Reasons for No Longer Teaching [or Using] the Normal Approximation Confidence Interval

by John N. Zorich, Jr.   www.JohnZorich.com

THIS IS v2 OF A SLIGHTLY MODIFIED MSWORD DOCUMENT VERSION OF MY PAPER PUBLISHED IN THE *MATHAMATYC EDUCATOR* (WHICH IS THE OFFICIAL JOURNAL OF THE AMERICAN MATHEMATICAL ASSOCIATION OF TWO YEAR COLLEGES). THE FIRST PAGE OF THE PUBLISHED 2021 VERSION CAN BE FOUND AT THE END OF THIS DOCUMENT; THE FULL AS-PUBLISHED VERSION CAN BE OBTAINED FROM:

https://amatyc.org/page/EducatorWinter2021

Introduction

A typical, modern-day introductory statistics textbook teaches "normal approximation" confidence interval formulas for use with "large" samples and "exact" formulas for "small" samples. However, decades ago, a Harvard University professor of statistics warned that "there is no safe general rule as to how large *n* must be for use of the normal approximation in computing confidence limits" (Cochran, 1977, p. 42). Related advice has been given more recently: "In situations where the choice of test statistic or confidence interval would make a difference to the inferences drawn, exact methods should be relied upon rather than normal approximations" (Fleiss, Levin, & Paik, 2003, p. 62).

What are the theoretical, practical, historical, and educational reasons for basing formula choice on sample size? What reasons might there be for no longer teaching the normal approximation confidence interval?

Introduction to Confidence Intervals

The term *confidence interval* was coined by J. Neyman in 1934 when he defined it as a range "in which we may assume are contained the values of the estimated characters of the population [i.e., the population parameters]" (Neyman, 1934, p. 562). Many different formulas were developed for calculating the confidence limits that mark the upper and lower boundaries of confidence intervals (e.g., Brown, Cai, & DasGupta, 2002; Newcombe, 1998); some of those formulas produce intervals that are surprisingly inaccurate.

Literature on this subject provides a basic criterion for evaluating the accuracy of confidence intervals (e.g., for a 95% confidence interval): If all possible samples are drawn from a population, then *no less than* 95% of the confidence intervals derived from them will contain the population parameter. That coverage-based criterion has been called the "fundamental property" of confidence intervals (Fleiss et al., 2003, p. 24).

For the sake of brevity, discussions in this article will focus on confidence intervals for the mean of a single sample from a normally distributed population and for the proportion of "successes" in a single sample from a binomial population.

Copyright 2021 by John N. Zorich Jr. ---- All rights reserved.

Confidence Intervals for Variable Data Averages

The equation for a single-sample variable-data confidence interval for the population average can be expressed as $m \pm C \cdot s / \sqrt{n}$, where *m* is the sample mean, *s* is the sample standard deviation, *n* is the sample size, and *C* is the coefficient taken from either a Normal-distribution *z*-table or from a Student's-distribution *t*-table. Some mid-20th-century introductory statistics textbooks that included detailed discussions of confidence intervals also included a *z*-table but not a *t*-table. In one such textbook, its *z*-table-only discussion of confidence intervals ended with this warning: "It must be borne in mind that, just as this [*z*-table-based] estimate of the confidence interval becomes more precise as *N* [sample size] increases, so inversely it becomes less acceptable as *N* decreases" (Treloar, 1938, p. 138). Another such textbook advised that "for small samples," the distribution of *C* in that confidence interval formula "is not normal and a probability table different from those based on the normal distribution is required" (Walker, 1943, p. 284). Instead of providing that required table (i.e., a *t*-table), Walker provided this: "At the close of this chapter are several references to such tables" (Walker, 1943, p. 284). Therefore, such textbooks could not be used to teach the exact ("more precise") *t*-table confidence interval, but rather only the approximate ("less acceptable") *z*-table interval.

Some introductory, mid-20th-century textbooks did not provide *t*-values between those for degrees of freedom (df) of 30 and infinity, where the *df* for infinity exactly equals the corresponding *z*-table value (e.g., Fisher, 1958, p. 174; Kenney, 1939, p. 138; Mendenhall, 1979, p. 535). Students who used such *t*-tables were, in effect, using the normal approximation for all sample sizes larger than 31. These days, a typical textbook provides *t*-values for *dfs* 1−30, 40, 60, 120, and infinity.

Many statistics textbooks, as well as the statistical software program *Minitab*®, advise that coefficient *C* be taken from a *z*-table when the standard deviation's true value (i.e., the parameter) is *known*, but be taken from a *t*-table when it is *unknown* (Minitab, 2017a). Some textbooks explain that although such advice is theoretically correct, it is "somewhat artificial," because *not* knowing the parameter "is by far the more realistic situation" (Glass & Stanley, 1970, p. 260).

For such realistic situations, most textbooks have long advised that "we can use [*z*-tables] … if the sample is large—say greater than 30.... If the sample is less than 30, we should use [*t*-tables]" (Duncan, 1986, p. 567). As can be seen in the variable-data confidence interval formula, the difference between the length of confidence intervals calculated with *z*- or *t*-tables is due solely to the difference in magnitude of *z*- and *t*-values, *z* always being smaller than *t* at a given confidence level (see Figure 1). At *n* = 30, there is a 4% reduction in interval length, a difference that has been called "very close" (Gravetter & Wallnau, 2000, p. 290) and "trivial" (Treloar, 1938, p. 138). At *n* = 120, there is still a 1% difference, which has been called "negligible" (Gravetter & Wallnau, 2000, p. 290). Those descriptors are misleading, because such differences are actually errors caused by using an approximate method rather than an exact one.

Copyright 2021 by John N. Zorich Jr. ---- All rights reserved.

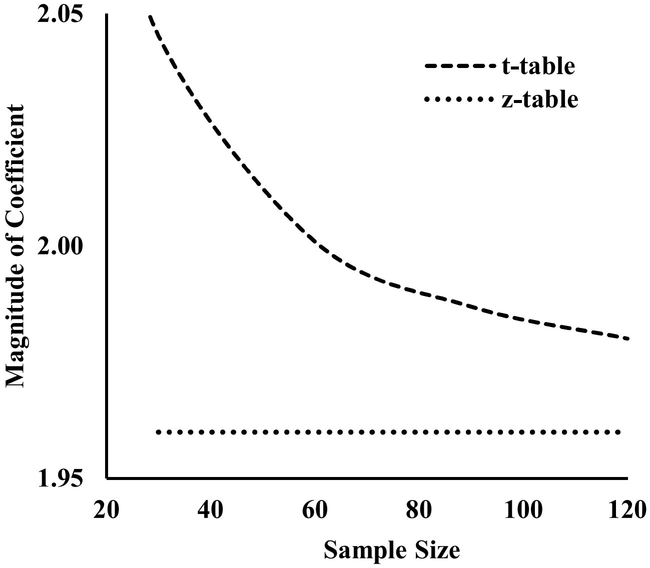

Figure 1. Comparison of magnitudes of *t*-table versus *z*-table coefficients, at $\alpha = .05$, 2-tailed; *z* coefficient is 4.17% smaller than the *t* coefficient at $n = 30$, and 1.02% smaller than at $n = 120$.

An important point is that approximate *z*-based intervals do not achieve their claimed level of confidence (= %-coverage). Figure 2 shows that at $n = 30$, coverage is about 94% for *z*-based intervals, rather than the claimed 95%; at $n = 120$, coverage is still less than 95%. In contrast, Figure 2 shows that *t*-based intervals exhibit at least 95%-coverage at all sample sizes. It is, therefore, valid to conclude that *z*-based intervals fail to meet the coverage-based criterion discussed previously, because they do not possess the "*fundamental property* that justifies the use of the term [95%] 'confidence'" (Fleiss et al., 2003, p. 24).

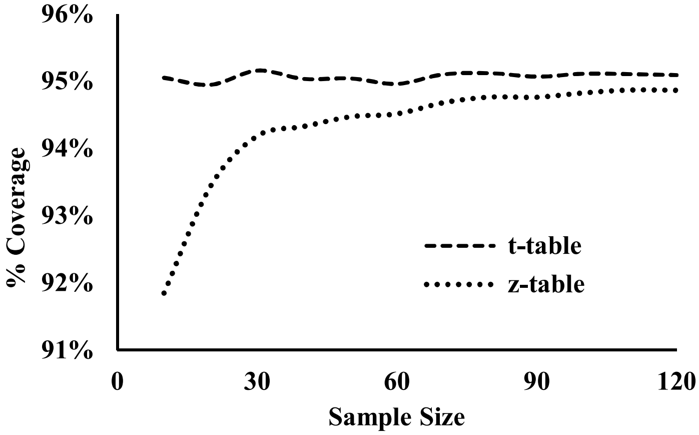

Figure 2. Average %-coverage, based on 100,000 samples for each of 12 sample sizes, from $n = 10, 20, 30, \ldots, 120$. Samples generated using Minitab's "Random Data, Normal" menu option, with parameter mean = 100 and parameter standard deviation = 10.

Copyright 2021 by John N. Zorich Jr. ---- All rights reserved.

Confidence Intervals for Binomial Data Proportions

Unfortunately, "except in special circumstances, no explicit formulas are available" for calculating single-sample binomial-proportion confidence limits, and, therefore, they "must be found by trial and error, by means of a formal iterative computation, or by special tables" (Fleiss et al., 2003, p. 22). Prior to the personal-computer era, those computations involved using published binomial probability tables, whose maximum sample size was limited in part by the capabilities of the computers available at that time.

In 1950, such tables published by Harvard University covered sample sizes up to 50; by 1953, that range was extended to 100, and then to 1000 by 1955 (Staff, 1955, pp. xx–xxi). Thus, in 1957, it was correctly stated that, regarding probabilities for large sample sizes not found in such tables, "to calculate these probabilities would be an almost insurmountable task. Therefore, some method of approximation must be used" (Bancroft, 1957, p. 106).

A highly accurate approximation method was developed based on the *F*-distribution, using the following formulas (Fleiss et al., 2003, p. 26):

$$\text{Lower Binomial Confidence Limit } = \frac{X}{X + F1(n - X + 1)}$$

$$\text{Upper Binomial Confidence Limit } = \frac{F2(X+1)}{(n-X) + F2(X+1)},$$

where

X = number of successes in the sample
n = sample size
$F1$ = F-table value for $df1 = 2(n - X + 1)$ and $df2 = 2X$
$F2$ = F-table value for $df1 = 2(X+1)$ and $df2 = 2(n - X)$
$df1$ = degrees of freedom for the numerator, and
$df2$ = degrees of freedom for the denominator.

Those formulas (or their mathematically equivalent rearrangements) are found in advanced statistical publications (e.g., Meeker & Escobar, 1998, p. 50) and are used by Minitab (2017b). Compared to the exact binomial method, such formulas are accurate to at least 9-decimal places at all sample sizes, all sample proportions, and all confidence levels (as determined by this author using Excel's F.INV and BINOM.DIST functions). Because of such extreme accuracy, *F*-table-based confidence intervals are described by some authors as being "exact" or the "same" as their binomial equivalents (respectively: Bowker & Lieberman, 1972, p. 467; Amstadter, 1971, p. 248). Henceforth, whenever this article refers to an "exact" binomial interval or limit, it is referring to the output of those formulas.

For many decades, most introductory statistics textbooks have included instructions on generic *F*-tests and have included *F*-tables that cover a sampling of *dfs* from 1 to infinity, but they have not included those formulas just given. Instead, such books have promoted what is here called the *Simple Normal* formula.

$$\text{Simple Normal Binomial Confidence Interval } = p \pm z\sqrt{\frac{p(1-p)}{n}},$$

where

p = proportion of successes in the sample
n = sample size
z = value taken from a normal-distribution z-table.

Copyright 2021 by John N. Zorich Jr. ---- All rights reserved.

The Simple Normal is "in virtually universal use" (Brown et al., 2002, p. 160) and is the "confidence [interval] expression most frequently used" (NIST, 2013). For example, Minitab uses it to calculate "Normal approximation" confidence limits (Minitab, 2017b).

When providing instructions for using the Simple Normal formula, textbooks typically include a warning that in order for the formula's output to be valid, sample size must meet a criterion. Some textbooks *explicitly* state that sample size must be "small," where small is defined as when "$n \leq .05N$" (e.g., Crossley, 2008, pp. 105–106), where n is sample size and N is population size. Conversely, the vast majority of textbooks *explicitly* state that sample size must be "large," where large is defined variously:

- "large" is when $np(1-p)$ is larger than 9, where p is the proportion of successes in the sample (Hájek & Dupač, 1967, p. 31),
- "large" is when np and $n(1-p)$
  a. both equal at least 20 (Treloar, 1938, p. 180),
  b. both equal at least 10 (Gravetter & Walnau, 2000, p. 201), or
  c. both equal at least 5 (Alder & Roessler, 1977, p. 117; Fleiss et al., 2003, p. 26).

The latter criterion, both equal at least 5, seems to be the most common.

A Simple Normal confidence interval can be significantly shorter than an exact binomial interval; consequently, the absolute position of its 95% confidence limits can differ from the exact ones by as much as 5 percentage points (see Figure 3). Therefore, its nominal claim of % confidence can be seriously incorrect; as seen in Figure 4, a supposed 95% confidence interval can have less than 90%-coverage (i.e., less than 90% confidence rather than the claimed 95%). All such differences represent errors that can be avoided by using an exact formula rather than an approximate one.

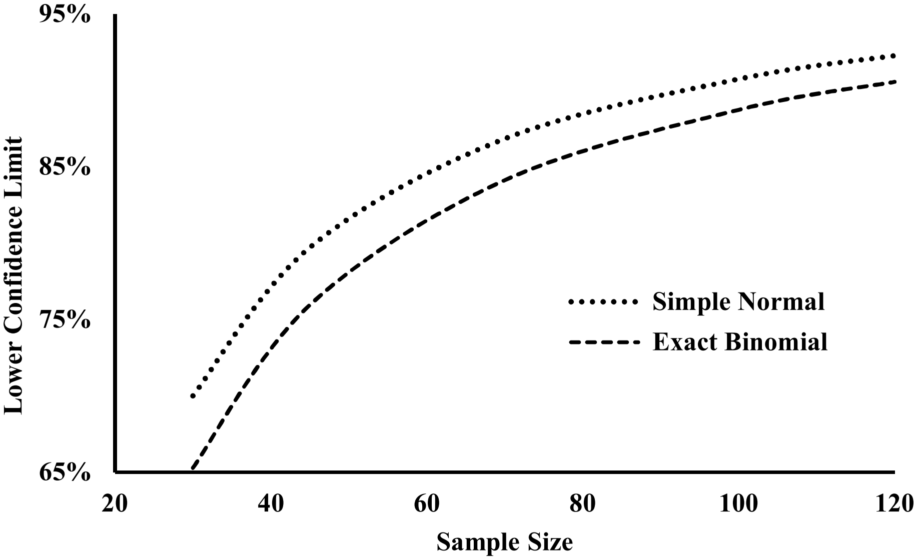

Figure 3. Lower 2-sided 1-proportion 95% confidence limit calculated using either the Simple Normal or the Exact Binomial. For each sample size n, sample proportion p was chosen so that the common cutoff criterion of $n(1-p)=5$ was met exactly.

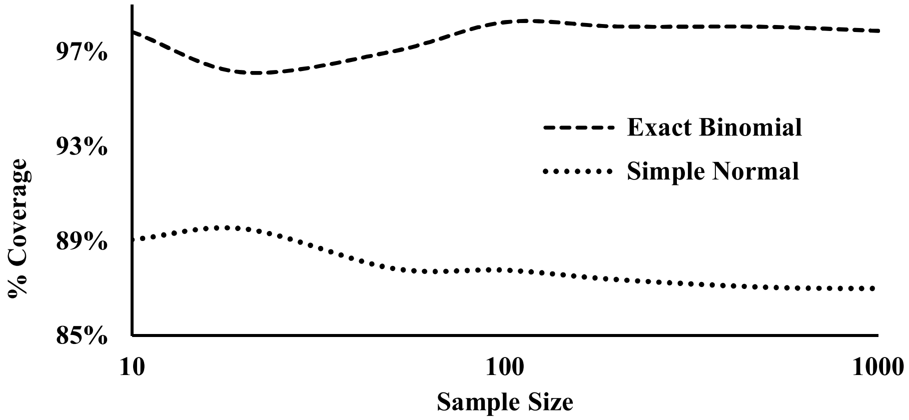

Figure 4. Average %-coverage of a 95% confidence interval based on 100,000 samples for each of 7 sample sizes ($n = 10, 20, 50, 100, 200, 500$, and 1000) generated using Minitab's "Random Data, Binomial" menu option, with parameter proportion p chosen so that the common cutoff criterion of $n(1 - p) = 5$ was met exactly in each case.

Also seen in Figure 4 is that the %-coverage of the Exact Binomial is approximately 97%. That 97% is consistent with the previously discussed *fundamental property* of a confidence interval, that is, to have *no less than* 95%-coverage.

Some combinations of sample size and sample proportion lead to Simple Normal confidence limits that are less than 0% or greater than 100%. Adherence to the previously mentioned cutoff criterion of 5 prevents such nonsense limits if the confidence level does not exceed 95%. However, when confidence = 99% is paired with $n = 100$ and $p = .95$, the upper limit is a nonsensical 101%, even though the $n(1 - p) \geq 5$ criterion has been met. In contrast, the F-based formulas previously discussed do not output nonsensical limits at any combination of sample size, sample proportion, and confidence level (as determined by this author using Excel's F.INV function).

An important related issue is the fact that, because a test of statistical significance is the inverse of a confidence interval calculation, introductory textbooks use a rearrangementof the Simple Normal formula to perform a simple-to-teach but *approximate* significance test on a sample proportion. However, the F-based formulas can similarly be used to perform an *exact* significance test (Fleiss et al., 2003, p. 38).

Summary and Recommendations

Other than tradition, there is no compelling theoretical, practical, historical, or educational reason for calculating confidence intervals using normal approximation formulas instead of exact ones. Although ease-of-calculation for large sample sizes was a valid concern many decades ago, that is no longer the case, given the present-day availability of statistical software, statistical functions in electronic spreadsheets, and textbook t- and F-tables with values for very large degrees of freedom. These days, normal approximation formulas are way past their expiration dates.

If we were to start teaching *only* exact methods for *all* sample sizes, students might be better prepared for life outside the classroom, where highly accurate methods are needed. For example, in the mid-to-late 20th century, automotive and semiconductor/high-tech product quality was being measured in defects per thousand and then defects per million, but in the 21st-century, that measure is becoming defects per *billion* (Kymal, 2017).

Two of the six recommendations that appear in the GAISE guidelines for teaching statistics in college are "focus on conceptual understanding" and "use technology" (Carver et al., 2016, p. 6). This present article stresses that accuracy itself is an important concept, one that can be taught by having students use technology to determine exact rather than approximate confidence intervals.

References

Alder, H. L, & Roessler, E. B. (1977). *Introduction to probability and statistics* (6th ed.). San Francisco, CA: W. H. Freeman & Co.

Amstadter, B. L. (1971). *Reliability mathematics: Fundamentals; practices; procedures*. New York, NY: McGraw Hill.

Bancroft, H. (1957). *Introduction to biostatistics*. New York, NY: Hoeber Medical Division of Harper & Row.

Bowker, A. H., & Lieberman, G. J. (1972). *Engineering statistics* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Brown, L. D., Cai, T. T., & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, 30(1), 160–201. doi:10.1214/aos/1015362189

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., … Wood, B. (2016). *Guidelines for assessment and instruction in statistics education (GAISE): College report 2016.* Retrieved from American Statistical Association website: http://www.amstat.org/education/gaise

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley & Sons.

Crossley, M. L. (2008). *The desk reference of statistical quality methods* (2nd ed.). Milwaukee, WI: ASQ Quality Press.

Duncan, A. J. (1986). *Quality control and industrial statistics* (5th ed.). Homewood, IL: Irwin.

Fisher, R. A. (1958). *Statistical methods for research workers* (13th ed.). New York: Hafner.

Fleiss, J. L., Levin, B, & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Gravetter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth.

Hájek, J., & Dupač, V. (1967). *Probability in science and engineering*. Prague, Czech Republic: Czechoslovak Academy of Sciences.

Kenney, J. F. (1939). *Mathematics of statistics (Part 1 & Part 2).* New York, NY: D. Van Nostrand Company.

Kymal, C. (2017, October 16). From percent rejects to parts per billion: Moving toward zero defects. *Quality Digest*. Retrieved from https://www.qualitydigest.com/print/30711

Meeker, W. Q., & Escobar, L. A. (1998). *Statistical methods for reliability data*. New York, NY: John Wiley & Sons.

Mendenhall, W. (1979). *Introduction to probability and statistics* (5th ed.). North Scituate, MA: Duxbury.

Minitab® 18, Minitab Statistical Software (Version 18.1). (2017a). Minitab menu sequence: Stat, Basic Statistics, One-Sample *z* for the Mean. Minitab Inc.

Minitab® 18, Minitab Statistical Software (Version 18.1). (2017b). Methods and formulas for 1 proportion. Minitab Inc. Retrieved October 18, 2018 from https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/1-proportion/methods-and-formulas/methods-and-formulas/

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*, 857–872. doi: **10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E**

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*, 558–625. doi: 10.2307/2342192

NIST/SEMATECH. (2013). *E-handbook of statistical methods* [e-book]. 7.2.4.1: Confidence intervals. doi: 10.18434/M32189

Staff of the Computation Laboratory. (1955). *Tables of the cumulative binomial probability distribution*. Cambridge, MA: Harvard University Press.

Treloar, A. E. (1938). *Elements of statistical reasoning*. New York, NY: John Wiley & Sons.

Walker, H. M. (1943). *Elementary statistical methods*. New York, NY: Henry Holt & Company.

John N. Zorich, Jr. (johnzorich@yahoo.com) received an MS degree in botany from the University of California, Davis, in 1979. He has worked in medical-device design and manufacturing as an independent statistical consultant and instructor since 1999. Annually since 2005, he has taught a course in introductory and applied statistics for the Biotechnology Center of Ohlone College (Newark, CA). Previously, he taught courses in applied statistics at Silicon Valley Polytechnic Institute and for the Silicon Valley ASQ Biomedical Division and is currently a periodic guest lecturer in applied statistics in the graduate program of Biomedical Engineering at San Jose State University. He currently resides near Houston, TX.

Reasons for No Longer Teaching the Normal Approximation Confidence Interval

John N. Zorich, Jr., *Ohlone College*

A typical, modern-day introductory statistics textbook teaches "normal approximation" confidence interval formulas for use with "large" samples and "exact" formulas for "small" samples. However, decades ago, a Harvard University professor of statistics warned that "there is no safe general rule as to how large *n* must be for use of the normal approximation in computing confidence limits" (Cochran, 1977, p. 42). Related advice has been given more recently: "In situations where the choice of test statistic or confidence interval would make a difference to the inferences drawn, exact methods should be relied upon rather than normal approximations" (Fleiss, Levin, & Paik, 2003, p. 62).

What are the theoretical, practical, historical, and educational reasons for basing formula choice on sample size? What reasons might there be for no longer teaching the normal approximation confidence interval?

Introduction to Confidence Intervals

The term *confidence interval* was coined by J. Neyman in 1934 when he defined it as a range "in which we may assume are contained the values of the estimated characters of the population [i.e., the population parameters]" (Neyman, 1934, p. 562). Many different formulas were developed for calculating the confidence limits that mark the upper and lower boundaries of confidence intervals (e.g., Brown, Cai, & DasGupta, 2002; Newcombe, 1998); some of those formulas produce intervals that are surprisingly inaccurate.

Literature on this subject provides a basic criterion for evaluating the accuracy of confidence intervals (e.g., for a 95% confidence interval): If all possible samples are drawn from a population, then *no less than* 95% of the confidence intervals derived from them will contain the population parameter. That coverage-based criterion has been called the "fundamental property" of confidence intervals (Fleiss et al., 2003, p. 24).

For the sake of brevity, discussions in this article will focus on confidence intervals for the mean of a single sample from a normally distributed population and for the proportion of "successes" in a single sample from a binomial population.

Confidence Intervals for Variable Data Averages

The equation for a single-sample variable-data confidence interval for the population average can be expressed as $m \pm C \cdot s/\sqrt{n}$, where m is the sample mean, s is the sample standard deviation, n is the sample size, and C is the coefficient taken from either a normal distribution z-table or from a student's distribution t-table. Some mid-20th-century introductory statistics textbooks that included detailed discussions of confidence intervals also included a z-table but not a t-table. In one such textbook, its z-table-only discussion of confidence intervals ended with this warning: "It must be borne in mind that, just as this [z-table-based] estimate of the confidence interval becomes more precise as N [sample size] increases, so inversely it becomes less acceptable as N decreases" (Treloar, 1938, p. 138). Another such textbook advised that "for small samples," the distribution of C in that confidence interval formula "is not normal and a probability table different from those based on the normal distribution is required" (Walker, 1943, p. 284). Instead of providing that required table (i.e., a t-table), Walker provided this: "At the close of this chapter are several references to such tables" (Walker, 1943, p. 284). Therefore, such textbooks could not be used to teach the exact ("more precise") t-table confidence interval, but rather only the approximate ("less acceptable") z-table interval.

Some introductory, mid-20th-century textbooks did not provide t-values between those for degrees of freedom (*df*) of 30 and infinity, where the *df* for infinity exactly equals the corresponding z-table value (e.g., Fisher, 1958, p. 174; Kenney, 1939, p. 138; Mendenhall, 1979, p. 535). Students who used such t-tables were, in effect, using the normal approximation for all sample sizes larger than 31. These days, a typical textbook provides t-values for *dfs* 1–30, 40, 60, 120, and infinity.

Many statistics textbooks, as well as the statistical software program *Minitab*®, advise that coefficient C be taken from a z-table when the standard deviation's true value (i.e., the parameter) is *known*, but be taken from a t-table when it is *unknown* (Minitab, 2017a). Some textbooks explain that although such advice is theoretically correct, it is "somewhat artificial," because *not* knowing the parameter "is by far the more realistic situation" (Glass & Stanley, 1970, p. 260).

For such realistic situations, most textbooks have long advised that "we can use [z-tables] … if the sample is large—say greater than 30.... If the sample is less than 30, we should use [t-tables]" (Duncan, 1986, p. 567). As can be seen in the variable-data confidence interval formula, the difference between the length of confidence intervals calculated with z- or t-tables is due solely to the difference in magnitude of z- and t-values, z always being smaller than t at a given confidence level (see Figure 1). At $n = 30$,