Stat 5421 Notes: Review of Baby Statistics

Charles J. Geyer

July 27, 2023

License

What is This?

Chapter 1 of the textbook (Agresti) already reviews bits of the theory of probability and statistics, but IMHO that chapter is not basic enough for some students. Hence this.

Sample, Population, Parameters, and Statistics

The Fundamental Scam of Baby Statistics

In most (all, AFAIK) intro statistics classes, the theory of probability is dumbed down to the theory of simple random sampling from a finite population. This makes much of the discussion of probability theory in such classes wrong or at least wrongheaded or incoherent.

In real probability theory (like that taught in theory classes like Stat 4101–4102 or 5101–5102 at the U of M) probabilities can be any numbers between zero and one, inclusive, not just rational numbers, as arise in finite population sampling. Also in real probability theory we allow infinite sample spaces (sets of possible outcomes of random processes), and that requires calculus for actually doing the math (which is why those theory classes require calculus). But this course requires neither calculus nor those theory courses as a prerequisite. So we will have to handwave some of the mathematical issues. But we won't dumb it down quite as much as baby statistics courses do.

So in statistics we are interested in random "samples" from finite "populations" except not really. This language actually makes no sense when we have infinite sample spaces. It just becomes a bad metaphor when we are "sampling" from an infinite "population".

Independent and Identically Distributed, Probability Distributions

So we replace "sample" and "population" with the notion of *independent and identically distributed* (IID) data. What corresponds to the "population" in the finite-population-sampling picture is a *probability model*, also called a *probability distribution.* You may have already met some of these in previous statistics courses. Those courses certainly used the *binomial distribution*, the *normal distribution*, the *t distribution* and the *F distribution.* But they may have been very sloppy in their discussion (dumbing things down to the point of actually being incorrect).

In this course, we will have no use for the t and F distributions. They only apply when the data are assumed to be normally distributed. But this course is about *discrete data*, also called *categorical data.* Normally distributed data is *continuous* not *discrete.* Hence no t and F. We will introduce some other distributions (chi-square, Poisson, multinomial, product multinomial, multivariate normal), but not on this page. They can wait until we start with Chapter 1 of Agresti.

A sequence of observations is IID if every random variable in the sequence has the same distribution (that's the *identically distributed* part) and the value of any single random variable or any set of random variables has nothing whatsoever to do with any other random variables (separately or collectively, that's the *independent* part). Here and everywhere else in probability and statistics *independent* means *stochastically independent.* We will eventually give a more rigorous definition of IID, but this will have to do for now.

The Sample is Not the Population

When teaching baby statistics, I repeat this section heading as often as possible. This is the most fundamental issue in statistics. Your data are *wrong.* They do not tell you what you want to know — no matter how much effort you put into collecting it.

And Statistics are Not the Parameters They Estimate

In statistics (the subject) we call properties of the probability model *parameters.* Usually the true probability model of the data is *unknown.* That is why we are doing statistics, we are trying to find out about the true unknown probability model (metaphorically, find out about the "population") by computing the analogous properties for a (hopefully) representative sample from the "population". Those quantities that we calculate from the sample (that do not depend on the unknown population parameters) are called *statistics* (singular *statistic*).

Thus we have *statistics* the academic subject whose practitioners are called statisticians. And we have *statistic* a technical term of this academic subject that refers to functions of random data only (that do not depend on unknown parameters).

Statistics of the second kind are used to *estimate* unknown parameters. For example, the sample mean (a *statistic*) is an estimate of the population mean (a *parameter*) and the sample standard deviation (a *statistic*) is an estimate of the population standard deviation (a *parameter*).

The "estimate" is there to remind you that statistics are not the parameters they estimate. The sample mean is not the population mean. The sample mean is a random quantity (because the data are random). The population is not random (unless you are a Bayesian, but we ignore that for now). Hence the sample mean (because it is random) has a very small probability (perhaps zero) of being exactly equal to the population mean.

The most we can hope for is that an estimate is *close* to the parameter it estimates.

Sampling Distributions

The Square Root Law

The *square root law* says that statistical precision varies like the square root of the sample size, more specifically for any estimator $\hat{\theta}_n$ we have

$$\operatorname{sd}(\hat{\theta}_n) \propto \frac{1}{\sqrt{n}}$$

where n is the sample size and the symbol $\propto$ means proportional to.

If we do not use the $\propto$ symbol, we can rewrite this as

$$\operatorname{sd}(\hat{\theta}_n) = \frac{\text{constant}}{\sqrt{n}}$$

where constant is some constant (not random) that we may or may not be able to calculate (and even if we do not know how to calculate it, sometimes the computer does know).

This is not a “law” except in baby statistics. There are required conditions for it to hold, which is another way to say there are exceptions to the rule (cases that do not satisfy the required conditions).

However, the square root law does hold (at least approximately) for a very wide variety of estimators and probability distributions, including most of those that arise in applied statistics and for every probability model we will consider in this course.

Except this is dumbed down to the point of being wrong (see next section).

Asymptotics and a More Precise Square Root Law

Asymptotic Theory, also called Large Sample Theory

It is very common in statistics that, no matter how much theory you know, you do not know the exact sampling distribution of an estimator. This will be true of most of the estimators we discuss in this course.

What we usually (not always) do have is *consistency* and *asymptotic normality* of estimators along with the square root law. Then we say we have a *consistent and asympotically normal* (CAN) estimator. And we write this is math as

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \text{Normal}(0, \tau^2)$$

where

- $\hat{\theta}_n$ is a statistic for sample size n,
- θ is the parameter is is supposed to estimate,
- $\xrightarrow{\mathcal{D}}$ indicates convergence in distribution (a concept of probability theory we will not precisely define, leaving that for theory courses),
- and τ^2 indicates a quantity which is the variance of the asymptotic normal distribution and which typically depends on unknown parameters.

This τ^2 need not have anything to do with the “population” variance (the variance of the true unknown probability distribution of the data). For example, we find out in theory courses that if $\hat{\theta}_n$ is the sample median for sample size n and θ is the “population” median, then

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \text{Normal}\left(0, \frac{1}{4f(\theta)^2}\right)$$

where f is the probability density function of the probability model (which means the this probability model is *continuous*) and we need to assume $f(\theta) > 0$ in order to not get divide by zero and we also need to assume that f is a continuous function (which is not implied by the probability distribution being continuous).

The only point of the example above is to show that the asymptotic variance need not have anything to do with the population variance. Otherwise it is irrelevant to this course and may be forgotten. We will see many other examples where the asymptotic variance has no obvious relationship to population variance, but are not as simple as the median (which is why we did not choose an example more relevent to this course).

Now we come to where the preceding section is dumbed down to the point of being wrong.

The *asymptotic variance* τ^2 is the variance of the *asymptotic distribution* — the distribution that the exact sampling distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ is getting closer and closer to as n goes to infinity.

The actual finite sample size estimator $\hat{\theta}_n$ need not have a variance hence not a standard deviation either.

As an example of this phenomenon, we know from intro stats that the usual estimator $\hat{\pi}_n$ of the usual parameter π of the binomial distribution has

$$\text{sd}(\hat{\pi}_n) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

so this estimator obeys the square root law *exactly* (not just approximately).

But for reasons to be discussed later (many times throughout the course, this being a major theme), we do not want to use this parameter but rather

$$\hat{\theta}_n = \operatorname{logit}(\hat{\pi}_n)$$

where the logit function (pronounced low-jit) is defined by

$$\operatorname{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log(\pi) - \log(1-\pi)$$

except when $\pi = 0$ or $\pi = 1$, where it is undefined. Or we can set

$$\begin{aligned} \operatorname{logit}(0) &= -\infty \\ \operatorname{logit}(1) &= +\infty \end{aligned}$$

because

$$\begin{aligned} \lim_{\pi \searrow 0} \operatorname{logit}(\pi) &= -\infty \\ \lim_{\pi \nearrow 1} \operatorname{logit}(\pi) &= +\infty \end{aligned}$$

And either way this $\hat{\theta}_n$ does not have variance or standard deviation (or has infinite variance and standard deviation, depending on how you want to think about it) *for all finite sample sizes.*

In fact (much more on this later in the course)

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \text{Normal}\left(0, \frac{1}{\pi(1-\pi)}\right)$$

So this is a CAN estimator, but it does not have a finite sample standard deviation. *This is the usual case for the estimators we use in this course.*

Sloppy versus Careful Asymptotics

Usually we will rewrite the first displayed formula of the preceding section as

$$\hat{\theta}_n \approx \text{Normal}\left(\theta, \frac{\tau^2}{n}\right)$$

where ≈ means approximately distributed as or something of the sort.

This is theoretically sloppy, because the right-hand side here cannot be a limit as n goes to infinity (such a limit cannot depend on n; that's the whole point of a limit, to get rid of the dependence on n). One will find in theory courses that one cannot use this sloppy notation when arguments get complicated (too sloppy). But the sloppy notation is good enough for most applied work and most of this course.

This makes sense because if

$$\sqrt{n}(\hat{\theta}_n - \theta)$$

had exactly (not approximately) the

$$\text{Normal}(0, \tau^2)$$

distribution, then $\hat{\theta}_n$ itself would have exactly (not approximately) the

$$\text{Normal}\left(\theta, \frac{\tau^2}{n}\right)$$

distribution by the linearity rule.

The Square Root Law Again

Note that the sloppy version of the asymptotics agrees with the square root law: if the asymptotic variance is τ^2/n, then the asymptotic standard deviation is the square root of that, which is $\tau/\sqrt{n}$. And that's the square root law (asymptotic version).

In this course, no estimators violate the square root law. So if your error of estimation (exact or approximate, theoretical or estimated) is not proportional to $1/\sqrt{n} = n^{-1/2}$, then you are making a mistake.

Summary

If $\hat{\theta}_n$ is as described in the binomial distribution example $\hat{\theta}_n = \operatorname{logit}(\hat{\pi}_n))$, then

- **(wrong)** the approximate or large-sample or asymptotic *variance* of $\hat{\theta}_n$ is $1/(n\pi(1-\pi))$ (this is infinitely wrong)
- **(right)** the approximate or large-sample or asymptotic *distribution* of $\hat{\theta}_n$ is the normal distribution having mean θ and variance $1/(n\pi(1-\pi))$.

Or even shorter **(wrong)** variance is approximated **(right)** the distribution is approximated.

There is No One True Way to Do Frequentist Statistics

In so-called frequentist statistics (which would be better called samplingdistributionist if English made compound words this way — like German does — because it is statistical inference derived from sampling distributions of statistics) there is no one right way to do anything.

- Any statistic you say is an estimator of a parameter θ is one. It may be completely ridiculuous, like the estimator that ignores the data and always says 42. It may be even more than completely ridiculous, like a negative estimate for a parameter that is known to be positive. Of course, some estimators are better than others. So we need theoretical statistics to tell us how to figure that out so we can use good ones rather than bad ones. But this course does not require theory as a prerequisite, so we will just have to tell you what theory says rather than go through proofs. In this course, we will usually use maximum likelihood estimation, which is asymptotically (but not exactly) best possible, meaning that its asymptotic normal distribution has the smallest variance possible for any estimation procedure but that does not tell that maximum likelihood is best at any finite sample size (and we have examples like James-Stein estimators to show that maximum likelihood need not be the best at any finite sample size in certain situations.
- Any interval you say is a confidence interval is one, even if it does not have the stated coverage probability, it has *approximately* the stated coverage probability if very bad approximation is allowed. In this course, because of discreteness, no confidence interval will be exact. So all will only have approximately the stated coverage probability (and the coverage probability will depend on the true unknown parameter value, which is unknown). Again, we know that confidence intervals associated with the maximum likelihood estimator are asymptotically best possible. But this is tricky for discrete data, because regardless of sample size, the approximation gets worse and worse as the parameter values get extreme (mean values go to the boundary of their parameter space), so the confidence intervals work well when the parameters are far from the boundary but not otherwise. More on this later.
- Similarly, for hypothesis tests. Any recipe for confidence intervals, implies a corresponding one for hypothesis tests and vice versa so when you have many confidence interval recipes you also have many hypothesis test recipes.

It would be nice (for you) if we dumbed this down and just said "learn this one" and allowed you to forget the others. But it would not be nice if later on you had to deal with the others in real life. So we will learn about all of the widely used procedures.

There is One True Way to Do Bayesian Statistics

The one true way to do Bayesian statistics is to use Bayes' rule (this may not make much sense if you have not been exposed to Bayesian statistics, but will eventually).

But different Bayesians will produce different inferences: if they use different prior distributions, then they will get different posterior distributions. But, unlike frequentist inference where some procedures are better than others. All of these Bayesian inferences are perfectly correct (assuming Bayes' rule was followed correctly and the prior was defensible). More on this when we get to Bayes.

No One Right Answer

But either way (frequentist or Bayesian, or "other" for that matter, not all statistical inference is either of these, although we will not cover anything else in this course) there will rarely be only one possible correct answer (unless we are very specific in telling you what to do). And there will never be only one possible correct answer in real life (if you ever do statistics in real life). Sorry about that. But that's just the way it is.