Confidence Intervals: The Average is Just Along for the Ride

Authored by:

Paul F. Doruska (CF #1213)
College of Natural Resources
University of Wisconsin-Stevens Point

Upon Request of the
Society of American Foresters Forest Science and Technology Board

---

You have performed a forest inventory using systematic random sampling and estimated there were 8.9 MBF of sawtimber volume per acre. Upon harvest, the stand cut out (using the same merchantability specifications employed during the inventory) at 9.7 MBF per acre. Hmmm……

The results from a questionnaire given to a random sample of park visitors indicated that 77% of the visitors were willing to pay an additional $3 to use a recreational area in that park. Upon subsequent implementation of the fee, only 59% of park visitors paid the fee and used the area. Hmmm……

An intern employed a random sampling protocol to monitor seedling survival in a large herbicide trial and determined that 81% of the seedlings survived, it turns out after all seedlings were later measured that the actual survival was 87%. Hmmm…..

A research study indicated a silvicultural treatment would increase diameter growth by 30%, enough to justify the cost of the treatment. Upon implementation of the treatment (under the same conditions as in the research study), growth only increased by 12.7%. Hmmm….

In each of the above cases, the results from properly designed and employed random samples (this was not explicitly stated but go ahead and assume it was the case) differed from the actual results. Sometimes actual results were larger than the estimated value, other times they were lower than the estimated value. Sometimes decisions were made based on the estimated values, other times decisions could be influenced by the estimated values. It sure would have been good to know the actual value to help in the decision making process, would it not? Why can the results from properly designed and employed random samples not give you the actual result? Should one be placing any faith in the results from random samples?

The reality is one almost has to put faith in the results of random sampling – it is simply impossible to measure everything all of the time. One must be careful though, into which results one places his/her faith. Most folks put faith in exclusively in the average (or mean). To repeat a concept stressed in an earlier submission in this series (Doruska 2009), a quote I attribute to Rich Oderwald (Virginia Tech) resurfaces.

> “If all you know is the average from a random sample, then here is what you really know: there is a 50% chance the correct (or true) average is larger than your number, a 50% chance the correct (or true) average is less than your number”

This is exactly what happened in each of the four brief examples that led off this discourse. The estimate did not equal the actual result. Why did this happen in these cases? Why does this happen in practically every case? The answer to that question has two components (a.) sampling uncertainty and (b.) variation. If you have already read the aforementioned series submission (Doruska 2009) entitled “Mean and Variation (The Hazards of Examining Just the Average)” , you have a sense for the term variation - if you have not read that submission, you might want to take a few minutes and peruse it as many of the concepts will appear in this article.

Now let us go back to each of the four examples that began this article, and delve deeper into the results from the sampling.

Remember that tract for which the inventory indicated had 8.9 MBF per acre but actually cut out at 9.7 MBF per acre. A closer look at the inventory results indicated that in addition to estimating the average of 8.9 MBF/acre one could be 80% sure (or confident) the interval [7.8 , 10.0] MBF/acre contained the true (but obviously unknown at the time of the inventory) result.

Remember that questionnaire? In addition to indicating 77% of visitors would pay the fee to visit the area, a closer look at the results indicated that one could by 95% confident the interval [55% , 99%] contained the true result.

How about that survival study, in which it was estimated that 81% of the seedlings survived, but it turned out that 87% actually did. What if other that concluding that 81% survived, one also added the statement that one could by 90% confident the interval [72% to 90%] contained the true survival percentage.

Lastly, what about that silvicultural experiment that indicated an increase on diameter growth of 30% would be obtained when only 12.7% actually resulted? Well, further analysis of the research project’s data indicated one could be 80% confident the interval [0% , 60%] contained the actual (but unknown at the time) increase. Take a close look at that last interval – are you asking yourself “did I need that experiment to make that statement?” I suspect you are thinking “no” and you are probably right – it did not take a research project to make that statement. (Note, I am <u>NOT IN ANY WAY</u> intending to slight silvicultural experiments here, I just needed one of the examples to make the previous point, and I happened to chose this one out of the four!)

First, notice that for each of the intervals described above, the average or result mentioned in the initial presentation of each scenario at the start of this article is the center of the interval. Also note that when initially presented, I suspect you were thinking things like “ hey that was pretty close” or “Whoa Nellie, that one was off”. Am I correct? Now, think about your career, have you been ever been in a similar situation, where the average or result estimated from a random sample did not match reality, leading you to say “Dadgummit, what the heck happened here”? I suspect you have.

Okay, now jump to the second presentation of each example, and the additional information provided. With that added information, namely the intervals, does the phrase "Dadgummit, what the heck happened here" apply? Does it not put the actual estimate in a slightly better perspective? In each case, the actual result was within the interval's range. Will this always be the case? Ideally, the answer will be yes, but in reality the answer will sometimes be no (more on that later…). What is perhaps more frightening, though, is that you will never know when this is the case unless you have an opportunity to obtain the actual answer.

So what changed one's perception of the results between presentations of the examples? Simply changing what results one examines. First time through, we had just a number. Now recall the quote

> "If all you know is the average from a random sample, then here is what you really know: there is a 50% chance the correct (or true) average is larger than your number, a 50% chance the correct (or true) average is less than your number."

Second time through we had what are called confidence intervals. Confidence intervals combine the estimate from the random sample, with those two additional components mentioned earlier (a.) sampling uncertainty and (b.) variation. In each of the statements presenting the interval, phrasing was included that stated one was "a certain percent confident" about the interval – that hints at the sampling uncertainty component, but what I ask one pays most attention to is the form of the interval:

[lower bound , upper bound]

Which under closer inspection in each case is actually :

[the average minus a specific number, the average plus that specific number]

and now we are getting somewhere. The intervals are always of the form: the average plus and minus a specific number. A "specific number" sounds kind of generic and boring, does it not? Let us give it a name. How about Paul? Erica? Ernie? Gingivitis? Those are probably not working for you in terms of confidence intervals[1], so let us go with this one: Error. The interval then becomes the average plus and minus error:

average minus error , average plus error]

I hope you are now beginning to see why I titled this article as I did. With confidence intervals, the average is just along for the ride – all it does is center the interval on a number line! All the action is with the error component, as that determines how wide the interval is and actually helps us determine whether the interval is actually useful to us (remember that silvicultural example from before – bingo!)

---

[1] Author's side note: Paul is my name, Erica is my wife's name, and Ernie and Gingivitis (Ginni for short) are the names of my dogs…

Said another way, at what point is the interval too wide to be useful to us? Do note that error is the half-width of the interval, so we are essentially asking at what point is the error from the interval unacceptable?

Still not with me? How about this analogy – think about a confidence interval as trying to wrap your arms around (or trying to grab) the range of values in which the average (as suggested by your sampling endeavor) can fall – are your arms close together or are they far apart? If they are close together (low error), I am thinking you are feeling pretty good about your sampling, if they are far apart (large error) things are not so good. Do keep in mind there is that percent confident phrasing to deal with at the same time – if your arms are close together because you are only 10% confident, that is not a good feeling; but if your arms are close together and you are 90% confident, I suspect you are feeling pretty good.

Hopefully one now realizes that this term called error is a big deal when summarizing sampling efforts. It is, but luckily for us, it is not difficult to calculate. Furthermore, the error term is where (a.) sampling uncertainty and (b.) variation are "injected" into our summarization process. It is not "injected" into the process by looking at just the mean.

Error is calculated as follows:

"A number representing how confident or sure you want to be" times "a measure of variation from the data"

At this point, please do not think "a number representing how confident or sure you want to be" IS how confident or sure you want to be. For example, if you wanted to be 90% confident, please be very aware that "a number representing how confident or sure you want to be" DOES NOT equal either 90 or 0.9 to represent 90%. The actual number to use requires the use of a statistical distribution (options include the normal, t, F, or χ^2 distributions[2]) with selection somewhat dependent on (a.) the nature of the attribute being estimated (is it an average? is it a proportion or percentage? is it a standard deviation? …) and (b.) certain aspects of the sampling endeavor. So if it is okay with you, I will just call that component "a number representing how confident or sure you want to be" in order to make it more understandable – do we have a deal?

There is one (and only one) property about "a number representing how confident or sure you want to be" that one needs to understand to grasp the main points of this article: and that property is as follows: *for any sampling endeavor, the more confident or sure you want to be, the larger the "number representing how confident or sure you want to be."* That was not so bad was it? What that property implies is that for a given sampling endeavor, if you want to be more sure about your interval (going from 70% to 80% to 90% to 95% sure, for example) the wider the interval must become (the farther apart your arms will go). If you want a specific example, allow me to use the t distribution (typically used when calculating confidence intervals for averages). With a reasonable sample size (say at least 25

[2] Note that "χ^2" is pronounces "kai square", where "kai" rhymes with "sky", and with χ being the Greek letter Chi

observations), the number representing how confident or sure you want to be" goes from roughly 1.0 to 1.3 to 1.7 to 2.0 as one goes from wanting to be 70% to 80% to 90% to 95% confident).

This makes logical sense, if you want to be more sure you have wrapped your arms about the actual answer, one should give oneself more "wiggle room" about an estimate – give oneself a larger error – widen out your arms. If one is willing to be less confident, less "wiggle room" is needed – giving oneself a smaller error – bring your arms in. However, you must realize that if you wish to be 70% confident, you are at the same time giving yourself a 30% chance the true answer (which you will probably never know) is outside of your interval. However, if you want to be 95% sure, you are leaving yourself just a 5% chance the true answer will be outside of that interval. Clearly, then, how sure you want to should play a role in the interval examined and truly should play a HUGE role in planning the sampling endeavor in the first place (more on that later).

Hopefully one now has the first part of the error term under control. Let us now move on to the second component, namely the "measure of variation from the data". The "measure of variation from the data" used is the standard error of the mean (Again, if you are unfamiliar with that term, I ask you read a previous submission in this series namely Doruska [2009] - "Mean and Variation [The Hazards of Examining Just the Average]".) For most applications the standard error of the mean is calculated as follows (and this form is quite useful for my purposes herein) :

$$\text{Standard error of the mean} = \frac{\text{standard deviation}}{\text{square root of sample size}}$$

Let us now examine the impact of both components of that formula on the interval.

The standard deviation is a measure of variation in the same units as the average and note that it appears in the numerator of the standard error of the mean formula. The more variable the data - the larger the standard deviation. As a result, the more variable the data in a sampling endeavor the wider the confidence interval calculated from it will be (the larger the error or the wider your arms will be), everything else being equal. The opposite also holds, the less variable the data, the smaller the standard deviation in that numerator will be (everything else being equal) and the more narrow the confidence interval will be (the lower the error, the closer your arms will be).

Moving on to the denominator, focus on the sample size more than the square root part. Would you agree that the larger the sample, the more you know about what you are sampling and the more sure about the results you should be? I think that is a fair statement. It would seem then the larger the sample size, the more narrow the interval (everything else being equal) should be. That is exactly what happens in the formula because the sample size is in its denominator. As sample size increases, the denominator increases. As the denominator increases, the standard error of the mean (the fraction calculation) decreases, and the confidence interval narrows, everything else being equal.

Wow, it is getting pretty deep, so let me summarize where we are at. Hopefully, we see that

(1.) a confidence interval is a calculation of the form:

Average plus and minus Error

(2.) Error is the half width of the interval and is calculated as:

"A number representing how confident or sure you want to be" times "a measure of variation from the data"

(3.) Both "a number representing how confident or sure you want to be "as well as "a measure of variation" are numbers that go up or down based on the desired level of confidence and/or characteristics about the sampling endeavor (how variable the data are and how many observations were taken)

And most importantly

(4.) All the action is in the Error – the mean is just along for the ride!

If you have followed all of that you basically have the confidence interval concept down pat. Good job. Allow me to introduce one more term here, as this is a good place to introduce it, though it is a term unrelated to our present discussion. That term is what I call percent error (and we will use it later):

$$\text{Percent Error} = \frac{\text{Error}}{\text{Average}} \times 100\%$$

That is, take the Error (the half width, the plus and minus number) from the interval, divide it by the average from the interval, then multiply by 100%. It basically makes the error relative to the average. So a percent error of 25% implies the error from the interval equals 25% of its average…

The last reality about confidence intervals to address is the notion of sampling uncertainty. Typically in any sampling scenario, a random sample of a certain size (number of observations) is taken, with the results (average, measures of variations, hopefully confidence intervals) calculated, summarized, and used when making decisions. But have you given any thought to the reality that if you repeated the sampling endeavor over and over again, you would get a different answer (average, measures of variations, confidence intervals) every single time? There is a comforting thought is it not? The answers you obtain are completely dependent on the random sample you happened to get… That is the reality of sampling uncertainty. If there is any one reason to examine a confidence interval rather than just a mean it is to give yourself a little wiggle room caused by sampling uncertainty (or obtaining the random sample you happened to obtain.)

We saw interpretations of confidence intervals earlier. I will repeat the interpretation of the first example, the MBF per acre example, here again.

"A closer look at the inventory results indicated that in addition to estimating the average of 8.9 MBF/acre, one could be 80% sure (or confident) the interval [7.8 , 10.0] MBF/acre contained the true (but obviously unknown at the time of the inventory) result."

What that “80% confident” implies is that given what you saw in this one sampling endeavor, if you were to repeat this sampling scenario 100 times you should expect about 80 of those sample averages to be in the interval [7.8 , 10.0] MBF/acre, but outside of it 20 times. If you repeated that sampling endeavor 1,000,000 times, you would expect the interval [7.8 , 10.0] MBF/acre to contain about 800,000 of those 1,000,000 averages, and not contain about 200,000 of them. Of course nobody can do this - repeating things 100 or 1,000,000 times, the sampling is done just once. Since the sampling is done just once I ask you one question. Are you more comfortable saying the tract contains 9.7 MBF per acre, or are you more comfortable saying you are 80% sure the interval [7.8 , 10.0] MBF/acre contains the actual MBF/acre in that tract?

When your interval is narrow and the confidence level is high, you probably feel pretty good about your estimate (recall your estimate is the center of the interval), but if your interval has to be fairly wide to have a lot of confidence in (perhaps too wide to even be meaningful to you) then you probably should not feel too good about your estimate and might be thinking – “Dadgummit, what the heck happened here?”. The interval gives you more information about your sampling endeavor.

Take a short break here as we are about to change perspectives – from after the sampling has been done to before the sampling embarks. Are you ready for the change? Okay, let us go.

Every aspect of what has been presented to this point in the article has pertained to “after the fact” or “after the sampling” calculations. But if you are buying into the use and interpretation of confidence intervals, you are probably realizing that you could end up with a confidence interval, *after the sampling has been completed*, that is useless to you. Think back to that silvilcultural treatment example – recall the sample average indicated a 30% increase but one was 80% confident the interval [0% , 60%] contained the actual increase. Are you now surprised that the actual answer was 12.7% ? Hopefully not.

Should the sampling endeavor, “before the fact”, be setup to give you a confidence interval “after the fact” that is meaningful and useful to you? Absolutely! Can this be done? Absolutely! Is this difficult to do? Not at all (but will require algebra – darn those math skills everyone in forestry had to learn at one time or another…). I will work through this using the confidence interval for the average as it is easiest to see how sample size determination works for that version. However, the concepts apply to other measures and more complex sampling designs as well.

Let us say the goal of a sampling endeavor is to estimate basal area (sq.ft.) per acre in a particulate stand. Rather than just getting the average from a random sample from your company’s standard practice of 1 plot for every 10 acres and going with that, you would feel much more comfortable about the decision you are about to make using the results if you were 90% sure about your results and have no more than 15% of error (or have a percent error [remember that new term from before] of no more than 15%). How many observations do you need to pull that off? You might get lucky with the company standard of 1 plot per 10 acres, or you might put in too few plots, or you might put in too many plots and waste time and money. Hmmm…, how does one figure this one out?

WARNING: formulas are about to appear!

Let us reexamine the term percent error from a confidence interval calculated after sampling is completed.

$$\text{Percent Error} = \frac{\text{Error from the "after the fact" interval}}{\text{Average from the "after the fact" interval}} \text{ times } 100\%$$

Now let us substitute (shortened a bit for space reasons) for the numerator of that fraction with information we saw earlier:

$$\text{Percent Error} = \left(\frac{\text{number indicating confidence} \times \text{measure of variation from the data}}{\text{Average from the "after the fact" interval}}\right) \text{ times } 100\%$$

Now let us substitute for the measure of variation from the data used in a confidence interval for a mean (recall this was the standard error of the mean)

$$\text{Percent Error} = \left(\frac{\text{number indicating confidence} \times \frac{\text{standard deviation from the data}}{\text{square root of sample size}}}{\text{Average from the "after the fact" interval}}\right) \text{ times } 100\%$$

Now let us perform a little algebra and we get:

$$\text{Percent Error} = \left(\frac{\text{number indicating confidence} \times \frac{\text{standard deviation from the data}}{\text{Average from the "after the fact" interval}}}{\text{square root of sample size}}\right) \text{ times } 100\%$$

A little more algebra and we get:

$$\text{Percent Error} = \left(\frac{\text{number indicating confidence} \times \left[\frac{\text{standard deviation from the data}}{\text{Average from the "after the fact" interval}} \times 100\%\right]}{\text{square root of sample size}}\right)$$

Are you still with me?

At this point note that the average from the “after the fact” interval is the average from the data, so both the average and the standard deviation in the previous formulation come from the data. Our words/phrasing just got a little crossed up there…

It turns out that another measure of variation one can calculate from data now appears in the formula – namely, the coefficient of variation (or CV). See Doruska (2009) in this publication series for background on that measure.

$$\text{Percent Error} = \frac{\text{number indicating confidence} \times \text{CV from the data}}{\text{square root of sample size}}$$

After some more algebra, we get:

$$\text{Square root of sample size} = \frac{\text{number indicating confidence} \times \text{CV from the data}}{\text{Percent Error}}$$

Then if we square both sides we get:

$$\text{Sample size} = \frac{(\text{number indicating confidence})^2 \times (\text{CV from the data})^2}{(\text{Percent Error})^2}$$

Recall we are trying to determine the sample size for a basal area per acre scenario. We then started with a formula that began with “Percent Error =”, then we did some substitution using what we know about confidence intervals and measures of variation, relied on our algebraic skills, and wound up with a formula that starts out with “Sample size =”. Well guess what, this is the formula that will allow us to estimate the sample size for the basal area per acre sampling endeavor (or any simple or systematic random sampling endeavor targeting an average)! Here’s how.

Remember we wanted to estimate basal area per acre with a percent error of no larger than 15%? No problem, substitute 15 for percent error in the formula.

Remember we wanted to be 90% sure about out answer (90% sure about the interval)? No problem, substitute the correct value for “number indicating confidence” for 90%. (The t distribution is the correct distribution for this example and as we saw before, the number from that distribution for 90% confidence is roughly 1.7). So place 1.7 into the formula where it belongs.

Making those substitutions into the formula yields:

Sample size = $\frac{(1.7)^2 \times (\text{CV from the data})^2}{(15)^2}$

The last step, shy of the actual calculation and getting our answer, is to substitute in the CV from the data. Hmmm…, we need the CV from our data, but remember we are trying to figure out how many observations should be in our data (the sample size). My hunch is you have an image of a dog chasing its tail at this point in your mind (by the way Ernie does that, Ginni does not – see the footnote from earlier!) – and you should. How can we determine the number of observations we should take for a scenario, when we seemingly need to have that data to determine the sample size we need? I like to think of that statement as the statistical version of "How much wood could a woodchuck chuck if a woodchuck could chuck wood"…

That is an Interesting question (the statistical version, more so than the woodchuck version…), with a very important answer. The answer is <u>you cannot determine an appropriate sample size unless you know a little bit about the type of data you are about to collect.</u> Read that again. Now read it one more time. Has it sunk in yet?

Do you buy a car without checking out/researching cars? Do you buy a house without investigating the house? Why the heck do folks conduct sampling endeavors without first investigating what they are about to sample??? Why do folks end up saying "Dadgummit, what the heck happened here?"

All one needs is a rough estimate of the CV of the data one is about to collect – results from a handful of observations will work (often called pilot data), results from a similar stand inventoried last year will work, an educated guess from an experienced forester will work… The only caveat to this process is to be sure the CV estimated or obtained is on the attribute of interest (BA/acre in this case) and was estimated or obtained using the same sampling technique you intend to use (i.e. if you intend to use 10 factor prisms, be sure the estimated CV comes from data collected with 10 factor prisms)

Okay, you now have this estimate of the CV for BA/acre in the stand you are about to enter and the estimate is 44%. The last step is to substitute 44 in for the CV.

The estimated sample size to reach your target then is:

Sample size = $\frac{(1.7)^2 \times (44)^2}{(15)^2}$ = 24.87 which rounds up to a sample size of 25 (Sample sizes are always rounded up, regardless of how small the decimal part is; 24.02 would also round up to 25.)

Does this mean you are guaranteed to get the answer you sought (90% sure about your results and have no more than 15% of error)? Unfortunately no, but it should be pretty darn close, close enough such that you should never have to utter "Dadgummit, what the heck happened here?".

I also call this calculation a reality check. If it takes a ridiculously large sample size to pull off what you want when all is said and done, something has to give (expectations need to change), or another sampling method (stratification perhaps) needs to be used. The reality check is to find this out before you sample rather than after you sample!

A few more thoughts and I promise I will be done …

The sample size formula described herein has three components to it:

(a) How sure you want to be (part of the numerator)
(b) An estimate of how variable the data you are about to collect are (part of the numerator), and
(c) A determination about how wide of an interval you can accept (part of the denominator).

and it was primarily derived from the Error part of the confidence interval we saw before (the average is just along for the ride…)

The more sure you want to be (everything else being equal), the more the numerator increases, the larger the needed sample size – this makes perfect sense. The more variable the data you are about to collect are (everything else being equal), the more the numerator increases, the larger the needed sample size – this makes perfect sense. Lastly, the more narrow you want your "after the fact" interval to be (everything else equal), the smaller the denominator gets, the larger the needed sample size – this makes perfect sense.

Be aware that this not a universal sample size formula, it applies only to simple random sampling and systematic random sampling efforts. Stratified random sampling and double sampling have different formulas, but the concepts examined herein apply to those formulas as well.

Confidence intervals are known as (1-α)% confidence intervals (read "one minus alpha percent"). What that implies for example, is that for a 90% confidence interval, α= 0.10 and for a 95% confidence interval, α= 0.05. If you are familiar with hypothesis testing, the α at which a hypothesis test is performed is basically the same α described here with confidence intervals. Myself, or another contributor, will likely describe the role of α in hypothesis testing in another entry in this series.

Lastly, I hope that I have given you reason to consider using confidence intervals (rather than just the average) and reason to consider planning a sample size rather than just going with what has been used in the past. I was ecstatic (well, as ecstatic as a biometrician like me can get) when I read the November 2009 issue of SAF's Forestry Source – specifically, the Wilent (2009) article entitles "Guidelines help family forest owners enter carbon market". The article details how stratification can be employed to determine the needed sample size in baseline inventories such that "there is 90 percent confidence that the resulting reported value is within 10% of the true mean." That latter part sure sounds familiar now, does it not?

The article also includes an example, similar to the one shown previously, worked by Steve Fairweather (President, Mason, Bruce, & Girard, Inc.) and includes the following quote from Steve: "This is a biometrician talking, but now we finally have somebody who is really insisting on a confidence interval. It's interesting to me that we have no real industry standards for forest inventory – until now". I agree wholeheartedly. After reading this entry into the series, perhaps you do as well.

Literature Cited

Doruska, P.F. 2009. The hazards of examining just the average. Society of American Foresters Forest Science and Technology Board Series on Statistics and Science. Topic No. 7. 11p. Available at: http://207.5.76.244/fp/statistics_and_science.cfm (verified Nov. 30, 2009)

Wilent. S. 2009. Guidelines help family forest owners enter carbon market. Forestry Source 14(11):1 and 4.