Using the Beta Distribution on Confidence Intervals for Proportions

Presented May 20th, 1999

1999 Quality and Productivity Research Conference

General Electric, Schenectady NY

Jim Rutledge
DATA VISION
Statistical Consulting and Training
2541 Far Hills Avenue #202
Dayton, Ohio 45419
(937) 293-0358
jim@data-vision.biz

Brad Warner
UNITED STATES AIR FORCE ACADEMY
HQ USAFA/DFMS
2354 Fairchild Drive Suite 6D2A
USAF Academy, Colorado 80840
(719) 333-4470
warnerba.dfms@usafa.af.mil

SUMMARY

Most introductory texts on statistics include approximate methods for finding confidence intervals for proportions. As a result, these approximate methods have become part of most 6-sigma training programs. Rules of thumbs are often given for when the approximate method is valid. For example, if n is the sample size and p is the sample proportion, the approximation is "good" as long as we have $np \geq 5$ and $n(1-p) \geq 5$. However, in many processes that have continuously been improved, the proportion of defectives is so small that it is impractical to comply with the rule of thumb. Besides, why should you use an approximate method when an exact one is readily available. This paper will discuss an easily applied exact method for calculating confidence intervals for proportions. The method is implemented using Microsoft EXCEL's built in BETAINV function. In addition, this methodology can be used to help determine the appropriate sample size needed with pass/fail type of data. Easily programmed (one liners) Microsoft EXCEL commands will be discussed. The author will make available a copy of his spreadsheet that implements the methods presented for both one-sided and two-sided confidence intervals.

1. The problem

Quality professionals are often interested in estimating the population proportion (π) for some process. For example, an auditor may look at 90 records and find 3 with errors. Since the 90 records are only a representative sample, it is best to build a confidence interval for the proportion of records with an error. The usual confidence interval employed is:

$$p \pm Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\ . \qquad (1)$$

The point estimate p is obtained from x/n where x is the number of events out of n trials. In our example, x = 3 is the number of bad records out of a sample of size n = 90. $Z_{1-\alpha/2}$ is the standard normal value associated with the (1-α / 2)100th percentile.

For our scenario we have:

$$\frac{3}{90} \pm 1.96\sqrt{\frac{\frac{3}{90}(1-\frac{3}{90})}{90}} = (-0.0038,\ 0.0704). \qquad (2)$$

Kiemele and Schmidt (1993) give the following rule of thumb for using equation (1): the approximation is good if $np \geq 5$ and $n(1-p) \geq 5$. In this situation the rule of thumb has been violated. In addition, the confidence interval has a negative number for the lower bound. This is disturbing because by definition, a proportion should be non-negative. What should the quality professional do in this situation? When they refer back to their original training material for the answer, they are unlikely to find a solution.

2. The Usual Hand Waiving

When the above rule of thumb does not hold, most introductory texts skirt the issue by saying something like, violations of this rule of thumb are uncommon, or they may say that methods to address this problem are beyond the scope of this text. Montgomery (1994) does not use one of these excuses but none the less dodges the bullet by saying to use special software. There is no indication of which software to use or how to use it.

> In situations where this approximation is inappropriate, particularly in cases where n is small, other methods must be used. … However, we prefer to use numerical methods based on the binomial probability mass function that are implemented in computer programs.

In all these situations, the quality professional is left with the feeling that this problem is beyond their capabilities and that they need to bring in a statistician. This need not be the case! Next, a method that most quality professionals could easily implement will be discussed.

3. A Possible Solution

Casella and Berger (1990) discuss how a confidence interval can always be obtained. Applying the theory they develop to a proportion, we have that a (1- α)100% confidence interval can be obtained by solving for π in the following two equations:

$$\text{P[number of “successes”} \le x \mid \pi] = \alpha / 2 \quad (3)$$

$$\text{P[number of “successes”} \ge x \mid \pi] = \alpha / 2 \,. \quad (4)$$

The upper confidence limit for π comes from (3) and the lower confidence limit comes from (4).

For our situation, the “numerical methods” Montgomery (19??) refers to involves solving the following equations for π:

$$P[X \le 3 \,|\pi] = \sum_{x=0}^{3} \binom{90}{x} \pi^x (1-\pi)^{90-x} = 0.025 \quad (5)$$

$$P[X \ge 3 \,|\pi] = \sum_{x=3}^{90} \binom{90}{x} \pi^x (1-\pi)^{90-x} = 1 - \sum_{x=0}^{2} \binom{90}{x} \pi^x (1-\pi)^{90-x} = 0.025 \quad (6)$$

In general, the solutions to the above binomial equations are not easily obtained. An efficient way to solve the above equations is to employ Newton’s method to find the roots of the equations. This requires specialized software that is neither readily available nor understandable to many practitioners. It is good news that an easier method exists. Using integration by parts, it can be shown that if the random variable X is distributed binomial(n, π), then $P(X \le x) = P(Y \le 1-\pi)$ where Y is beta(n-x, x+1). Similarly, $P(X \ge x) = P(W \ge 1-\pi)$, where W is beta(n-x+1, x). Thus (3) and (4) become:

$$P(Y \le 1-\pi) = \alpha / 2 \quad (7)$$

$$P(W \ge 1-\pi) = \alpha / 2 \,. \quad (8)$$

We know by definition that $P[Y \le beta_{\alpha/2}(n-x, x+1)] = \alpha/2$. Setting $1-\pi$ equal to $beta_{\alpha/2}(n-x, x+1)$ and solving for π will give the upper confidence limit. Likewise, the lower confidence limit can be found by setting $1-\pi$ equal to $beta_{1-\alpha/2}(n-x+1, x)$ and solving for π. Thus a (1-α)100% confidence interval for π is:

$$1 - beta_{1-\alpha/2}(n-x+1, x) \le \pi \le 1 - beta_{\alpha/2}(n-x, x+1) \qquad (9)$$

where $beta_{\alpha}(a, b)$ is the beta value associated with the α th percentile of the beta distribution with parameters "a" and "b." If x=0 then 0 should be used for the lower bound. Likewise, if x=n then take 1 as the upper bound. For our example the interval is:

$$1 - beta_{.975}(88,\ 3) \le \pi \le 1 - beta_{.025}(87,\ 4)$$

Solving this formula is easy in Microsoft EXCEL! This procedure will be discussed next.

4. Using Excel to Obtain the Interval Estimate

Microsoft EXCEL comes equipped with the inverse beta function (BETAINV). This function can easily be used to find the upper and lower bounds of our confidence interval. The lower bound is given by the EXCEL function:

1-BETAINV(.975,88,3).

The upper bound is given by:

1-BETAINV(.025,.87,4).

Using EXCEL, the following interval was obtained:

$$0.0069 \le \pi \le 0.0943.$$

It should be noted that the general form of this confidence interval would not change. It is easy to write an EXCEL spreadsheet where all you need to specify is the level of confidence (1-α), the number of events (x), and the sample size (n). Extending this method to one-sided intervals is easy, simply don't divide α by 2 in equation (9).

5. Sample Size

Another nice feature of this confidence interval is that it can be used to help the quality professional determine appropriate sample sizes for studies. Consider the quality professional who is trying to decide how many missiles to test launch in order to prove that the missile reliability is at least 80% with 90% confidence. The quality professional knows from previous history that failures are rare. Using the one-sided lower bound for the proportion of successful launches involves the following EXCEL function:

1-BETAINV(.1,n-x, x+1).

The .1 is used because we want to be (1-.1)100% = 90% confident. The quality professional suspects that x (the number of successful launches) will be equal to n (the number of missiles tested). Using different values for n and x the quality engineer can play "what if?" As an example, see the following table that can easily be generating using equation (??) in Microsoft EXCEL.

Successful Launches (x)	Number of Launches (n)	90% lower confidence bound on the proportion of missiles that will work. 1-BETAINV(.1,n-x, x+1).
9	9	.77
10	10	.79
11	11	.81
11	12	.71

6. Summary

If the normal approximation to the binomial distribution is not valid, it is still possible to obtain confidence intervals for the population proportion. This paper has discussed how to build such confidence intervals. In addition an Excel spreadsheet (CI_P.XLS) has been designed to perform this analysis. A copy of this spreadsheet can be obtained from the author.

REFERENCES

Casella, G. and Berger, R. (1990). Statistical inference. *Wadsworth & Brooks/Cole Statistics/Probability Series*, Pacific Grove, California.

Kiemele, M. and Schmidt, S. (1993). Basic statistics: Tools for continuous improvement. *Air Academy Press*, Colorado Springs, Colorado.

Montgomery, D. and Runger, G. (1994). Applied statistics and probability for engineers. *John Wiley and Sons*, New York, New York.