Chapter 8 - Lecture 2 (2)
Large-Sample Confidence Intervals for a Population Proportion

Yuan Huang

February 15, 2013

1. Review: general population with σ^2 unknown

2. CI for proportion: traditional interval

3. CI for proportion: score interval

Review

Without normality assumption, when n is large and σ^2 is unknown, we still get approximate z CI for the population mean μ.

An approximate $100(1-\alpha)\%$ confidence interval for μ is:

$$(\bar{x} - z_{\alpha/2}s/\sqrt{n}, \bar{x} + z_{\alpha/2}s/\sqrt{n})$$

The above CI is valid when CLT is valid.

For estimating proportion, each random variable is bernoulli variable, with 1, if event happens and 0 otherwise. The data we collect will be a sequence of 0 and 1. The estimator is

$$\hat{p} = \bar{X} = \frac{\text{Number of nonzero}}{\text{Sample size}}$$

For the proportion, we know that to apply CLT, we have to check $np > 10, n(1-p) > 10$. And then we plug in the $s = \sqrt{\hat{p}(1-\hat{p})}$ to get the traditional interval.

The **traditional way** of approximate $100(1-\alpha)\%$ confidence interval for p is:

$$
\begin{aligned}
& P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < z_{\alpha/2}\right) \approx 1-\alpha \\
\rightarrow \quad & (\hat{p} - z_{\alpha/2} s/\sqrt{n}, \hat{p} - z_{\alpha/2} s/\sqrt{n}) \\
\rightarrow \quad & (\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}, \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n})
\end{aligned}
$$

To get the score interval (default in book),

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \approx 1-\alpha$$

$$\rightarrow \quad P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)}/\sqrt{n}} < z_{\alpha/2}\right) \approx 1-\alpha$$

Resolve p will get a (score) interval for a population proportion p with a confidence level approximately $(1-\alpha)$ CI is :

$$\frac{\hat{p}+\frac{z_{\alpha/2}^2}{2n}}{1+\frac{z_{\alpha/2}^2}{n}} \pm z_{\alpha/2}\frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z_{\alpha/2}^2}{4n^2}}}{1+\frac{z_{\alpha/2}^2}{n}}$$

Example 8.8

The article "Repeatability and Reproducibility for Pass/Fail Data" (*J. Testing Eval.*, 1997: 151–153) reported that in $n = 48$ trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Let p denote the long-run proportion of all such trials that would result in ignition. A point estimate for p is $\hat{p} = 16/48 = .333$. A confidence interval for p with a confidence level of approximately 95% is

$$\frac{.333 + 1.96^2/96}{1 + 1.96^2/48} \pm 1.96 \frac{\sqrt{(.333)(.667)/48 + 1.96^2/(4 \cdot 48^2)}}{1 + 1.96^2/48}$$
$$= .346 \pm .129 = (.217, .475)$$

The traditional interval is

$$.333 \pm 1.96\sqrt{(.333)(.667)/48} = .333 \pm .133 = (.200, .466)$$

These two intervals would be in much closer agreement were the sample size substantially larger. ■

Figure: Modern Mathematical Statistics with Applications, 2rd, P397