( \frac{45}{316} \) = 0.142465065
- \( p_{2\text{bar}} \) = \( \frac{49}{385} \) = 0.12739843

### Step 2 (Calculation of Standard Error)

- Standard Error = 0.00326564

### Step 3 (Calculation of Confidence Interval)

- \( 1 - \text{SE} = 0.99673436 \)
- \( Z \text{ Test Statistic} = 1.40339099 \)
- \( p_{1\text{bar}} - p_{2\text{bar}} \)
- \( p_{1\text{bar}} + p_{2\text{bar}} \)

### Step 4 (Critical Value (Hurdle))

- Critical Value = 1.644853627
- P-value = 0.08252273

### Step 5 (Decision)

- Because our test statistic is not past our hurdle and the p-value is bigger than our alpha, we fail to reject the null hypothesis.
- The evidence does not suggest that there has been a significant reduction in errors.
- We do run a 5% risk that the null was not rejected even though the alternative was true.

### Analysis

**Advantage of z method:**
- Can do two-tail test or one tail test.

**Disadvantage of z method:**
- Can only compare two proportions (could compare more but would have build up of Type I error)
An RMD wanted to check the accuracy of patient contact information data entry. It took independent random samples of records and rechecked the accuracy of the data.

To no errors, data accurate
Yes = Yes, Errors were found

Is there a significant reduction in errors from 2013 to 2014?

\( p_1 = \text{proportion for population 1} \)
\( p_2 = \text{proportion for population 2} \)
\( p_{\text{pool}} = \text{sample proportion for simple random sample from population 1} \)
\( p_{\text{pool}} = \text{sample proportion for simple random sample from population 2} \)

<table>
<thead>
<tr>
<th>Variables</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count/Row</td>
<td>COUNTA</td>
<td>COUNTB</td>
</tr>
<tr>
<td>Count/No.</td>
<td>COUNT</td>
<td>COUNT</td>
</tr>
<tr>
<td>Yes/No count</td>
<td>COUNTA</td>
<td>COUNTB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculations for Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha = 0.05 )</td>
</tr>
<tr>
<td>( \text{Confidence Coefficient Level} = 0.95 )</td>
</tr>
<tr>
<td>( \text{Degrees of Freedom} = \text{df} )</td>
</tr>
<tr>
<td>( \text{Value of Error} = t_{0.025, df} )</td>
</tr>
<tr>
<td>( \text{Pooled Standard Error} = \text{SE} )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculations for Hypothesis Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0: p_1 = p_2 )</td>
</tr>
<tr>
<td>( H_1: p_1 \neq p_2 )</td>
</tr>
<tr>
<td>( \text{Test Statistic} = z )</td>
</tr>
<tr>
<td>( \text{Critical Value (df)} = \text{NORM.S.INV}(0.05) )</td>
</tr>
<tr>
<td>( \text{p-value} = \text{NORM.S.DIST}(</td>
</tr>
</tbody>
</table>

|因为在零假设是有效的，我们需要接受它。 |
|因为误差是重要的，所以我们不能接受它。 |
|我们不能接受零假设。 |

**Advantages of z** method:
- Can do paired or one tail test.
- Can only compare two proportions.

**Disadvantages of z** method:
- Can do paired or one tail test.
- Can only compare two proportions (though more than one).
Hypothesis Test to check Equality of 2 or more population proportions using chi-square

\[ \chi^2 \] New Distribution

Chi-square ("kie" like "pie") Distribution

\[ \chi^2 \] "Chi-Square" Test Statistic

* using sample data, \( \chi^2 \) test statistic can be used to test whether 2 or more population proportions are equal.

* \( \chi^2 \) can be used to check Equality of 2 pop. proportions, but only for a 2-tail Test.

* To calculate the \( \chi^2 \) test statistic and p-value for \( \chi^2 \) we will compare "Observed Frequencies" for our proportions to the calculated "Expected Frequencies" For our proportions in a multiple step process that uses remember: "Cross Tab" tables (chapter 2) also know as "Contingency tables".

* The \( \chi^2 \) test statistic calculations for testing 2 or more pop. proportions in this section is similar to the calculations for a "Test of Independence" in the next section. The difference is that "Test of Independence" has only 1 population, whereas "testing 2 or more pop. proportions" has multiple populations.

* If you are checking Equality of 2 pop. proportions & you do Z method & \( \chi^2 \) method, then: \((Z \text{ test statistic})^2 = \chi^2\)
Variables for \( \chi^2 \) Hypothesis Test of Equality of 2 or more population proportions

\[ p_1 = \text{population 1 proportion} \]
\[ p_2 = \text{population 2 proportion} \]
\[ \vdots \]
\[ p_k = \text{population K proportion} \]

\( K = \# \text{ of populations} \)

\( r = \# \text{ of rows in "cross tab" table or } \# \text{ of categorical variables} \)

\[ \overline{p}_1 = \bar{p}_{1i} = \text{sample proportion from pop. 1} \]
\[ \overline{p}_2 = \bar{p}_{2i} = \text{sample proportion from pop. 2} \]
\[ \overline{p}_k = \bar{p}_{ki} = \text{sample proportion from pop. K} \]

\( \text{df} = \text{Degrees of Freedom} \)

\( f_{ij} = \text{observed frequency for Row i, column j in "cross tab" table} \)

\( E_{ij} = \text{expected frequency for Row i, column j} \)

\( \chi^2 = \text{calculated Test statistic} \)

\( \chi^2_{\alpha} = \text{critical value (Hurdle)} \)

\( p\text{-value} = P(\geq \chi^2) \)
Example of Hypothesis Test for Equality of 2 or more Population Proportions

Question: Executive of HMO wanted to check the accuracy of patient contact information. A sample of patient records was taken for each of the years: 2011, 2012, 2013, 2014. Are error rates the same for each year?

<table>
<thead>
<tr>
<th>Errors/Years</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>39</td>
<td>43</td>
<td>45</td>
<td>41</td>
<td>168</td>
</tr>
<tr>
<td>No</td>
<td>304</td>
<td>223</td>
<td>271</td>
<td>341</td>
<td>1139</td>
</tr>
<tr>
<td>Total</td>
<td>343</td>
<td>266</td>
<td>316</td>
<td>382</td>
<td>1307</td>
</tr>
</tbody>
</table>

Sample sizes
- 343
- 266
- 316
- 382

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Overall P(yes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>( \frac{39}{343} = 0.114 )</td>
<td>( \frac{43}{266} = 0.162 )</td>
<td>( \frac{45}{316} = 0.142 )</td>
<td>( \frac{41}{382} = 0.107 )</td>
<td>( \frac{168}{1307} = 0.129 )</td>
</tr>
</tbody>
</table>

* If each \( \bar{p} \) (sample proportion) was equal to 0.129, then there would be zero difference (All Equal).
* Remember, some difference (sampling error) is acceptable. We need to check for significant difference.
The essence of how we will check to see if there is a significant difference in the proportions is by comparing:

- Observed Frequencies
- Expected Frequencies

<table>
<thead>
<tr>
<th>Errors / Years</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>39</td>
<td>43</td>
<td>45</td>
<td>41</td>
<td>168</td>
</tr>
<tr>
<td>No</td>
<td>309</td>
<td>223</td>
<td>271</td>
<td>341</td>
<td>1139</td>
</tr>
<tr>
<td>Total</td>
<td>348</td>
<td>266</td>
<td>316</td>
<td>382</td>
<td>1307</td>
</tr>
</tbody>
</table>

\[ \text{Remember, overall } p(\text{yes}) = \frac{168}{1307} = 0.129 \]

\[ \text{observed frequency of "Yes" in 2013} = 45 \]

\[ \text{Sample size in 2013} = 316 \]

\[ 2013 \overline{p}_3 = \frac{45}{316} = 0.142 \]

If \( p(\text{yes in 2013}) \) were to equal 0.129 then:

\[ \left\{ \begin{array}{l}
\text{Sample size in 2013} \\
0.129 = 316 \times 0.129 = 40.6
\end{array} \right. \]

\[ \text{check} \frac{40.6}{316} = 0.129 \]
5 steps for Hypothesis Test to check Equality of 2 or more population proportions

**Step 1**

**Null & Alternative Hypothesis:**

- **H₀:** \( p_1 = p_2 = \cdots = p_k \)
- **Hₐ:** Not all population proportions are equal

**Step 2a**

**Alpha** \( \alpha = \text{Risk that we reject } H₀ \text{ when it was TRUE.} \)

**Step 2b**

Select Random Samples from each of the populations & create "Cross Tab" table with observed Frequencies.

Example:

"observed Frequencies" = \( f_{ij} \)

\( i = \text{Row Number} = \text{Categorical variable number} \)

\( j = \text{Column Number} = \text{Population number} \)

<table>
<thead>
<tr>
<th>( j )</th>
<th>Pop.1</th>
<th>Pop.2</th>
<th>Pop.3</th>
<th>Pop.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( f_{11} )</td>
<td>( f_{12} )</td>
<td>( f_{13} )</td>
<td>( f_{14} )</td>
</tr>
<tr>
<td>2</td>
<td>( f_{21} )</td>
<td>( f_{22} )</td>
<td>( f_{23} )</td>
<td>( f_{24} )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Errors / Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>39</td>
<td>43</td>
<td>45</td>
<td>41</td>
<td>168</td>
</tr>
<tr>
<td>No</td>
<td>309</td>
<td>223</td>
<td>271</td>
<td>341</td>
<td>1,139</td>
</tr>
<tr>
<td>Total</td>
<td>343</td>
<td>266</td>
<td>316</td>
<td>382</td>
<td>1,307</td>
</tr>
</tbody>
</table>

\( f_{13} = 45 = \text{observed frequency} \text{ Row 1, Column 3} \)
Step 3: Assume $H_0$ TRUE and compute Expected Frequencies $E_{ij}$

$E_{ij} = "Expected\ Frequencies"$

$i = Row\ Number = categorical\ Variable\ Number$

$j = Column\ Number = population\ Number$

$$E_{ij} = \frac{\text{Row } i \text{ Total}}{\text{Sum of all Sample Sizes}} \times (\text{Column } j \text{ Total})$$

<table>
<thead>
<tr>
<th>Observed</th>
<th>(j=1)</th>
<th>(j=2)</th>
<th>(j=3)</th>
<th>(j=4)</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Errors/Year</td>
<td>2011</td>
<td>2012</td>
<td>2013</td>
<td>2014</td>
<td>Total</td>
</tr>
<tr>
<td>Yes</td>
<td>39</td>
<td>43</td>
<td>45</td>
<td>41</td>
<td>168</td>
</tr>
<tr>
<td>No</td>
<td>304</td>
<td>223</td>
<td>271</td>
<td>341</td>
<td>1139</td>
</tr>
<tr>
<td>Total</td>
<td>343</td>
<td>266</td>
<td>316</td>
<td>382</td>
<td>1307</td>
</tr>
</tbody>
</table>

$E_{13} = \frac{168}{1307} \times 316 \approx 0.129 \times 316 = 40.6$

$P(\text{Yes}) = \frac{168}{1307} \\
\text{Sample size for year 2013} = 316$

Expected value = Expected count for "Yes" in 2013 if proportion same for each year

Completed Table of Expected Frequencies:

<table>
<thead>
<tr>
<th>Errors/Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>44.1</td>
<td>34.2</td>
<td>40.6</td>
<td>49.1</td>
<td>168</td>
</tr>
<tr>
<td>No</td>
<td>298.9</td>
<td>231.8</td>
<td>275.4</td>
<td>332.9</td>
<td>1139</td>
</tr>
<tr>
<td>Total</td>
<td>343</td>
<td>266</td>
<td>316</td>
<td>382</td>
<td>1307</td>
</tr>
</tbody>
</table>
check to see if all $e_{i,j} \geq 5$

**Expected Frequencies**

- Step 4a

<table>
<thead>
<tr>
<th>Errors/Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>44.1</td>
<td>34.2</td>
<td>40.6</td>
<td>49.1</td>
</tr>
<tr>
<td>No</td>
<td>298.9</td>
<td>231.8</td>
<td>275.4</td>
<td>332.9</td>
</tr>
</tbody>
</table>

TRUE All $e_{i,j} \geq 5$

**Because All $e_{i,j} \geq 5$, we can calculate:**

$\chi^2 = \sum \sum \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}}$

*Example for Row i column j:

$\frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}} = \frac{(f_{13} - e_{13})^2}{e_{13}} = \frac{(45 - 40.62)^2}{40.62} = 0.4727$

**Full Table of calculation Answers for $\chi^2$:**

<table>
<thead>
<tr>
<th>Errors</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.587</td>
<td>2.169</td>
<td>0.4727</td>
<td>1.337</td>
</tr>
<tr>
<td>No</td>
<td>0.087</td>
<td>0.3347</td>
<td>0.0697</td>
<td>0.1971</td>
</tr>
</tbody>
</table>

Adding all these numbers: $\chi^2 = 5.35$

(calculated Test statistic)
Step 4

Remaining Calculations:

\[ \alpha = \alpha = 0.05 \]

\[ k = \# \text{ of populations} = 4 \text{ years} \]

\[ df = 4 - 1 = 3 \]

\[ \chi^2_{\alpha} = \text{critical value} = \text{CHISQ.INV.RT}(\alpha, df) \]

\[ = \text{CHISQ.INV.RT}(0.05, 3) = 7.815 \]

\[ \text{P-value} = \text{CHISQ.TEST(} \text{actual range, expected range)} \]

\[ = \text{CHISQ.TEST(} \text{Table of observed frequencies, Table of expected frequencies)} \]

\[ = 0.1476 \]

\[ \text{P-value} = \text{CHISQ.DIST.RT}(X, df) \]

\[ \text{calculated } \chi^2 \text{ test statistic} = 5.35 \]

\[ = \text{CHISQ.DIST.RT}(5.35, 3) = 0.1476 \]

Step 5

Rules same as with previous Hypothesis Tests:

P-value \( \leq \alpha \), then Reject \( H_0 \) & Accept \( H_a \)

Critical value \( \geq \) Calculated Test Statistic, Reject.

Conclusion:

Statistical evidence does not suggest a significant difference between pop. proportions.
Question: Executive at HMO wanted to check Patient Record Data accuracy. An independent and random sample of medical records was taken for each of the four years.

Step 1: \[ H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4 \]
Chi-Square method can only do a two-tail test
Because of the way the calculations are set up, it will always be a test on the upper end (test statistic is always \( \geq 0 \))

Step 2: \[ \alpha = 0.05 \]
\[ i = \text{row} \]
\[ j = \text{column} \]

Step 3: Observed Frequencies: \( f_{ij} \)

<table>
<thead>
<tr>
<th>Errors in Form/Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>168</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>304</td>
<td>43</td>
<td>45</td>
<td>41</td>
<td>1139</td>
</tr>
<tr>
<td>Total</td>
<td>472</td>
<td>118</td>
<td>45</td>
<td>41</td>
<td>1307</td>
</tr>
</tbody>
</table>

Sample Proportions of Error Rate:

<table>
<thead>
<tr>
<th>Errors in Form/Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Overall Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.113</td>
<td>0.161</td>
<td>0.142</td>
<td>0.107</td>
<td>0.1285</td>
</tr>
<tr>
<td>No</td>
<td>0.887</td>
<td>0.839</td>
<td>0.858</td>
<td>0.893</td>
<td>0.8716</td>
</tr>
</tbody>
</table>

Expected Frequencies:

<table>
<thead>
<tr>
<th>Errors in Form/Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>( e_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>44.09</td>
<td>74.19</td>
<td>40.61</td>
<td>40.10</td>
<td>( 0.1285 \times 1139</td>
</tr>
<tr>
<td>No</td>
<td>289.91</td>
<td>231.81</td>
<td>275.38</td>
<td>332.85</td>
<td>( 0.8716 \times 1139</td>
</tr>
<tr>
<td>Total</td>
<td>334</td>
<td>245</td>
<td>316</td>
<td>374</td>
<td></td>
</tr>
</tbody>
</table>

Deviation Squared:

<table>
<thead>
<tr>
<th>Errors in Form/Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>( \sum (f_{ij} - e_{ij})^2 / e_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>25.89</td>
<td>77.59</td>
<td>19.20</td>
<td>65.64</td>
<td>1.3368</td>
</tr>
<tr>
<td>No</td>
<td>25.89</td>
<td>77.59</td>
<td>19.20</td>
<td>65.64</td>
<td>1.3368</td>
</tr>
<tr>
<td>Total</td>
<td>51.78</td>
<td>155.18</td>
<td>38.40</td>
<td>131.28</td>
<td>2.6736</td>
</tr>
</tbody>
</table>

Deviations Squared/Expected Frequencies: Final Calculations to add to get Chi-Square

<table>
<thead>
<tr>
<th>Errors in Form/Year</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>5.35</td>
<td>5.35</td>
<td>5.35</td>
<td>5.35</td>
<td>21.40</td>
</tr>
<tr>
<td>No</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.36</td>
</tr>
<tr>
<td>Total</td>
<td>5.44</td>
<td>5.44</td>
<td>5.44</td>
<td>5.44</td>
<td>21.76</td>
</tr>
</tbody>
</table>

Step 4: Test Statistic \( \chi^2 = \frac{\sum (f_{ij} - e_{ij})^2}{e_{ij}} \)

<table>
<thead>
<tr>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.40</td>
</tr>
</tbody>
</table>

SUM of ALL \( \chi^2 = 21.40 \)

Because \( p \)-value is bigger than \( \alpha \) and the test statistic is not past our critical value, we fail to reject \( H_0 \).
It is reasonable to assume that there is not a difference between the population proportions.
The statistical evidence suggests that there is not a significant difference between the population proportions, that is the error rate for each one of the years is not significantly different. There is a 5% risk that we have concluded that there is a difference, when in fact there is a difference.
Chi-Square method can only do a two-tail test (that is, it uses "Equal" and Alternative uses "Not Equal")

Chi-Square can only do a two-tail test (that is, it uses "Equal" and Alternative uses "Not Equal")

Because of the way calculations are set up, it will always be a test on the upper end (test statistic is always \( > 0 \))

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

**Desired Frequencies \( f_i \):**

<table>
<thead>
<tr>
<th>Year</th>
<th>Sex</th>
<th>Unemployed</th>
<th>Employed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>No</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2013</td>
<td>No</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2015</td>
<td>No</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2017</td>
<td>Yes</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

**Observed Frequencies \( f_i \):**

<table>
<thead>
<tr>
<th>Year</th>
<th>Sex</th>
<th>Unemployed</th>
<th>Employed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>No</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2013</td>
<td>No</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2015</td>
<td>No</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2017</td>
<td>Yes</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

**Sample Proportion of Error Rates:**

\[
\hat{p} = \frac{\text{Number of Errors}}{\text{Total Observations}}
\]

**Expected Frequencies:**

\[
E_{ij} = \frac{R_i \times C_j}{N}
\]

**Deviations Squared:**

\[
\sum (O - E)^2
\]

**Sum of All \( (O_i - E_i)^2 / E_i \):**

Because \( p \)-value is bigger than \( \alpha \) and the test statistic \( \chi^2 \) is not past our critical value, we fail to reject \( H_0 \).

It is reasonable to assume that there is no difference between the population proportions.

The statistical evidence suggests that there is no significant difference between the population proportions, that is the error rates at the different offices is not significantly different.

There is a 5% risk that we have concluded that there is a difference, when in fact there is a difference.
Test of Independence of 2 Categorical Variables using $\chi^2$ chi-square ("kie" like pie) (p. 20)

* using Sample Data, we test if 2 categorical variables sampled from 1 population are independent (not dependent or associated).

* Example:

1. Researchers want to determine if there is a difference in "Hiring Plans over the Next Year" based on the "Type of Firm" Private Firm or Public Firm. (2 categorical variables)

2. Human Resource Executives were surveyed and asked about their hiring plans over next year. The survey looked like this:

   1. what are your hiring plans over next year? select one:
      - Hiring ___
      - Not Hiring ___
      - Lay off workers ___

   2. what sort of firm do you manage? select one:
      - Private ___
      - Public ___

   one possible survey result might look like this:

   1. what are your hiring plans over next year? select one:
      - Hiring X
      - Not Hiring ___
      - Lay off workers ___

   2. what sort of firm do you manage? select one:
      - Private ___
      - Public X
From the random sample, our survey results would look like:

<table>
<thead>
<tr>
<th>Hiring Plans over Next Year</th>
<th>Firm Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiring</td>
<td>Public</td>
</tr>
<tr>
<td>Hiring</td>
<td>Private</td>
</tr>
<tr>
<td>Not Hiring</td>
<td>Public</td>
</tr>
<tr>
<td>Lay off Workers</td>
<td>Private</td>
</tr>
</tbody>
</table>

2 categorical variables from 1 population of executives

If the objective of the study is to determine if there is a difference in hiring over next year based on firm type when we cross-tabulate to get observed frequencies, we must:

1. "Must put "Response Variable" (Row Variable)"
2. "Hiring plans over next year"

Explanatory variable (column variable) "Firm Type"

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Private</th>
<th>Public</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiring</td>
<td>40</td>
<td>32</td>
<td>72</td>
</tr>
<tr>
<td>Not Hiring</td>
<td>16</td>
<td>32</td>
<td>48</td>
</tr>
<tr>
<td>Lay off Employees</td>
<td>16</td>
<td>46</td>
<td>62</td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>110</td>
<td>182</td>
</tr>
</tbody>
</table>
If we look at observed frequencies & proportions for just the "Response Variable":

<table>
<thead>
<tr>
<th>Hiring Plans over Next Year</th>
<th>Observed Frequencies</th>
<th>Overall Sample Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiring</td>
<td>72</td>
<td>40%</td>
</tr>
<tr>
<td>Not Hiring</td>
<td>48</td>
<td>26%</td>
</tr>
<tr>
<td>Lay off Employees</td>
<td>62</td>
<td>34%</td>
</tr>
<tr>
<td>Total</td>
<td>182</td>
<td>100%</td>
</tr>
</tbody>
</table>

we can think of checking Independence of 2 categorical variables this way: "Do the proportions for the "response variable" significantly differ when a second "explanatory variable" is added?"

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Private</th>
<th>Public</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiring</td>
<td>56%</td>
<td>29%</td>
<td>40%</td>
</tr>
<tr>
<td>Not Hiring</td>
<td>22%</td>
<td>29%</td>
<td>26%</td>
</tr>
<tr>
<td>Lay off workers</td>
<td>22%</td>
<td>42%</td>
<td>34%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

* is there a significant change? * if there was zero change, they would all be the same as overall sample proportion.
6 ways we can ask Independence Question:

* Is the Response variable "Hiring Plans over Next Year" independent of the Explanatory variable "Firm Type"?

* Is the Response variable "Hiring Plans over Next Year" dependent on the Explanatory variable "Firm Type"?

* Is there an association between the Response variable "Hiring Plans over Next Year" and the Explanatory Variable "Firm Type"?

7 possible conclusions:

1. If 2 categorical variables are independent:
   * "Hiring Plans over Next Year" will not depend on "Firm Type" and the proportions for Hiring, Not Hiring and Laying off workers will be same for Public & Private Firms.

2. If 2 categorical variables are NOT Independent:
   * Proportions for Hiring, Not Hiring & Lay off workers should differ by Firm.
   * After our Chi-square test we will have evidence that "Hiring Plans over Next Year" variable is associated with or dependent on "Firm Type".
   * We can gain more insights by comparing Sample Proportions and making charts.
chi-square Test will involve some formulas as last section. We will compare "observed frequencies" to Expected Frequencies in a "cross tabulated table" sometimes called a "contingency table".

* Assume Ho TRUE & compute Expected Frequencies

\[ e_{ij} = \frac{\text{Row } i \text{ Total}}{\text{Sample size}} \times \frac{\text{Column } j \text{ Total}}{\text{Sample size}} \]

* IF ALL \( e_{ij} \geq 5 \) then:

\[ \chi^2 = \sum \sum \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \]
4. chi-square Test for Independence of 2 categorical variables (contingency Table Test)

1. **Null & Alternative:**
   - Ho: The Response Variable (Row variable) is independent of the Explanatory Variable (column variable).
   - Ha: The Response Variable (Row variable) is not independent of the Explanatory Variable (column variable).

2. Select alpha. Select random sample from one population and collect data for both variables for every element in sample. Explanatory variable is the column variable & the response variable is the row variable for a cross tabulated frequency table w/ observed frequencies = \( f_{ij} \), where \( i = \text{row} \) and \( j = \text{column} \).

3. Assume Ho TRUE & compute expected frequencies:
   \[
   E_{ij} = \frac{\text{Row } i \text{ Total}}{\text{Sample size}} \times \frac{\text{Column } j \text{ Total}}{\text{Sample size}}
   \]

4. If all \( E_{ij} \geq 5 \), then compute chi-square test statistic:
   \[
   \chi^2 = \sum \sum \frac{(f_{ij} - E_{ij})^2}{E_{ij}}
   \]

5. **Rejection Rules**
   - Reject Ho & Accept Ha if \( p\text{-value} \leq \alpha \)
   - Reject Ho & Accept Ha if \( \text{Test statistic} \chi^2 \geq \text{critical value} \chi^2 \)
   \[
   df = (r-1) \times (c-1), \text{ where } r = \# \text{ Rows} \quad \text{c = \# Columns}
   \]

6. IF variables are NOT independent, investigate association by calculating Pbars or creating chart.
Survey asked human resource executives about their firms hiring plans over the next year. The survey offered three nominal "Hiring Over Next Year" categories: Hiring, Not Hiring, Lay off employees. The survey also asked whether the firm was public or private. We will call this nominal variable "Firm Type".

At the 0.05 alpha level, is the "Hiring Over Next Year" Nominal Variable independent of the "Firm Type" Nominal Variable?

Categorical variables, or more specifically, nominal variables are:
- "Firm Type" will be the "Explanatory Variable" = Column Variable
- "Hiring Over Next Year" is the "Response Variable" = Row Variable

Determine ways to ask questions about independence:
- Is the "Hiring Over Next Year" variable independent of the "Firm Type" variable?
- Does the "Firm Type" variable influence/affection bias the "Hiring Over Next Year" variable?
- Is the "Hiring Over Next Year" variable dependent on the "Firm Type" variable?

Conduct a Test of Independence or Contingency Table Test to determine if the "Hiring Over Next Year" variable is independent of the "Firm Type" variable.

Create "Cross Tab" or "Contingency Tables" for the Observed frequencies and the Expected frequencies.

**Step 1:**
- **H0**: the "Hiring Over Next Year" variable is independent of the "Firm Type" variable.
- **H1**: the "Hiring Over Next Year" variable is NOT independent of the "Firm Type" variable.

**Step 2:**
- Alpha (α) = 0.05

**Step 3:**
- Observed Frequencies:

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Hiring</th>
<th>Not Hiring</th>
<th>Lay Off Workers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private</td>
<td>32</td>
<td>12</td>
<td>8</td>
<td>52</td>
</tr>
<tr>
<td>Public</td>
<td>11</td>
<td>11</td>
<td>8</td>
<td>30</td>
</tr>
</tbody>
</table>

**Step 4:**
- For proportions:

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Private</th>
<th>Public</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiring</td>
<td>0.56</td>
<td>0.29</td>
<td>0.40</td>
</tr>
<tr>
<td>Not Hiring</td>
<td>0.40</td>
<td>0.71</td>
<td>0.66</td>
</tr>
<tr>
<td>Lay Off</td>
<td>0.44</td>
<td>0.43</td>
<td>0.46</td>
</tr>
<tr>
<td>Total</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Expected Frequencies:**

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Private</th>
<th>Public</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiring</td>
<td>10.45</td>
<td>43.51</td>
<td>54.96</td>
</tr>
<tr>
<td>Not Hiring</td>
<td>29.01</td>
<td>29.01</td>
<td>58.02</td>
</tr>
<tr>
<td>Lay Off</td>
<td>12.52</td>
<td>12.52</td>
<td>25.04</td>
</tr>
<tr>
<td>Total</td>
<td>42.00</td>
<td>42.00</td>
<td>84.00</td>
</tr>
</tbody>
</table>

**Step 6:**
- All cells ≥ 5?
- TRUE

**Calculations for Chi-Square Test Statistic:**

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Private</th>
<th>Public</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiring</td>
<td>4.564</td>
<td>3.017</td>
<td>7.581</td>
</tr>
<tr>
<td>Not Hiring</td>
<td>0.472</td>
<td>0.309</td>
<td>0.781</td>
</tr>
<tr>
<td>Lay Off</td>
<td>3.864</td>
<td>3.864</td>
<td>7.728</td>
</tr>
</tbody>
</table>

**Chi-Square Test Statistic (χ²):**

\[
χ^2 = 11.44735 + 13.38791495 = 24.83527
\]

**df = (r-1)(c-1)**

\[\text{df}=2\]

**p-value:**

0.001238

**Chi-Square Critical Value:**

5.991465

**Step 5:**
- Because the p-value is much smaller than our alpha, and because the calculated test statistic is much bigger than our critical value, we reject H0 and accept H1.

The sample evidence suggests that the "Hiring Over Next Year" variable is **NOT** independent of the "Firm Type" variable.

It is reasonable to assume that the proportions for Hiring Over Next Year Over are different for Private and Public Firms.

It is reasonable to assume that the "Hiring Over Next Year" variable is dependent on the "Firm Type" variable.

It is reasonable to assume that the "Firm Type" variable influences/affects the "Hiring Over Next Year" variable.

We do run a 5% risk of reject H0 when in fact it was true.
Survey asked human resource executives about their firms hiring plans over the next year. The survey offers three nominal 'Hiring Over Next Year' categories: Hiring, Not hiring, Lay off employees. The survey also asks whether the firm was public or private. We will call this nominal variable, 'Firm Type'.

At the 0.05 alpha level, is the 'Hiring Over Next Year' Nominal Variable Independent of the 'Firm Type' Nominal Variable?

Categorical variables are more specifically, Nominal Variables are:

- 'Firm Type' will be the 'Explanatory Variable' or 'Column Variable'.
- 'Hiring Over Next Year' is the 'Response Variable' or 'Row Variable'.

Determine whether the 'Firm Type' variable independent of the 'Hiring Over Next Year' variable does the 'Firm Type' variable influence affect the 'Hiring Over Next Year' variable?

Is the 'Hiring Over Next Year' variable dependent on the 'Firm Type' variable?

Conduct a x2 Test of Independence for Contingency Table Test to determine if the 'Hiring Over Next Year' variable is independent of the 'Firm Type' variable.

Create Cross Tab or Contingency Tables for the 'Observed Frequencies' and the 'Expected Frequencies'.

**H0:** the 'Hiring Over Next Year' variable independent of the 'Firm Type' variable.  
**H1:** the 'Hiring Over Next Year' variable not independent of the 'Firm Type' variable.

**Step 1:**

1. **Observed Frequencies**

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Hiring</th>
<th>Not Hiring</th>
<th>Lay Off Workers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private</td>
<td>223</td>
<td>79</td>
<td>52</td>
<td>354</td>
</tr>
<tr>
<td>Public</td>
<td>123</td>
<td>124</td>
<td>56</td>
<td>303</td>
</tr>
<tr>
<td>Total</td>
<td>346</td>
<td>203</td>
<td>108</td>
<td>657</td>
</tr>
</tbody>
</table>

2. **Expected Frequencies**

<table>
<thead>
<tr>
<th>Firm Type</th>
<th>Hiring</th>
<th>Not Hiring</th>
<th>Lay Off Workers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private</td>
<td>177.5</td>
<td>71.5</td>
<td>35</td>
<td>284</td>
</tr>
<tr>
<td>Public</td>
<td>68.5</td>
<td>122.5</td>
<td>43</td>
<td>234</td>
</tr>
<tr>
<td>Total</td>
<td>246</td>
<td>194</td>
<td>78</td>
<td>522</td>
</tr>
</tbody>
</table>

**Step 2:**

1. **Chi-Square Test Statistics**

   - **Private**
     - Hiring: 12.528
     - Not Hiring: 5.789
     - Lay Off Workers: 1.307
     - Total: 3.841
   - **Public**
     - Hiring: 5.789
     - Not Hiring: 12.528
     - Lay Off Workers: 1.307
     - Total: 3.841
   - **Total**
     - Hiring: 18.317
     - Not Hiring: 18.317
     - Lay Off Workers: 2.614
     - Total: 7.522

**Step 3:**

1. **Calculations for Chi-Square Test**

   - **Private**
     - Hiring: \( \chi^2 = 12.528 \)
     - Not Hiring: \( \chi^2 = 5.789 \)
     - Lay Off Workers: \( \chi^2 = 1.307 \)
     - Total: \( \chi^2 = 3.841 \)
   - **Public**
     - Hiring: \( \chi^2 = 5.789 \)
     - Not Hiring: \( \chi^2 = 12.528 \)
     - Lay Off Workers: \( \chi^2 = 1.307 \)
     - Total: \( \chi^2 = 3.841 \)
   - **Total**
     - Hiring: \( \chi^2 = 18.317 \)
     - Not Hiring: \( \chi^2 = 18.317 \)
     - Lay Off Workers: \( \chi^2 = 2.614 \)
     - Total: \( \chi^2 = 7.522 \)

Because the p-value is much smaller than our alpha, and because the calculated test statistic is much bigger than our critical value, we reject H0 and accept H1.

The sample evidence suggests that the "Hiring Over Next Year" variable is NOT independent of the "Firm Type" variable.

It is reasonable to assume that the proportions for Hiring Over Next Year are different for Private and Public Firms. It is reasonable to assume that the "Hiring Over Next Year" variable is dependent on the "Firm Type" variable. It is reasonable to assume that the "Firm Type" variable influences affects the "Hiring Over Next Year" variable.

We do not have a 95% level of reject H0 when in fact it was TRUE.
5. **Chi-square Distribution & Chi-square Test Statistic**

   can be used in 3 situations:

2. **Hypothesis Test for 2 or More Population Proportions**

   used to test whether 2 or more pop. proportions are equal.

   - **Expected Frequencies under Assumption**: $E_{ij} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Sum of all sample sizes}}$
   - **Chi-square Test Statistic if All $E_{ij} \geq 5$**:
     $$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{(f_{ij} - E_{ij})^2}{E_{ij}}$$

   - Critical value for each combination of 2 proportions - done if test says there is a difference

3. **Test of Independence "Contingency Table Test"**

   used to test if 2 categorical variables sampled from 1 population are independent. Does a set of proportions from the response variable (row variable) significantly differ when a second explanatory variable (column variable) is added?

   - **df** = $(r-1) \times (c-1)$
   - **Expected Frequency** for category $i$:
     $$E_{ij} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Sample Size}}$$

3. **Goodness of Fit Test** (for nominal data)

   used to determine whether a random variable has a specific probability distribution. Check to see if observed frequencies for a multinomial probability distribution are equal to the expected frequencies.

   - **Chi-square Test Statistic if All $E_i \geq 5$**:
     $$\chi^2 = \sum_{i=1}^{K} \frac{(f_i - E_i)^2}{E_i}$$

   - **df** = $K - 1$
Independence Defined:

1. Probability Theory (Chapter 4):
   If occurrence of one event does not affect the probability of the other event.

2. Probability Distributions (Chapter 5 on...)
   Random variables are independent if the occurrence of one does not affect the probability distribution of the other.