Statistics Unit 4

Topic: Interval estimates and Bootstrap sampling

- Date: @March 21, 2024
- Lane: *Statistics*

What it covers

- Review of interval estimation, expressing errors in statistics

Requires

Random variables, expectation, variance,

Required by

Significance testing

Why use sampling to estimate errors?

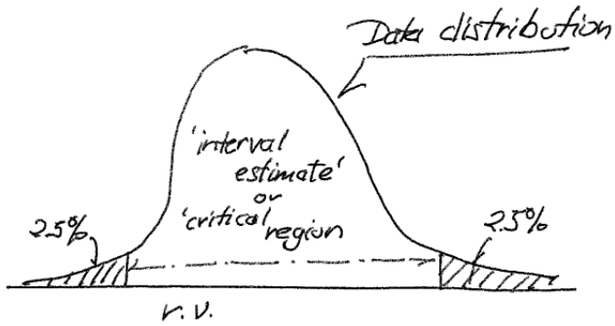
Any inference comes with degree of uncertainty. One way to understand statistics is as techniques to quantify this uncertainty, by measuring the error inherent in the data used to make the inference. Traditionally, when computation was the scarce resource (i.e. before computers) elegant mathematical approximations were developed — taking advantage of the tendency for probability distributions to converge to normal distributions. However much of what was done analytically can now be done computationally, avoiding the assumption of normality, but introducing different concerns when large data and computation come into play.

A brute force way to estimate errors is simply to re-run the experiment and retake the measurement. When data is limitless this is possible and desirable. But until recently this was never the premise in statistics. The most interesting development in the last couple decades are methods that extract the errors in a summary statistic by re-using the same data used to create the statistic. This apparent magic is called “the bootstrap.” First we consider using interval estimates, then consider how “bootstrapping” is used.

Statistics: measuring the error of estimates.

Any uncertain measurement is defined as an *expectation* —- an equivalent certain value that can be used in place of the uncertain measurement — that comes with the nice mathematical properties of expectation. Corresponding to each “ideal” expected value is a invented statistic, or maybe a few, that approximate the desired expected value. A “statistic” can be any function that summarizes the data as a number. There’s a principle, usually referring to the limit that as the number of measurements — the *sample size* — gets large, the statistic approximates the desired value better. The textbook example is the arithmetic average used to estimate the mean value of a measurement’s distribution. The next question in statistics is, then, how good an approximation is any measurement? Estimation of the statistic’s error around the expected value is a property of the statistic. In the simplest case there’s an associated statistic for the erro: For an estimate of the mean, the error estimate is called the *standard error,* or simply, the standard deviation of the mean.

By reducing a the estimation of a measurement to an estimate (the mean) and it’s standard error, one implicitly approximates its distribution — a likelihood of the data — as a normal distribution. It follows that the probability covered by an interval around the mean can be determined as the probability of a normal distribution with that mean and standard deviation over that interval.



The standard convention for how large a probability such an interval should cover is 95% — so there is a 20 to 1 chance of the interval containing the desired measurement.

1. Why summarize uncertainty by intervals?

Without an error estimate, the point estimate of the mean has little value.

2. Why not just model the entire likelihood distribution of P(data | actual value)?

The use of just two parameters leads naturally to the normal distribution assumption

Two interpretations - “frequentist” versus “Bayesian”

The “frequentist” argument: The sampling interval is the basis for *inference,* for making a determination about the adequacy of the estimate. In short if the interval contains the desired value for the measurement, then it is evidence for the measurement having that value. This may sound like a convoluted way of stating the value of evidence, since it avoids making any claims about the probability of the actual measurement value, or about any hypothesis associated with it. Hence the use of the term “confidence interval” and not “probability interval.” This is the basis for all (frequentist) statistical arguments. Whether observed data applies to one “hypothesis” or another reduces to where an interval falls relative to a desired value.

The Bayesian alternative: One may ask, “what about just computing the probability distribution of the desired value, and from that making an inference about the probability of an hypothesis? Wouldn’t that be more to the point? That is exactly what a Bayesian method does, at the cost of radically different assumptions, starting with the presumption that it makes sense to apply probability to an unobserved value. However, when such assumptions can be justified, this makes sense. Nevertheless, a student of statistics has to understand the basis for the alternative frequentist methods. And there are times when both the frequentist “sampling distribution” methods and Bayesian “inverse probability” methods both apply.

“Bootstrap” resampling

By interval estimation we can, for example, create an interval for a point in a data sample. It answers the question, if we were to draw a new point from the same source, where is it likely to fall? What if we want to know about the variation of the sample statistic itself? This “second order” question is about finding the “sampling” variation in the mean and variance statistics. One way this could be answered is if we could repeat the entire experiment many times with fresh data. Instead we can get reliable estimates by re-sampling the existing data.

Bootstrap sampling extends the ability to create estimation intervals for a summary statistic when one must work solely with the data at hand. At first, this appears like magic. The point is that each time we create a new “resample” of the same size as the original data by sampling with replacement, we get a slightly different dataset, and these “resamples’ approximate the underlying distribution from which or statistic was drawn.

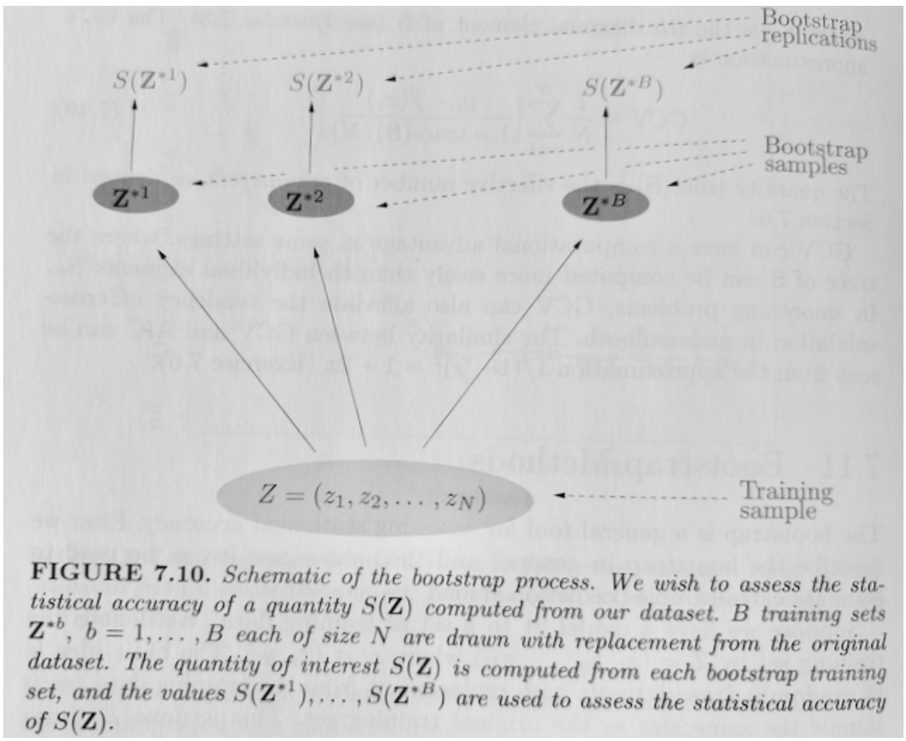


FIGURE 7.10. *Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity $S(\mathbf{Z})$ computed from our dataset. B training sets $\mathbf{Z}^{*b}$, $b = 1, \ldots, B$ each of size N are drawn with replacement from the original dataset. The quantity of interest $S(\mathbf{Z})$ is computed from each bootstrap training set, and the values $S(\mathbf{Z}^{*1}), \ldots, S(\mathbf{Z}^{*B})$ are used to assess the statistical accuracy of $S(\mathbf{Z})$.*

(Illustration from T. Hastie, R Tibshirani, J. Friedman, “The Elements of Statistical Learning” (2001) p. 231

Bootstrapping predictions

How widely can bootstrap be applied? In short, any process that results in a summary value of a set of data can be “bootstrapped” by repeating the process on the bootstrap samples and observing the variation in the results. Regression is a good case. To find the variation in a regression prediction, take the dataset that the regression was learned from and create B resamples, on which one re-learns the regression, each which generates a modified prediction. The distribution of these modified predictions gives us the error estimate we are looking for. This could be applied likewise to any parameter of the regression, such as the regression coefficients.

Bootstrapping gives us a universal tool for estimating error intervals regardless of the statistic or “underlying” distribution to which it is applied. In fact it is a powerful tool that can be used to estimate errors of any process or system that is a function of data, in the fields of optimization, dynamics, or control.

Readings

Each textbook spends a chapter on Linear Regression and related topics. These sections apply to the lectures in the rest of the course:

Evans Chapter 6.4.2

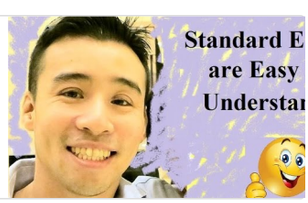
ISLP_python, Ch.5.2

References

Video on Sampling Variation - Good explanation on re-sampling to estimate sample errors

Simplest Explanation of the Standard Errors of Regression Coefficients - Statistics Help
A simple tutorial explaining the standard errors of regression coefficients. This is a step-by-step explanation of the meaning and importance of the standard error.

https://www.youtube.com/watch?v=1oHe1a3JqHw



References

- Gareth James , Daniela Witten , Trevor Hastie, Robert Tibshirani Introduction to Statistical Learning (2023)https://www.statlearning.com/resources-python (”ISLP” 2023)

- Bradley Efron, Trevor Hastie “Computer Age Statistical Inference” (”CASI” 2021) An advanced, readable text addressed to those already conversant in the advances in statistics brought on by computer science and machine learning. See Chapter 10 on the Bootstrap.