Categorical Data

Motivation

Frequently, a population is composed of objects of two types, A and B, say. Let \( p \) be the true population proportion of objects of type A. Sample \( n \) objects. Let \( Y \) be the number of A in the sample.

**Question:** What is the distribution of \( Y \)?

**Answer:** \( Y \) is binomial with \( n \) trials and success probability \( p \).

Definitions

There are two sample proportions of interest:

- \( \hat{p} = \frac{Y}{n} \) - usual sample proportion
- \( \tilde{p} = \frac{Y + 1}{n + 4} \) - Wilson-Adjusted sample proportion

Suppose a coin is flipped 10 times, yielding 4 heads, and you want to estimate the probability of heads. With the usual sample proportion, your estimate would be \( \hat{p} = 4/10 = .4 \). With the Wilson-Adjusted sample proportion, your estimate would be \( \tilde{p} = 6/14 = 3/7 = .43 \). There is a generalization here: the Wilson-adjusted sample proportion tends to shrink its “guess” of \( p \) toward .5.

Inference about the true value of \( p \).

A Recall the approximation \( \hat{p} \sim N(p, \sqrt{p(1-p)/n}) \) if \( np \geq 5 \) and \( n(1-p) \geq 5 \). Let

\[
SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n} \quad \text{and} \quad SE_{\tilde{p}} = \sqrt{\tilde{p}(1-\tilde{p})/(n + 4)}
\]

Then the 100(1 - \( \alpha \))% CI for \( p \) is given by

- \( \hat{p} \pm z_{\alpha/2}SE_{\hat{p}}, \) or
- \( \tilde{p} \pm z_{\alpha/2}SE_{\tilde{p}} \)

Note: The Wilson-Adjusted CI is more likely to contain \( p \) when the assumptions are not well satisfied (i.e. either \( np < 5 \) or \( n(1-p) < 5 \)) or if \( n \) is small.
Example 1. Scientists wish to establish the proportion of plants of a certain population which are female. \( n = 200 \) were gender-typed, giving \( y = 132 \) females. Find the 95% CI for \( p \) using both the Wilson-adjusted and usual sample proportion formulas.

Solution. \( z_{0.025} = 1.96 \) from the normal table. Then \( \hat{p} = 132/200 = .66 \), and

\[
SE_{\hat{p}} = \sqrt{.66(.34)/200} = .033.
\]

So the CI is given by \( .66 \pm 1.96 \times .033 = (.60, .73) \).

Similarly, \( \hat{p} = 134/204 = .66 \). Also \( SE_{\hat{p}} = .033 \). To the CI are the same (or close enough given rounding error).

Planning a study to estimate \( p \)

Goal: choose \( n \) large enough to guarantee that the total CI length is less than \( a \), some positive constant.

Solution: Assume \( \alpha = .05 \) and the “usual” sample proportion \( \hat{p} \). The total CI length is then twice the error margin, or

\[
2EM = 2z_{0.025}SE_{\hat{p}} = 2(1.96)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

Normally, we would set the right-side above to be less than or equal to \( a \) and then solve for \( n \). But in this case, we have a “chicken or egg” problem: that is, we don’t have a pilot study that provides an estimate of \( \sigma \) and, furthermore, the standard error depends on the quantity we’re going to end up with after we do the study we’re planning.

We need a trick. Let \( f(x) = x(1-x) \) for \( 0 < x < 1 \). A graph of this function looks like
In other words, no matter what, \( f(x) \leq 1/4 \). Carefully examine the standard error of \( \hat{p} \): we can say that

\[
SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq \sqrt{\frac{1}{4n}} = \frac{1}{2\sqrt{n}}
\]

Then, go back to the original problem:

\[
2EM = 2z_{.025}SE_{\hat{p}} \\
\leq \frac{1.96}{\sqrt{n}} \\
< a.
\]

Solve for \( n \), giving the following equation for the sample size needed to ensure the total CI length is less than \( a \):

\[
\frac{1.96^2}{a^2} \leq n
\]

This works for any value of \( p \). On the other hand, what if we want to plan a study that ensures the minimum CI length for \( p \) is less than \( a \), but for the Wilson-adjusted sample
proportion? The following formula is then needed:

\[ \frac{1.96^2}{a^2} - 4 \leq n. \]

Example 2. A gambler wishes to find a 95% CI for \( p \), the probability of heads in a coin toss. The gambler needs the total length of the CI not to exceed .03. Find \( n \) to guarantee this (use the Wilson-adjusted proportion formula).

Solution. The formula with \( a = .03 \) gives

\[ (1.96/.03)^2 - 4 = 4264.44, \]

which should be rounded up to \( n = 4265 \). This works regardless of the value of the coin.

Transition: 2 categories are rarely enough

Chapter 9.4: The Chi-Square Goodness of Fit test

Summary:

- Assume a categorical random variable with \( k \) categories
- Let \( n \) be the sample size
- Let \( p_1, \ldots, p_k \) be the hypothesized proportion of objects of each type in the population (note \( p_1 + \ldots + p_k = 1 \)).

Under the hypothesized proportions, the expected number of objects of each type in the sample is

\[ e_i = np_i \quad i = 1, \ldots, k \]

Let the observed number of objects of each type in the sample be denoted by

\[ o_i \quad i = 1, \ldots, k. \]

Define the Chi-Squared Statistic

\[ \chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}. \]

Let

- \( H_0 \) : the true population proportions are given by the hypothesized \( p_1, \ldots, p_k \)
• $H_A$: otherwise. Then, when $H_0$ is true,

$$\chi^2 \sim \chi^2_{k-1}$$

where $\chi^2_{k-1}$ is the Chi-squared distribution with $k - 1$ degrees of freedom (this corresponds to a table in the back of the textbook).

• Reject $H_0$ when the p-value is less than $\alpha$. To find the p-value, calculate the area under the $\chi^2_{k-1}$ curve to the right of the observed value of the test statistic $\chi^2$. (We don’t deal with directionality in the case of the Chi-Square tests: always the p-value is the area to the right).

**Example 3.** Suppose an infinite pool with 3 colors of balls in it: (R)ed, (B)lue, and (G)reen. Take a random sample of size $n = 100$ from the pool. Hypothesize that $p_R = p_B = .33$ and $p_G = .34$; that is, that the pool is composed of equal parts each color. Suppose that $o_R = 31, o_B = 60, o_G = 9$.

**Solution.** Let

• $H_0 : p_R, p_B, p_G$ are as given in the question.

• $H_A :$ otherwise.

Then, calculate the expected counts $e_i$:

$$e_R = np_R = 100(.33) = 33$$
$$e_B = np_B = 100(.33) = 33$$
$$e_G = np_G = 100(.34) = 34$$

Next, calculate the observed value of the test statistic:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(31 - 33)^2}{33} + \frac{(60 - 33)^2}{33} + \frac{(9 - 34)^2}{34}$$

$$= 40.6$$

The degrees of freedom is $k - 1 = 3 - 1 = 2$. To calculate the p-value, note that the largest value on the $\chi^2_2$ curve is 18.42, which has area .0001 to the right of it (Table 9). Since $40.6 > 18.42$, the p-value is less than .0001. There is extremely significant evidence that the pool is not composed of equal parts each color.

□