STAT COE-Report-20-2014

A Practical Guide for Interpreting Confidence Intervals

Best Practice

Authored by:
Jennifer Kensler, PhD STAT COE
Luis A. Cortes, STAT COE

24 December 2014

The goal of the STAT T&E COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT.

STAT T&E Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

This best practice introduces confidence intervals, which use observed data to obtain an interval estimate of an unknown population parameter such as a mean or proportion. Confidence intervals are a key tool of inferential statistics and appear in many contexts. While there are many types of confidence intervals, the underlying principles are the same. This best practice provides an overview of confidence intervals and introduces key concepts and terminology. It highlights the importance of understanding and correctly interpreting a confidence interval as well as common errors and misunderstandings. A simple example of calculating a confidence interval for a proportion illustrates the concepts presented in the context of Department of Defense (DoD) testing.

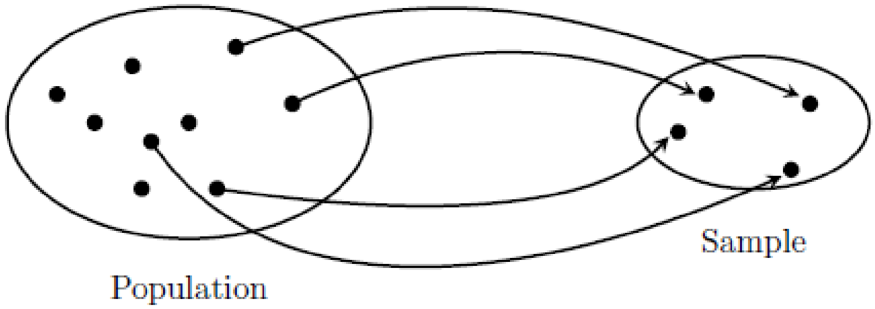
Keywords: Confidence Interval, Statistical Inference, Confidence Level, Sample Size, Margin of Error

Introduction

Statistical Inference

Often one is interested in drawing conclusions about a population, but examining the entire population is usually impractical or impossible. Statistical inference involves using information obtained from a sample to draw conclusions about a larger population. Statistical inference is key to having rigorous and adequate DoD tests because we are often interested in future performance of a system under similar conditions. Since we do not know what the future holds we are dependent on statistical inference to make statements about future performance. Therefore, the sample must be representative of the population; this objective is usually achieved by obtaining a randomly selected sample. Figure 1 illustrates that a sample is a subset of the population. A confidence interval, one form of statistical inference, uses data observed from a sample to estimate a population parameter. The data one observes will be different depending on which individuals of the population the sample captures. Confidence intervals address this random sampling “error” (i.e. variation) and allow one to estimate the value of a single parameter or function of parameters.

Figure 1. Sample vs. population



An Illustrative Example

The Navy wants to estimate the probability that a missile will hit its target. The population contains all missiles, and the proportion of missiles in the entire population that hit their target is denoted by π. The proportion of missiles in the sample that hit their target is denoted by $\hat{p}$ (read "p-hat"). While $\hat{p}$ will be used to estimate π, the Navy knows that $\hat{p}$ will not be exactly π. Therefore, the Navy wants to use the sample data to come up with an interval estimate likely to contain π. This best practice uses the simple missile example to illustrate the concept of a confidence interval. Usually a continuous response variable (e.g. miss distance from target) is preferable to a binary response variable (e.g. hit or miss). For a discussion on the advantages and disadvantages of continuous and binary response variables, see Ortiz (2013). Furthermore, it is often necessary to examine the performance of the missile across the operating environment instead of at a single point.

DoD Testing

Throughout the acquisition lifecycle many questions must be answered regarding the performance and suitability of a system. A common objective is to summarize the sample data with a mean value and to estimate the "real but unknown" population mean. What is the mean miles between system failures? What is the mean target location error? What is the maximum range of a weapon? What is the mean detection range of a radar? Confidence intervals are a method for answering these questions.

Foundations of Confidence Intervals

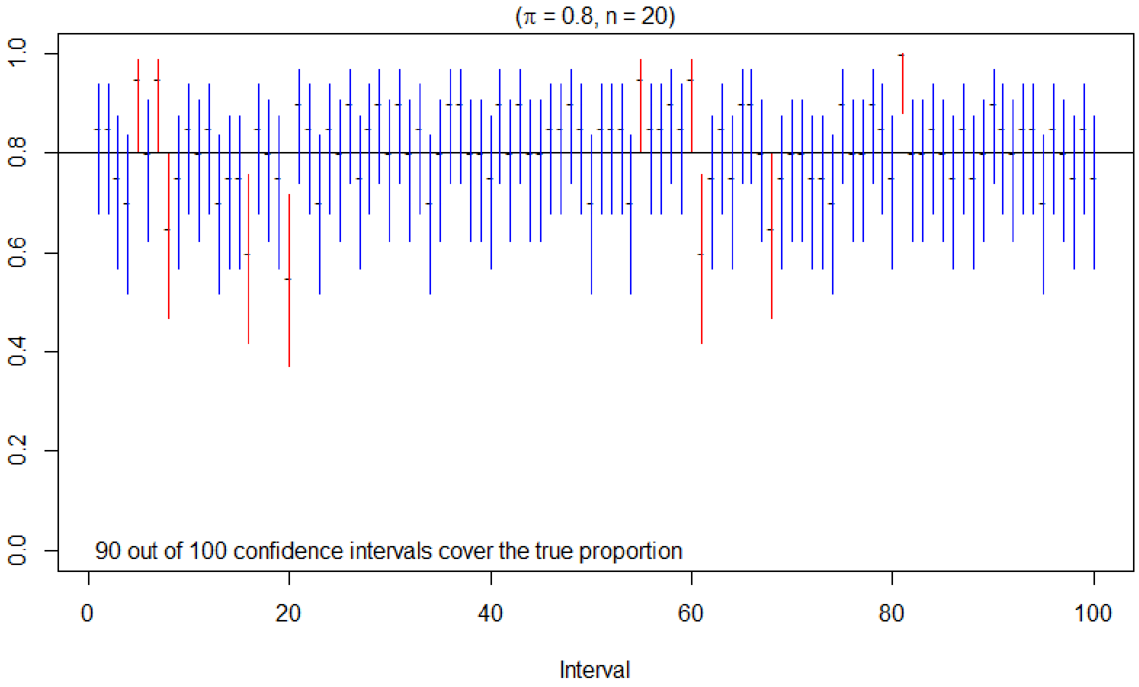
Confidence Interval Fundamentals

Estimation is a key objective of many statistical analyses. A point estimate is a single numerical value used to estimate a population parameter. For example, the sample proportion, $\hat{p}$, is a point estimate used to estimate the population proportion, π. A point estimate is our best guess estimate for the population parameter; however, we know that the point estimate will not be exactly the population parameter. A confidence interval provides an estimate of the population parameter and the accompanying confidence level indicates the proportion of intervals that will cover the parameter. In other words, a confidence interval provides a range of values that would contain the true population parameter for a specified confidence level.

A $100(1-\alpha)\%$ confidence interval is an interval estimate where if we could repeat the process of interval estimation an infinite number of times the intervals would contain the true value of the parameter $100(1-\alpha)\%$ of the time. For example, a 95% confidence interval means that in the long run 95% of confidence intervals constructed in this manner will contain the true parameter. However, we cannot know whether the interval estimate we calculated is one of the intervals that contains the true parameter or one of the intervals that does not.

Figure 2 shows 90% confidence intervals for 100 samples (each with 20 observations) drawn from a binomial population with $\pi = 0.8$. Note that 90 of the intervals cover the population proportion of 0.8.

Figure 2. 90% Confidence Intervals



Note that the probability refers to the method, not the individual interval. A 90% confidence interval ***does not*** mean there is a 90% probability that the parameter is in the interval. The interval either covers the parameter or it does not. So the probability that the parameter is in the interval is either 1 or 0. Consider a telecommunications system that correctly receives 99% of messages, thus there is a 99% probability that a message you send will be received. However, after you send the message there is no longer any randomness. The message was either received or not received. In the case of the confidence interval the data sample is what is random. Hence, once the data is collected and the confidence interval constructed, it no longer makes sense to talk about the probability of a parameter being in the interval. The interval either covers the parameter or it does not. Salsburg (2001) explains "Note that, to Neyman, the probability associated with the confidence interval was not the probability that we are correct. It was the frequency of correct statements that a statistician who uses his method will make in the long run. It says nothing about how 'accurate' the current estimate is."

A confidence interval is valid or accurate if it contains the population parameter. On the other hand, the narrower the confidence interval the more precise it is. Figure 3 illustrates the effect of changing the confidence level to 80%. Compared to the intervals in Figure 2, the intervals in Figure 3 are more precise (narrower), but less accurate (more intervals fail to cover 0.8). Note that a higher confidence level requires a wider interval in order to cover the parameter more often.

Figure 3. 80% Confidence Intervals

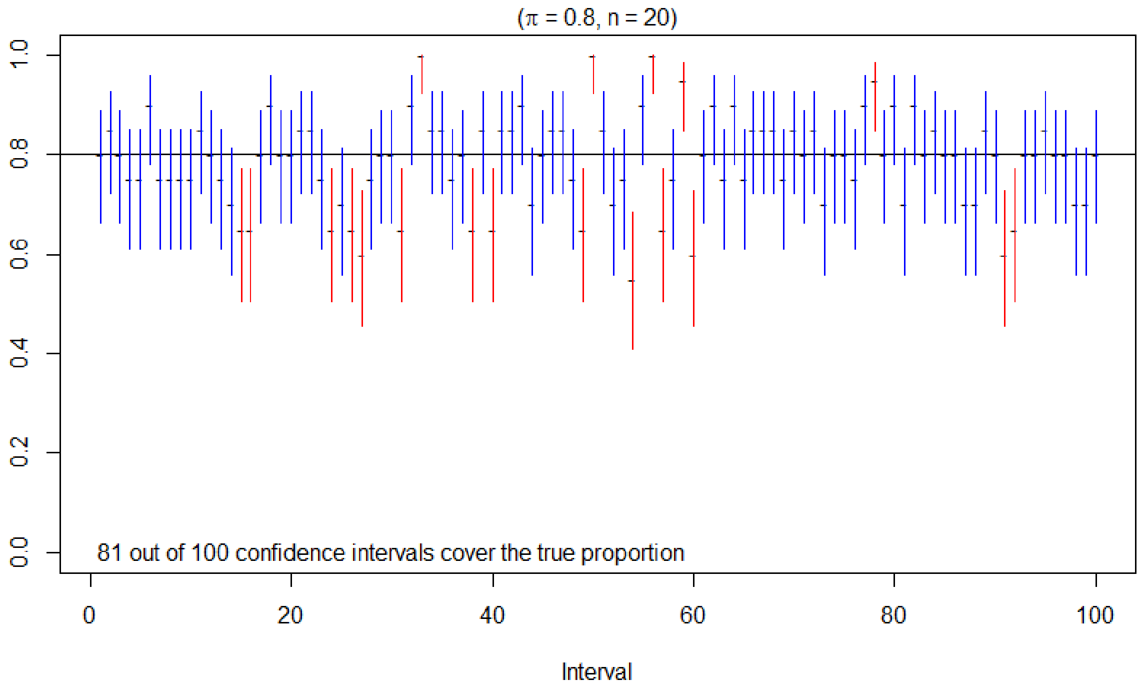


Figure 4. 90% Confidence Intervals

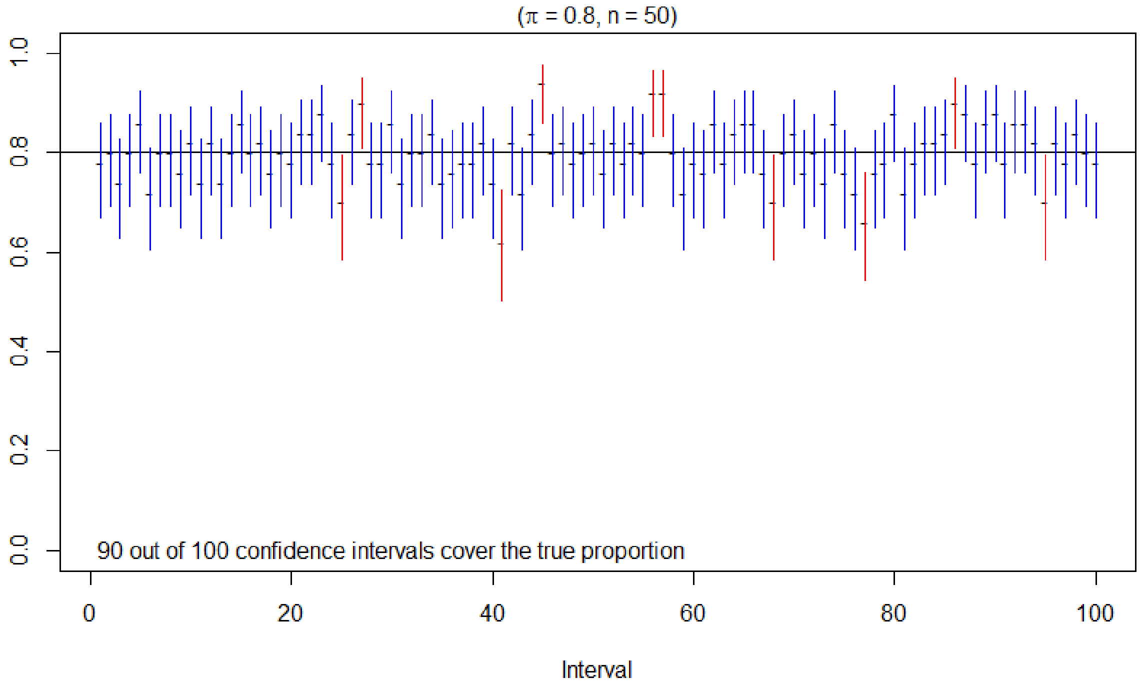


Figure 4 shows 90% confidence intervals from samples of size 50. The intervals in Figure 4 are more precise that the intervals in Figure 2, but have the same level of accuracy. The narrower intervals reflect the increased information from the larger sample.

Sample Size for Confidence Intervals

One may be interested in estimating a parameter with a confidence interval of a certain precision. That is, we may want to control the width of the confidence interval. In the case of a proportion, quantities that affect the width of the confidence interval include the confidence level, sample size, and the sample proportion. Using the confidence level, desired interval width, and a planning value for the proportion; the necessary sample size can be calculated.

Calculating a Confidence Interval

Planning

The Navy wants to estimate the proportion of missiles that hit their targets. Currently, the information the Navy has suggests that a missile has an 80% probability of hitting the target, this value is used as the planning value for the population proportion. The navy decides to use a 90% confidence level and decides to compare the sample sizes required to obtain various interval widths (Table 1).

Table 1. Interval Half-Width	Sample Size
0.05	172
0.10	43
0.15	18
0.20	10

The sample size, n, required to obtain a Wilson score $100(1-\alpha)\%$ confidence interval with a half-width e (two-sided confidence interval for continuous variables) is

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}} \left[\pi_0\,(1-\pi_0\,) - 2e^2 + \sqrt{\pi_0^2(1-\pi_0)^2 - 4e^2\pi_0(1-\pi_0) + e^2}\right]}{2e^2} \right\rceil$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution and π_0 is the planning value of the proportion. If no information is known about proportion, then $\pi_0 = \frac{1}{2}$ should be used.

The interval with the half-width of 0.05 is ruled out because the sample size of 172 is too expensive. The interval with a half-width of 0.20 is ruled out because it is too wide to be useful. Given their current budget, the Navy decides to use a sample size of 18 to obtain an interval with a half-width of 0.15.

Analyzing the Data

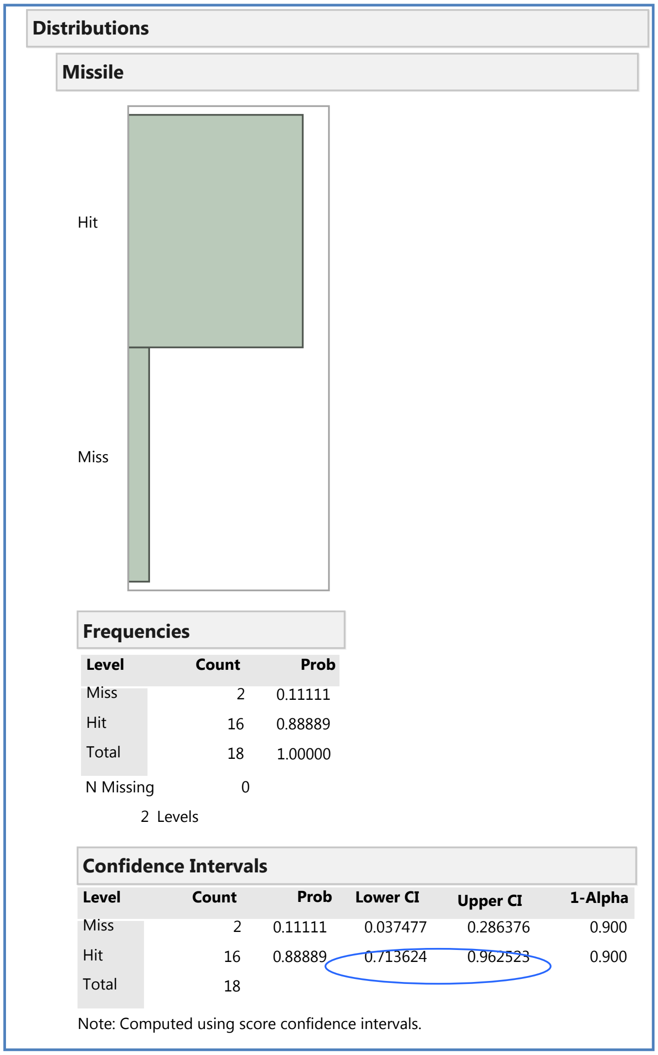
The Navy runs the test and observes 13 of 15 missiles hitting the target. Figure 5 shows the statistical output from JMP (Cary, NC).

The sample proportion of hits is $\hat{p} = \frac{16}{18} = 0.89$. The 90% Wilson score confidence interval for the population proportion of hits is (0.71, 0.96). The Wilson score $100(1-\alpha)\%$ confidence interval for a population proportion (two-sided confidence interval for binomial variables) is

$$\hat{p}\left(\frac{n}{n+z_{1-\frac{\alpha}{2}}^{2}}\right)+\frac{1}{2}\left(\frac{z_{1-\frac{\alpha}{2}}^{2}}{n+z_{1-\frac{\alpha}{2}}^{2}}\right)\pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n+z_{1-\frac{\alpha}{2}}^{2}}\left[\hat{p}(1-\hat{p})\left(\frac{n}{n+z_{1-\frac{\alpha}{2}}^{2}}\right)+\frac{1}{4}\left(\frac{z_{1-\frac{\alpha}{2}}^{2}}{n+z_{1-\frac{\alpha}{2}}^{2}}\right)\right]}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution, $\hat{p}$ is the sample proportion, and n is the sample size.

Figure 5. Statistical output



Distributions

Missile

Frequencies

Level	Count	Prob
Miss	2	0.11111
Hit	16	0.88889
Total	18	1.00000

N Missing 0

2 Levels

Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Miss	2	0.11111	0.037477	0.286376	0.900
Hit	16	0.88889	0.713624	0.962523	0.900
Total	18				

Note: Computed using score confidence intervals.

Conclusion

Confidence intervals allow us to take information from a sample and use it to form an interval estimate for a population parameter or function of parameters. In DoD testing, confidence intervals are often calculated for almost every performance measure (such as mean time between failures, proportions, etc.) required for the evaluation. Up front planning scopes the sample size needed to obtain an interval of a certain width. This planning helps avoid the situation where not enough data is collected resulting in a confidence interval too wide to be useful. Regardless of the particular interval being calculated, it is important to correctly interpret the confidence interval.

References

Agresti, A. (2002). *Categorical Data Analysis* (second ed.). Hoboken, New Jersey: John Wiley & Sons.

Ortiz, F. (2013). *Dealing with Categorical Data Types in a Designed Experiment Part I: Why You Should Avoiding Using Categorical Data Types.* Retrieved December 30, 2014, from STAT in T&E Center of Excellence: http://www.afit.edu/docs/Dealing%20with%20Categorical%20Data%20(Part%201%20of%203).pdf

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* New York: W.H. Freeman and Company.

Sullivan, M. (2004). *Statistics Informed Decisions Using Data.* Upper Saddle River, New Jersey: Pearson Education.