Journal of Scientific Research & Reports
6(5): 329-339, 2015, Article no.JSRR.2015.158
ISSN: 2320-0227

SCIENCEDOMAIN *international*
www.sciencedomain.org

Comparison of Threshold Rules for a Normal Approximation to a Binomial Distribution

Steven T. Garren[1*]

[1] *Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA 22807, USA.*

Article Information
DOI: 10.9734/JSRR/2015/16653
Editor(s):
(1) Narcisa C. Apreutesei, Technical University of Iasi, Romania.
Reviewers:
(1) Anonymous, Ghana.
(2) Cliff Richard Kikawa, Mathematics and Statistics, Tshwane University of Technology, South Africa.
Complete Peer review History:
http://www.sciencedomain.org/review-history.php?iid=966&id=22&aid=8469

Original Research Article

Received: 9th February 2015
Accepted: 26th February 2015
Published: 16th March 2015

Abstract

A commonly used rule for determining if a Binomial(n, p) distribution may be reasonably approximated by a normal distribution is whether or not np and $n(1-p)$ are at least some constant, such as 10. Two competing rules, one based on the binomial variance and the other based on the coefficient of variation, are considered when constructing confidence intervals and performing hypothesis testing, both using and not using a continuity correction. Under one criterion the rule based on the coefficient of variation is found to be the best in terms of coverage probabilities, and under another criterion the rule based on the binomial variance is found to be the best.

Keywords: Binomial distribution; coverage probabilities; coefficient of variation; normal approximation.
2010 Mathematics Subject Classification: 62F03; 62F25

**Corresponding author: E-mail: garrenst@jmu.edu*

1 Introduction

In elementary statistics courses, a commonly used threshold rule for approximating a binomial(n, p) distribution with a normal distribution is based on the minimum *mean* of binomial(n, p) and binomial$(n, 1-p)$ distributions, for sample size n and success probability p. More specifically, this classical binomial-*mean* rule is

$$\min\{np,\ n(1-p)\} \geq \kappa_m, \qquad (1.1)$$

for some positive constant κ_m, where κ_m is often selected to be an easily memorable number such as 5, 10 or 15; *cf.,* [1,2,3,4]. For example, if a coin is tossed n times where the probability of heads is p, then X, the number of heads, is a binomial(n, p) random variable. Thus, the normal approximation to the binomial distribution is deemed reasonable if np (the average number of heads) and $n(1-p)$ (the average number of tails) are at least κ_m, under this rule (1.1).

Using coverage probabilities, the normal distribution reasonably approximates the binomial distribution based on the binomial-*mean* rule (1.1) [5]. The normal distribution is a reasonable approximation to the binomial distribution when the cumulative distribution functions are relatively close to one another, as demonstrated with coverage probabilities [6,7].

Herein, we consider two new rules, which are competitors to the binomial-*mean* rule (1.1). One competitor is the binomial-*variance* rule; *i.e.,*

$$np(1-p) \geq \kappa_v, \qquad (1.2)$$

where κ_v is a positive constant. This rule (1.2) is based on the concept that $np(1-p)$ is the variance of a binomial(n, p) distribution, and variance is a measure of spread. Requiring a large enough variance ensures a large enough sample size.

The other competing rule is based on the coefficient of variation, $np/\sqrt{np(1-p)}$ and $n(1-p)/\sqrt{np(1-p)}$, or equivalently requiring that

$$\min\{np/(1-p),\ n(1-p)/p\} \geq \kappa_c, \qquad (1.3)$$

where κ_c is a positive constant. This rule (1.3) is based on the concept that the coefficient of variation is the ratio of the mean (*i.e.,* np) to the standard deviation (*i.e.,* $\sqrt{np(1-p)}$, which is the squareroot of the variance) of a binomial(n, p) distribution. Requiring a large enough coefficient of variation ensures a large enough sample size.

Coverage probabilities using these three rules are compared under two opposing criteria. The *equal-centers* criterion is based on selecting the three values of κ such that the three minimum sample sizes are the same when $p = 0.5$. The *equal-tails* criterion is based on selecting the three values of κ such that the three minimum sample sizes are the same in the limit as $p \downarrow 0$ and $p \uparrow 1$.

2 Methodology

The nominal levels are set at $\alpha = 0.05$. Considered are both one-sided confidence intervals, $\hat{p} - z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}$, and two-sided confidence intervals, $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$, as well as the option for including or excluding a continuity correction, resulting in a set of four graphs. This continuity correction is based on adding or subtracting 0.5 from X in the sample proportion $\hat{p} = X/n$. Since the coverage probabilities are determined under both the *equal-centers* criterion (shown in Figure 1) and the *equal-tails* criterion (shown in Figure 3), then using confidence intervals produces two sets of four graphs.

Here are additional details regarding the construction of the continuity correction. Since a binomial(n, p) random variable (*e.g.,* the number of heads when a coin is tossed n times, where p is the probability of heads) can take on only integers, then an approximation to the normal distribution (which is continuous) typically can be improved by adjusting halfway between these integers; *i.e.,* a correction of 0.5. For example, if $n = 100$, then the probability of obtaining *at least* 43 heads is approximated by $P(Y > 42.5)$, where Y is a normal random variable with mean np and variance $np(1-p)$. Similarly, the probability of obtaining *no more than* 43 heads is approximated by $P(Y < 43.5)$. This improvement based on the continuity correction is a well-known result [8,9,10].

Additional calculations are performed based on hypothesis tests, by replacing the standard

error $\sqrt{\hat{p}(1-\hat{p})/n}$ by $\sqrt{p(1-p)/n}$, noting the null value of p with hypothesis testing. These calculations for hypothesis testing produce another two sets of four graphs (shown in Figures 2 and 4), for a grand total of 16 graphs when $\alpha = 0.05$. The *R*-code for producing these graphs is shown in the appendix.

Setting $\kappa_m = 5$, the minimum value of n is 10, under the binomial-*mean* rule. Then, for each value of $n \geq 10$, we determine the minimum value (without loss of generality), $p_{\min}$, of p satisfying the binomial-*mean* rule. As an intense grid search, we consider 10,000 values of p^* uniformly sequenced between $p_{\min}$ and 0.5. The minimum coverage probability over the valid values of p^*, as based on binomial probabilities, is plotted as a function of n.

When the *equal-centers* criterion is used, we have $\kappa_v = \kappa_m/2$ and $\kappa_c = 2\kappa_m$, so the minimum valid value of n is 10 for all three rules with $\kappa_m = 5$. When the *equal-tails* criterion is used, we then have $\kappa_m = \kappa_v = \kappa_c$, so the minimum valid values of n are 10, 20, and 5, respectively, for $\kappa_m = 5$.

One might naively argue that when generating confidence intervals, the three rules should be based on $\hat{p}$ rather than p. For example, consider the binomial-*mean* rule with $\kappa_m = 5$ and $n = 10$. The only value of X which satisfies this rule is $X = 5$. The two-sided confidence interval without the continuity correction is $(0.19, 0.81)$. Therefore, if $p < 0.19$ or $p > 0.81$, then the coverage probability (which is based on a minimum for all values of X satisfying the binomial-*mean* rule) is zero. Comparing the three rules based on $\hat{p}$ is hence unreasonable. Moreover, $P(X = 5)$ is only 0.02 when $n = 10$ and $p = 0.19$ or $p = 0.81$. Therefore, we decided that our three rules should be based on p rather than $\hat{p}$.

3 Results and Discussion

With nominal coverage levels of 95%, the empirical coverage probabilities, as based on the minimum among all $p^* \in [p_{\min}, 0.5]$, sometimes tend to be too low, especially when the continuity correction is not used, as we will note from our graphs. The large values of n allow values of p to be close to zero or unity. If one is concerned about how well the normal distribution approximates the binomial distribution for large or small values of p, then the *equal-centers* criterion should be used when comparing the rules. On the other hand, if one is concerned about how well the normal distribution approximates the binomial distribution for values of p near 0.5, then the *equal-tails* criterion should be used when comparing the rules.

Herein, a rule is judged to be superior to another rule, when the former rule has a significantly uniformly higher coverage probability than the latter rule, where the comparison is made within the graphs. Under the *equal-centers* criterion, when confidence intervals are being constructed, the coefficient-of-variation rule (1.3) is clearly superior to the binomial-*mean* rule (1.1), which is superior to the binomial-*variance* rule (1.2), as illustrated in Figure 1. This superiority holds for the one-sided and two-sided cases, both with and without the continuity correction.

Under this same *equal-centers* criterion, when hypothesis testing is being performed, the coefficient-of-variation rule and the binomial-*mean* rule are almost equivalent, except for the two-sided case without the continuity correction, where the coefficient-of-variation rule is slightly superior, as illustrated in Figure 2. Furthermore, the coefficient-of-variation rule and the binomial-*mean* rule are almost equivalent, except for the one-sided case without the continuity correction, where the binomial-*mean* rule is somewhat superior.

Under the *equal-tails* criterion, when one-sided confidence intervals are being constructed, the binomial-*variance* rule (1.2) is clearly superior to the binomial-*mean* rule (1.1), which is superior to the coefficient-of-variation rule (1.3), for small sample sizes ($n < 35$), as illustrated in Figure 3. This superiority of the binomial-*variance* rule also holds for two-sided confidence intervals with the continuity correction for small and moderate sample sizes ($n < 100$), but the binomial-*variance* and binomial-*mean* rules are almost equivalent for two-sided confidence intervals without the continuity correction for $n > 30$.

Under this same *equal-tails* criterion, when hypothesis testing is being performed, all three rules are pretty much equivalent, so none of the rules should be considered superior over the others.

Moreover, in Figures 2 and 4, each of the red, black, and green curves falls above the purple horizontal line. Therefore, when the continuity correction is used with hypothesis testing, all three rules produce coverages exceeding the nominal 95% coverage.

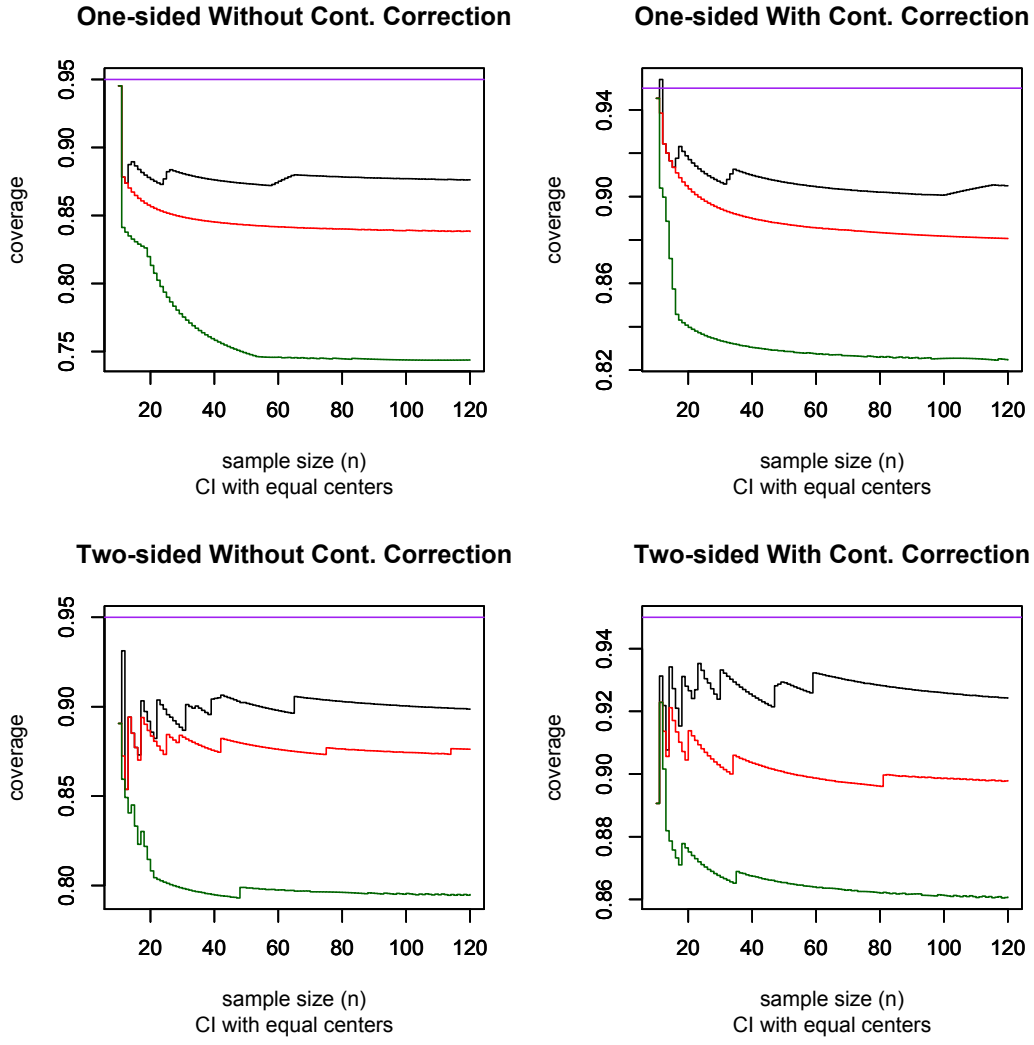

Figure 1. Coverage probabilities, based on the equal-centers criterion for 95% confidence intervals. The **red** curve represents the **binomial-$mean$** rule (1.1). The **green** curve represents the **binomial-$variance$** rule (1.2). The **black** curve represents the **coefficient-of-variation** rule (1.3).

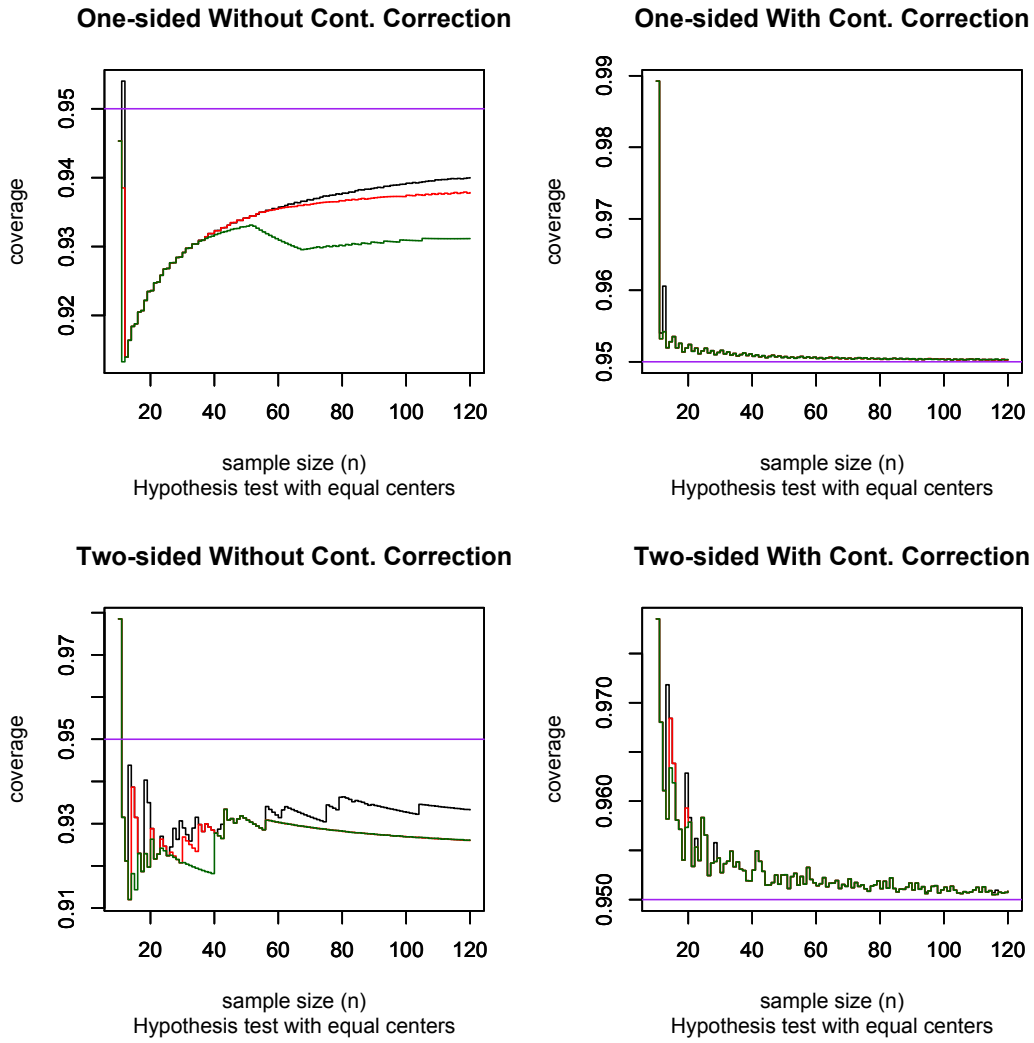

Figure 2. Coverage probabilities, based on the equal-centers criterion for hypothesis tests of level 0.05. The **red** curve represents the **binomial-$mean$** rule (1.1). The **green** curve represents the **binomial-$variance$** rule (1.2). The **black** curve represents the **coefficient-of-variation** rule (1.3).

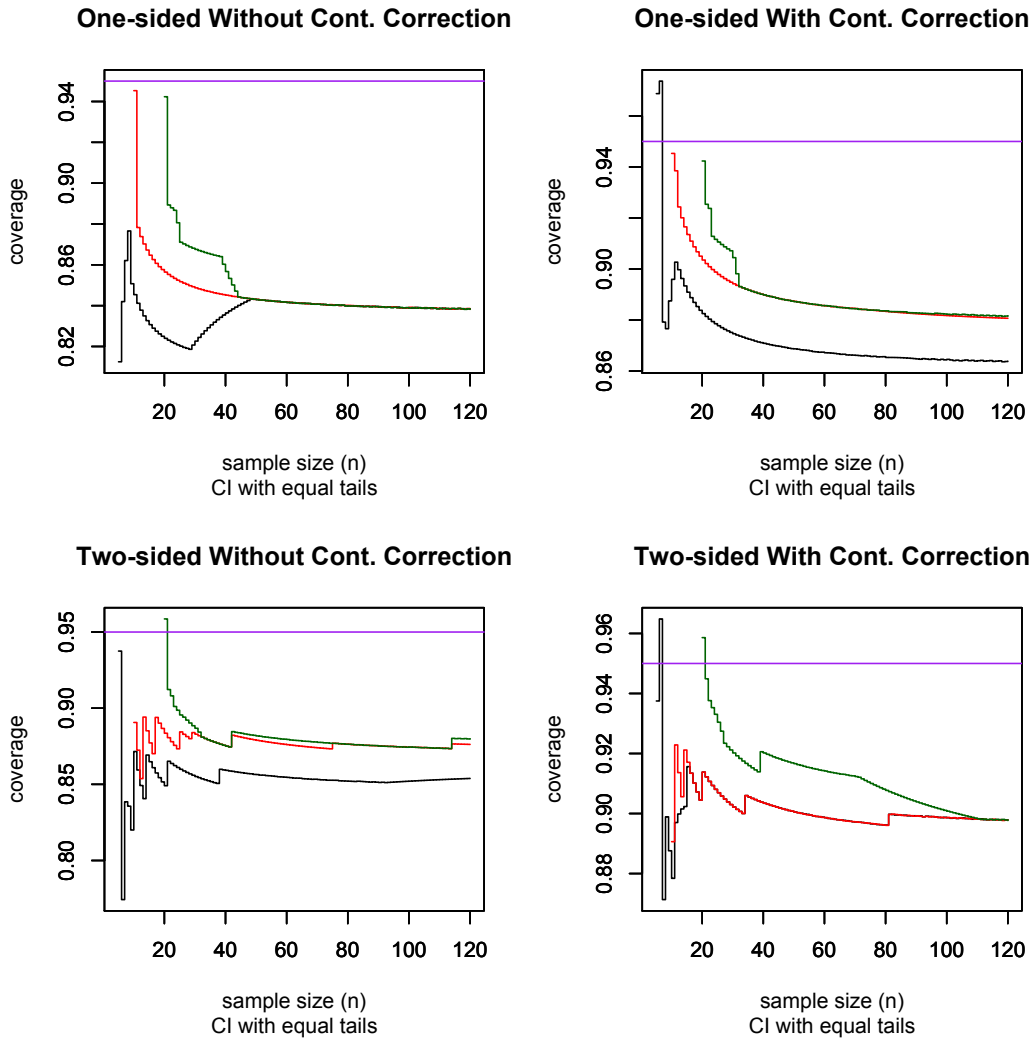

Figure 3. Coverage probabilities, based on the equal-tails criterion for 95% confidence intervals. The **red** curve represents the **binomial-$mean$** rule (1.1). The **green** curve represents the **binomial-$variance$** rule (1.2). The **black** curve represents the **coefficient-of-variation** rule (1.3).

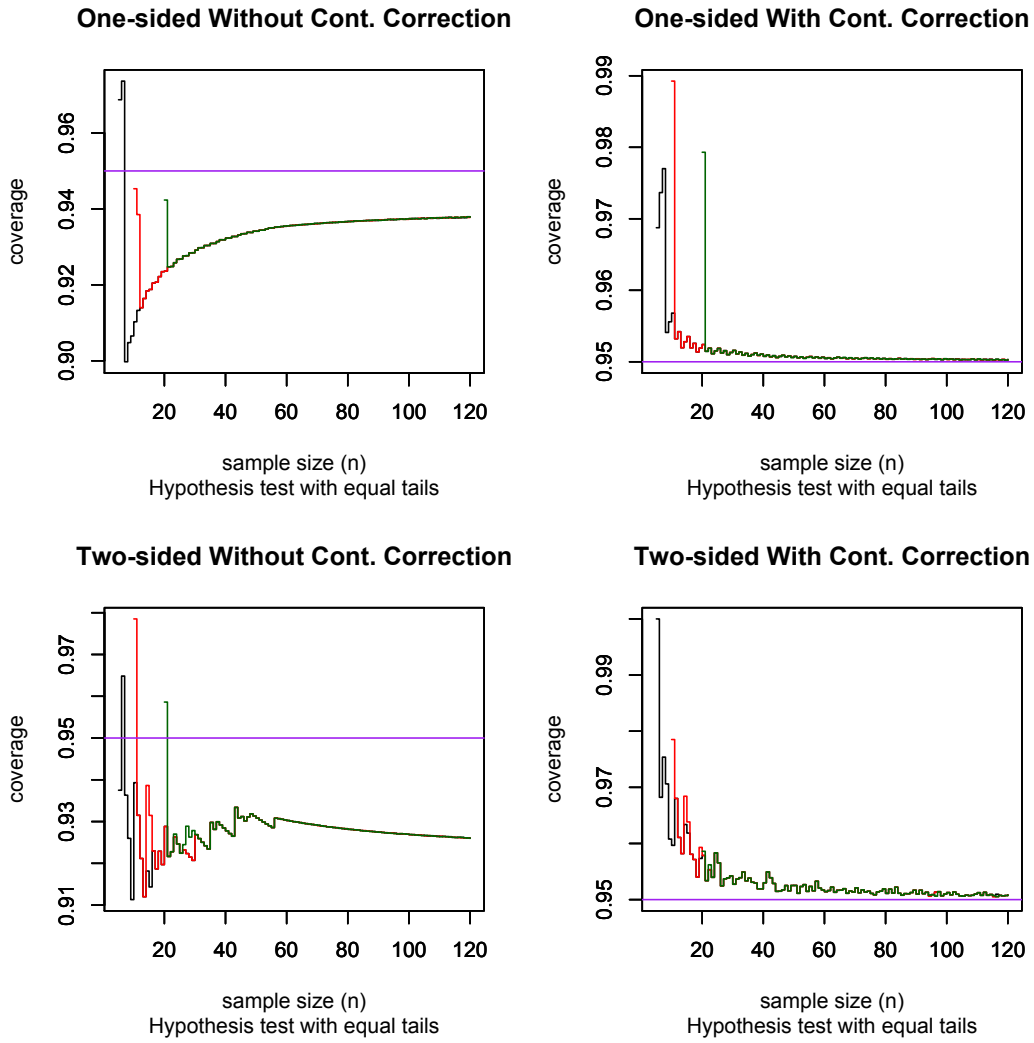

Figure 4. Coverage probabilities, based on the equal-tails criterion for hypothesis tests of level 0.05. The **red** curve represents the **binomial-$mean$** rule (1.1). The **green** curve represents the **binomial-$variance$** rule (1.2). The **black** curve represents the **coefficient-of-variation** rule (1.3).

4 Conclusion

If one is concerned about coverage probabilities when p is near zero or unity when constructing **confidence intervals,** then based on the *equal-centers* criterion, the clear winner is the coefficient-of-variation rule, for the one-sided and two-sided cases, both with and without the continuity correction, as shown in Figure 1. Coming in second place is the binomial-*mean* rule, which greatly improves over the binomial-*variance* rule. When performing **hypothesis tests** under the *equal-centers* criterion, the coefficient-of-variation rule and the binomial-*mean* rule are almost equivalent across the board, and are somewhat superior to the binomial-*variance* rule when not using the continuity correction, as shown in Figure 2.

If one is concerned about coverage probabilities when p is near 0.5 when constructing **confidence intervals,** then based on the *equal-tails* criterion, the winner is the binomial-*variance* rule (1.2), especially for the smaller sample sizes, as shown in Figure 3. However, when performing **hypothesis tests** under the *equal-tails* criterion, all three rules are nearly equivalent, as shown in Figure 4.

Acknowledgements

The author is very thankful to two anonymous reviewers and an editor, whose helpful suggestions added greatly to the quality of this paper.

Competing Interests

The author declares that no competing interests exist.

References

1. Agresti A, Franklin C. Statistics, The Art and Science of Learning from Data, 3rd ed. Pearson, New York; 2013.

2. Moore DS, Notz WI, Fligner MA. The Basic Practice of Statistics, 6th ed. W. H. Freeman, New York; 2011.

3. Navidi W, Monk B. Essential Statistics. McGraw-Hill, New York; 2014.

4. Peck R, Olsen C, Devore J. Introduction to Statistics and Data Analysis, 5th ed. Cengage, Boston; 2015.

5. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical Science. 2001;16(2):101-117.

6. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician. 1998;52(2):119-126.

7. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine. 1998;17:857-872.

8. Devore JL. Probability and statistics for engineering and the sciences, 8th ed., Brooks/Cole, Boston; 2012.

9. Feller W. On the normal approximation to the binomial distribution. The Annals of Mathematical Statistics. 1945;16(4):319-329.

10. Reed JF. Better binomial confidence intervals. Journal of Modern Applied Statistical Methods. 2007;6(1):153-161.

Appendix

R-code for producing Figures 1, 2, 3, and 4

```
p.minimum <- function( n.seq=c(10:50, 1000, 1e4, Inf),
                    mu=5, rule=c("classical","variance","coef_of_var") )   {
   p.min = NULL
   for ( n in n.seq )   {
      if ( is.infinite(n) )  p=0
      if ( is.finite(n) )   {
         if ( rule[1]=="coef_of_var" )    {
            p = ifelse ( n >= mu, mu / ( n + mu ), NA )   }
         if ( rule[1]=="classical" )  {
            p = ifelse( n >= 2 * mu, mu / n, NA )   }
         if ( rule[1]=="variance" )    {
            p = ifelse( n >= 4 * mu, ( 0.5 - sqrt( 1 - 4 * mu / n )/2 ), NA )  }
      }
   p.min = c( p.min, p )    }
   return( p.min )
}


binom.error <- function( n, p, mu=NULL, rule = c("classical","variance",
"coef_of_var"),
               method=c("hyptest","confint"), two.sided=TRUE, correct=TRUE, level=0.95 )
{

   prob0 =   ifelse( two.sided, 0.5 + level/2, level )
   x = 0:n ;  phat = x/n
   if ( method[1]=="confint" )    me = qnorm( prob0 )*sqrt( phat*(1-phat)/n )
   if ( method[1]=="hyptest" )   me = qnorm( prob0 )*sqrt( p*(1-p)/n )
   indicator = 1
   if (two.sided)  indicator = ( (x-0.5*correct)/n-me <= p )
   coverage.prob = sum(  dbinom(x,n,p) * indicator * ( p <= (x+0.5*correct)/n+me )  )
   the.error = level - coverage.prob
   return( the.error )
}


binom.error.vec <- function( n.seq, mu, rule,
   method=c("hyptest","confint"), two.sided=TRUE, correct=TRUE, level=0.95,
 length.out=10000 )
{
   p.min = p.minimum( n.seq=n.seq, mu=mu, rule=rule )
   max.error = rep( NA, length(n.seq) )
   for ( i in 1:length(n.seq) )   {
      if ( is.na( p.min[i] ) )  max.error[i]=NA
      else  {
         max.error[i] = -Inf
         if (p.min[i]>0)   p.seq = seq( p.min[i], 0.5, length.out=length.out )
```

```
        if (p.min[i]==0)  p.seq = 0
        for ( p in p.seq )   {
           the.error = binom.error( n=n.seq[i], p=p, mu=mu, rule=rule, method=method,
                                    two.sided=two.sided, correct=correct, level=level )
           max.error[i] = max(  c( max.error[i], c( the.error ) )  ) } } }
   return( list( p.min=p.min, max.error=max.error ) )
}


all.three.rules <- function( mu=5, method, two.sided=TRUE, correct=TRUE, level=0.95,
                     length.out=10000, equal.tails=TRUE, n.max=100 )   {
   rule = c( "classical", "variance", "coef_of_var" )
   if ( equal.tails)  { mu = rep( mu, length(rule) ) ;
   n.seq = ceiling(mu[1]):n.max }
   if (!equal.tails)  { mu = c( mu, mu/2, mu*2 )     ;
   n.seq = ceiling(mu[1]*2):n.max }
   output1 = binom.error.vec( n.seq=n.seq, mu=mu[1], rule=rule[1], method=method,
                two.sided=two.sided, correct=correct, level=level,
                length.out=length.out )
   output2 = binom.error.vec( n.seq=n.seq, mu=mu[2], rule=rule[2], method=method,
                two.sided=two.sided, correct=correct, level=level,
                length.out=length.out )
   output3 = binom.error.vec( n.seq=n.seq, mu=mu[3], rule=rule[3], method=method,
                two.sided=two.sided, correct=correct, level=level,
                length.out=length.out )
   return( list( "================================================================",
     mu=mu, rule=rule, two.sided=two.sided, correct=correct, level=level,
     length.out=length.out,
     two.sided=two.sided, correct=correct, level=level, method=method,
     frame=data.frame( n.seq=n.seq, p.min.CV=output3$p.min,
           p.min.class=output1$p.min,
           p.min.var=output2$p.min, error.CV=output3$max.error,
           error.class=output1$max.error,
           error.var=output2$max.error ),
     "================================================================" )  )
}


graph <- function( binom.output, level=0.95, main=NULL, sub=NULL )
{
   x  = binom.output$frame[,"n.seq"]
   y1 = level - binom.output$frame[,"error.CV"]
   y2 = level - binom.output$frame[,"error.class"]
   y3 = level - binom.output$frame[,"error.var"]
   xlim.range = range( x, na.rm=TRUE )
   ylim.range = range( c( level, y1[is.finite(y1)], y2[is.finite(y2)],
   y3[is.finite(y3)] ),
       na.rm=TRUE )
   plot( x[is.finite(y1)], y1[is.finite(y1)], xlab="sample size (n)", ylab="coverage",
       xlim=xlim.range, ylim=ylim.range, type="s", main=main, sub=sub )
   par( new=TRUE ); plot( x[is.finite(y2)], y2[is.finite(y2)], xlab="", ylab="",
```

```
       xlim=xlim.range, ylim=ylim.range, col="red", type="s" )
   par( new=TRUE ); plot( x[is.finite(y3)], y3[is.finite(y3)], xlab="", ylab="",
       xlim=xlim.range, ylim=ylim.range, col="darkgreen", type="s" )
   abline(h=level, col="purple")
}


four.graphs <- function( mu=5, method="hyptest", level=0.95, n.max=120,
length.out=10000,
                         equal.tails=TRUE )  {
   par( mfrow = c(2,2) )
   if ( method == "confint" &  equal.tails )   sub = "CI with equal tails"
   if ( method == "confint" & !equal.tails )   sub = "CI with equal centers"
   if ( method == "hyptest" &  equal.tails )   sub = "Hypothesis test with equal
tails"
   if ( method == "hyptest" & !equal.tails )   sub = "Hypothesis test with equal
centers"
   graph(  all.three.rules( mu=mu, method=method,  two.sided=FALSE, correct=FALSE,
           level=level, n.max=n.max, length.out=length.out, equal.tails=equal.tails ),
           level=level, main="One-sided Without Cont. Correction", sub=sub )
   graph(  all.three.rules( mu=mu, method=method,  two.sided=FALSE, correct=TRUE,
           level=level, n.max=n.max, length.out=length.out, equal.tails=equal.tails ),
           level=level, main="One-sided With Cont. Correction", sub=sub )
   graph(  all.three.rules( mu=mu, method=method,  two.sided=TRUE, correct=FALSE,
           level=level, n.max=n.max, length.out=length.out, equal.tails=equal.tails ),
           level=level, main="Two-sided Without Cont. Correction", sub=sub )
   graph(  all.three.rules( mu=mu, method=method,  two.sided=TRUE, correct=TRUE,
           level=level, n.max=n.max, length.out=length.out, equal.tails=equal.tails ),
           level=level, main="Two-sided With Cont. Correction", sub=sub )
}


four.graphs( mu=5, method="confint", level=0.95, n.max=120, length.out=10000,
equal.tails=FALSE )
four.graphs( mu=5, method="hyptest", level=0.95, n.max=120, length.out=10000,
equal.tails=FALSE )
four.graphs( mu=5, method="confint", level=0.95, n.max=120, length.out=10000,
equal.tails=TRUE )
four.graphs( mu=5, method="hyptest", level=0.95, n.max=120, length.out=10000,
equal.tails=TRUE )
```


Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=966&id=22&aid=8469