

Confidence intervals are the most valuable statistical tools available to decision makers. However, for a variety of reasons, confidence intervals are not used as frequently as they should. This article answers two questions that are often misunderstood:

- Why are point estimates useless for making decisions?
- What is the best confidence level?

This article does not discuss how to calculate confidence intervals, since widely available software automates this task. Formulas and calculation methods are well documented in many books such as *Montgomery* (2008) or *Sleeper* (2006).

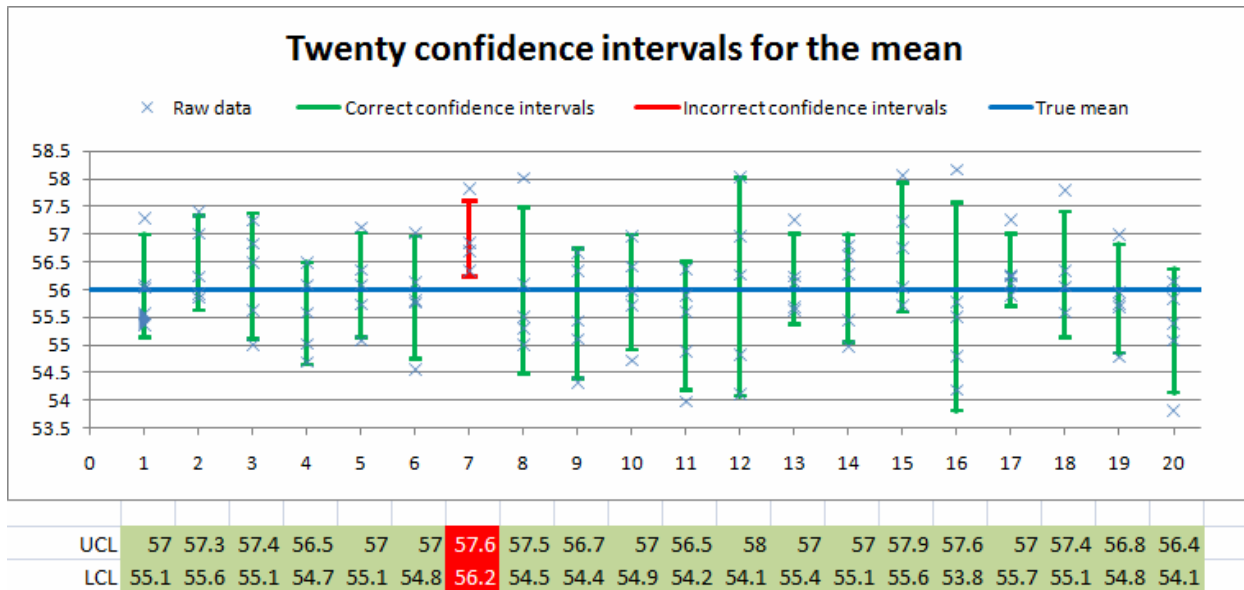
### Why are Point Estimates Useless for Making Decisions?

**Example 1:** Suppose I measure the hardness of five steel parts, and the measurements are 57, 55, 57, 56 and 55. The mean of these measurements is 56. Usually, we expect this value to represent something about a larger population of parts. The population mean  $\mu$  cannot be known with certainty, but the sample mean, 56, is our best estimate of the population mean, based only on these five parts. The number 56 is a **point estimate** of the population mean.

The mean hardness of the population,  $\mu$ , could be any real number representing a feasible value, like 56.12 or 53.85, but the mean hardness is almost certainly *not* equal to 56.00. This is an important point, which consumers of statistics often misunderstand. If I say to you, “ $\mu$  is 56,” that would be a lie, because  $\mu$  refers to the unknowable population mean. But if I say ambiguously, “The mean is 56,” I might think about the sample, while you might think about the population. To make a business decision, we all need to know something about the population, not the sample. Without additional information, the point estimate of 56 is a useless number. The population mean  $\mu$  is probably close to 56, but how probable and how close?

Instead, if I said to you, “I am 95 percent confident that  $\mu$  is between 54.76 and 57.24,” this would be a true statement. An **interval estimate** or **confidence interval** is a range of values that contains the true population parameter value with a known probability. The probability that the interval contains the true value, 95 percent in this example, is the **confidence level**. If I make such statements about  $\mu$  frequently, 95 times out of 100,  $\mu$  will be within the range I specify; and 5 times out of 100  $\mu$  will be above or below the range. This assumes that my math is correct, and that the assumptions behind the confidence interval are valid.

Figure 1 illustrates this situation in a different way. Suppose for a moment that the true population has a normal distribution with mean  $\mu = 56$  and standard deviation  $\sigma = 1$ , which we cannot know for sure in real life. Making these assumptions, suppose I collected 20 samples of size  $n = 5$ , and calculated a 95 percent confidence interval for  $\mu$  from each sample. The graph shows 20 such intervals with green or red lines. Nineteen of the 20 intervals contain the value 56, so these are correct. One interval does not contain 56, so it is incorrect.



**Figure 1: Simulation of 20 of the 95 Percent Confidence Intervals**

Since this spreadsheet is a random simulation, recalculating it leads to different results. Sometimes, all 20 intervals are correct, and sometimes fewer than 19 are correct. In the long run, 95 percent of these interval estimates contain the true value of the mean  $\mu$ . This Excel file is available from the author for anyone who wants to play with it.

**Example 2:** Here's another example near and dear to our Six Sigma hearts. Suppose the CEO has decreed that we need  $C_{PK}$  to exceed 1.50 for all critical characteristics. If I measure a sample of parts and announce " $C_{PK}$  is 1.63," this sounds like good news. But then you ask a really good question: "How large is the sample size?" If you discover the sample size was only three, should you be worried? What if you discover the sample size was 300?

We have to make a decision about the capability of the population, but once again, the point estimate is not enough information by itself to make this decision. It is another useless number.

Instead, suppose I said, "I am 95 percent confident that  $C_{PK}$  is at least 1.52." Or I could say, "I am 97 percent confident that  $C_{PK}$  is at least 1.50." Either of these would be a true statement; and since sample size is used to make these calculations, they provide all information necessary to make the business decision.

These one-sided confidence intervals are often called **lower confidence bounds**, because the upper limit of each confidence interval is infinity. In the case of  $C_{PK}$ , we usually don't care how large it is, so a lower confidence bound is more appropriate than a two-sided confidence interval.

Because they are single numbers, point estimates are almost always above or below the parameters they are supposed to estimate. Without additional information, point estimates are useless for making decisions. But confidence interval estimates are very

likely to be true, and the confidence level specifies and controls the probability that the interval estimates are true. Since properly applied confidence intervals incorporate sample size and other tested assumptions, these are reliable tools to make business decisions.

### What is the Best Confidence Level?

Most confidence intervals use a confidence level of 95 percent or 90 percent, but these levels are not right for every situation. To pick the most appropriate confidence level, we need to think about the business decision to be made, and the potential effects of making a bad decision.

When using a confidence interval to make a decision with two choices, there is always a decision value. If the decision value is inside the interval, one choice will be made, but if the decision value is outside the interval, the opposite choice will be made.

In the example of hardness, suppose the ideal value of hardness is 56, and we must decide whether to adjust the hardening process. If the decision value of 56 is outside the interval, we adjust, but if 56 is inside the interval, we leave the process alone.

In the  $C_{PK}$  example, the decision value is 1.50. If 1.50 is below the lower confidence bound, outside the interval, we conclude that the population  $C_{PK} > 1.50$  and we choose to accept the process for production. But if 1.50 is above the lower confidence bound, then we do not know whether  $C_{PK} > 1.50$  and we choose to reject the process, until the capability is improved.

Since nothing can be known with absolute certainty, there are two possible errors in making this decision process. Table 1 describes these two errors as they might affect the hardness or  $C_{PK}$  decisions.

Errors with confidence intervals	Hardness example: Adjust process to 56 or not?	$C_{PK}$ example: Accept that $C_{PK} > 1.50$ ?	Controlling the probability of error
Type I error: The decision value should be inside the interval, but it falls outside.	Mean $\mu=56$ , but 56 is outside the interval. Result: needless adjustment.	True $C_{PK} \leq 1.50$ , but the lower confidence bound $> 1.50$ . Result: the process is accepted with poor capability.	The probability of a type I error, $\alpha$ , is 1 minus the confidence level.
Type II error: The decision value should be	Mean $\mu \neq 56$ , but the interval includes 56.	True $C_{PK} > 1.50$ , but the lower confidence bound $< 1.50$ .	The probability of a type II error, $\beta$ , depends on the sample

outside the interval, but it falls inside.	Result: No adjustment when it is needed.	Result: The process is rejected when it is acceptable.	size and how far the population is away from the decision value.
--	--	--	--

**Table 1: Two Types of Errors with Confidence Intervals**

A type I error occurs when the correct decision would be indicated by the decision value falling inside the confidence interval, but random variation causes the decision value to fall outside. Depending on the situation, type I errors may be called “false alarms” or “producer’s risk,” but these labels do not always apply. The probability of a type I error, represented by  $\alpha$ , is always controlled to be 1 minus the confidence level.

When calculating 95 percent confidence intervals for mean hardness, there is a 5 percent probability of falsely concluding that the mean hardness is too high or too low. When calculating 95 percent lower confidence bounds for  $C_{PK}$ , there is a 5 percent probability of accepting the process with a bad  $C_{PK}$  of less than 1.50. If this 5 percent probability is too small or too large, the confidence level can be adjusted up or down as required.

A type II error occurs when the correct decision would be indicated by the decision value falling outside the interval, but random variation causes the decision value to fall inside. Depending on the situation, type II errors may be called “missed detections” or “consumer’s risk.” The probability of a type II error, represented by  $\beta$ , depends on the sample size and the location of the true population value. Calculating  $\beta$  is usually very difficult, and requires sample size calculations for hypothesis tests. It is often easier to estimate  $\beta$  by simulation. Also, because  $\beta$  depends on the true population values, it is impossible to control  $\beta$  after the dataset is collected. Sample size is the only control which can change  $\beta$ .

The two risks,  $\alpha$  and  $\beta$ , can be traded for each other. With constant sample size, increasing  $\alpha$  decreases  $\beta$ , and vice versa. By examining the effects of the two decision errors in your specific situation, you can be smarter about choosing confidence levels. If the potential impact of a type I error is very serious, consider increasing the confidence level to decrease  $\alpha$ . If the potential impact of a type II error is very serious, consider increasing the sample size or decreasing the confidence level to decrease  $\beta$ .

**Example 3:** The manufacturer of a new milling machine claims to reduce standard deviation of part dimensions by at least a factor of two. We would be willing to buy the machine if we can prove the new standard deviation is less than half of the old. To test this claim, we mill a set of parts on the new machine and estimate an upper confidence bound for  $\sigma$ , the standard deviation of the parts. Since the old standard deviation  $\sigma_0$  is known, the decision value is  $\sigma_0/2$ . If the upper confidence bound  $U_\sigma > \sigma_0/2$ , we will reject the machine, but if  $U_\sigma < \sigma_0/2$ , this proves the manufacturer’s claim and we will purchase the machine.

A Type I error happens if the machine is not as good as claimed ( $\sigma > \sigma_0/2$ ), but the upper confidence bound  $U_\sigma < \sigma_0/2$  so we decide to buy the machine, which is not as good as we expect.

A Type II error happens if the machine is as good as claimed ( $\sigma > \sigma_0/2$ ), but the upper confidence bound  $U_\sigma > \sigma_0/2$  so we falsely reject the machine.

In this example, we put the burden of proof on the manufacturer to prove the machine is good enough. We can directly control  $\alpha$ , the probability of buying an inadequate machine by setting the confidence level. We could be brutally hard on the manufacturer by demanding 99.9 percent confidence, but we decide that 90 percent confidence is reasonable. This means that we are accepting a 10 percent probability of buying an inadequate machine.

After the data is collected, we can calculate the exact confidence level that sets  $U_\sigma = \sigma_0/2$ . Subtracting this confidence level from 1 gives the actual  $\alpha$ , the probability that the machine does not meet the manufacturer's claims.

**Example 4:** Fatigue is a potentially deadly failure mode in aircraft parts. During the design process, fatigue safety factors (FSF) are estimated from simulation and testing. To prevent fatigue failures, the true value of FSF must be greater than 1.

After testing a set of parts, we calculate  $L_{FSF_{01}}$ , a lower confidence bound for the 1<sup>st</sup> percentile of FSF. In other words, we will have a specified confidence that 99 percent of the parts will have FSF in excess of 1. If  $L_{FSF_{01}} < 1$ , we will choose to reject the design as inadequate, but if  $L_{FSF_{01}} > 1$ , we will choose to accept the design as reliable. What confidence level is appropriate here?

A type I error happens if the 1<sup>st</sup> percentile of FSF is  $< 1$ , so the design is inadequate, but  $L_{FSF_{01}} > 1$  leads to accepting an inadequate design. This would be bad.

A type II error happens if the 1<sup>st</sup> percentile of FSF is  $> 1$ , so the design is adequate, but  $L_{FSF_{01}} < 1$  leads to rejecting the design. This could result in needless added cost, weight and project delays.

Since human safety is involved in the type I error, its probability must be small and controlled. The team may decide to calculate a 99 percent confidence interval, setting  $\alpha = 1$  percent.

It is tempting to add lots more 9s to this example to make the parts *really* safe. But beyond a certain point, the added cost would be unacceptable, and it would impair functionality. A brick of steel may never fatigue, but it will not fly far either. A properly performed failure modes effects analysis (FMEA) can help to balance the competing failure risks and identify the most serious ones. This knowledge helps avoid costly over-design, while preventing the most likely and most serious failures. For more information on FMEA, see AIAG (2008).

When thoughtfully applied, confidence intervals and confidence bounds are powerful tools for making decisions. Unlike point estimates, confidence intervals incorporate the

risks inherent in small sample sizes, and control the risks of making bad decisions. Of the two types of decision making risks, the risk of Type I errors,  $\alpha$ , can be directly controlled, since  $\alpha = 1$  minus the confidence level. The risk of type II errors,  $\beta$ , is indirectly controlled through the sample size.

As a final caution, all confidence interval methods make assumptions about the data. When practical, these assumptions should be verified. Invalid assumptions might change the actual  $\alpha$  and  $\beta$  away from what you think they are. Also, not every statistic has easily calculated confidence intervals. Notably, I am not aware of any major statistical software product that calculates confidence intervals for gage R&R statistics. These formulas are massive, but they are all worked out by Burdick, Borror and Montgomery (2005).

Now that you know more about confidence intervals, don't hesitate to use them. Remember, a point estimate by itself is just another useless number. Have confidence in your analysis!

## References

AIAG (2008): *Potential Failure Modes Effects Analysis*, 4<sup>th</sup> Edition, Automotive Industry Action Group, [www.aiag.org](http://www.aiag.org)

Burdick, R. K., Borror, C. M., and Montgomery, D. C. (2005) *Design and Analysis of Gauge R&R Studies: Making Decisions with Confidence Intervals in Random and Mixed Anova Models* (ASA-SIAM Series on Statistics and Applied Probability)

Montgomery, D. C. (2008) *Introduction to Statistical Quality Control*, 6<sup>th</sup> Edition, Wiley

Sleeper, A. D. (2006) *Design for Six Sigma Statistics: 59 Tools for Diagnosing and Solving Problems in Six Sigma Initiatives*, McGraw-Hill