

Figure 3: Poisson vs. normal approximations for large sample sizes.

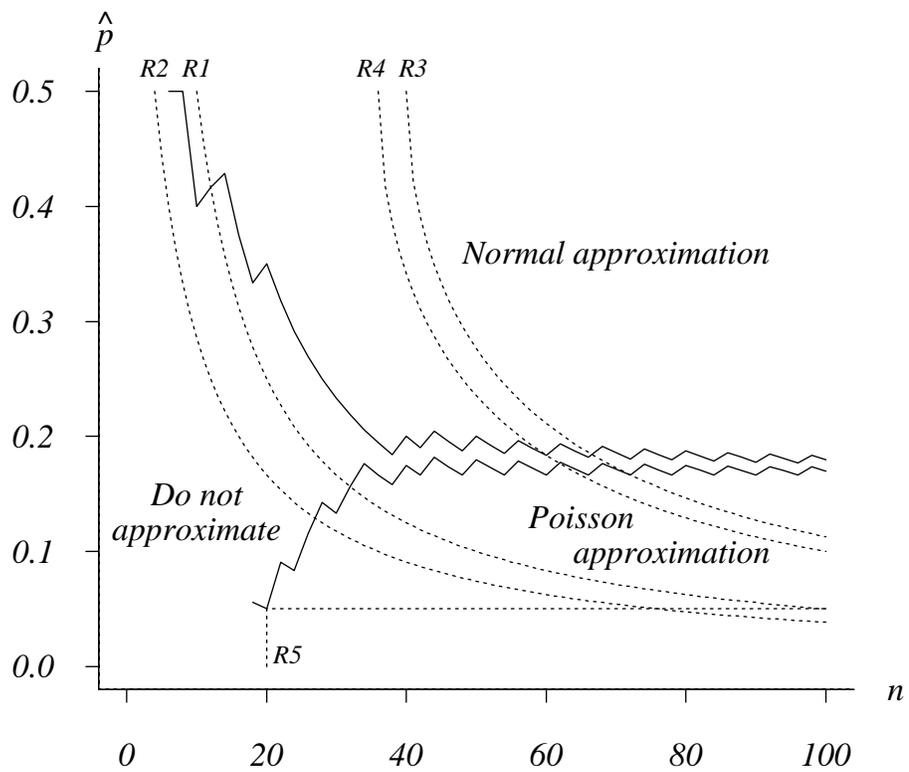


Figure 2: Approximation methods for confidence limits (maximum error: 0.04).

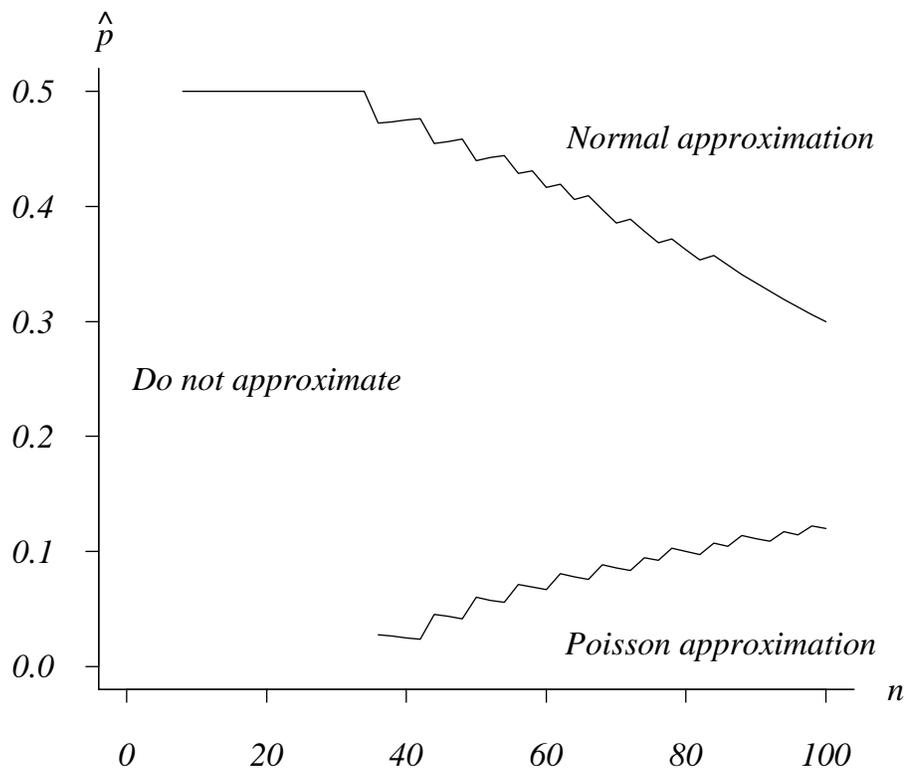


Figure 1: Approximation methods for confidence limits (maximum error: 0.01).

- [17] R.E. Walpole and R.H. Myers, *Probability and Statistics for Engineers and Scientists*, Fifth Edition, Macmillan, 1993.
- [18] S. Wolfram, *Mathematica: A System for Doing Mathematics by Computer*, Second Edition, Addison-Wesley, 1991.

- [5] J.E. Freund, *Mathematical Statistics*, Fifth Edition, Prentice-Hall, 1992.
- [6] B.K. Ghosh, "A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter", *Journal of the American Statistical Association*, Volume 74, Number 368, 1979, 894-900.
- [7] R.N. Goldman and J.S. Weinberg, *Statistics: An Introduction*, Prentice-Hall, 1985.
- [8] R.V. Hogg and E.A. Tanis, *Probability and Statistical Inference*, Fourth Edition, 1993.
- [9] L.L. Lapin *Probability and Statistics for Modern Engineering*, Brooks/Cole, 1983.
- [10] R.J. Larsen and M.L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, Second Edition, Prentice-Hall, 1986.
- [11] R.F. Ling, "Just Say No to Binomial (and other Discrete Distributions) Tables", *American Statistician*, Volume 46, Number 1, 1992, pp. 53-54.
- [12] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*, Third Edition, MacMillan, 1992.
- [13] S.Ross, *A First Course in Probability*, Third Edition, Macmillan, 1988.
- [14] M. Schader and F. Schmid, "Two Rules of Thumb for the Approximation of the Binomial Distribution by the Normal Distribution", *American Statistician*, Volume 43, Number 1, 1989, pp. 23-24.
- [15] R.L. Scheaffer and J.T. McClave, *Probability and Statistics for Engineers*, Third Edition, PWS-Kent, 1990.
- [16] K.S. Trivedi, *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, Prentice-Hall, 1982.

- For sample sizes larger than 150, the absolute error of either upper and lower confidence limit is less than 0.01 if the appropriate approximation technique is used. Figure 3 should be consulted for specific guidance as to whether the binomial or Poisson approximation is appropriate.
- Introductory probability and statistics textbooks targeting statistics and mathematics majors would benefit from including the use of the F distribution to find p_L and p_U . Also, more of these texts should include the use of the Poisson approximation to the binomial distribution for determining interval estimates for p . These confidence limits only require a table look-up associated with the chi-square distribution and are very accurate for large n and small p .

Acknowledgment

This research was supported by the Institute for Computer Applications in Science and Engineering (ICASE). Their support is gratefully acknowledged. Helpful comments from Pam Burch, Herb Multhaup and Bruce Schmeiser are gratefully acknowledged.

References

- [1] A.D. Aczel, *Complete Business Statistics*, Second Edition, Irwin, 1993.
- [2] C. Blyth, "Approximate Binomial Confidence Limits", *Journal of the American Statistical Association*, Volume 81, Number 395, 1986, pp. 843-855.
- [3] G. Casella and R. Berger, *Statistical Inference*, Wadsworth and Brooks/Cole, 1990.
- [4] H. Chen, "The Accuracy of Approximate Intervals for a Binomial Parameter", *Journal of the American Statistical Association*, Volume 85, Number 410, 1990, pp. 514-518.

corresponding to the normal approximation rule $n\hat{p}(1 - \hat{p}) \geq 10$

- the rule labeled “R4” is a plot of $\hat{p} = \frac{1}{2} - \frac{\sqrt{n(n-36)}}{2n}$ on the range $[36, 100]$ corresponding to the normal approximation rule $n\hat{p}(1 - \hat{p}) > 9$
- the rule labeled “R5” is a plot of $n \geq 20$ and $\hat{p} \leq 0.05$ or $n \geq 100$ and $n\hat{p} \leq 10$ corresponding to the guideline for using the Poisson approximation.

The n, \hat{p} combinations falling above the dotted curves for rules R1, R2, R3, and R4 correspond to those that would be used if the rules of thumb were followed. Clearly, rules R3 and R4 are significantly more conservative than R1 and R2.

Figure 3 is a continuation of Figure 2 for sample sizes larger than $n = 100$. Note that the vertical axis has been modified and the horizontal axis is logarithmic. The curve in the figure represents the largest value of \hat{p} where the Poisson approximation to the binomial is superior to the normal approximation to the binomial. Since this relationship is linear, a rather unwieldy rule of thumb for n between 100 and 10,000 is: use the normal approximation over the Poisson approximation if $\hat{p} > \frac{5.2 - \log_{10} n}{18.8}$.

4 Conclusions

Although there are a number of different variations of the calculations that have been conducted here (e.g., one-sided confidence intervals, different significance levels, different definitions of error), there are three general conclusions:

- The traditional advice from most textbooks of using the normal and Poisson approximations to the binomial for the purpose of computing confidence intervals for p should be tempered with a statement such as: “the Poisson approximation should be used when $n \geq 20$ and $p \leq 0.05$ if the analyst can tolerate an error that may be as large as 0.04” (see Figure 2).

on each Bernoulli trial is arbitrary, we only consider the range $0 < \hat{p} \leq \frac{1}{2}$. Figures 1, 2 and 3 have mirror images for the range $\frac{1}{2} \leq \hat{p} < 1$.

Figure 1 contains a plot of n versus \hat{p} for $n = 2, 4, \dots, 100$ and considers the range $0 < \hat{p} \leq \frac{1}{2}$ for a maximum error of 0.01. Thus if the actual error for a particular (n, \hat{p}) pair is greater than 0.01, the point lands in the “Do not approximate” region. If one of the two approximations yields an error of less than 0.01, then the pair belongs to either the “Normal approximation” or “Poisson approximation” regions, depending on which yields a smaller error. Not surprisingly, the normal approximation performs better when the point estimate is closer to $\frac{1}{2}$ and the Poisson approximation performs better when the point estimate is closer to 0. Both approximations perform better as n increases. In order to avoid any spurious discontinuities in the regions, the calculations were made for even values of n . The edges of the region are not smooth because of the discrete natures of n and \hat{p} . The boundary of the approximation regions are those (n, \hat{p}) pairs where the error is less than 0.01. If the horizontal axis were extended, the normal and Poisson regions would meet at approximately $n = 150$. Mathematica [18] was used for the comparisons because of its ability to hold variables to arbitrary precision.

If the maximum error is relaxed to 0.04, then there are more cases where the approximations perform adequately. Figure 2 is analogous to Figure 1 but considers an error of 0.04. This figure also contains the rules of thumb associated with the normal and Poisson approximations to the binomial distribution. In particular,

- the rule labeled “R1” is a plot of $\hat{p} = 5/n$ on the range $[10, 100]$ corresponding to the normal approximation rule $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$
- the rule labeled “R2” is a plot of $\hat{p} = \frac{4}{4+n}$ on the range $[4, 100]$ corresponding to the normal approximation rule $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ falling in the interval $(0, 1)$
- the rule labeled “R3” is a plot of $\hat{p} = \frac{1}{2} - \frac{\sqrt{n(n-40)}}{2n}$ on the range $[40, 100]$

or

$$1 - \sum_{k=0}^{y-1} \frac{(np_{PL})^k e^{-np_{PL}}}{k!} = \alpha/2.$$

The left-hand side of this equation is the cumulative distribution function for an Erlang random variable with parameters np_{PL} and y (denoted by $E_{np_{PL},y}$) evaluated at one. Consequently,

$$P[E_{np_{PL},y} \leq 1] = \alpha/2$$

Since $2np_{PL}E_{np_{PL},y}$ is equivalent to a χ^2 random variable with $2y$ degrees of freedom, this reduces to

$$P[\chi_{2y}^2 \leq 2np_{PL}] = \alpha/2$$

or

$$p_{PL} = \frac{1}{2n} \chi_{2y,1-\alpha/2}^2.$$

By a similar line of reasoning, the upper limit based on the Poisson approximation to the binomial distribution is

$$p_{PU} = \frac{1}{2n} \chi_{2(y+1),\alpha/2}^2.$$

This approximation works best when p is small (e.g., reliability applications where the probability of failure p is small).

3 Comparison of the Approximate Methods

There are a multitude of different ways to compare the approximate confidence intervals with the exact values. We have decided to compute the error of an approximate two-sided confidence interval as the maximum error

$$\max\{|p_L - \tilde{p}_L|, |p_U - \tilde{p}_U|\}$$

where \tilde{p}_L and \tilde{p}_U are the approximate lower and upper bounds, respectively. This error is computed for all combinations of n and \hat{p} . Since the definition of “success”

```

fcrit = Quantile[FRatioDistribution[2 * y, 2 * (n - y + 1)], alpha/2]
pl = 1 / ( 1 + (n - y + 1) / ( y * fcrit) )

```

```

fcrit = Quantile[FRatioDistribution[2 * (y + 1), 2 * (n - y)], 1 - alpha/2]
pu = 1 / ( 1 + (n - y) / ( (y + 1) * fcrit) )

```

This method is significantly faster than the approach using the binomial distribution, but encounters difficulty with determining the F ratio quantiles for some combinations of n and y .

The first approximate confidence interval is based on the normal approximation to the binomial. The random variable $\frac{Y-np}{\sqrt{np(1-p)}}$ is asymptotically standard normal. Thus an approximate confidence interval for p is

$$\frac{Y}{n} - z_{\alpha/2} \sqrt{\frac{\frac{Y}{n}(1 - \frac{Y}{n})}{n}} < p < \frac{Y}{n} + z_{\alpha/2} \sqrt{\frac{\frac{Y}{n}(1 - \frac{Y}{n})}{n}}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ fractile of the standard normal distribution. This approximation works best when $p = \frac{1}{2}$ (e.g., political polls). It allows confidence limits that fall outside of the interval $[0, 1]$. One should also be careful when $Y = 0$ or $Y = n$ since the confidence interval will have a width of 0.

The second approximate confidence interval is based on the Poisson approximation to the binomial (see, for example, Trivedi [16], page 498). This confidence interval does not appear as often in textbooks as the first approximate confidence interval. The random variable Y is asymptotically Poisson with parameter np . Therefore, the exact lower bound p_L satisfying

$$\sum_{k=y}^n \binom{n}{k} p_L^k (1 - p_L)^{n-k} = \alpha/2$$

can be approximated with a Poisson lower limit p_{PL} which satisfies

$$\sum_{k=y}^{\infty} \frac{(np_{PL})^k e^{-np_{PL}}}{k!} = \alpha/2$$

Since this probability is equal to $\alpha/2$ for a two-sided confidence interval,

$$F_{2y, 2(n-y+1), 1-\alpha/2} = \frac{(n-y+1)p_L}{y(1-p_L)}$$

or

$$p_L = \frac{1}{1 + \frac{n-y+1}{yF_{2y, 2(n-y+1), 1-\alpha/2}}}.$$

In a similar fashion,

$$p_U = \frac{1}{1 + \frac{n-y}{(y+1)F_{2(y+1), 2(n-y), \alpha/2}}}.$$

The next paragraph discusses numerical issues associated with determining these bounds.

The Mathematica (see [18]) code for solving the binomial equations numerically is

```
pl = FindRoot[
  Sum[Binomial[n, k] * p ^ k * (1 - p) ^ (n - k), {k, y, n}] == alpha/2,
  {p, y / n} ]
```

```
pu = FindRoot[
  Sum[Binomial[n, k] * p ^ k * (1 - p) ^ (n - k), {k, 0, y}] == alpha/2,
  {p, y / n} ]
```

for a given n , y and α . This code works well for small and moderate sized values of n . Some numerical instability occurred for larger values of n , so the well known relationship (Larsen and Marx [10], page 101) between the successive values of the probability mass function $f(x)$ of the binomial distribution

$$f(x) = \frac{(n-x+1)p}{x(1-p)} f(x-1) \quad x = 1, 2, \dots, n$$

was used to calculate the binomial cumulative distribution function. The Mathematica code for determining p_L and p_U using the F distribution is

the lower limit p_L satisfies

$$\sum_{k=y}^n \binom{n}{k} p_L^k (1 - p_L)^{n-k} = \alpha/2$$

where y is the observed value of the random variable Y and α is the nominal coverage of the confidence interval (see, for example, [10], page 279). For $y = 1, 2, \dots, n - 1$, the upper limit p_U satisfies

$$\sum_{k=0}^y \binom{n}{k} p_U^k (1 - p_U)^{n-k} = \alpha/2.$$

This confidence interval requires numerical methods to determine p_L and p_U and takes longer to calculate as n increases. This interval will be used as a basis to check the approximate bounds reviewed later in this section. A figure showing the coverage probabilities for bounds of this type is shown in Blyth [2]. Following a derivation similar to his, a faster way to determine the lower and upper limits can be determined. Let W_1, W_2, \dots, W_n be iid $U(0, 1)$ random variables. Let Y be the number of the W_i 's that are less than p . Hence Y is binomial with parameters n and p . Using a result from page 233 of Casella and Berger [3], the order statistic $W \equiv W_{(y)}$ has the beta distribution with parameters y and $n - y + 1$. Since the events $Y \geq y$ and $W < p$ are equivalent, $P[Y \geq y]$ (which is necessary for determining p_L) can be calculated by

$$\begin{aligned} P(Y \geq y) &= P(W < p) \\ &= \frac{\Gamma(n+1)}{\Gamma(y)\Gamma(n-y+1)} \int_0^p w^{y-1} (1-w)^{n-y} dw. \end{aligned}$$

Using the substitution $t = \frac{(n-y+1)w}{y(1-w)}$ and simplifying yields

$$\begin{aligned} P(Y \geq y) &= \frac{\Gamma(n+1)}{\Gamma(y)\Gamma(n-y+1)} \left(\frac{n-y+1}{y}\right)^{n-y+1} \int_0^{\frac{(n-y+1)p}{y(1-p)}} \frac{t^{y-1}}{\left(\frac{n-y+1}{y} + t\right)^{n+1}} dt \\ &= P[F_{2y, 2(n-y+1)} < \frac{(n-y+1)p}{y(1-p)}]. \end{aligned}$$

Poisson approximations to the binomial distribution. Determining a confidence interval for p when the sample size is large using approximate methods is often needed in simulations with a large number of replications and in polling.

Computing probabilities using the normal and Poisson approximations is not considered here since work has been done on this problem. Ling [11] suggests using a relationship between the cumulative distribution functions of the binomial and F distributions to compute binomial probabilities. Ghosh [6] compares two confidence intervals for the Bernoulli parameter based on the normal approximation to the binomial distribution. Schader and Schmid [14] compare the maximum absolute error in computing the cumulative distribution function for the binomial distribution using the normal approximation with a continuity correction. They consider the two rules for determining whether the approximation should be used: np and $n(1 - p)$ are both greater than 5, and $np(1 - p) > 9$. Their conclusion is that the relationship between the maximum absolute error and p is approximately linear when considering the smallest possible sample sizes to satisfy the rules.

Concerning work done on confidence intervals for p , Blyth [2] has compared five approximate one-sided confidence intervals for p based on the normal distribution. In addition, he uses the F distribution to reduce the amount of time necessary to compute an exact confidence interval. Using an arcsin transformation to improve the confidence limits is considered by Chen [4].

2 Confidence Interval Estimators for p

Two-sided confidence interval estimators for p can be determined with the aid of numerical methods. One-sided confidence interval estimators are analogous. Let $p_L < p < p_U$ be an “exact” (see [2]) confidence interval for p . For $y = 1, 2, \dots, n - 1$,

1 Introduction

There is conflicting advice concerning the sample size necessary to use the normal approximation to the binomial distribution. For example, a sampling of textbooks recommend that the normal distribution be used to approximate the binomial distribution when:

- np and $n(1 - p)$ are both greater than 5 (see [1], page 211, [5], page 245, [7], page 304, [9], page 148, [16], page 497, [17], page 161)
- $p \pm 2\sqrt{\frac{p(1-p)}{n}}$ lies in the interval $(0, 1)$ (see [15], page 242, [12], page 299)
- $np(1 - p) \geq 10$ (see [13], page 171)
- $np(1 - p) > 9$ (see [1], page 158).

Many other textbook authors give no specific advice concerning when the normal approximation should be used. To complicate matters further, most of this advice concerns using these approximations to compute probabilities. Whether these same rules of thumb apply to confidence intervals is seldom addressed. The Poisson approximation, while less popular than the normal approximation to the binomial, is useful for large values of n and small values of p . The same sampling of textbooks recommend that the Poisson distribution be used to approximate the binomial distribution when $n \geq 20$ and $p \leq 0.05$ or $n \geq 100$ and $np \leq 10$ (see [8], page 177, [5], page 204).

Let X_1, X_2, \dots, X_n be iid Bernoulli random variables with unknown parameter p and let $Y = \sum_{i=1}^n X_i$ be a binomial random variable with parameters n and p . The maximum likelihood estimator for p is $\hat{p} = \frac{Y}{n}$, which is unbiased and consistent. The interest here is in confidence interval estimators for p . In particular, we want to compare the approximate confidence interval estimators based on the normal and

A Comparison of Approximate Interval Estimators for the Bernoulli Parameter

Lawrence Leemis
Department of Mathematics
College of William and Mary
Williamsburg, VA 23187-8795

Kishor S. Trivedi
Department of Electrical Engineering
Duke University, Box 90291
Durham, NC 27708-0291

The goal of this paper is to compare the accuracy of two approximate confidence interval estimators for the Bernoulli parameter p . The approximate confidence intervals are based on the normal and Poisson approximations to the binomial distribution. Charts are given to indicate which approximation is appropriate for certain sample sizes and point estimators.

KEY WORDS: Confidence interval, Binomial distribution, Bernoulli distribution, Poisson distribution.