Section 6.5 – Inference Methods for Proportions

Relative to Inference for a Population Proportion $p$

Consider the activity of inspecting components (widgets) to see if each does or does not have a characteristic of interest. Now imagine a population. Its elements consist of zeros and ones. Each zero identifies a component that does not have a characteristic of interest, and each one (1) identifies a component that does have the characteristic. The result of each inspection is viewed as a number selected at random from this population.

Define a random variable $X$ on this population. $X$ can assume value either 0 or 1. The probability function of $X$ is

$P(X = x) = f(x) = \begin{cases} 
 0 & \text{if } x = 0 \\
 1 - p & \text{if } x = 1 \\
 p & \text{if } x = 1
\end{cases}$

where $p$ is the proportion of 1’s in the population. Using equations (5.1) and (5.2), we compute parameters $\mu$ and $\sigma^2$ for this population as follows:

$EX = \Sigma x \cdot f(x)$
$EX = 0 \cdot (1 - p) + 1 \cdot p$
$EX = p = \mu$

$VARX = (0 - p)^2 \cdot f(0) + (1 - p)^2 \cdot f(1)$
$VARX = p^2 (1 - p) + (1 - p)^2 \cdot p$
$VARX = p(1 - p) = \sigma^2$

Now consider sampling random variables $X_1, X_2, \ldots, X_n$ each defined on this population. The random variable $n \hat{p} = \Sigma X_i$ is a Binomial $(n, p)$ random variable. Here $\hat{p}$ is the proportion of 1’s in the sample and it is the estimate of $p$. We will be interested in estimating $p$ from large samples only. Let $\bar{X} = 1/n \Sigma X_i$ be the sample mean random variable. For large $n$ the central limit theorem says $\bar{X}$ is approximately normally distributed $\bar{X} \approx N(\mu, \sigma^2/n)$. Since $\bar{X}$ estimates $p$, $\bar{X} \equiv \hat{p}$. Thus we have

$\hat{p} \equiv \bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \equiv N\left(p, \frac{p(1 - p)}{n}\right)$

so
\[
Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1).
\]

Since we do not know \(\sigma\) this is not useful in practice. However when we use \(\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})}\) then we have the practically useful (we are saying we know \(\sigma\) is \(\hat{\sigma}\))

\[
Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0, 1).
\]

(Recall we are assuming large \(n\).) Using this result we can make inferences (confidence intervals and significance tests) about the parameter \(p\) using the random sample estimator \(\hat{p} = 1/n \Sigma X_i\).

**SINGLE SAMPLE CASE**

**For Large \(n\) (5 \(\leq\) \(n\) \(p\) \(\leq\) \(n\) - 5)**

\[
Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1) \tag{1}
\]

and (using \(\hat{p}\) in place of \(p\))

\[
Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0, 1) \tag{2}
\]

Using (2) an approximate 100(1 - \(\alpha\))% C.I. for \(p\) is

\[
\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.
\]

To test:

\[H_0 : p = \# \text{ vs. } H_{a} : p \neq \#\]

the test statistic is (using (1))

\[
Z = \frac{\hat{p} - \#}{\sqrt{\frac{\#(1-\#)}{n}}}
\]

Examples 16 and 17 in Section 6.5 illustrates this.
**TWO SAMPLE CASE**

For Large $n_1$ and $n_2$

\[ Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0, 1) \]  \hspace{1cm} (3)

and

\[ Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \approx N(0, 1) \]  \hspace{1cm} (4)

using (4) a 100(1 – $\alpha$)% C.I. for $p_1 - p_2$ is (its approximate)

\[ \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \]

To Test:

\[ H_0 : p_1 - p_2 = 0 \hspace{0.5cm} \text{vs.} \hspace{0.5cm} H_a : p_1 - p_2 \neq 0 \]

The test is (as always) based on assuming $H_0$ true, i.e., $p_1 = p_2 = p$, so the test statistic is obtained by pooling the two samples to estimate the common value $p$ as $\hat{p}$

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

Then in (3) substitute zero for $p_1 - p_2$ in the numerator and $p$ for $p_1$ and $p_2$ in the denominator. This gives

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

Then using $\hat{p}$ to estimate $p$ gives the test statistic

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

See Example 18 in Section 6.5 for an application.