Reasonable Confidence Limits For Binomial Proportions

THIS IS THE ORIGINAL MSWORD DOCUMENT SUBMITTED TO MD&DI,
FROM WHICH THEY CREATED THE PUBLISHED VERSION.
BOTH VERSIONS ARE INCLUDED HERE; THE WORD VERSION IS SHOWN FIRST,
AND THE PUBLISHED VERSION IS SHOWN NEXT (STARTING AFTER PAGE 26).
TABLES & CHARTS (STARTING ON PAGE 12) IN THIS MSWORD VERSION ARE EASIER TO READ THAN THOSE IN THE PUBLISHED VERSION.

UPDATE October 21, 2013: In this VERSION 3 of the MS Word document, the beta-distribution formulas for Reasonable Confidence Limits have been corrected (there was an error in Version 2) and changed to more closely match the format of the formulas found in the source article by Krishnamoorthy (the beta-formulas in the published MD&DI version of this document are correct but relatively cumbersome); and a relevant "NOTE" has been added below those formulas in this document.

SUMMARY

This article adds an entry to the long list of binomial confidence intervals introduced since 1934. The importance of the new entry is demonstrated by comparison to the two most commonly used intervals: the "Wald" and the "Exact". Those two intervals are so wide that their confidence limits can be statistically significantly different from the sample proportion on which basis they were calculated. In contrast, the interval introduced in this article is as wide as possible without its limits differing significantly from the sample proportion. Thus the decision to use this new interval may be easier to defend to regulatory bodies such as the FDA; use of any other interval may appear to be arbitrary or unreasonable. The proposed name for the new entry is: a "Reasonable Confidence Interval for a Binomial Proportion", the extreme values of which would then be "Reasonable Confidence Limits". Also discussed is the simultaneous use of two intervals (the Exact and the Reasonable).

KEY WORDS: binomial confidence interval; binomial confidence limit; proportion; Wald; normal approximation; Exact; Reasonable; Score Method, zone of uncertainty.

INTRODUCTION

Confidence intervals are serious business! Recently, the FDA told a medical-device start-up company that, in regards to the company's proposed clinical trial, "The equivalence of the device to the predicate can be demonstrated if the confidence interval for the difference in the mean values for the tested parameter excludes a difference larger than 20% from the predicate." Unfortunately, the company could not meet that FDA mandate because in order to reduce the width of the confidence interval so that it would achieve that exclusion, a much larger number of patients was required than the company could afford.

The term "confidence interval" was coined and first published by J. Neyman in 1934, when he applied it to binomial as well as variables data. He described confidence intervals as ranges "in which we may assume are contained the values of the estimated characters of the population."[1] As applied to a "proportion...of individuals in the sample" (which in this article will be represented by "P_S") that has been derived from an unknown "proportion...of individuals in the population" ("P_P", in this article) whose "distribution...is then a binomial", he defined the confidence interval for P_P as having the form $P_L \leq P_S \leq P_U$, where P_L is what we now call the "lower confidence limit", P_U is the "upper confidence limit", and P_P is assumed with a specified level of "confidence" to be somewhere in the interval[2] $P_L - P_U$. A few months later, the first rigorous method for calculating such an interval was published.[3]

There is only one generally accepted method for calculating confidence intervals for variables data; that method involves "t" tables and the standard error of the mean, as described in any basic statistics text book. However, there are many methods in use for calculating confidence intervals for binomial data, each such method resulting in different confidence limits and different interval widths.[4, 5]

The reason that there are many binomial methods is partly historical and partly theoretical. The historical part is that, prior to the widespread availability of computers, binomial calculations were difficult. For that reason, simpler-to-calculate alternatives were developed; as was said several decades ago, "To calculate [*large-sample binomial*] probabilities would be an almost insurmountable task. Therefore, some method of approximation must be used."[6] The theoretical part is the lack of agreement on criteria for judging an interval (we'll discuss this in more detail later in this article). That disagreement was present from the birth of the interval concept[2, 7] in 1934: Neyman's description for the interval was the equivalent of $P_L \leq P_S \leq P_U$, whereas Clopper & Pearson's description was $P_L < P_S < P_U$; notice the use of ≤ vs. <.

Some of those many binomial confidence interval methods involve formulaic calculations: for example, what is commonly referred to as the "Wald" formula uses a "binomial standard deviation" coupled with a Standard Normal Distribution Z-table to calculate an approximate confidence interval; Wald intervals are based upon the fact that when sample size is large or P_S is near 50%, the Normal distribution is a "reasonable" model of the Binomial, even if such an application is "not completely accurate."[8] Other methods involve trial-and-error: for example, in order to calculate each confidence limit for what is commonly called the "Exact binomial confidence interval", repeated attempts must be made to determine the proportion that yields a cumulative binomial histogram probability of exactly half the chosen significance level (we'll discuss this in more detail, below).

It has been said recently that the Wald interval is "in virtually universal use."[9] Similarly, the National Institute for Standards and Technology (NIST) "*e-Handbook of Statistical Methods*" website states that the Wald formula is the "confidence *[interval]* expression most frequently used."[10] And Wald's was the only method included in a binomial-calculation-spreadsheet issued in 1998 by the CDRH division of the FDA, for general use

in handling clinical trials data.[11] On the other hand, the Exact interval is the sole method provided in some mainstream statistical software programs (*e.g.*, Statgraphics™)[12], and the Exact is the only non-Z-table method for calculating binomial proportion confidence limits that is mentioned on the NIST website[10]; indeed, some statisticians (*e.g.*, Agresti and Coull) refer to the Exact method as the "gold standard."[13]

The implicit if not explicit focus of interest in a binomial confidence interval is its largest and smallest values, *i.e.*, P_U and P_L. For example, a clinical trial might be considered successful only if the lower confidence limit on the outcome success rate is larger than a protocol-specified value. Because of such focus, this paper introduces a new criterion for comparing the validity of confidence interval methods (since 1934, many other criteria have been proposed[14]). The new criterion is this: Are the confidence limits of the interval "reasonable"? If it is unreasonable to conclude that P_L and P_U could be P_P, then it is unreasonable to use the confidence interval method that generated them. And being "reasonable" is what even J. L. Fleiss has urged: "A confidence interval for a statistical parameter is a set of values that are...reasonable candidates for being the true underlying value"[15] (underlining not in the original text).

REASONABLE CONFIDENCE INTERVALS AND LIMITS

Before we define what it means to be "reasonable", let's explain the basis of the definition. A test of reasonableness is equivalent to performing a binomial test of significant difference using what J. L. Fleiss has called the "traditional statistical approach" for an "inference for a single proportion". It involves "calculating the probability, assuming the null hypothesis holds, of obtaining the outcome that actually occurred, plus the probabilities of all other outcomes as extreme as, or more extreme than, the one that was observed; and rejecting the null hypothesis in favor of the alternative hypothesis if the sum of all these probabilities — the so-called *p-value* — is less than or equal to a predetermined level, denoted by α, called the *significance level.*"[16]

Based on that approach, let's call a confidence interval "reasonable" only if such a test results in a conclusion of "not statistically different" when the test compares the observed Sample proportion (P_S) to either of the two most extreme values in the interval, namely the upper and lower confidence limits (P_U and P_L, respectively). In terms a bit more mathematical, let's define "unreasonable" as follows (for sample size = N, observed number of successes in that sample = K, and K/N = P_S): In regards to the probability distribution histograms derived from limits P_L and P_U, if either of the distribution tails in which K is found represents a cumulative probability of occurrence of less than or equal to $\alpha/2$, we conclude that P_S is statistically significantly different from the limit that generated that distribution, and that therefore the confidence interval $P_L - P_U$ is unreasonable.

In that definition, we use $\alpha/2$ rather than α because we are performing a 2-sided test twice. The first test determines whether a random sample proportion differs from P_L; the second test determines whether that random sample proportion differs from P_U; in each case, we

focus on tails that represent $\alpha/2$ of the distribution (see Figures 1A and 1B). That is the standard way to approach such tests.[10, 17, 18, 19]

It is important to note that the test of significance just described is performed using the probability distribution histograms generated from the two confidence limits (P_L and P_U) rather than being performed using the single probability distribution histogram generated from the observed Sample proportion (P_S). Such an approach is used because P_L and P_U are together considered to be the "null hypothesis"; based upon the "new criterion" described above, the question we are trying to answer is this: Is it reasonable to assume that the random sample proportion P_S could have been obtained from a population that had a proportion equal to either P_L or P_U (*i.e.*, could either of them be Fleiss's "true underlying value", P_P)?[15, 17, 18]

Let's examine the "reasonableness" of Wald limits, after we first explain how to calculate them. As demonstrated on the NIST website[20], calculation of the upper and lower confidence limits of a Wald ("Normal Approximation") binomial confidence interval uses the following formula: $P_S +/- Z\alpha/2 \times SD_{PS}$, where P_S is the observed Sample proportion, $Z\alpha/2$ is the two-tailed value from a normal distribution Z-table at the chosen α significance level, and SD_{PS} is the binomial standard deviation for the observed proportion, calculated as the square root of $P_S(1-P_S)/N$. The limits of such an interval can be calculated as shown below, using Microsoft® Excel™ (subscript "W" indicates the "Wald" method, "Normsinv" and "Sqrt" are MS Excel functions that output Z-table values and square roots respectively, " * " is the MS Excel symbol for "multiply", and the other terms are as defined above):

$$P_{UW} = P_S + \text{Normsinv}(1-\alpha/2) * \text{Sqrt}\left(P_S * (1-P_S) / N\right)$$

$$P_{LW} = P_S - \text{Normsinv}(1-\alpha/2) * \text{Sqrt}\left(P_S * (1-P_S) / N\right)$$

Because the "normal approximation" assumption becomes less valid as the sample size becomes smaller and/or as P_S departs farther from 0.500 (50%), the Wald calculation is typically restricted to situations where both of the following are true: $N(P_S) > 5$ and $N(1 - P_S) > 5$. Even the FDA's spreadsheet[11] includes the warning "minimum [*N(PS), N(1 – PS)*]...must be > 5 to use normal approximation".

Let's examine the reasonableness of Wald confidence limits for the following situation: $\alpha = 5\%$ (and therefore Confidence = $1 - \alpha = 95\%$), Sample size = N = 100, Successes = K = 10, $P_S = K/N = 10/100 = 0.10$, and $N(P_S) = 10$. The resulting limits are: $P_{LW} = 0.041201116$ and $P_{UW} = 0.158798884$; how reasonable are they?

Let's focus just on P_{LW}, the lower limit. The probability distribution for a population whose proportion equals that P_{LW} is shown in Figure 2. Notice that the probability of occurrence of the observed Sample result or a more extreme result (*i.e.*, $K \geq 10$) is less than $\alpha/2 = 2.5\%$ (in fact, it is 0.8%); we therefore conclude that P_S is statistically significantly different from P_{LW}. Based upon our definition of reasonableness, it is therefore unreasonable to conclude that $P_{LW} = P_P$. If it is unreasonable to conclude that

$P_{LW} = P_P$ (at $\alpha = 5\%$), then it is unreasonable to consider $P_{LW} - P_{UW}$ to be the $1 - \alpha = 95\%$ confidence interval. As evidenced by this example, Wald intervals and confidence limits can be unreasonable.

Let's now examine the reasonableness of Exact limits, after we first explain how to calculate them. What are sought (by trial and error) are two proportions, one larger and one smaller than the observed sample proportion (P_S); each must have a cumulative binomial probability of "exactly" $\alpha/2$ for obtaining the observed Sample result or a more extreme value (*i.e.*, 0 to K, or K to N). An MS Excel spreadsheet can be used to calculate the limits as accurately as (for example) Statgraphics, to at least a billionth of a probability unit (= 9 places to the right of the decimal point); how to do so is described on the NIST website.[10] The following is a generic example of applying the NIST/MSExcel method (α, N, K, P_P, P_S, P_U, and P_L are as defined above; subscript "E" indicates the "Exact" method; and P is a proportion we seek):

P_{UE} = the value of P ($P > P_S$) needed to ensure that the MS Excel function Binomdist(K, N, P, True) outputs a probability value of $\alpha/2$ precisely (to the desired number of significant digits).

P_{LE} = the value of P ($P < P_S$) needed to ensure that the MS Excel function Binomdist(K−1, N, P, True) outputs a probability value of $1 - \alpha/2$ precisely (to the desired number of significant digits).

Those formulas, applied to the situation we evaluated previously (N = 100, K = 10, P_S = 0.10, and $\alpha = 5\%$), result in the following Exact limits: $P_{LE} = 0.049004689$ and $P_{UE} = 0.176222598$. The probability distribution for a population whose proportion equals that P_{LE} is shown in Figure 3. Notice that the probability of occurrence of the observed Sample result or a more extreme result (*i.e.*, $K \geq 10$) is precisely equal to $\alpha/2 = 2.5\%$; the entire histogram bar representing K=10 (the value observed in the Sample) is found in the 2.5% "tail" of the distribution; we therefore conclude that $P_S = K/N$ is statistically significantly different from P_{LE}. Based upon our definition of reasonableness, it is therefore unreasonable to conclude that $P_{LE} = P_P$. If it is unreasonable to conclude that $P_{LE} = P_P$ (at $\alpha = 5\%$), then it is unreasonable to consider $P_{LE} - P_{UE}$ to be the 95% confidence interval. As evidenced by this example, Exact confidence intervals and limits can be unreasonable.

Let's now introduce formulas for the "Reasonable" confidence limits being introduced in this article. The limits of a "Reasonable binomial confidence interval" are defined as follows (where the subscript "R" indicates the "Reasonable" method):

P_{UR} = the value of P ($P > P_S$) needed to ensure that the MS Excel function Binomdist(K−1, N, P, True) outputs a probability value of $\alpha/2$ precisely (to the desired number of significant digits).

P_{LR} = the value of P ($P < P_S$) needed to ensure that the MS Excel function Binomdist(K, N, P, True) outputs a probability value of $1-\alpha/2$ precisely (to the desired number of significant digits).

Notice that the first term in the MS Excel functions for "Reasonable" limits is changed by a value of 1 from its corresponding "Exact" function (in the P_U definitions, the change is from K to K–1, and in the P_L formulas it is from K–1 to K). That was done to ensure that P_S is not statistically significantly different from either P_{LR} or P_{UR}. In effect, those two formulas identify the widest possible confidence interval such that the observed Sample proportion is not statistically significantly different from any point in the interval, most especially the highest and lowest points, namely P_{UR} and P_{LR} (the meaning of "statistically significantly different" is discussed further, later in this article).

If those Reasonable method formulas are applied to the situation we evaluated previously (N = 100, K = 10, P_S = 0.10, and α = 5%), they result in the following Reasonable limits: P_{LR} = 0.056207020 and P_{UR} = 0.163982255. The probability distribution for a population whose proportion equals that P_{LR} is shown in Figures 4A and 4B. Notice that, in Figure 4A, the probability of occurrence of the observed Sample result or a more extreme result (*i.e.*, $K \geq 10$) equals approximately 5.5 %; in Figure 4B, the entire histogram bar representing K=10 is found in the $(1 - \alpha/2)$% body+lower_tail (*i.e.*, in the lower 97.5% of the distribution); therefore, the observed Sample proportion ($P_S = K/N = 0.10$) is not statistically significantly different from P_{LR}, and therefore P_{LR} could possibly be P_P. Similarly, as seen in Table 1, it is reasonable to conclude that P_{UR} could possibly be P_P, since the observed Sample result or a more extreme result (*i.e.*, $K \leq 10$) is equal to approximately 4.9% and the probability of $K \leq 9$ is precisely 2.5%. Therefore, because both P_{LR} and P_{UR} could possibly be P_P (at α = 5%), it is indeed reasonable to consider $P_{LR}-P_{UR}$ to be the 95% confidence interval.

At first glance, the Wald method has an advantage over Reasonable and Exact ones: Calculation of Wald limits can be performed without trial-and-error and therefore can be automated using simple computer applications such as MS Excel functions. Upon further investigation, we discover that the Beta-distribution formulas that have been used to approximate Exact limits without using trial-and-error[21] can be modified to also approximate Reasonable limits. As shown in Table 2, the following MS Excel formulas approximate the output of the Reasonable binomial formulas shown above, to at least a millionth of a probability unit (= 6 places to the right of the decimal point) (α, N, K, P_{UR}, and P_{LR} are as defined above, and "BetaInv" is an MS Excel function):

$$P_{UR\,(BETA)} = \text{BetaInv}\ (\ 1 - \alpha/2,\ K\ ,\ N - K + 1\ )$$

$$P_{LR\,(BETA)} = \text{BetaInv}\ (\ \alpha/2,\ K+1\ ,\ N - K\ )$$

NOTE: When either K = 0 (P_S = 0.000) or K = N (P_S = 1.000), those two beta-formulas produce error messages or nonsense results (see next paragraph, below).

Let's now turn our attention to the most extreme values of P_S: When P_S equals 0.000 precisely, no method can calculate a Lower Confidence Limit (P_L), because proportions (*i.e.*, probabilities) lower than zero are undefined; and when P_S equals 1.000 precisely, no method can calculate P_U, because proportions above unity are likewise undefined. Not surprisingly, the P_L for $P_S = 1.000$ and the P_U for $P_S = 0.000$ are called "one-sided limits".

Such one-sided limits can be calculated by the Exact method but not by Wald or Reasonable ones. Wald limits for $P_S = 0.000$ or 1.000 cannot be calculated because in either case the binomial standard deviation itself equals 0.000, no matter what the sample size is (recall that SD_{PS} = square root of $P_S(1-P_S)/N$). Similarly, Reasonable limits cannot be calculated because if K = 0 ($P_S = 0.000$), then the P_{UR} binomdist formula value "K−1" is meaningless; and if K = N ($P_S = 1.000$), then the P_{LR} binomdist formula always results in a probability of 1.000 and therefore can never equal the sought-after "value of P ($P < P_S$)".

On the other hand, Reasonable and Exact methods (but not the Wald method) share the following advantage: They can calculate two-sided confidence limits for any P_S greater than zero and less than unity, no matter how small or large (*e.g.*, if sample size is a million and K = 1, then $P_S = 0.000001 = 1.000E\text{-}6$; and $P_{LR} = 0.242E\text{-}6$).

A common criterion for evaluation of confidence interval methods is "coverage". That term refers to the percentage of time the interval can be expected to include P_P, the "true underlying value". Typically, that percentage is determined experimentally[4, 5, 14] by generating confidence intervals for thousands of random samples drawn from populations of known proportions (*i.e.*, known P_P's), and then determining what percentage of those intervals contain the corresponding P_P. As seen in Figure 5, Reasonable confidence limits are completely contained within Exact ones, and therefore Reasonable intervals can be expected to have slightly less "coverage"; experimental results support that conclusion (see Table 3).

The distinctiveness of Reasonable Intervals and their Limits is not apparent when the limits are plotted in the still-common manner introduced in 1934 by Neyman[22] (see also Figure 6). However, alternate plotting methods (*e.g.*, see Figure 7) clearly demonstrate that Reasonable Intervals are narrower than Wald or Exact ones. As a result, only Reasonable Intervals are narrow enough to have Upper and Lower limits that are truly "reasonable". As Neyman insisted: "confidence intervals should be as narrow as possible."[23]

RECOMMENDATIONS

With variables data, a test of significance is mathematically equivalent to use of a confidence interval[24]; *i.e.*, a borderline value being compared to the sample result is considered significantly different if it is outside the sample's confidence interval, but non-significant if inside. Likewise, a borderline binomial proportion is classically viewed as being inside or outside the sample proportion's confidence interval. But such a view is misleading for proportions, since they are based upon counts; because counts can take only discrete or "discontinuous" values, the observed result's probability distribution

histogram bar appears as if it spans the border between significance and non-significance, when the histogram is based upon a borderline value.

For example: In the case of N=100 and observed result K=10, we've already discussed (see Figure 3) that probability distribution histograms based upon $P \leq P_{LE} = 0.0490$ result in the histogram bar for K=10 being in the upper 2.5% tail of the distribution. Similarly, we've discussed (see Figure 4B) that if the histogram is based upon $P \geq P_{LR} = 0.0562$, then K=10 is in the 97.5% body+lower_tail of the distribution. The problem is that histograms based upon P values between P_{LE} and P_{LR} result in the observed-K histogram-bar being neither fully in the 2.5% upper tail nor fully in the 97.5% body+lower tail (see Figures 8A and 8B). In such cases, how do we objectively conclude significance or non-significance? Either conclusion could be viewed as unreasonably subjective and arbitrary. A solution, introduced in this article, is to always use not one but two confidence intervals: the Exact and the Reasonable.

Before explaining that solution, a brief history lesson is useful (in the following discussion, "Lδ" is the likelihood of obtaining the observed result "δ", assuming the null hypothesis is true). The concept of "statistical significance" has undergone much change since the precursors to modern tests of significance were developed in the 1800s; in that century, δ was not considered significant unless it was extremely unlikely.[25] As decades passed, the requirement for significance became less extreme. By the 1930's it could be said that "it is conventional among certain workers to adopt the following rule: If $L\delta \geq 0.05$, δ is not significant; if $L\delta \leq 0.01$, δ is significant; if $0.05 > L\delta > 0.01$, our conclusions about δ are doubtful and we cannot say with much certainty whether the deviation is significant or not until we have additional information"[26] (the original text uses a different symbol than "L" and does not include underlining). In effect, the region between 0.01 and 0.05 was considered a "zone of uncertainty" (a term not in the original text).

On the basis of that history lesson, perhaps the best solution to the problem of borderline proportions is to consider the range of values between P_{LE} and P_{LR}, and between P_{UR} and P_{UE}, to be "zones of uncertainty" (see Figure 9). Using that approach, and assuming your sample size has the "power" to detect a clinically significant difference, if the Null Hypothesis proportion (P_{NH}) being compared to the study result (P_S) is outside the study result's Exact confidence interval (*i.e.*, $P_{NH} \leq P_{LE}$ or $P_{UE} \leq P_{NH}$), you can claim statistical significance. On the other hand, if P_{NH} is inside the corresponding Reasonable interval (i.e., $P_{LR} \leq P_{NH} \leq P_{UR}$), you can claim statistical non-significance. However, if P_{NH} is in either of the zones of uncertainty (that is, either $P_{LE}<P_{NH}<P_{LR}$ or $P_{UR}<P_{NH}<P_{UE}$), then your results are statistically inconclusive.

FDA NOTES

A search of the website FDA.gov finds recent submissions, panel opinions, and guidance documents that include a variety of binomial confidence interval methods, the most common being the Exact and the "Score" (the Score formula is an elaborated classic Wald, from the perspective that it uses a binomial standard deviation and not one but a

few copies of a normal distribution Z-table value; Score confidence limits are found midway between Exact and Reasonable ones). In a 2007 "Statistical Guidance" document, the FDA recommends "score confidence intervals, and alternatively, exact (Clopper-Pearson) confidence intervals."[27] Exact vs. Score methods were compared briefly in a 2009 FDA publication: "An advantage with the Score method is that...it can be calculated directly. Score confidence bounds tend to yield narrower confidence intervals than Clopper-Pearson [*i.e.*, *Exact*] confidence intervals, resulting in a larger lower confidence bound."[28] It is interesting to note that that recommendation was based on ease-of-calculation rather than on theoretical correctness.

On the other hand, in 2003, Medtronic Neurological received approval for its Active Dystonia Therapy ("deep brain stimulation system") by using the classic Wald method almost exclusively: "Exact 95% confidence intervals were used when the # (%) of patients was 0 (0%) because the normal approximation to the binomial does not provide a confidence interval. In every other case, the normal approximation to the binomial was used to calculate confidence intervals" even though in very many of those cases the "N(Ps) > 5" criterion (mentioned above) was not met[29] (underlining was not in the original text).

The new methods proposed in this article have not yet been used in any regulatory submission to the FDA, Health Canada, or a Notified Body, as far as this author knows.

CONCLUSION

The most commonly used binomial confidence intervals can have confidence limits that are statistically significantly different from the random sample proportion from which they were derived. Therefore, such confidence limits are unreasonable to consider as being the proportion of the population from which the random sample was taken. A more defensible choice of confidence intervals is presented in this article, namely a "Reasonable Confidence Interval for a Binomial Proportion", the extreme values of which are "Reasonable Confidence Limits". Although such an interval is slightly narrower than other intervals and thus offers less "coverage", it is more reasonable to use because it is the widest possible range that contains no values that are statistically significantly different from the sample proportion on which basis the interval was calculated. Alternatively, an even more reasonable approach is to use a combination of Exact and Reasonable limits, coupled with the concept of "zones of uncertainty".

REFERENCES

1. J. Neyman, "On the Two Different Aspects of the Representative Method," *Journal of the Royal Statistical Society*, 97, no. 4 (1934): p. 562.

2. Neyman (1934): p. 589.

3. C. Clopper and E. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, no. 4 (1934): p. 409.

4. L. D. Brown, *et. al.*, "Confidence Intervals for a Binomial Proportion and Asymptotic Expansions," The Annals of Statistics, 30, no. 1 (2002): 160–201.

5. R. G. Newcombe, "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics in Medicine*, 17 (1998): 857-872

6. H. Bancroft, *Introduction to Biostatistics* (New York: Hoeber Medical Division of Harper & Row, 1957): p. 106.

7. Clopper (1934): p. 404.

8. H. Motulsky, Intuitive Biostatistics (New York, NY: Oxford University Press, 1995): p. 18.

9. Brown (2002): p. 160.

10. *NIST/SEMATECH e-Handbook of Statistical Methods*, last updated: 7/18/2006, http://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

11. FDA Website Document, "*Two Group Confidence Interval & Power Calculator*" version 1.6 (Issued March 26, 1998).

12. Statgraphics™ Centurion XV, version 15.2.12 (Copyright 1982-2007 by StatPoint, Inc.)

13. L. D. Brown, *et. al.*, "Interval Estimation for a Binomial Proportion," *Statistical Science*, 16, no. 2 (2001): p. 117.

14. M. D. deB. Edwardes, "The Evaluation of Confidence Sets With Application to Binomial Intervals," Statistica Sinica, 8, (1998): pp. 393-409.

15. J. L. Fleiss, *et. al.*, Statistical Methods for Rates and Proportions, 3rd ed. (Hoboken, NJ: John Wiley & Sons, 2003): p. 22.

16. Fleiss (2003): pp. 18-19.

17. J. L. Phillips Jr., *How to Think About Statistics*, revised ed. (New York, NY: W. H. Freeman & Co., 1992): pp. 62-64.

18. W. Mendenhall, *Introduction to Probability and Statistics*, 5th ed. (North Scitueate, MA: Duxbury Press,1979): pp. 231-232.

19. *NIST*: http://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm

20. *NI*ST: http://www.itl.nist.gov/div898/handbook/prc/section2/prc24.htm

21. K. Krishnamoorthy, *Handbook of Statistical Distributions with Applications* (Boca Raton, FL: Taylor & Francis Group, 2006): p. 38.

22. Neyman (1934): p. 590.

23. Neyman (1934): p. 563.

24. Motulsky (1995): pp. 106-117.

25. S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900* (Cambridge, MA: Belknap Press of Harvard University Press, 1986): pp. 300ff.

26. J. F. Kenney, *Mathematics of Statistics (Part One & Part Two)* (New York, NY: D. Van Nostrand Company, 1939): Part Two, p. 117.

27. FDA Website Document, "*Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests, Draft Guidance*" (Issued March 13, 2007): p. 23.

28. FDA Website Document, *"Assay Migration Studies for In Vitro Diagnostic Devices, Draft Guidance"* (Issued January 5, 2009): p. 31.

29. FDA Website Document, *"Summary Of Safety And Probable Benefit, Humanitarian Device Exemption (HDE) Number: H020007"* (Issued April 15, 2003): p. 9.

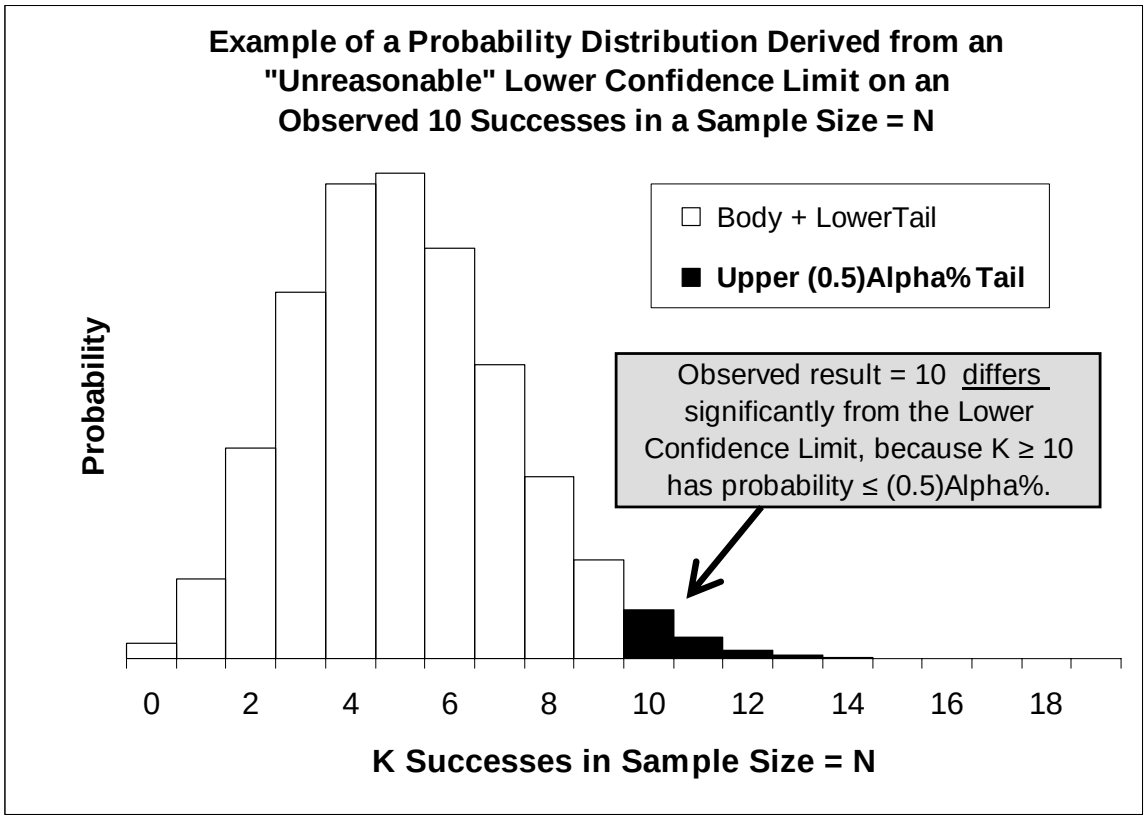
Example of a Probability Distribution Derived from an "Unreasonable" Lower Confidence Limit on an Observed 10 Successes in a Sample Size = N
Body + LowerTail
Upper (0.5)Alpha% Tail
Observed result = 10 differs significantly from the Lower Confidence Limit, because K ≥ 10 has probability ≤ (0.5)Alpha%.
Probability
0
2
4
6
8
10
12
14
16
18
K Successes in Sample Size = N

Figure 1A

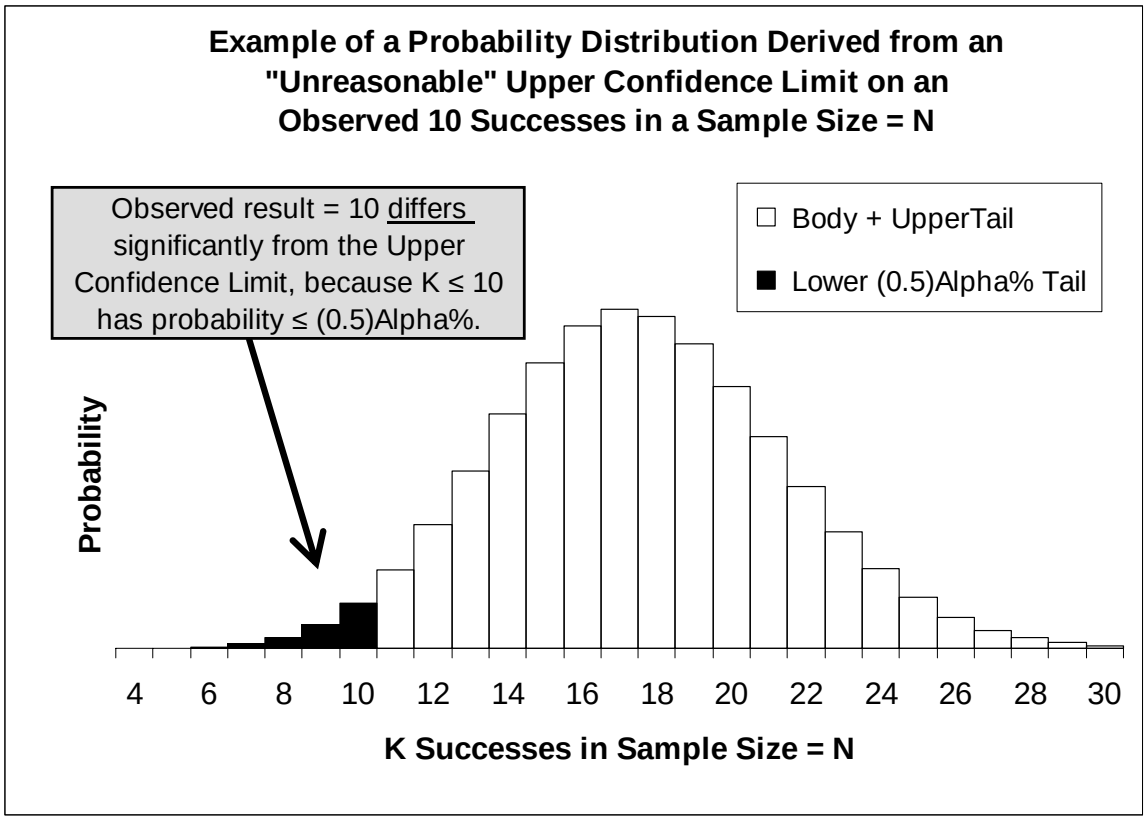
Example of a Probability Distribution Derived from an "Unreasonable" Upper Confidence Limit on an Observed 10 Successes in a Sample Size = N
Observed result = 10 differs significantly from the Upper Confidence Limit, because K ≤ 10 has probability ≤ (0.5)Alpha%.
Body + UpperTail
Lower (0.5)Alpha% Tail
Probability
4 6 8 10 12 14 16 18 20 22 24 26 28 30
K Successes in Sample Size = N

Figure 1B

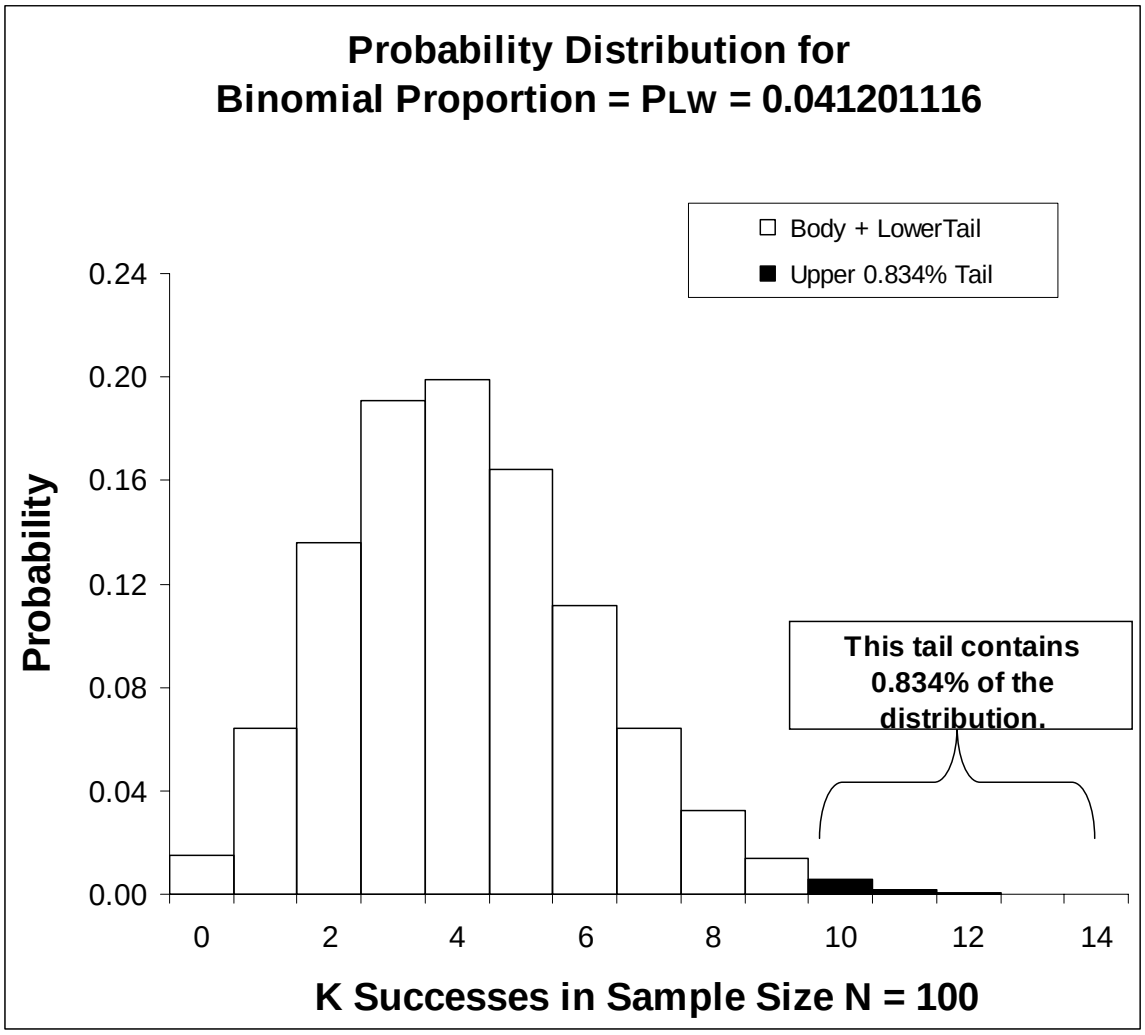
Probability Distribution for
Binomial Proportion = PLW = 0.041201116
Body + LowerTail
Upper 0.834% Tail
0.24
0.20
0.16
0.12
0.08
0.04
0.00
Probability
This tail contains
0.834% of the
distribution.
0
2
4
6
8
10
12
14
K Successes in Sample Size N = 100

Figure 2

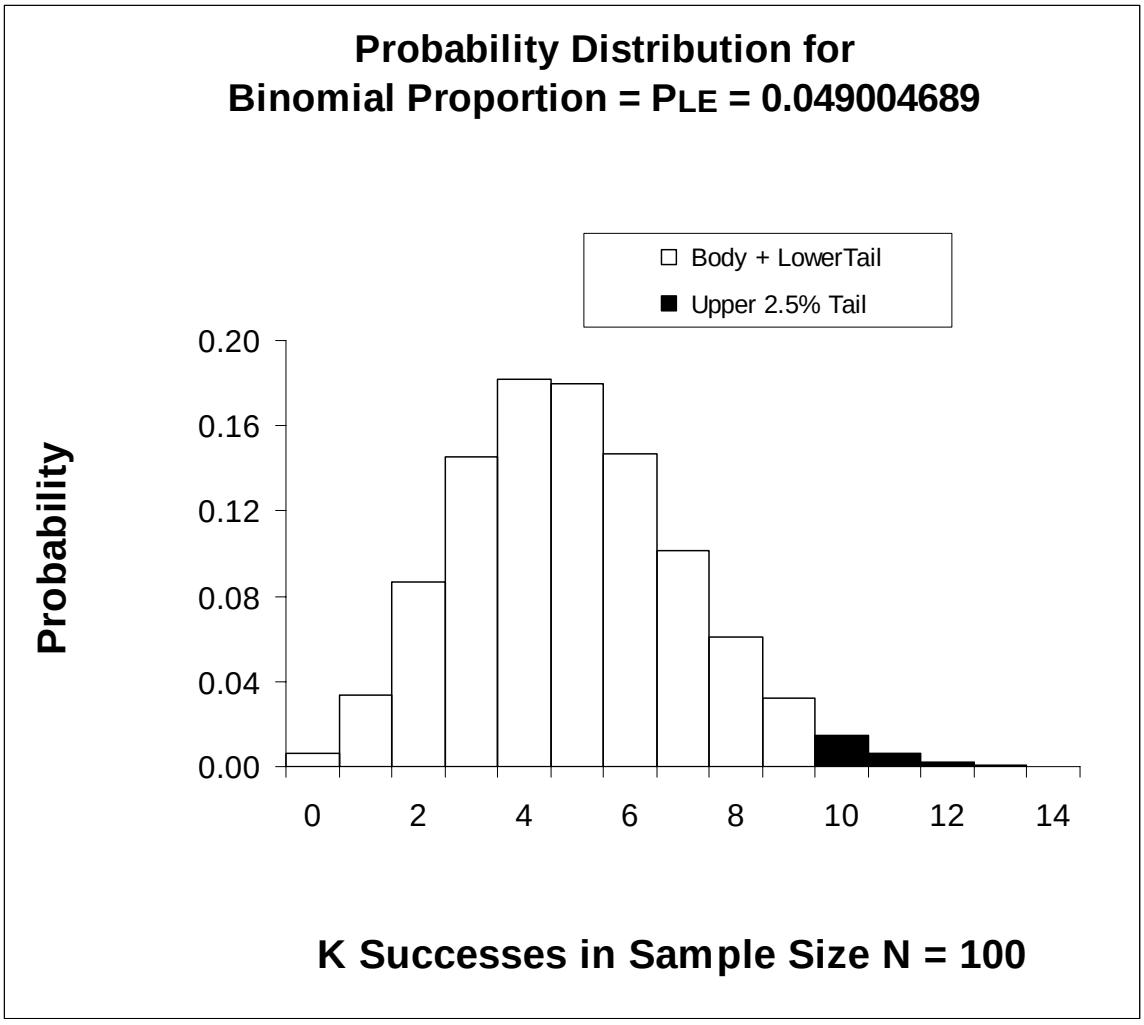
Probability Distribution for
Binomial Proportion = PLE = 0.049004689
Body + LowerTail
Upper 2.5% Tail
Probability
0.20
0.16
0.12
0.08
0.04
0.00
0
2
4
6
8
10
12
14
K Successes in Sample Size N = 100

Figure 3

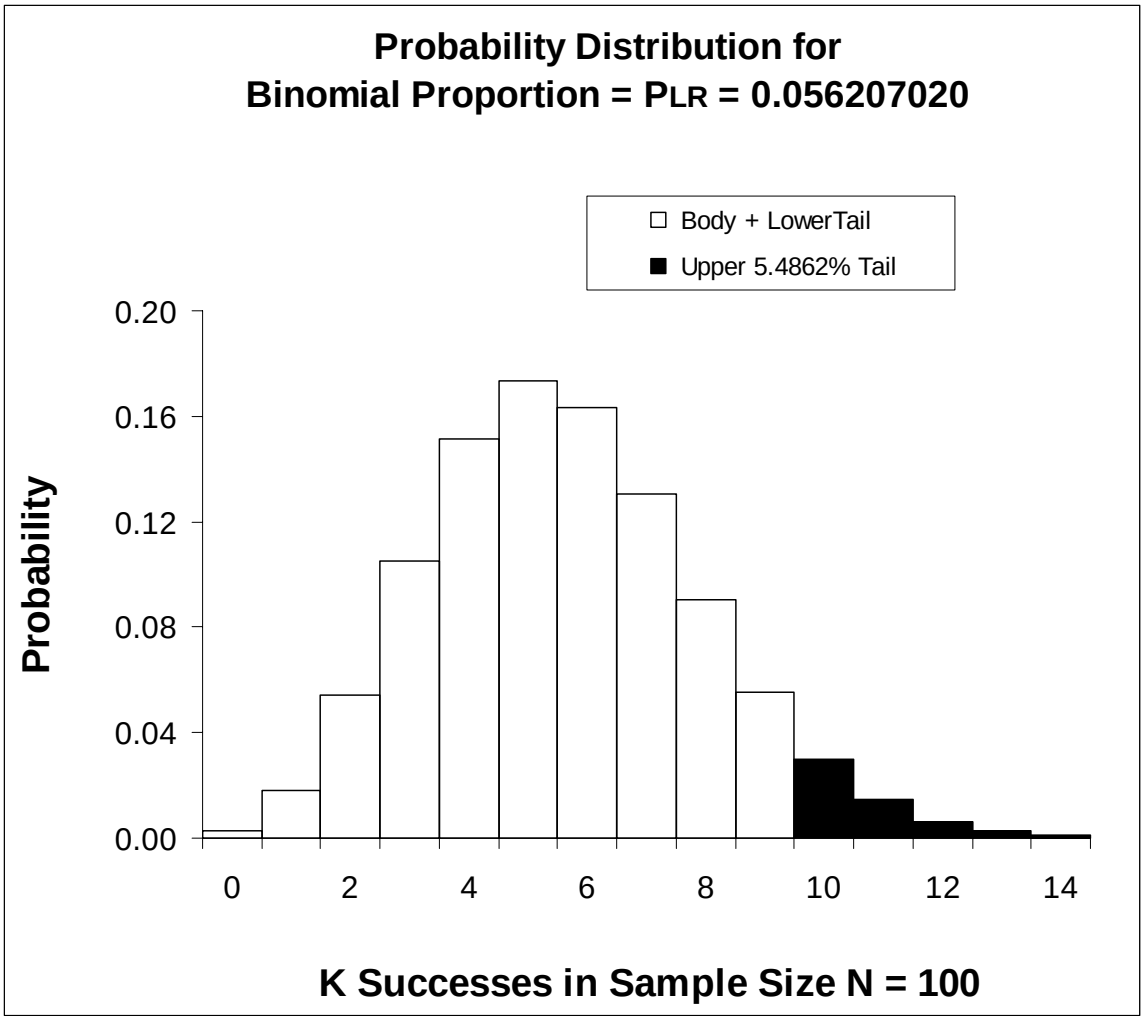
Probability Distribution for
Binomial Proportion = PLR = 0.056207020
Body + LowerTail
Upper 5.4862% Tail
Probability
0.20
0.16
0.12
0.08
0.04
0.00
0
2
4
6
8
10
12
14
K Successes in Sample Size N = 100

Figure 4A

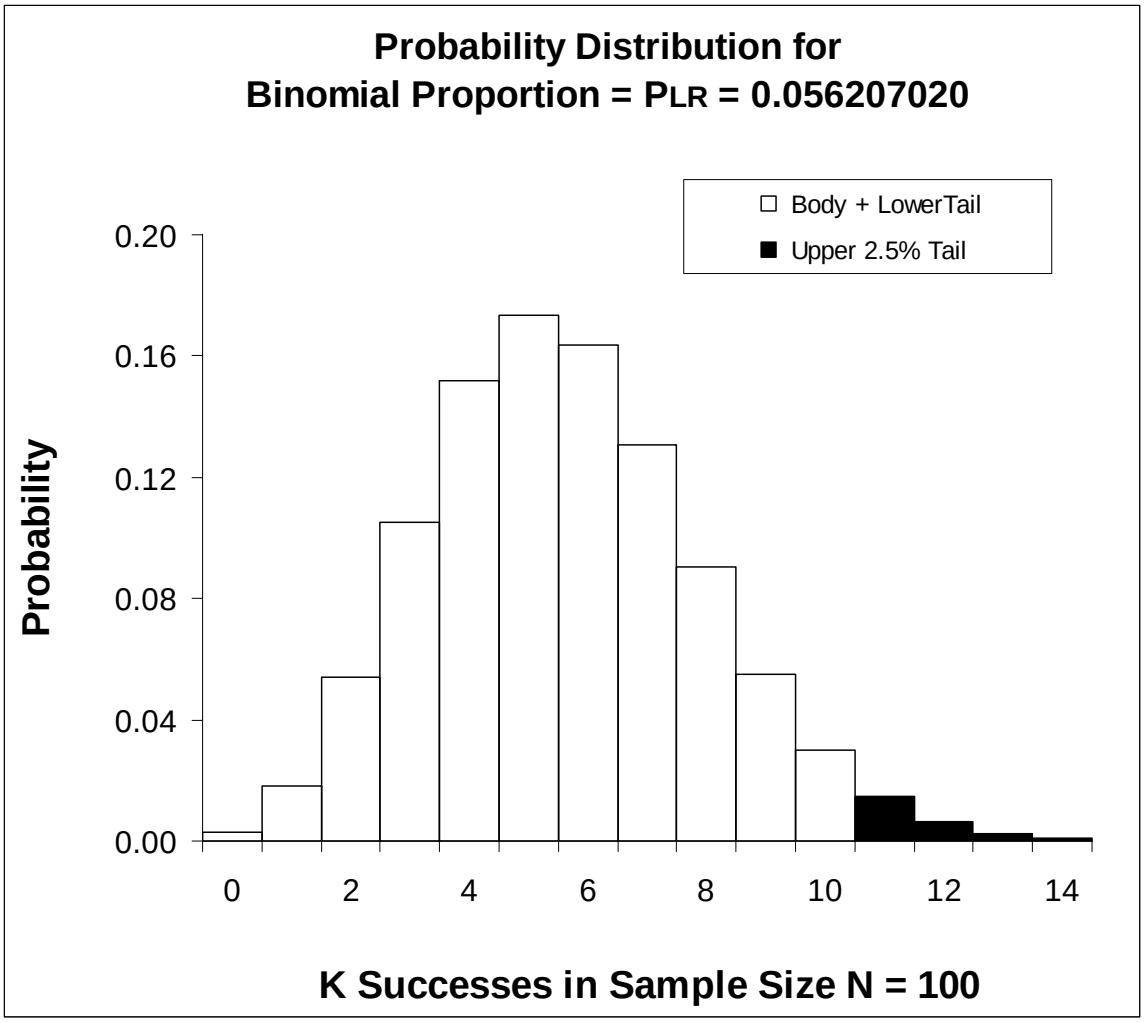
Probability Distribution for
Binomial Proportion = PLR = 0.056207020
Body + LowerTail
Upper 2.5% Tail
Probability
0.20
0.16
0.12
0.08
0.04
0.00
0
2
4
6
8
10
12
14
K Successes in Sample Size N = 100

Figure 4B

Table 1	Assuming a population has Pxx proportion of successes, this table lists the probability for the listed range of K successes, in a sample of size N = 100.			
	K ≤ 9	**K ≤ 10**	**K ≥ 10**	**K ≥ 11**
Wald:				
P_{UW} = 0.158798884	0.033816 ≈ 3.4 %	0.064560 ≈ 6.5 %		
P_{LW} = 0.041201116			0.008340 ≈ 0.8 %	0.002809 ≈ 0.3 %
Exact:				
P_{UE} = 0.176222598	0.011762 ≈ 1.2 %	0.025000 = 2.5 %		
P_{LE} = 0.049004689			0.025000 = 2.5 %	0.009978 ≈ 1.0 %
Reasonable:				
P_{UR} = 0.163982255	0.025000 = 2.5 %	0.049303 ≈ 4.9 %		
P_{LR} = 0.056207020			0.054862 ≈ 5.5 %	0.025000 = 2.5 %

Table 2		Reasonable 95% Confidence Limits for K Number of Successes in Sample Size N = 100		
K	**K / N**	**Identified using...**	**Lower Limit**	**Upper Limit**
1	0.01	Binomial distribution trial+error:	0.002431337	0.036216693
		Beta distribution formula:	0.002431333	0.036216736
10	0.10	Binomial distribution trial+error:	0.056207020	0.163982255
		Beta distribution formula:	0.056206942	0.163982391
25	0.25	Binomial distribution trial+error:	0.177394438	0.335735489
		Beta distribution formula:	0.177394390	0.335735321
50	0.50	Binomial distribution trial+error:	0.408036329	0.591963671
		Beta distribution formula:	0.408036232	0.591963768
75	0.75	Binomial distribution trial+error:	0.664264511	0.822605562
		Beta distribution formula:	0.664264679	0.822605610
90	0.90	Binomial distribution trial+error:	0.836017745	0.943792980
		Beta distribution formula:	0.836017609	0.943793058
99	0.99	Binomial distribution trial+error:	0.963783307	0.997568663
		Beta distribution formula:	0.963783264	0.997568667

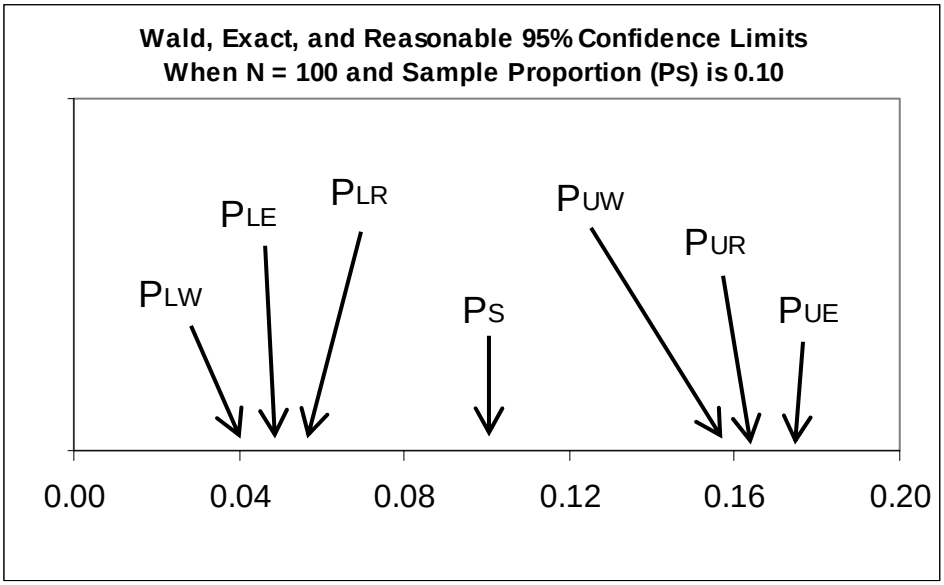
Wald, Exact, and Reasonable 95% Confidence Limits
When N = 100 and Sample Proportion (PS) is 0.10
PLW
PLE
PLR
PS
PUW
PUR
PUE
0.00
0.04
0.08
0.12
0.16
0.20

Figure 5

Table 3	COVERAGE Percentage of 95% confidence intervals that include Pp. Each percentage is based upon 10,000 random samples of N=100, drawn from a population of proportion = Pp, using Statgraphics Centurion XV. The Beta distribution was used to calculate Exact and Reasonable intervals.		
Pp	**Wald**	**Exact**	**Reasonable**
0.10	93.4 %	95.8 %	90.8 %
0.30	95.0 %	96.4 %	93.6 %
0.50	93.9 %	96.2 %	93.9 %

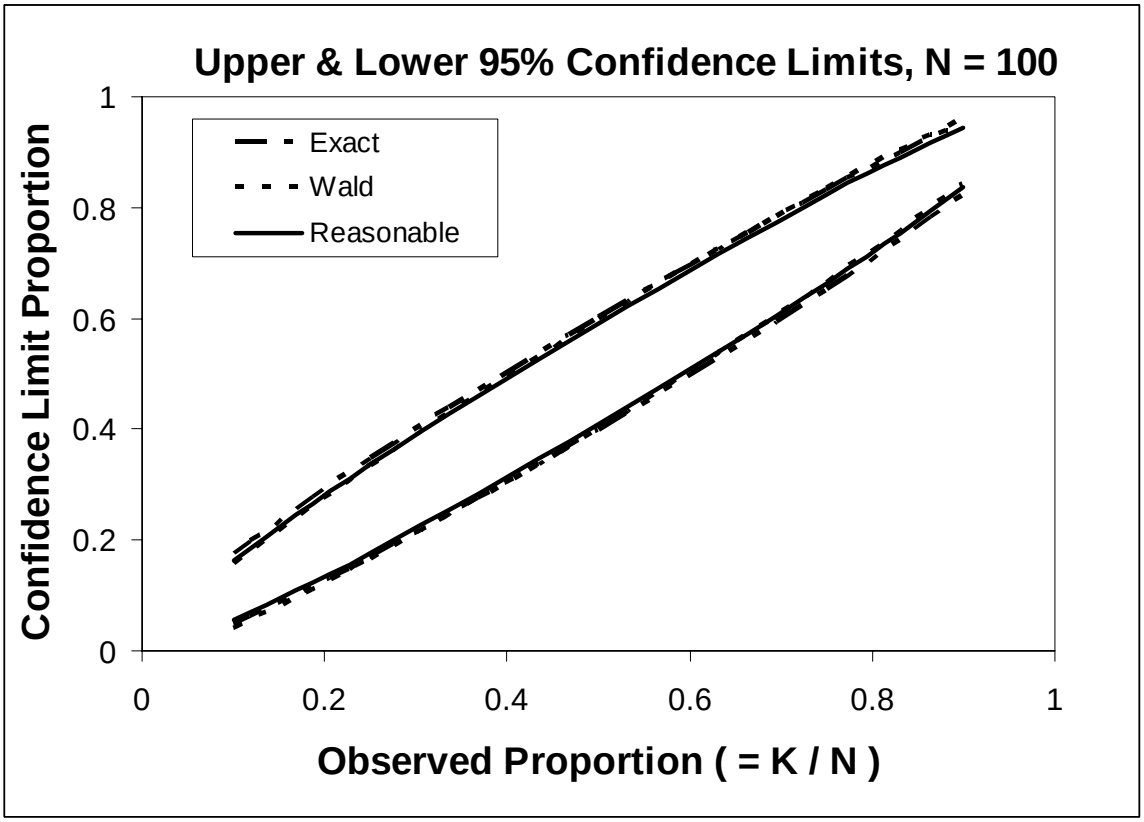
Upper & Lower 95% Confidence Limits, N = 100
Confidence Limit Proportion
Observed Proportion (= K / N)
Exact
Wald
Reasonable
0
0.2
0.4
0.6
0.8
1

Figure 6

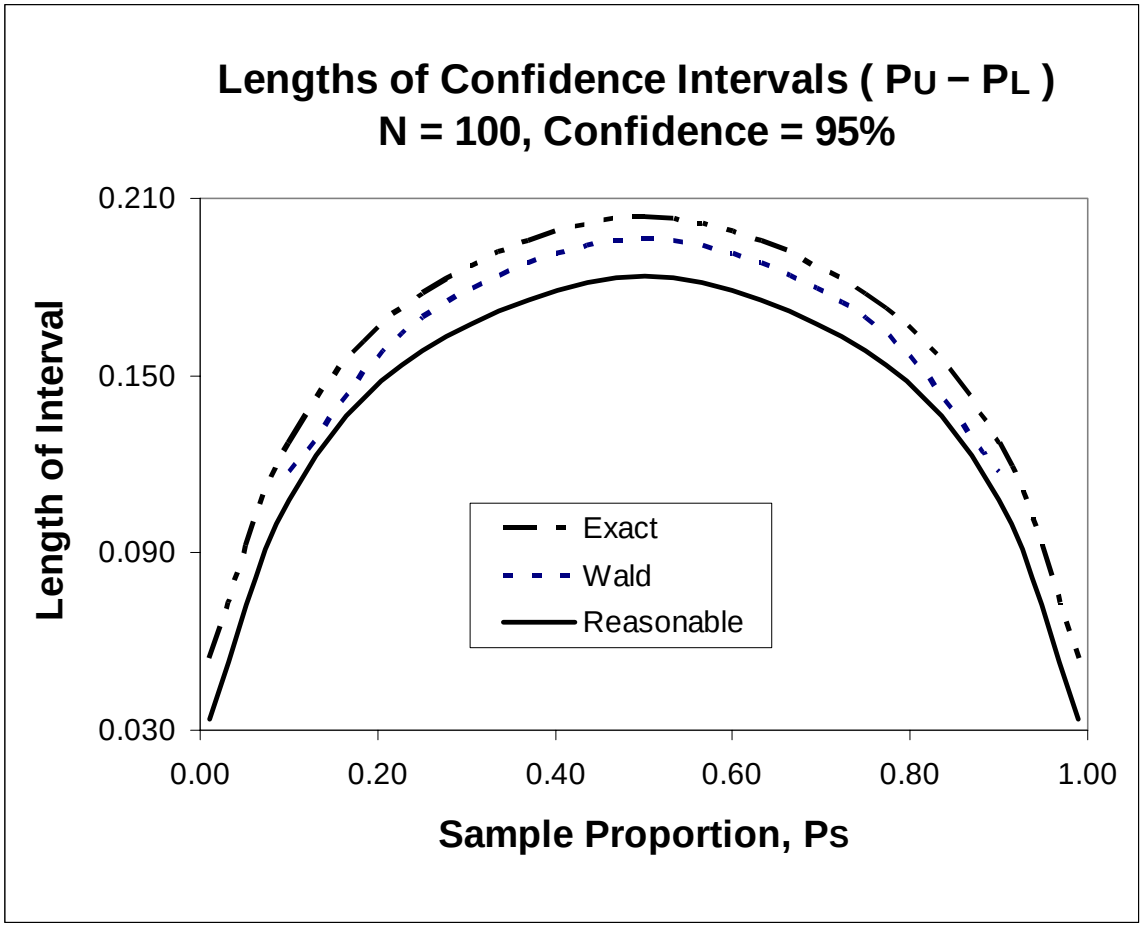
Lengths of Confidence Intervals (PU − PL)
N = 100, Confidence = 95%
Length of Interval
0.210
0.150
0.090
0.030
0.00
0.20
0.40
0.60
0.80
1.00
Sample Proportion, Ps
Exact
Wald
Reasonable

Figure 7

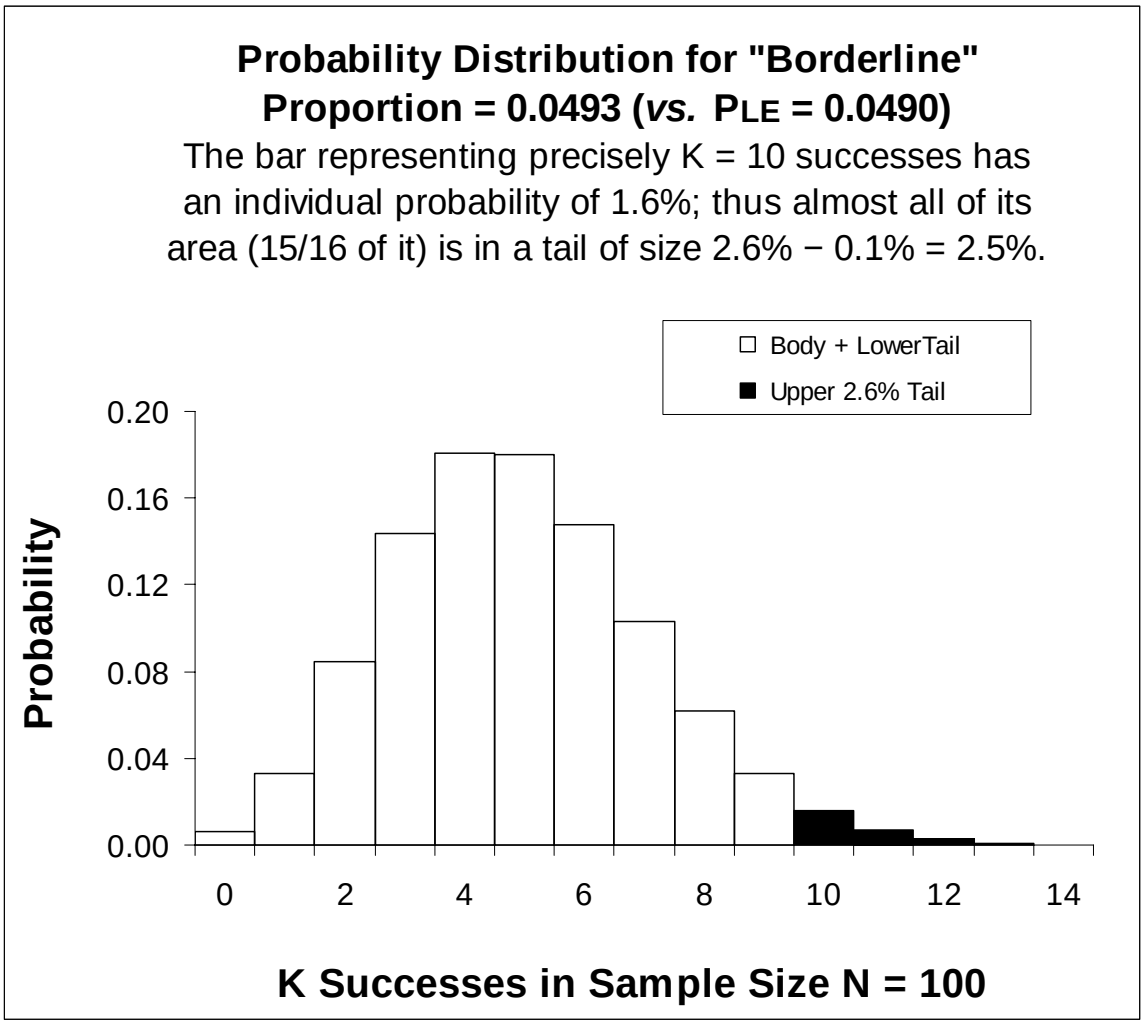
Probability Distribution for "Borderline"
Proportion = 0.0493 (vs. PLE = 0.0490)
The bar representing precisely K = 10 successes has an individual probability of 1.6%; thus almost all of its area (15/16 of it) is in a tail of size 2.6% − 0.1% = 2.5%.
Body + LowerTail
Upper 2.6% Tail
Probability
0.20
0.16
0.12
0.08
0.04
0.00
0
2
4
6
8
10
12
14
K Successes in Sample Size N = 100

Figure 8A

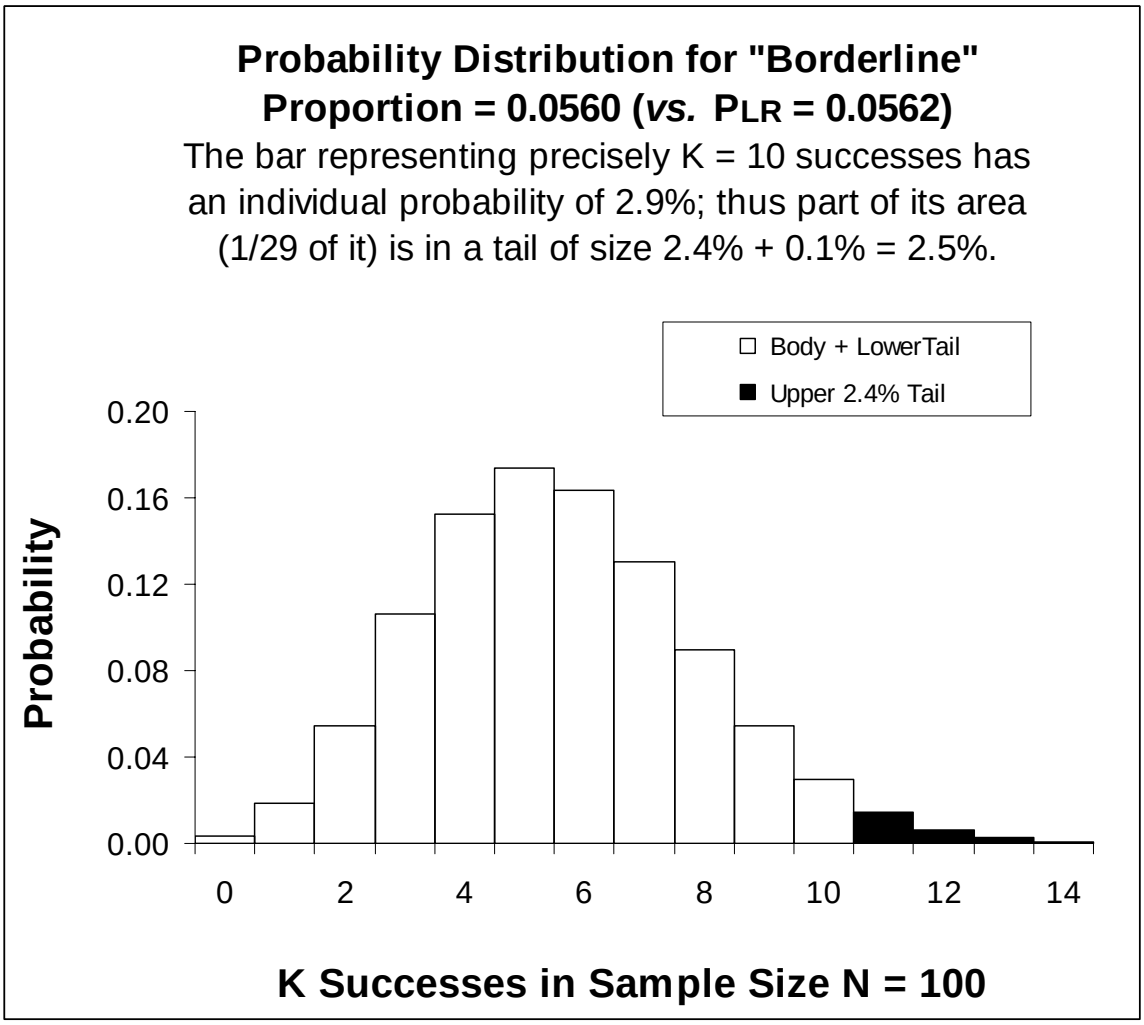
Probability Distribution for "Borderline"
Proportion = 0.0560 (vs. PLR = 0.0562)
The bar representing precisely K = 10 successes has an individual probability of 2.9%; thus part of its area (1/29 of it) is in a tail of size 2.4% + 0.1% = 2.5%.
Body + LowerTail
Upper 2.4% Tail
Probability
0.20
0.16
0.12
0.08
0.04
0.00
0
2
4
6
8
10
12
14
K Successes in Sample Size N = 100

Figure 8B

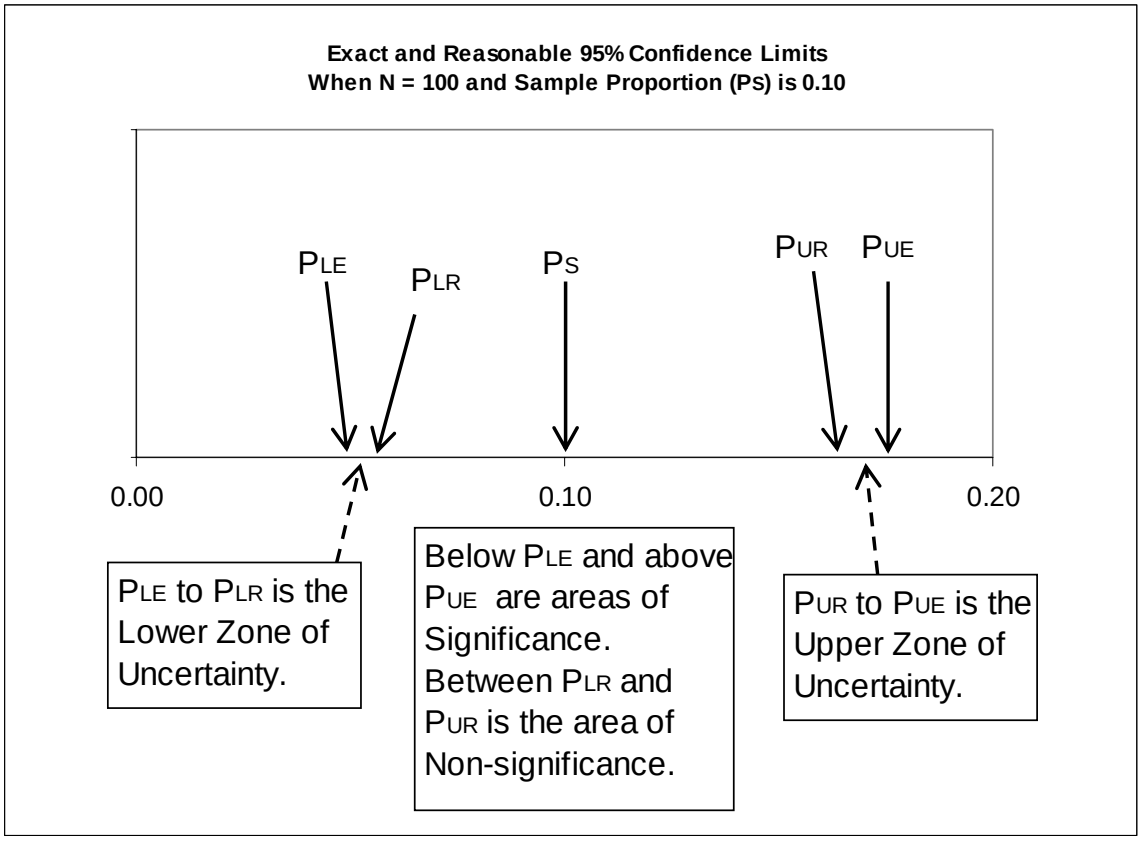
Exact and Reasonable 95% Confidence Limits
When N = 100 and Sample Proportion (PS) is 0.10
PLE
PLR
PS
PUR
PUE
0.00
0.10
0.20
PLE to PLR is the Lower Zone of Uncertainty.
Below PLE and above PUE are areas of Significance. Between PLR and PUR is the area of Non-significance.
PUR to PUE is the Upper Zone of Uncertainty.

Figure 9

Published on *MDDI Magazine* (http://www.mddionline.com)

New Interval Offers Confidence—Without Limits

By: John Zorich
August 9th, 2010

Reasonable confidence limits for binomial proportions could be easier to defend to regulatory bodies, including FDA. Reasonable limits are not statistically different from the sample proportion.

Confidence intervals are serious business. Recently, FDA told a medical device start-up company that, in regard to the company's proposed clinical trial, "The equivalence of the device to the predicate can be demonstrated if the confidence interval for the difference in the mean values for the tested parameter excludes a difference larger than 20% from the predicate." Unfortunately, the company could not meet that FDA mandate because in order to reduce the width of the confidence interval so that it would achieve that exclusion, a much larger number of patients was required than the company could afford to evaluate.

The term confidence interval was coined and first published by J. Neyman in 1934, when he applied it to binomial as well as variables data. He described confidence intervals as ranges "in which we may assume are contained the values of the estimated characters of the population."[1] As applied to a "proportion . . . of individuals in the sample" (which in this article will be represented by P_S) that has been derived from an unknown "proportion . . . of individuals in the population" (P_P in this article) whose "distribution . . . is then a binomial," he defined the confidence interval for P_P as having the form $P_L \leq P_S \leq P_U$, where P_L is the lower confidence limit, P_U is the upper confidence limit, and P_P is assumed with a specified level of confidence to be somewhere in the interval $P_L - P_U$.[2] A few months later, the first rigorous method for calculating such an interval was published.[3]

There is only one generally accepted method for calculating confidence intervals for variables data; that method involves t tables and the standard error of the mean, as described in any basic statistics textbook. However, there are many methods in use for calculating confidence intervals for binomial data, each such method resulting in different confidence limits and different interval widths.[4,5]

The reason that there are many binomial methods is partly historical and partly theoretical. The historical part is that, prior to the widespread availability of computers, binomial calculations were difficult. For that reason, simpler-to-calculate alternatives were developed; as was said several decades ago, “To calculate [large-sample binomial] probabilities would be an almost insurmountable task. Therefore, some method of approximation must be used.”6 The theoretical part is the lack of agreement on criteria for judging an interval (this is discussed in more detail later in this article). That disagreement was present from the birth of the interval concept in 1934: Neyman’s description for the interval was the equivalent of $P_L \le P_S \le P_U$, whereas Clopper & Pearson’s description was $P_L < P_S < P_U$; notice the use of ≤ versus <.[2,7]

Some of those many binomial confidence interval methods involve formulaic calculations. For example, what is commonly referred to as the Wald formula uses a binomial standard deviation coupled with a standard normal distribution Z-table to calculate an approximate confidence interval. Wald intervals are based on the fact that when sample size is large or when P_S is near 50%, the normal distribution is a reasonable model of the binomial, even if such an application is “not completely accurate.”[8] Other methods involve trial and error. For example, in order to calculate each confidence limit for what is commonly called the Exact binomial confidence interval, repeated attempts must be made to determine the proportion that yields a cumulative binomial histogram probability of exactly half the chosen significance level (this is discussed in more detail later).

It has been said recently that the Wald interval is “in virtually universal use.”[9] Similarly, the National Institute for Standards and Technology (NIST) “e-Handbook of Statistical Methods” Web site states that the Wald formula is the “confidence [interval] expression most frequently used.”[10] And Wald’s was the only method included in a binomial calculation spreadsheet issued in 1998 by CDRH for general use in handling clinical trials data.[11] By contrast, the Exact interval is the sole method provided in some mainstream statistical software programs (e.g., Statgraphics),[12] and the Exact is the only non-Z-table method for calculating binomial proportion confidence limits that is mentioned on the NIST Web site;[10] indeed, some statisticians (e.g., Agresti and Coull) refer to the Exact method as the gold standard.[13]

The implicit if not explicit focus of interest in a binomial confidence interval is its largest and smallest values, i.e., P_U and P_L. For example, a clinical trial might be considered successful only if the lower confidence limit on the outcome success rate is larger than a protocol-specified value. Because of such focus, this article introduces a new criterion for comparing the validity of confidence interval methods (since 1934, many other criteria have been proposed).[14] The new criterion is this: Are the confidence limits of the interval reasonable? If it is unreasonable to conclude that P_L and P_U could be P_P, then it is unreasonable to use the confidence interval method that generated them. And being reasonable is what even J. L. Fleiss has urged: “A confidence interval for a statistical parameter is a set of values that are . . . reasonable candidates for being the true underlying value.”[15]

Reasonable Confidence Intervals and Limits

Before defining what it means to be reasonable, it is important to understand the basis of the definition. A test of reasonableness is equivalent to performing a binomial test of significant difference using what Fleiss has called the "traditional statistical approach" for an "inference for a single proportion." It involves "calculating the probability, assuming the null hypothesis holds, of obtaining the outcome that actually occurred, plus the probabilities of all other outcomes as extreme as, or more extreme than, the one that was observed; and rejecting the null hypothesis in favor of the alternative hypothesis if the sum of all these probabilities—the so-called p-value—is less than or equal to a predetermined level, denoted by α, called the significance level."[16]

Based on that approach, a confidence interval is reasonable only if such a test results in a conclusion of "not statistically different" when the test compares the observed sample proportion (P_S) to either of the two most extreme values in the interval, namely the upper and lower confidence limits (P_U and P_L, respectively). In terms a bit more mathematical, unreasonable is defined as follows (for sample size = N, observed number of successes in that sample = K, and $K/N = P_S$): In regard to the probability distribution histograms derived from limits P_L and P_U, if either of the distribution tails in which K is found represents a cumulative probability of occurrence of less than or equal to α/2, we conclude that PS is statistically significantly different from the limit that generated that distribution, and that, therefore, the confidence interval $P_L - P_U$ is unreasonable.

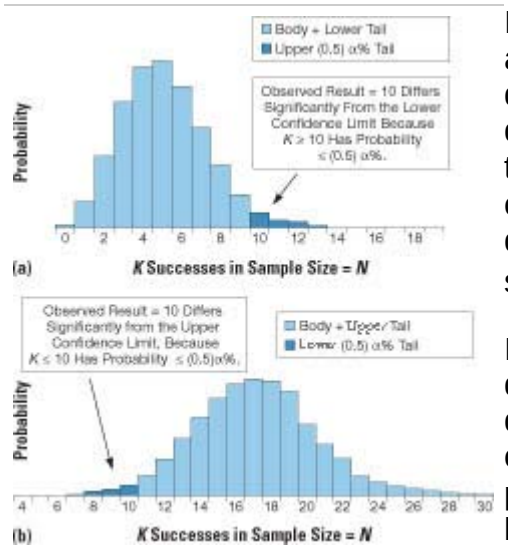

Figure 1. Example of a probability distribution derived from an "unreasonable" lower (a) and upper (b) confidence limit on an observed 10 successes in a sample size = N.

In that definition, we use α/2 rather than α because we are performing a two-sided test twice. The first test determines whether a random sample proportion differs from P_L; the second test determines whether that random sample proportion differs from P_U; in each case, we focus on tails that represent α/2 of the distribution (see Figures 1a and 1b). That is the standard way to approach such tests.[10,17–19]

It is important to note that the test of significance just described is performed using the probability distribution histograms generated from the two confidence limits (P_L and P_U) rather than being performed using the single probability distribution histogram generated from the observed sample proportion (P_S). Such an approach is used because P_L and P_U are together considered to be the "null hypothesis;" based on the new criterion described above, the question to answer is this: Is it reasonable to assume that the random sample proportion P_S could have been obtained from a population that had a proportion equal to either P_L or P_U (i.e., could either of them be Fleiss's "true underlying value," P_P)?[15,17,18]

This next discussion examines the reasonableness of Wald limits and explains how to calculate them. As demonstrated on the NIST Web site,[20] calculation of the upper and

lower confidence limits of a Wald (normal approximation) binomial confidence interval uses the following formula:

$P_S \pm Z\alpha/2 \times SDP_S$,

where P_S is the observed sample proportion, $Z\alpha/2$ is the two-tailed value from a normal distribution Z-table at the chosen α significance level, and SDP_S is the binomial standard deviation for the observed proportion, calculated as the square root of $P_S(1 - P_S)/N$. The limits of such an interval can be calculated as shown below, using Microsoft Excel (subscript *W* indicates the Wald method, *Normsinv* and *Sqrt* are MS Excel functions that output Z-table values and square roots respectively, asterisk (*) is the MS Excel symbol for multiply, and the other terms are as defined earlier):

$P_{UW} = P_S + Normsinv(1– \alpha/2) * Sqrt\ (P_S * (1 - P_S)/N)$
$P_{LW} = P_S - Normsinv(1– \alpha/2) * Sqrt\ (P_S * (1 - P_S)/N)$

Because the normal approximation assumption becomes less valid as the sample size becomes smaller or as PS departs farther from 0.500 (50%), the Wald calculation is typically restricted to situations in which both of the following are true: $N(P_S) > 5$ and $N(1 – P_S) > 5$. Even FDA's spreadsheet includes the warning "minimum [$N(P_S)$, $N(1 – P_S)$]. . . must be > 5 to use normal approximation."[11]

Let's examine the reasonableness of Wald confidence limits for the following situation: α = 5% (and therefore confidence = 1 – α = 95%), sample size = N = 100, successes = K = 10, $P_S = K/N$ = 10/100 = 0.1, and $N(P_S)$ = 10.
The resulting limits are P_{LW} = 0.041201116 and PUW = 0.158798884. The question is: how reasonable are they? Let's focus just on P_{LW} (the lower limit). The probability distribution for a population whose proportion equals that P_{LW} is shown in Figure 2. Notice that the probability of occurrence of the observed sample result or a more extreme result (i.e., $K \geq 10$) is less than α/2 = 2.5% (in fact, it is 0.8%); it can be concluded that P_S is statistically significantly different from P_{LW}. Based on our definition of reasonableness, it is therefore unreasonable to conclude that $P_{LW} = P_P$. If it is unreasonable to conclude that $P_{LW} = P_P$ (at α = 5%), then it is unreasonable to consider $P_{LW} - P_{UW}$ to be the 1 – α = 95% confidence interval. As evidenced by this example, Wald intervals and confidence limits can be unreasonable.

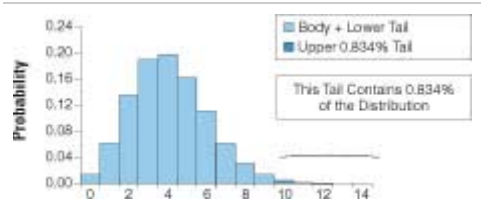

Figure 2. Probability distribution for binomial proportion. Binomial proportion = P_{LW} = 0.041201116.

To examine the reasonableness of Exact limits, it is essential to know how to calculate them. What are sought (by trial and error) are two proportions, one larger and one smaller than the observed sample proportion (P_S); each must have a cumulative binomial probability of "exactly" α/2 for obtaining the observed sample result or a more extreme value (i.e., 0 to K, or K to N). An MS Excel spreadsheet can be used to calculate the limits as accurately as, for example, Statgraphics, to at least a billionth of a probability unit (equal to 9 places to the right of the decimal point); how to do so is described on the NIST Web site.[10] The following is a generic example of applying the NIST/MS Excel method (α, N, K, P_P, P_S, P_U, and P_L are as defined above; subscript *E* indicates the Exact method; and P is a proportion sought):

P_{UE} = the value of P ($P > P_S$) needed to ensure that the MS Excel function Binomdist(K, N, P, True) outputs a probability value of α/2 precisely (to the desired number of significant digits).

P_{LE} = the value of P ($P < P_S$) needed to ensure that the MS Excel function Binomdist(K – 1, N, P, True) outputs a probability value of 1 – α/2 precisely (to the desired number of significant digits).

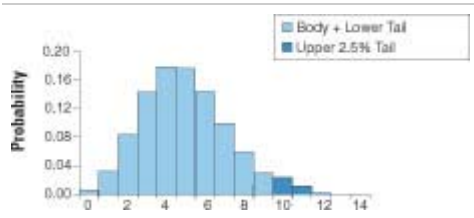

Figure 3. Probability distribution for binomial proportion. Binomial proportion = P_{LE} = 0.049004689.

Those formulas, applied to the situation evaluated previously (N = 100, K = 10, P_S = 0.1, and α = 5%), result in the following Exact limits: P_{LE} = 0.049004689 and P_{UE} = 0.176222598. The probability distribution for a population whose proportion equals that P_{LE} is shown in Figure 3. Notice that the probability of occurrence of the observed sample result or a more extreme result (i.e., $K \geq 10$) is precisely equal to α/2 = 2.5%; the entire histogram bar representing K = 10 (the value observed in the sample) is found in the 2.5% tail of the distribution; it can therefore be concluded that $P_S = K/N$ is statistically significantly different from P_{LE}. Based on our definition of reasonableness, it is therefore unreasonable to conclude that $P_{LE} = P_P$. If it is unreasonable to conclude that $P_{LE} = P_P$ (at α = 5%), then it is unreasonable to consider $P_{LE} - P_{UE}$ to be the 95% confidence interval. As evidenced by this example, Exact confidence intervals and limits can be unreasonable.

Two formulas for the Reasonable confidence limits are being introduced in this article; the first uses the binomial distribution and the second uses the beta distribution. The limits of a “reasonable binomial confidence interval” are defined as follows (where the subscript R indicates the Reasonable method):

P_{UR} = the value of P ($P > P_S$) needed to ensure that the MS Excel function Binomdist(K − 1, N, P, True) outputs a probability value of α/2 precisely (to the desired number of significant digits).

P_{LR} = the value of P ($P < P_S$) needed to ensure that the MS Excel function Binomdist(K, N, P, True) outputs a probability value of 1 – α/2 precisely (to the desired number of significant digits).

Notice that the first term in the MS Excel functions for Reasonable limits is changed by a value of 1 from its corresponding Exact function (in the P_U definitions, the change is from K to K – 1, and in the P_L formulas it is from K – 1 to K). That was done to ensure that P_S is not statistically significantly different from either P_{LR} or P_{UR}. In effect, those two formulas identify the widest possible confidence interval such that the observed Sample proportion is not statistically significantly different from any point in the interval, most especially the highest and lowest points, namely P_{UR} and P_{LR} (the meaning of statistically significantly different was discussed briefly earlier, and it will be discussed in more detail later in this article).

If those Reasonable method formulas are applied to the situation evaluated previously (N = 100, K = 10, P_S = 0.10, and α = 5%), they result in the following Reasonable limits: P_{LR} = 0.056207020 and P_{UR} = 0.163982255. The probability distribution for a population whose proportion equals that P_{LR} is shown in Figures 4a and 4b (p.79). Notice that, in Figure 4a,

the probability of occurrence of the observed sample result or a more extreme result (i.e., $K \geq 10$) equals approximately 5.5 %; in Figure 4b, the entire histogram bar representing $K = 10$ is found in the (1 − α/2)% body + lower_tail (i.e., in the lower 97.5% of the distribution); therefore, the observed sample proportion ($P_S = K/N = 0.10$) is not statistically significantly different from P_{LR}, and therefore P_{LR} could possibly be P_P. Similarly, as seen in Table I, it is reasonable to conclude that P_{UR} could possibly be P_P, because the observed sample result or a more extreme result (i.e., $K \leq 10$) is equal to approximately 4.9% and the probability of $K \leq 9$ is precisely 2.5%. Therefore, because both P_{LR} and P_{UR} could possibly be P_P (at α = 5%), it is indeed reasonable to consider $P_{LR} - P_{UR}$ to be the 95% confidence interval.

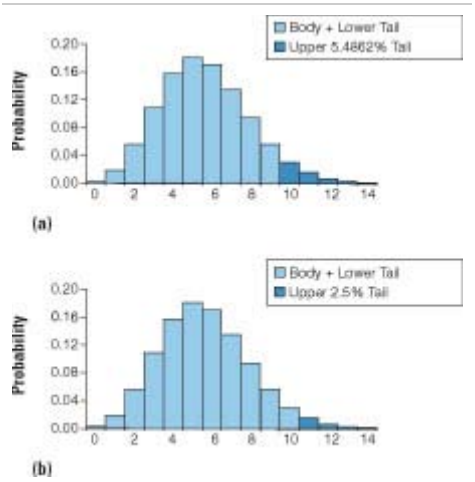

Figure 4. Probability distribution for binomial proportion. Binomial proportion = P_{LR} = 0.056207020 (a) and Binomial proportion = P_{LR} = 0.056207020 (b).

At first glance, the Wald method has an advantage over Reasonable and Exact ones: Calculation of Wald limits can be performed without trial and error and therefore can be automated using simple computer applications such as MS Excel functions. Upon further investigation, it can be seen that the beta-distribution formulas that have been used to approximate Exact limits [21] without using trial and error can be modified to also approximate Reasonable limits. As shown in Table II, the following MS Excel formulas approximate the output of the Reasonable binomial formulas shown above, to at least a millionth of a probability unit (equal to six places to the right of the decimal point) (α, N, K, P_{UR} , and P_{LR} are as defined earlier, and BetaInv is an MS Excel function):

Interval Method	Population Sample	$K \leq 9$	$K \leq 10$	$K \geq 10$	$K \geq 11$
Wald	P_{UW} = 0.158798884	0.033816 = 3.4 %	0.064560 = 6.5 %	—	—
	P_{LW} = 0.041201116	—	—	0.008340 = 0.8 %	0.002809 = 0.3 %
Exact	P_{UE} = 0.176222598	0.011762 = 1.2 %	0.025000 = 2.5 %	—	—
	P_{LE} = 0.049004689	—	—	0.025000 = 2.5 %	0.009978 = 1.0 %
Reasonable	P_{UR} = 0.163982255	0.025000 = 2.5 %	0.049303 = 4.9 %	—	—
	P_{LR} = 0.056207020	—	—	0.054862 = 5.5 %	0.025000 = 2.5 %

Table I. Assuming a population has P_{XX} proportion of successes, this table lists the probability for the listed range of K successes, in a sample of size N = 100.

P_{UR} (Beta) = 1 − BetaInv (α/2, $N - K + 1$, K)
P_{LR} (Beta) = 1 − BetaInv (1 − α/2, $N - K$, $K + 1$)

K	K / N	Identification Method	Lower Limit	Upper Limit
1	0.01	Binomial distribution trial+error:	0.002431337	0.036216693
		Beta distribution formula:	0.002431333	0.036216736
10	0.10	Binomial distribution trial+error:	0.056207020	0.163982255
		Beta distribution formula:	0.056206942	0.163982391
25	0.25	Binomial distribution trial+error:	0.177394438	0.335735489
		Beta distribution formula:	0.177394390	0.335735321
50	0.50	Binomial distribution trial+error:	0.408036329	0.591963671
		Beta distribution formula:	0.408036232	0.591963768
75	0.75	Binomial distribution trial+error:	0.664264511	0.822605562
		Beta distribution formula:	0.664264679	0.822605610
90	0.90	Binomial distribution trial+error:	0.836017745	0.943792980
		Beta distribution formula:	0.836017609	0.943793058
99	0.99	Binomial distribution trial+error:	0.963783307	0.997568663
		Beta distribution formula:	0.963783264	0.997568667

The most extreme values that P_S can take are 0.0 and 1.0. When P_S equals 0.0 precisely, no method can calculate a lower confidence limit (P_L), because proportions (i.e., probabilities) lower than zero are undefined; and when P_S equals 1.0 precisely, no method can calculate P_U, because

Table II. Reasonable 95% confidence limits for *K* number of successes in sample size *N* = 100.

proportions above unity are likewise undefined. Not surprisingly, the P_L for P_S = 1.0 and the P_U for P_S = 0.0 are called one-sided limits.

Such one-sided limits can be calculated by the Exact method but not by Wald or Reasonable ones. Wald limits for P_S = 0.0 or 1.0 cannot be calculated because in either case the binomial standard deviation itself equals 0.0, no matter what the sample size is (recall that SDP_S = square root of $P_S(1 - P_S)/N$). Similarly, Reasonable limits cannot be calculated because if K = 0 (P_S = 0.000), then the P_{UR} formula value $K - 1$ is meaningless; and if $K = N$ (P_S = 1.000), then the PLR formula always results in a probability of 1.000 and therefore can never equal the sought-after value of P ($P < P_S$).

On the other hand, Reasonable and Exact methods (but not the Wald method) share the following advantage: they can calculate two-sided confidence limits for any P_S greater than zero and less than unity, no matter how small or large (e.g., if sample size is a million and K = 1, then P_S = 0.000001 = 1.000E-6; and P_{LR} = 0.242E-6).

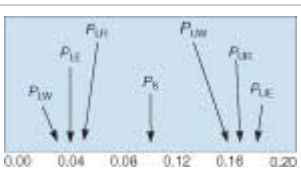

Figure 5. Wald, Exact, and Reasonable 95% confidence limits when $_N$ = 100 and sample proportion (P_S) = 0.10.

A common criterion for evaluation of confidence interval methods is *coverage*. That term refers to the percentage of time the interval can be expected to include P_P (the "true underlying value"). Typically, that percentage is determined experimentally[4,5,14] by generating confidence intervals for thousands of random samples drawn from populations of known proportions (i.e., known P_Ps), and then determining what percentage of those intervals contain the corresponding P_P. As seen in Figure 5, Reasonable confidence limits are completely contained within Exact ones, and therefore Reasonable intervals can be expected to have slightly less coverage; experimental results support that conclusion (see Table III).

P_P	Wald	Exact	Reasonable
0.10	93.4 %	95.8 %	90.8 %
0.30	95.0 %	96.4 %	93.6 %
0.50	93.9 %	96.2 %	93.9 %

Table III. Percentage of 95% confidence intervals that include P_P. Each percentage is based on 10,000 random samples of N = 100, drawn from a population of proportion = P_P, using Statgraphics Centurion *XV*. The beta distribution was used to calculate Exact and Reasonable intervals.

The distinctiveness of Reasonable intervals and their limits is not apparent when the limits are plotted in the still-common manner introduced in 1934 by Neyman (see Figure 6).[22] However, alternative plotting methods (see Figure 7) clearly demonstrate that Reasonable intervals are narrower than Wald or Exact ones. As a result, only Reasonable intervals are narrow enough to have upper and lower limits that are truly reasonable. As Neyman insisted, "Confidence intervals should be as narrow as possible."[23]

Recommendations

With variables data, a test of significance is mathematically equivalent to use of a confidence interval; i.e., a borderline value being compared with the sample result is considered significantly different if it is outside the sample's confidence interval, but nonsignificant if inside.[24] Likewise, a borderline binomial proportion is classically viewed as being inside or

outside the sample proportion's confidence interval. But such a view is misleading for proportions because they are based on counts. Because counts can take only discrete values, the observed result's probability distribution histogram bar appears as if it spans the border between significance and nonsignificance when the histogram is based on a borderline value.

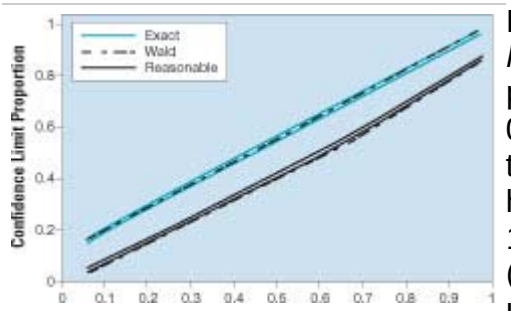

Figure 6. Upper and lower 95% confidence limits (N = 100).

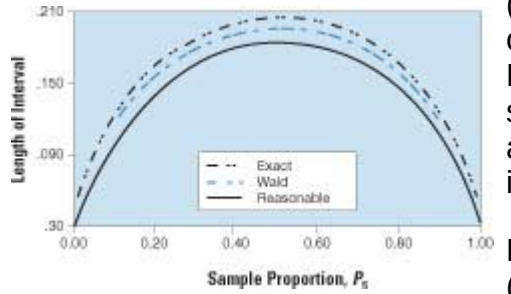

Figure 7. Lengths of confidence intervals ($P_U - P_L$). N = 100 and confidence = 95%.

For example, as discussed previously, in the case of N = 100 and observed result K = 10 (see Figure 3), probability distribution histograms based on $P \le P_{LE}$ = 0.0490 result in the histogram bar for K = 10 being in the upper 2.5% tail of the distribution. Similarly, if the histogram is based upon $P \ge P_{LR}$ = 0.0562, then K = 10 is in the 97.5% body + lower_tail of the distribution (see Figure 4b). The problem is that histograms based upon P values between P_{LE} and P_{LR} result in the observed-K histogram-bar being neither fully in the 2.5% upper tail nor fully in the 97.5% body + lower_tail (see Figures 8a and 8b). In such cases, how do you objectively conclude significance or non-significance? Either conclusion could be viewed as unreasonably subjective and arbitrary. A solution, introduced in this article, is to always use not one but two confidence intervals: the Exact and the Reasonable.

Before explaining that solution, a brief history is useful (in the following discussion, "$L\delta$" is the likelihood of obtaining the observed result δ, assuming the null hypothesis is true). The concept of statistical significance has undergone much change since the precursors to modern tests of significance were developed in the 1800s. In that century, δ was not considered significant unless it was extremely unlikely.[25] As decades passed, the requirement for significance became less extreme. By the 1930s, it could be said that "it is conventional among certain workers to adopt the following rule: If $L\delta \ge 0.05$, δ is not significant; if $L\delta \le 0.01$, δ is significant; if $0.05 > L\delta > 0.01$, our conclusions about δ are doubtful, and we cannot say with much certainty whether the deviation is significant or not until we have additional information"[26] (the original text uses a different symbol than L). In effect, the region between 0.01 and 0.05 was considered a zone of uncertainty (a term not in the original text).

On the basis of that background, perhaps the best solution to the problem of borderline proportions is to consider the range of values between P_{LE} and P_{LR}, and between P_{UR} and P_{UE}, to be zones of uncertainty (see Figure 9). Using that approach, and assuming the sample size has the power to detect a clinically significant difference, if the null hypothesis proportion (P_{NH}) being compared with the study result (P_S) is outside of the study result's Exact confidence interval (i.e., $P_{NH} \le P_{LE}$ or $P_{UE} \le P_{NH}$), you can claim statistical significance. If P_{NH} is inside the corresponding Reasonable interval (i.e., $P_{LR} \le P_{NH} \le P_{UR}$), you can claim statistical nonsignificance. However, if P_{NH} is in either of the zones of

uncertainty (i.e., either $P_{LE} < P_{NH} < P_{LR}$ or $P_{UR} < P_{NH} < P_{UE}$), then the results are statistically inconclusive.

FDA Notes

A search of the Web site FDA.gov finds recent submissions, panel opinions, and guidance documents that include a variety of binomial confidence interval methods, the most common being the Exact and the Score (the Score formula is an elaborated classic Wald, from the perspective that it uses a binomial standard deviation and not one but a few copies of a normal distribution Z-table value; Score confidence limits are found midway between Exact and Reasonable ones). In a 2007 statistical guidance document, FDA recommends "score confidence intervals, and alternatively, Exact (Clopper-Pearson) confidence intervals."[27] Exact versus Score methods were compared briefly in a 2009 FDA publication. It noted, "An advantage with the Score method is that... it can be calculated directly. Score confidence bounds tend to yield narrower confidence intervals than Clopper-Pearson [i.e., Exact] confidence intervals, resulting in a larger lower confidence bound."[28] It is interesting to note that that recommendation was based on ease of calculation rather than on theoretical correctness.

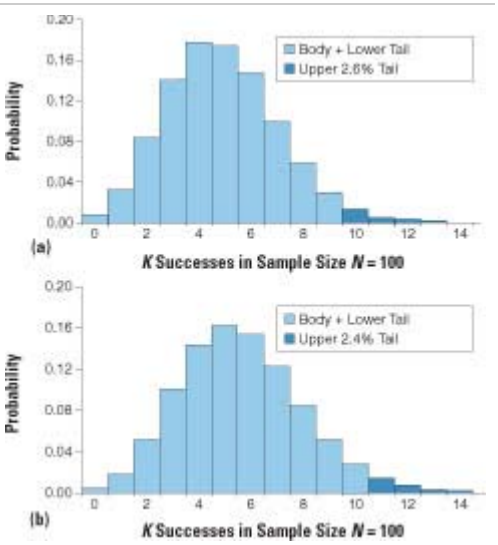

Figure 8. Probability distribution for "borderline." Proportion = 0.0493 (vs. P_{LE} = 0.0490) (a); proportion = 0.560 (vs. P_{LR} = 0.562) (b). The bar representing K = 10 successes has an individual probability of 1.6%; thus almost all of its area (15/16) is in a tail of size 2.6% – 0.1% = 2.5% (a). The bar representing K = 10 successes has an individual probability of 2.9%; thus part of its area (1/29) is in a tail of size 2.4% + 0.1% = 2.5% (b).

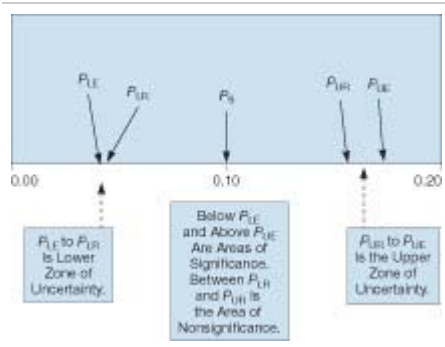

Figure 9. Exact and Reasonable 95% confidence limits when N = 100 and sample proportion (P_S) is 0.10.

In 2003, Medtronic Neurological received approval for its Active Dystonia Therapy (deep brain stimulation system) by using the classic Wald method almost exclusively. "Exact 95% confidence intervals were used when the # (%) of patients was 0 (0%) because the normal approximation to the binomial does not provide a confidence interval. In every other case, the normal approximation to the binomial was used to calculate confidence intervals" even though in very many of those cases the $N(P_S) > 5$ criterion (mentioned above) was not met.[29]

The new methods proposed in this article have not yet been used in any regulatory submission to FDA, Health Canada, or a notified body, as far as the author has been able to ascertain.

Conclusion

The most commonly used binomial confidence intervals can have confidence limits that are statistically significantly different from the random sample proportion from which they

were derived. Therefore, such confidence limits are unreasonable to consider as being the proportion of the population from which the random sample was taken. A more defensible choice of confidence intervals is presented in this article, namely a Reasonable confidence interval for a binomial proportion, the extreme values of which are Reasonable confidence limits.

Although such an interval is slightly narrower than other intervals and thus offers less coverage, it is more reasonable to use because it is the widest possible range that contains no values that are statistically significantly different from the sample proportion on whose basis the interval was calculated. Alternatively, an even more reasonable approach is to use a combination of Exact and Reasonable limits, coupled with the concept of zones of uncertainty.

References

1. J Neyman, “On the Two Different Aspects of the Representative Method,” Journal of the Royal Statistical Society 97, no. 4 (1934): 562.
2. Neyman (1934): 589.
3. C Clopper and E Pearson, “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial,” Biometrika 26, no. 4 (1934): 409.
4. LD Brown et al., “Confidence Intervals for a Binomial Proportion and Asymptotic Expansions,” The Annals of Statistics, 30 no. 1 (2002): 160–201.
5. RG Newcombe, “Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods,” Statistics in Medicine 17 (1998): 857–872.
6. H Bancroft, Introduction to Biostatistics (New York: Hoeber Medical Division of Harper & Row, 1957): 106.
7. Clopper (1934): 404.
8. H Motulsky, Intuitive Biostatistics (New York: Oxford University Press, 1995): 18.
9. Brown (2002): 160.
10. NIST/SEMATECH e-Handbook of Statistical Methods, last updated: 7/18/2006; available from Internet: www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm [1].
11. Two Group Confidence Interval & Power Calculator, version 1.6 (Rockville, MD: FDA, March 26, 1998).
12. Statgraphics Centurion XV, version 15.2.12 (Warrenton, VA: StatPoint Inc., 1982–2007).
13. LD Brown et al., “Interval Estimation for a Binomial Proportion,” Statistical Science 16, no. 2 (2001): 117.
14. MD deB Edwardes, “The Evaluation of Confidence Sets With Application to Binomial Intervals,” Statistica Sinica 8, (1998): 393–409.
15. JL Fleiss et al., Statistical Methods for Rates and Proportions, 3rd ed. (Hoboken, NJ: Wiley, 2003): 22.
16. Fleiss (2003): 18–19.
17. JL Phillips Jr., How to Think About Statistics, revised ed. (New York: Freeman, 1992): 62–64.
18. W Mendenhall, Introduction to Probability and Statistics, 5th ed. (North Scitueate, MA: Duxbury Press, 1979): 231–232.
19. NIST: www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm [2].
20. NIST: www.itl.nist.gov/div898/handbook/prc/section2/prc24.htm [3].
21. K Krishnamoorthy, Handbook of Statistical Distributions with Applications (Boca Raton, FL: Taylor & Francis Group, 2006): 38.

22. Neyman (1934): 590.
23. Neyman (1934): 563.
24. Motulsky (1995): 106–117.
25. SM Stigler, The History of Statistics: The Measurement of Uncertainty before 1900 (Cambridge, MA: Belknap Press of Harvard University Press, 1986): 300ff.
26. JF Kenney, Mathematics of Statistics (Part One & Part Two) (New York: D. Van Nostrand Co., 1939): Part Two, 117.
27. Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests, Draft Guidance (Rockville, MD: FDA, March 13, 2007): 23.
28. Assay Migration Studies for In Vitro Diagnostic Devices, Draft Guidance (Rockville, MD: FDA, January 5, 2009): 31.
29. Summary of Safety and Probable Benefit, Humanitarian Device Exemption (HDE) Number: H020007 (Rockville, MD: FDA, April 15, 2003): 9.

John Zorich is an independent consultant and contractor in the areas of regulatory compliance and statistical methods.

Published in MD&DI [4], August 2010, Volume 32, No. 8 [5]

- Previous story: Piezo Motors and Actuators: Medical Device Performance [6]
- Next story: Risk Analysis: Beyond Probability and Severity [7]

Author:
John Zorich

Feature Risk Management

Links:
[1] http://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm
[2] http://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm
[3] http://www.itl.nist.gov/div898/handbook/prc/section2/prc24.htm
[4] http://www.mddionline.com/epublish/3
[5] http://www.mddionline.com/epublish/3/22
[6] http://www.mddionline.com/article/piezo-motors-and-actuators-medical-device-performance
[7] http://www.mddionline.com/article/risk-analysis-beyond-probability-and-severity