Confidence Intervals for a Single Population Proportion

Intervals Discussed in the Text

Confidence intervals for a single population proportion are discussed in the text, Petruccelli, Nandram and Chen, on pp. 253-256. Two types of interval are considered there: (1) An **Exact Confidence Interval**, and (2) a **Classical Large-Sample Interval**. The latter is also known as a **Wald Interval**, after statistician **Abraham Wald**. The formula for a level L Wald interval is

$$\hat{p} \pm z_{(1+L)/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where n, $\hat{p}$, and $z_{(1+L)/2}$ are the sample size, sample proportion, and $(1+L)/2$ quantile of the standard normal distribution.

Both types of intervals have some difficulties with small samples. Due to the discrete nature of the binomial distribution, the exact interval tends to be **conservative**. This means its true coverage probability tends to be higher than advertized. For example, level 0.95 intervals may contain the true population parameter p 98% of the time. While this is not as bad as the intervals being **liberal** (meaning the true coverage probability tends to be lower than advertized), it means that exact intervals are **inefficient** (meaning they tend to be wider than if the true coverage were 0.95).

Still, the conservative performance of the exact intervals is preferable to the performance of the Wald intervals in small samples. Recall that the Wald intervals are based on the Central Limit Theorem normal approximation to the binomial distribution. Because that approximation is poor for small sample sizes, the Wald intervals are very liberal, giving true coverage probabilities substantially lower than the advertized value.

The Score Interval

Interestingly, there is an approximate confidence interval that has better small-sample performance than either the exact or Wald intervals. This interval is known as the **Score Interval**. While the theory behind it is beyond the scope of the text, its small-sample performance is enough of an improvement over that of the exact and Wald intervals, that it should be used in preference to them.[1]

The endpoints of a level L Score Interval are obtained from the formula $a \pm z_{(1+L)/2}b$, where a and b are given by

$$a = \hat{p}\left(\frac{n}{n+z^2_{(1+L)/2}}\right) + 0.5\left(\frac{z^2_{(1+L)/2}}{n+z^2_{(1+L)/2}}\right),$$

$$b = \sqrt{\frac{1}{n+z^2_{(1+L)/2}}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z^2_{(1+L)/2}}\right) + 0.25\left(\frac{z^2_{(1+L)/2}}{n+z^2_{(1+L)/2}}\right)\right]}$$

An Approximate Score Interval

When a computer routine is available, computing the score interval is painless, but its complicated formula is inconvenient when using only a hand calculator. In this case, a simple and effective approximation is obtained as follows:

1. Compute adjusted y, n, and p:

$$\tilde{y} = y + 0.5z^2_{\frac{(1+L)}{2}},\ \tilde{n} = n + z^2_{\frac{(1+L)}{2}},\ \tilde{p} = \tilde{y}/\tilde{n},$$

where y is the observed number of successes in the sample.

[1] In large samples, all three intervals should be substantially the same.

2. The approximate score interval is obtained by substituting $\tilde{p}$ for $\hat{p}$ and $\tilde{n}$ for n in the Wald confidence interval formula:

$$\tilde{p} \pm z_{(1+L)/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},$$

Example

Consider Example 5.6, pp. 254-256 of the text. There it is shown that an exact level 0.95 confidence interval for p is (0.028,0.096), and a level 0.95 Wald interval is (0.023,0.087).

To compute the score interval, we use the fact that $z_{0.975} = 1.96$, $n = 200$ and $\hat{p} = 0.055$. Then

$$a = 0.055\left(\frac{200}{200+1.96^2}\right) + 0.5\left(\frac{1.96^2}{200+1.96^2}\right) = 0.0634,$$

$$b = \sqrt{\frac{1}{200+1.96^2}\left[0.055(1-0.055)\left(\frac{200}{200+1.96^2}\right) + 0.25\left(\frac{1.96^2}{200+1.96^2}\right)\right]} = 0.0165$$

The resulting interval is then $0.0634 \pm (1.96)(0.0165) = (0.031, 0.096)$.

To compute the approximate score interval, note that the number of the 200 spools of wire that are defective is $y = 11$, so that

$$\tilde{y} = 11 + 0.5(1.96^2) = 12.9208,\ \tilde{n} = 200 + 1.96^2 = 203.8416,$$

and

$$\tilde{p} = \frac{12.9208/203.8416}{=}0.0634,$$

and the interval is

$$0.0634 \pm 1.96\sqrt{\frac{0.0634(1-0.0634)}{203.8416}} = (0.0299, 0.0968).$$

Confidence Intervals for the Difference of Two Population Proportions

Intervals Discussed in the Text

To estimate the difference $p_1 - p_2$ of proportions in two independent populations, the text in chapter 5 (pp. 266-268) presents only a Wald large-sample interval, which does not perform well for small samples. The closest analogue to the exact interval is the bootstrap interval, discussed in chapter 11, but beyond the scope of the chapter 5 material.

An Approximate Score Interval

The score interval is very complicated for the two sample case. There is, however, an analogue of the approximate score interval, which works well for small samples. It is constructed as follows:

1. Compute the adjusted values of y_1, y_2, n_1 and n_2:

$$\tilde{y}_1 = y_1 + 0.25z^2_{(1+L)/2},\ \tilde{n}_1 = n_1 + 0.5z^2_{(1+L)/2},$$

$$\tilde{y}_2 = y_2 + 0.25z^2_{(1+L)/2}\ \tilde{n}_2 = n_2 + 0.5z^2_{(1+L)/2},$$

and the adjusted estimates of p_1 and p_2:

$$\tilde{p}_1 = \tilde{y}_1/\tilde{n}_1,\ \ \tilde{p}_2 = \tilde{y}_2/\tilde{n}_2,$$

where n_1 and n_2 are the numbers of observations, and y_1 and y_2 are the observed number of successes in samples 1 and 2, respectively.

2. The approximate score interval is obtained by substituting $\tilde{p}_1$, $\tilde{p}_2$, $\tilde{n}_1$ and $\tilde{n}_2$ for the corresponding quantities in the Wald confidence interval formula:

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{(1+L)/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2}}$$

Example

Consider Example 5.10, pp. 267-268 of the text. There, $n_1 = 100$, $y_1 = 13$, $n_2 = 200$ and $y_2 = 34$, giving a Wald level 0.90 interval (-0.11,0.03). From these figures, we obtain $\tilde{y}_1 = 13 + .025 \times 1.645^2 = 13.6765$, $\tilde{n}_1 = 100 + 0.5 \times 1.645^2 = 101.3530$, $\tilde{y}_2 = 34 + .025 \times 1.645^2 = 34.6765$, $\tilde{n}_2 = 200 + 0.5 \times 1.645^2 = 201.3530$. Then $\tilde{p}_1 = 13.6765/101.3530 = 0.135$, and $\tilde{p}_2 = 34.6765/201.3530 = 0.172$. Plugging these into the formula given above, we obtain the level 0.90 confidence interval for $p_1 - p_2$ as

$$0.135 - 0.172 \pm \sqrt{\frac{0.135(1-0.135)}{101.3530} + \frac{0.172(1-0.172)}{201.3530}},$$

an interval that, to two decimal places, is identical to the Wald interval.

References

- Agresti, Alan, and Brent A. Coull (1998), "Approximate is Better Than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119-126.
- Agresti, Alan, and Brian Caffo (2000), "Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures," *The American Statistician*, 54, 280-288.