Chapter 13 (continued)

Tests Concerning Proportions

1. Estimation of Proportions

Assume that we are sampling a binomial population with unknown parameter \( \theta \) (parameter of Binomial distribution is \( p \)), the maximum likelihood estimate of \( \theta \) is \( \hat{\theta} = \frac{X}{n} \) for a random sample of size \( n \). For large \( n \) the binomial distribution can be estimated with a normal distribution so that

\[
Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \sim N(0,1)
\]

This relation enables us to construct a \((1 - \alpha)\times 100\%\) confidence interval for \( \theta \) as follows:

\[
\Pr\left(-z_\alpha < \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} < z_\alpha \right) = 1 - \alpha
\]

\[
\hat{\theta} = \frac{X}{n} \Rightarrow X = n\hat{\theta}
\]

\[
\Pr\left(-z_\alpha < \frac{n\hat{\theta} - n\theta}{\sqrt{n\theta(1-\theta)}} < z_\alpha \right) = 1 - \alpha \Rightarrow \Pr\left(\hat{\theta} - \frac{z_\alpha \sqrt{\hat{\theta}(1-\hat{\theta})}}{\sqrt{n}} < \theta < \hat{\theta} + \frac{z_\alpha \sqrt{\hat{\theta}(1-\hat{\theta})}}{\sqrt{n}} \right) = 1 - \alpha
\]

The estimation of Differences between proportions:

Assume \( X \) and \( Y \) are two independent populations having Binomial distributions with parameters \( \hat{\theta}_1, \hat{\theta}_2 \), respectively. We would like to estimate the difference between the proportions of these two populations. Let \( \hat{\theta}_1 = \frac{X}{n} \) denote the sample proportion from a random sample of size \( n \) from population \( X \), and \( \hat{\theta}_2 = \frac{Y}{m} \) denote the sample proportion from a random sample of size \( m \) from population \( Y \).

The sampling distribution of \( \hat{\theta}_1 - \hat{\theta}_2 \) is approximately Normal with mean, and variance

\[
E[\hat{\theta}_1 - \hat{\theta}_2] = \theta_1 - \theta_2, \quad Var[\hat{\theta}_1 - \hat{\theta}_2] = \frac{\theta_1(1-\theta_1)}{n} + \frac{\theta_2(1-\theta_2)}{m}, \text{ respectively.}
\]
\[ Z = \frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{m}}} \sim N(0,1) \]

Therefore, \((1 - \alpha)x100\%\) confidence interval for \(\theta_1 - \theta_2\) is

\[
\Pr \left( \frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{m}}} < \theta_1 - \theta_2 < \right) = 1 - \alpha.
\]

**Tests concerning proportions:**

For large values of \(n\) we use the normal approximation to the binomial distribution to test

\[ H_0: \ \theta = \theta_0 \text{ against alternatives } H_A: \ \theta \neq \theta_0, \ \theta > \theta_0 \text{ or } \theta < \theta_0. \]

For small \(n\): Given \(\alpha\) and \(H_A: \ \theta < \theta_0\), we find the critical value \(X_c\) as \(\Pr(X \leq X_c|\theta = \theta_0) = \alpha\)

\[
\sum_{k=0}^{X_c} \binom{n}{k} \theta_0^k (1-\theta_0)^{n-k} = \alpha.
\]

For large \(n\): The test statistics \(z = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}} = \frac{x-n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \sim N(0,1)\)

We compare the value of \(z\) with \((\pm)z_{\alpha}\) or \((\pm)z_{\alpha/2}\) depending on the alternative hypothesis to reject or fail to reject the null hypothesis.

**Comparison of two population proportions:**

Let \(X\) denote a Binomial population having \((n, p_1)\) and \(Y\) denote a Binomial population having \((m, p_2)\). We concern the comparison of two population proportions in terms of their difference:

\[ p_1 - p_2 \]. Random samples from both populations are taken and the point estimator of unknown parameters yield: \(\hat{p}_1 - \hat{p}_2\).

\[ E[\hat{p}_1 - \hat{p}_2] = p_1 - p_2 \text{ is unbiased estimator of the difference}. \]

\[ \text{Var}[\hat{p}_1 - \hat{p}_2] = \frac{p_1q_1}{n} + \frac{p_2q_2}{m} \]. Then \(z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n} + \frac{p_2q_2}{m}}} \sim N(0,1)\). Here one can denote \(p_1 - p_2\) also as \(\theta_1 - \theta_2\).

Hypothesis Testing: \(H_0: \ \theta_1 = \theta_2 \text{ vs. } H_A: \ \theta_1 < \theta_2 \text{ or } H_A: \ \theta_1 > \theta_2 \text{ or } H_A: \ \theta_1 \neq \theta_2\).

This is equivalent of saying: \(H_0: \ \theta_1 - \theta_2 = 0\).
The test statistics for large \( n \) and \( m \) is:
\[
\frac{\bar{\theta}_1 - \bar{\theta}_2 - (\theta_1 - \theta_2)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \theta_i(1-\theta_i) + \frac{1}{m} \sum_{j=1}^{m} \theta_j(1-\theta_j)}}
\]
and compare with the critical value of \( Z_\alpha \) or \( Z_{\alpha/2} \).

Tests concerning differences among \( k \)-proportions:

\( H_0: \theta_1 = \theta_2 = \cdots = \theta_k = \theta_0 \) vs. \( H_A: \) at least one does not equal to \( \theta_0 \).

\[
z_i = \frac{X_i-n_i \hat{\theta}_0}{\sqrt{n_i \hat{\theta}_0(1-\hat{\theta}_0)}} \sim N(0,1)
\]

\[
z_i^2 = \left( \frac{X_i-n_i \hat{\theta}_0}{\sqrt{n_i \hat{\theta}_0(1-\hat{\theta}_0)}} \right)^2 \sim \chi^2 \Rightarrow \chi^2 = \sum z_i^2 \sim \chi^2_k.
\]

When \( \theta_0 \) is not specified and the hypothesis is:

\( H_0: \theta_1 = \theta_2 = \cdots = \theta_k = \theta_0 \) vs. \( H_A: \) at least one does not equal to each other, then,

\[
\hat{\theta}_{POOLED} = \hat{\theta} = \frac{\sum X_i}{\sum n_i}, \Rightarrow \chi^2 = \sum \frac{(X_i-n_i \hat{\theta})^2}{n_i \hat{\theta}(1-\hat{\theta})} \sim \chi^2_{k-1}
\]

Renaming the entries, given the cell frequency of \( f_{ij} = X_{ij} \), and \( e_{ij} = n_i \hat{\theta} \),

\[
\Rightarrow \chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{(f_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{k-1}.
\]

Decision: reject the null hypothesis when \( \chi^2 > \chi^2_{(k-1),\alpha} \).

Exercise 1: If 26 of 200 tires of Brand A failed to last 30,000 miles, whereas the corresponding figures for tires of brands B, C, and D were 23, 15, and 32, test the null hypothesis that there is no difference in the quality of the four kinds of tires at the 0.05 level of significance.

The analysis of an \( r \times c \) Table:

\( H_0: \theta_{1j} = \theta_{2j} = \cdots = \theta_{rj} \) vs. \( H_A: \) at least one differs in column \( j = 1, \ldots, c \) where \( i = 1, \ldots, r \) corresponds to the rows.

\[
\Rightarrow \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}.
\]

Decision: reject the null hypothesis when \( \chi^2 > \chi^2_{(r-1)(c-1),\alpha} \).

Goodness of Fit Test: This test is used to determine that the data set follow a statistical distribution.

\[
\Rightarrow \chi^2 = \sum_{i=1}^{n} \frac{(f_i-e_i)^2}{e_i} \sim \chi^2_{m-k-1},
\]

where \( m \) is the number of classes and \( k \) is the number of parameters.
**Exercise 2:** Tests of the fidelity and the selectivity of 190 radios produced results shown in the following table:

<table>
<thead>
<tr>
<th>Selectivity</th>
<th>Fidelity</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Average</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>7</td>
<td>12</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>35</td>
<td>59</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>15</td>
<td>13</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Use the 0.01 level of significance to test the null hypothesis that the fidelity is independent of selectivity.

**Exercise 3:** Each day, Monday through Saturday, a baker bakes there large chocolate cakes, and those not sold on the same day are given away to a food bank. Use the data shown below to test at the 5% level of significance whether they may be looked upon as values of a binomial random variable:

<table>
<thead>
<tr>
<th>No. of cakes sold</th>
<th>Number of days</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>228</td>
</tr>
</tbody>
</table>