Binomial Confidence Intervals

Gene Quinn

Estimating Proportions

Very often the quantity of interest in an estimation problem is the *proportion* p of the population under study that has a certain characteristic.

Estimating Proportions

Very often the quantity of interest in an estimation problem is the *proportion* p of the population under study that has a certain characteristic.

In fact, a huge industry has evolved to provide such estimates, particularly in the areas of public opinion polling and market research.

Estimating Proportions

Examples are very plentiful. Typical objectives would be to estimate the proportion of the population that:

- is likely to vote for a certain candidate
- supports a certain government policy
- has health insurance
- is employed
- is in the market for a new car
- watches a certain television show

Estimating Proportions

As we have seen, if $\{x_1, x_2, \ldots, x_n\}$ represents a vector of outcomes of independent Bernoulli trials each with probability of success p, and X represents the number of successes in n trials, the maximum likeklihood and method of moments estimator for p is

$$\hat{p} = \frac{X}{n} = \frac{\text{number of successes}}{\text{number of trials}}$$

Estimating Proportions

Interval estimates for p are nearly always constructed using an approximation based on the fact that when n is large and the actual population proportion p is not very close to zero or one, $\hat{p}$ has an approximately normal distribution:

$$\hat{p} \sim N\left(\frac{X}{n}, \frac{p(1-p)}{n}\right) \quad \text{(approximately)}$$

Estimating Proportions

Interval estimates for p are nearly always constructed using an approximation based on the fact that when n is large and the actual population proportion p is not very close to zero or one, $\hat{p}$ has an approximately normal distribution:

$$\hat{p} \sim N\left(\frac{X}{n}, \frac{p(1-p)}{n}\right) \quad \text{(approximately)}$$

A commonly used rule of thumb states that this approximation is valid when $np > 10$.

DeMoivre-Laplace Limit Theorem

The approximation is based on the following theorem:

Theorem (DeMoivre-Laplace) Let X be a binomial random variable based on n independent Bernoulli trials each with probability of success p. Then for any numbers a and b,

$$\lim_{n\to\infty} P\left(a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2}\, dz$$

DeMoivre-Laplace Limit Theorem

The approximation is based on the following theorem:

Theorem (DeMoivre-Laplace) Let X be a binomial random variable based on n independent Bernoulli trials each with probability of success p. Then for any numbers a and b,

$$\lim_{n\to\infty} P\left(a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2}\, dz$$

DeMoivre-Laplace is actually a special case of a more general theorem known as the *central limit theorem*, which is the justification for the wide use of the normal distribution with real-world data.

Confidence Intervals for Proportions

Theorem Let k be the number of successes in n independent Bernoulli trials with each with (unknown) probability of success p.

An approximate $100(1-\alpha/2)\%$ confidence interval for p is given by:

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}, \quad \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}\right)$$

Confidence Intervals for Proportions

Theorem Let k be the number of successes in n independent Bernoulli trials with each with (unknown) probability of success p.

An approximate $100(1-\alpha/2)\%$ confidence interval for p is given by:

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}, \quad \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}\right)$$

As before, $z_{\alpha/2}$ can be obtained from a standard normal table or from a spreadsheet:

$$z_{\alpha/2} \; = \; \text{NORMSINV}(1-\alpha/2)$$

Confidence Intervals for Proportions

As with the confidence interval for a population mean, the interpretation of the 95% confidence interval for a proportion is as follows:

If we repeated the experiment of conducting n *independent trials many times, and constructed a* 95% *confidence interval for each repetition, then on average* 95% *or 19 out of 20 of the resulting intervals will contain the true population proportion* p.

Confidence Intervals for Proportions

As with the confidence interval for a population mean, the interpretation of the 95% confidence interval for a proportion is as follows:

If we repeated the experiment of conducting n independent trials many times, and constructed a 95% confidence interval for each repetition, then on average 95% or 19 out of 20 of the resulting intervals will contain the true population proportion p.

(This *is not* the same as stating that the probability that p lies within the confidence interval is 95%. Because we are assuming that p is a parameter, and not a random variable, we do not associate probabilities with values of p.

Confidence Intervals for Proportions

Example: A rating survey contacts $1,000$ households during a certain time slot and finds that 145 are viewing a certain television program that airs in this time slot.

Construct a 95% confidence interval for the percentage of households that watched the program.

Confidence Intervals for Proportions

Example: A rating survey contacts $1,000$ households during a certain time slot and finds that 145 are viewing a certain television program that airs in this time slot.

Construct a 95% confidence interval for the percentage of households that watched the program.

From previous examples we know that $z_{\alpha/2} =$ NORMSINV$(1 - .05/2) = 1.96$, and we are given that $n = 1000$ and $k = 145$, so the approximate 95% confidence interval for p is:

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}}, \quad \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{\frac{k}{n}(1-\frac{k}{n})}{n}} \right)$$

Confidence Intervals for Proportions

Substituting the numbers for this example, the approximate 95% confidence interval is:

$$\left(\frac{145}{1000} - 1.96\sqrt{\frac{\frac{145}{1000}(1-\frac{145}{1000})}{1000}}, \quad \frac{145}{1000} + 1.96\sqrt{\frac{\frac{145}{1000}(1-\frac{145}{1000})}{1000}}\right)$$

or

$$\left(0.145 - 1.96\sqrt{\frac{0.145(1-0.145)}{1000}}, \quad 0.145 + 1.96\sqrt{\frac{0.145(1-0.145)}{1000}}\right.$$

$$= \; (0.123, \quad 0.167)$$

Confidence Intervals for Proportions

Substituting the numbers for this example, the approximate 95% confidence interval is:

$$\left(\frac{145}{1000} - 1.96\sqrt{\frac{\frac{145}{1000}(1-\frac{145}{1000})}{1000}}, \quad \frac{145}{1000} + 1.96\sqrt{\frac{\frac{145}{1000}(1-\frac{145}{1000})}{1000}}\right)$$

or

$$\left(0.145 - 1.96\sqrt{\frac{0.145(1-0.145)}{1000}}, \quad 0.145 + 1.96\sqrt{\frac{0.145(1-0.145)}{1000}}\right)$$

$$= \;(0.123, \quad 0.167)$$

Confidence Intervals for Proportions

Example 2: A rating survey contacts $1,000$ households during a certain time slot and finds that 145 are viewing a certain television program that airs in this time slot.

Construct a 99% confidence interval for the percentage of households that watched the program.

Confidence Intervals for Proportions

Example 2: A rating survey contacts $1,000$ households during a certain time slot and finds that 145 are viewing a certain television program that airs in this time slot.

Construct a 99% confidence interval for the percentage of households that watched the program.

In this case $z_{\alpha/2} =$ NORMSINV$(1 - .01/2) = 1.96$, $n = 1000$ and $k = 145$. The approximate 99% confidence interval for p is:

$$\left(0.145 - 2.58\sqrt{\frac{0.145(1-0.145)}{1000}}, \quad 0.145 + 2.58\sqrt{\frac{0.145(1-0.145)}{1000}}\right.$$

$$= \ (0.116, \quad 0.174)$$