1 Comparing Two population Proportion (Large sample)(p464)

<table>
<thead>
<tr>
<th></th>
<th>Forgot Key Point</th>
<th>Remember it</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial A</td>
<td>43</td>
<td>37 (X)</td>
<td>80 ($n_1$)</td>
</tr>
<tr>
<td>Commercial B</td>
<td>52</td>
<td>18 (Y)</td>
<td>70 ($n_2$)</td>
</tr>
<tr>
<td>Total</td>
<td>95</td>
<td>55</td>
<td>150</td>
</tr>
</tbody>
</table>

Let $p \in [0, 1]$, such that

\[ p_1 = \text{success rate (Remember key point of A)} \]
\[ p_2 = \text{success rate (Remember key point of B)} \]

Researcher’s interest:

C1) Commercial A is better than Commercial B?
\[ p_1 - p_2 > 0 \]
C2) Commercial A is worth than Commercial B?
\[ p_1 - p_2 < 0 \]
C3) Commercial A is different from Commercial B?
\[ p_1 - p_2 \neq 0 \]
Parameter of interest: $p_1 - p_2$

Estimator (your guess): $\hat{p}_1 - \hat{p}_2$ where $\hat{p}_1 = \frac{X}{n_1}$, $\hat{p}_2 = \frac{Y}{n_2}$.

Test Statistic

$$Z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})}\sqrt{1/n_1 + 1/n_2}} \quad (1)$$

where

$$\hat{p} = \frac{X + Y}{n_1 + n_2} \quad (2)$$

This is called **Pooled estimator**. It is same as $p_1$ under $H_0$. Hence it is same as $p_2$ under $H_0$.

C1: Reject $H_0$ if $z_{\alpha} < Z_{obs}$

C2: Reject $H_0$ if $z_{\alpha} > Z_{obs}$

C3: Reject $H_0$ if $z_{\alpha/2} < |Z_{obs}|$

$100(1 - \alpha)\%$ CI of $p_1 - p_2$: (p465)

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \quad (3)$$
Researchers interest is whether \( p_1 - p_2 = 0 \) or not.

1) Write the hypothesis.
2) Give the estimate of \( \text{Var}(\hat{p}_1 - \hat{p}_2) \) under \( H_0 \).
3) Test statistic?
4) Conclusion?
5) P-value?
6) Find 95% CI.

## 2 Ch 11.4 Chi-Square Test (p475)

We learned testing of 2 input categories \( \times \) 2 output categories. **Chi-Square Test** is for testing of \( k \) input categories \( \times \) \( m \) output categories (more general form). Let’s take a look at the example below. This is 2 Commercial comparison but now the output (answer) has 3 categories. (Remember, we are treating categorical output, not continuous.). Each sample size is already determined.

<table>
<thead>
<tr>
<th></th>
<th>Forget</th>
<th>Somehow Remember</th>
<th>Remember</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commrcl A</td>
<td>19</td>
<td>24</td>
<td>37</td>
<td>80 ( (n_1) )</td>
</tr>
<tr>
<td>Commrcl B</td>
<td>24</td>
<td>28</td>
<td>18</td>
<td>70 ( (n_2) )</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>52</td>
<td>55</td>
<td>150 ( (n) )</td>
</tr>
</tbody>
</table>

This is called 2 \( \times \) 3 **contingency table**.
We define parameters as follows.

<table>
<thead>
<tr>
<th></th>
<th>Forget</th>
<th>Somehow Remember</th>
<th>Remember</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commrcl A</td>
<td>$p_{A1}$</td>
<td>$p_{A2}$</td>
<td>$p_{A3}$</td>
<td>1</td>
</tr>
<tr>
<td>Commrcl B</td>
<td>$p_{B1}$</td>
<td>$p_{B2}$</td>
<td>$p_{B3}$</td>
<td>1</td>
</tr>
</tbody>
</table>

where, $p_{A1} + p_{A2} + p_{A3} = 1$, and $p_{B1} + p_{B2} + p_{B3} = 1$.

Researchers interest:
Commercial A’s effect = Commercial B’s effect?

Using mathematical notation, it can be stated as:

$H_0 : p_{A1} = p_{B1}$ AND $p_{A2} = p_{B2}$ AND $p_{A3} = p_{B3}$.

$H_1 : H_0$ is not true.

Q) Write $H_1$ more explicit way, using set operation. (We learned it long time ago!)

We do this hypothesis test with level of significance $\alpha = 0.05$. In this chapter, there is only one $H_1$. As usual, next step is decide test statistic.

Test Statistic: $\chi^2$ (chi-square statistic)

Step1) Guess population proportion of
\( p_1 = \text{probability people choose "Forget"} \)
\( p_2 = \text{probability people choose "Somehow Remember"} \)
\( p_3 = \text{probability people choose "Remember"} \)

under \( H_0 \).

Using the data, we find

\[
\hat{p}_1 = \frac{43}{150} \quad \hat{p}_2 = \frac{100}{150} \quad \hat{p}_3 = \frac{27}{150}
\]  \tag{4}

Step2) Find expected Frequency. Each entry is named as \( E_1, E_2, \ldots, E_6 \).

<table>
<thead>
<tr>
<th></th>
<th>Forget</th>
<th>Somehow</th>
<th>Remember</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>80(\hat{p}_1 = 22.93)</td>
<td>80(\hat{p}_2 = 27.73)</td>
<td>80(\hat{p}_3 = 29.33)</td>
<td>80</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td>70</td>
</tr>
</tbody>
</table>

Step3) Get

\[
\chi^2_{obs} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \]  \tag{5}

where \( O_i \) is actual observation for cell \( i \). See the table below. Each cell’s entry has ”\( O_i(E_i) \)” (p477 Table 7(a),(b))

\[
\chi^2_{obs} = \]
<table>
<thead>
<tr>
<th></th>
<th>Forget</th>
<th>Somehow</th>
<th>Remember</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>19 (22.93)</td>
<td>24 (27.73)</td>
<td>37 (29.33)</td>
</tr>
<tr>
<td>B</td>
<td>24 (20.07)</td>
<td>28 (24.27)</td>
<td>18 (25.67)</td>
</tr>
</tbody>
</table>

Reject \( H_0 \) if \( \chi^2_{df=(r-1)(c-1), \alpha} < \chi^2_{obs} \).

Note: \( \chi^2_{df=(r-1)(c-1), \alpha} = qchisq(1 - \alpha, df = (r - 1)(c - 1)) \).

\( r= \) number of row  
\( c= \) number of column

In our case, \( df=(2-1)(3-1)=2 \). Hence  
\( qchisq(0.95, df = 2) = 5.99 \). Since \( \chi^2_{obs} = 6.816 \), we

[Picture of \( \chi^2 \) distribution]
Example 11 (p479) Survey is undertaken to determine the incidence of alcoholism in different professional groups. **Each sample size is already determined.**

Remember, each entry is $O_i$ (Observed frequency).

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>NonAlcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clergy</td>
<td>32</td>
<td>268</td>
<td>300</td>
</tr>
<tr>
<td>Educators</td>
<td>51</td>
<td>199</td>
<td>250</td>
</tr>
<tr>
<td>Executives</td>
<td>67</td>
<td>233</td>
<td>300</td>
</tr>
<tr>
<td>Merchants</td>
<td>83</td>
<td>267</td>
<td>350</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>233</strong></td>
<td><strong>976</strong></td>
<td><strong>1200</strong></td>
</tr>
</tbody>
</table>

Q) Write down hypothesis test by using the following table and do $\alpha = 0.05$ test.

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>NonAlcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clergy</td>
<td>$p_{A1}$</td>
<td>$p_{A2}$</td>
<td>1</td>
</tr>
<tr>
<td>Educators</td>
<td>$p_{B1}$</td>
<td>$p_{B2}$</td>
<td>1</td>
</tr>
<tr>
<td>Executives</td>
<td>$p_{C1}$</td>
<td>$p_{C2}$</td>
<td>1</td>
</tr>
<tr>
<td>Merchants</td>
<td>$p_{D1}$</td>
<td>$p_{D2}$</td>
<td>1</td>
</tr>
</tbody>
</table>

$H_0$:  

$H_1$: Not $H_0$

You can simplify $H_0$ (Hint: $p_{A1} + p_{A2} = 1$).
Get
\[ \hat{p}_1 = \frac{233}{1200} \]

\[ \hat{p}_2 = \]

Using formula \( E_i \) (expected observation under \( H_0 \)) is added.

\[ E_1 = 300 \times \frac{233}{1200} = 58.25 \]
\[ E_2 = 300 \times \left(\frac{976}{1200}\right) = 241.75 \]
\[ E_3 = 250 \times \frac{233}{1200} = 48.54 \]

etc.

<table>
<thead>
<tr>
<th></th>
<th>Alcoholic</th>
<th>NonAlcoholic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clergy</td>
<td>32 (58.25)</td>
<td>268 (241.75)</td>
<td>300</td>
</tr>
<tr>
<td>Educators</td>
<td>51 (48.54 )</td>
<td>199 (201.46 )</td>
<td>250</td>
</tr>
<tr>
<td>Executives</td>
<td>67 (58.25 )</td>
<td>233 (241.75 )</td>
<td>300</td>
</tr>
<tr>
<td>Merchants</td>
<td>83 (67.96 )</td>
<td>267 (282.04 )</td>
<td>350</td>
</tr>
<tr>
<td>Total</td>
<td>233</td>
<td>976</td>
<td>1200</td>
</tr>
</tbody>
</table>
After calculation, you get
\[ \chi^2_{obs} = (32 - 58.25)^2/58.25 + (268 - 241.75)^2/241.75 + ... = 20.59. \]

Cutoff value (Critical point):

\[ \chi^2_{df=3, \alpha=0.05} = qchisq(0.95, df = 3) = 7.81. \]

Conclusion:
3 Ch 11.5 Contingency Table with Neither Margin Fixed (p487)

When two traits are observed for each element or unit of a random sample, the data can be simultaneously classified with respect to these traits. We then obtain a two-way contingency table in which neither set totals is fixed so both are random.

Example 12 (p487) A random sample of 500 persons is questioned regarding political affiliation and attitude toward a tax reform program. Below is the frequency table.

<table>
<thead>
<tr>
<th></th>
<th>Favor</th>
<th>Indifferent</th>
<th>Opposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>138</td>
<td>83</td>
<td>64</td>
<td>285($n_1$)</td>
</tr>
<tr>
<td>Republican</td>
<td>64</td>
<td>67</td>
<td>84</td>
<td>215($n_2$)</td>
</tr>
<tr>
<td>Total</td>
<td>202</td>
<td>150</td>
<td>148</td>
<td>500(n)</td>
</tr>
</tbody>
</table>

Note: $n$ is fixed but $n_1$ and $n_2$ are **NOT FIXED**. In other word, you don’t know these numbers before the experimentation. This is biggest difference between Ch 11.4 and Ch 11.5. We define parameters as follows.
\[ p_{D1} = \text{Prob( Democrat and Favor)} \] (cell probability)
\[ p_D = \text{Prob( Democrat)} \] (row marginal prob.)
\[ p_1 = \text{Prob(Favor)} \] (column marginal prob.)

Q) Relation between \( p_{R1} \) and \( p_R \)?

Q) Relation between \( p_{D2} \) and \( p_2 \)?

Q) Relation between \( p_{D1} \) and \( p_{R3} \)?

Researchers interest: Two classifications
Classification 1: ”Democrat/ Republican”
Classification 2: ”Favor/ Indifference/ Opposed”
are independent or not. Using mathematical notation, it can be expressed as

\[ H_0 : \]
\[ p_{D1} = p_{DP1} \quad \text{AND} \]
\[ p_{D2} = p_{DP2} \quad \text{AND} \]
\[ p_{D3} = p_{DP3} \quad \text{AND} \]
\[ p_{R1} = p_{RP1} \quad \text{AND} \]
\[ p_{R2} = p_{RP2} \quad \text{AND} \]
\[ p_{R3} = p_{RP3} \]

\[ H_1 : H_0 \text{ is not true.} \]

Textbook (p489) expressed it by

\[ H_0 : \text{Each cell probability is the product of the corresponding pair of marginal probabilities.} \]

Get Test statistics:

Step1) Guess marginal probability

\[ \hat{P}_D = \frac{285}{500} \quad \hat{P}_R = \frac{215}{500} \quad (6) \]

\[ \hat{P}_1 = \frac{202}{500} \quad \hat{P}_2 = \frac{202}{500} \quad \hat{P}_3 = \frac{202}{500} \quad (7) \]

Step2) Find expected Frequency. Each entry is named as \( E_1, E_2, ..., E_6 \).

Finally you get (Table 13b p490)
<table>
<thead>
<tr>
<th></th>
<th>Favor</th>
<th>Indifferent</th>
<th>Opposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dem</td>
<td>500(\hat{p}_D\hat{p}_1) = 115.14</td>
<td>500(\hat{p}_D\hat{p}_2) = 85.50</td>
<td>500(\hat{p}_D\hat{p}_3) = 84.36</td>
</tr>
<tr>
<td>Rep</td>
<td>500(\hat{p}_R\hat{p}_1) = 86.86</td>
<td>500(\hat{p}_R\hat{p}_2) = 64.50</td>
<td>500(\hat{p}_R\hat{p}_3) = 63.64</td>
</tr>
</tbody>
</table>

Step3)

\[
\chi^2_{obs} = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}
\]  \hspace{1cm} (8)

In our case, \(\chi^2_{obs} = 22.153\)

Reject \(H_0\) with \(\alpha = 0.05\) test, if \(\chi^2_{df=(r-1)(c-1),\alpha} < \chi^2_{obs}\)

\[
\chi^2_{df=(2-1)(3-1),0.05} = \chi^2_{df=2,0.05} = qchisq(0.95, df = 2) = 5.99.
\]

Hence, Conclusion: