Estimation

Kevin Shi

May 2025

Contents

1 Introduction

For this set of notes I will be using this online textbook, as well as chapter 9 of *Intro to Probability and Statistics* by Bertsekas and Tsitsiklis.

2 Point Estimation

The underlying idea of all point estimation is that there is some underlying parameter θ that is an unknown constant, as opposed to the Bayesian framework where θ is random. Generally, we state that an estimator $\hat{\Theta}$ for true parameter θ takes the following form:

$$\hat{\Theta} = g(X) \tag{1}$$

where $X = (X_1, \ldots, X_N)$ are some observations whose distribution depends on θ, and for some function g. Further, I will follow with some important definitions.

Definition 1 (Estimation Error)**.** The *estimation error* of an estimator $\hat{\Theta}$ is defined as $\bar{\Theta} = \hat{\Theta} - \theta$.

Definition 2 (Bias)**.** The *bias* of an estimator $\hat{\Theta}$ is defined as the expected value of the estimation error, or $b_\theta(\hat{\Theta}) = \mathbb{E}_\theta[\bar{\Theta}] = \mathbb{E}_\theta[\hat{\Theta}] - \theta$.

As **unbiased estimator** is an estimator where the bias is 0 for every possible value of θ, or equivalently when the expected value of the estimator is equal to the true value of the parameter.

As a slight extension, we might consider how bias evolves as a function of the number of observations N. In this case, if we are concerned with the number of observations, we can write an estimator as $\hat{\Theta}_N$. With an emphasis placed on number of observations, we then provide more definitions.

Definition 3 (Asymptotically Unbiased)**.** An estimator is *asymptotically unbiased* if $\lim_{N\to\infty} b_\theta(\hat{\Theta}_N) = 0$ for every possible value of θ, or equivalently $\lim_{N\to\infty} \mathbb{E}_\theta[\hat{\Theta}_N] = \theta$.

Definition 4 (Consistency)**.** We call an estimator *consistent* if it the sequence of estimators $\hat{\Theta}_N$ converges to the true value of θ in probability, for every value of θ. Formally, we have: $P(|\hat{\Theta}_N - \theta| \geq \varepsilon) \to 0$ as $n \to \infty$ and for all $\varepsilon > 0$.

We are also interested in the expected estimation error, which is computed using the mean squared error (MSE):

$$\mathbb{E}[\bar{\Theta}^2] = \mathbb{E}[(\hat{\Theta} - \theta)^2] = \mathbb{E}[\hat{\Theta}^2] - 2\theta\mathbb{E}[\hat{\Theta}] + \theta^2 \tag{2}$$

$$= (\mathrm{Var}(\hat{\Theta}) + \mathbb{E}[\hat{\Theta}]^2) - 2\theta\mathbb{E}[\hat{\Theta}] + \theta^2 \tag{3}$$

$$= \mathrm{Var}(\hat{\Theta}) + (\mathbb{E}[\hat{\Theta}] - \theta)^2 = \mathrm{Var}(\hat{\Theta}) + b(\hat{\Theta})^2 \tag{4}$$

Where I have omitted the subscript θ as it is pretty clear we are taking the variance and expectation with respect to it. From this derivation we can see what is often referred to as the bias-variance tradeoff, but this is somewhat of a misconception! From this expression it's not like there *needs* to be a tradeoff between bias and variance, merely that MSE rises whenever bias or variance increase. In fact, it is possible to have models with 0 bias and 0 variance. This is mainly an empirical result observed in parameterized models (such as regression) that observes that fewer parameters tend to underfit (increased bias) while higher amounts of parameters tend to overfit to the data/noise (increased variance). However, this has been challenged by recent literature, such as in Neal (2019), which shows it does not hold universally. Still, it is a good heuristic to have, even if it must be qualified. I would like to dive deeper into Neal's paper at some later date.

2.1 Maximum Likelihood

Next I will take some notes on maximum likelihood. Essentially this framework is set up as follows:

- Assume we are given some data $\mathbf{X} = [X_1, \ldots, X_N]^T$.
- We make the modeling assumption that each X depends on some parameters θ.
- We have a conditional PMF $p(\mathbf{X}|\theta)$ or PDF $f(\mathbf{X}|\theta)$. This is what we call our likelihood function.
- Naturally, we try to maximize the conditional over the parameters, or in other words: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}(p(\mathbf{X}|\theta))$

It is important to note that the likelihood function is actually a function in θ, *not* in $\mathbf{X}$. The data is known to us and given, while it is our task to determine the correct setting of parameters. Therefore this likelihood is, in general, not true density or mass function, as marginalizing over the deterministic parameters θ does not necessarily sum up to 1.

In the setting where we have iid data samples, we can factor the likelihood to make computations easier.

As an example, let's try to estimate the parameter θ of a Bernoulli random variable. We are given a dataset with N elements, $\mathbf{X} = [X_1, \dots, X_N]^T$ of coin flips. Our likelihood function is: $p(\mathbf{X}|\theta)$. Then:

$$p(\mathbf{X}|\theta) = \prod_{i=1}^{N} p(X_i|\theta) \tag{5}$$

$$= \theta^h (1-\theta)^{N-h} \tag{6}$$

Where h is the number of heads (or successes) observed. We then take the derivative of this log-likelihood wrt θ and set to 0, giving:

$$h\theta^{h-1}(1-\theta)^{N-h} - (N-h)\theta^h(1-\theta)^{N-h-1} = 0 \tag{7}$$

$$\rightarrow h\theta^{h-1}(1-\theta)^{N-h} = (N-h)\theta^h(1-\theta)^{N-h-1} \tag{8}$$

$$\rightarrow h(1-\theta) = (N-h)\theta \tag{9}$$

$$\rightarrow h = \theta N \rightarrow \hat{\theta} = \frac{h}{N} \tag{10}$$

Which agrees with our intuition that the probability of a success of a Bernoulli trial is best estimated by the number of successes observed divided by the total number of trials.

2.2 Estimators for Mean and variance

A natural estimator for the mean μ of a distribution (with variance σ^2) is simply the sample mean of data drawn from said distribution, $\hat{\Theta} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. We can verify this is an unbiased estimator:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[X_i\right] = \mu \tag{11}$$

As the estimator is unbiased, it is also trivially asymptotically unbiased. To show consistency, we use Chebyshev's inequality:

$$P(|\hat{\Theta} - \theta| \geq \varepsilon) = P(|\hat{\Theta} - \mathbb{E}[\hat{\Theta}]| \geq \varepsilon) \leq \frac{\text{Var}(\hat{\Theta})}{\varepsilon^2} \tag{12}$$

$$= \frac{1}{\varepsilon^2}\text{Var}\left(\frac{1}{n}\left(\sum_{i=1}^{N} X_i\right)\right) = \frac{\sigma^2}{n\varepsilon^2} \tag{13}$$

Which verifies the consistency of the estimator $\hat{\Theta}$, as the upper bound on the probability goes to 0 as the number of data points increases. This is also an application of the weak law of large numbers.

We might think again to estimate the variance as $\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i])^2$. However, we don't actually know $\mathbb{E}[X_i]$, and so we might use the sample mean instead, giving us

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2 &= \frac{1}{N}\sum_{i=1}^{N}(X_i^2-2X_i\bar{X}+\bar{X}^2) && (14)\\
&= \frac{1}{N}\sum_{i=1}^{N}X_i^2+\frac{1}{N}\sum_{i=1}^{N}\bar{X}^2-\frac{2\bar{X}}{N}\sum_{i=1}^{N}X_i && (15)\\
&= \frac{1}{N}\sum_{i=1}^{N}X_i^2+\bar{X}^2-2\bar{X}^2 && (16)\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}X_i^2\right)-\bar{X}^2=\frac{1}{N}\left(\sum_{i=1}^{N}X_i^2-N\bar{X}^2\right) && (17)
\end{aligned}$$

Taking the expected value of this estimator, and by the linearity of expectations, we get:

$$\begin{aligned}
\mathbb{E}[\hat{\Theta}] &= \frac{1}{N}\left(\sum_{i=1}^{N}\mathbb{E}[X_i^2]-N\mathbb{E}[\bar{X}^2]\right) && (18)\\
&= \frac{1}{N}\left(\sum_{i=1}^{N}(\sigma^2+\mu^2)-N(\mathrm{Var}(\bar{X})+\mu^2)\right) && (19)\\
&= \frac{1}{N}\left(N\sigma^2-N\frac{\sigma^2}{N}\right) && (20)\\
&= \frac{1}{N}\left(N\sigma^2-\sigma^2\right)=\frac{N-1}{N}\sigma^2 && (21)
\end{aligned}$$

And we can see that this estimator is biased. However, we can clearly see that it is asymptotically unbiased, because as $N\to\infty$, $\frac{N-1}{N}\sigma^2\to\sigma^2$. In order to get an unbiased estimator, we simply multiply our original estimator by $\frac{N}{N-1}$. The unbiased variance estimator is also easily shown to be consistent by Chebyshev's inequality.

For the unbiased estimator, we can (less trivially) show that the estimator is also consistent by showing that it converges to $\frac{N-1}{N}\sigma^2$ in probability, and then recognizing that $\frac{N-1}{N}\sigma^2$ itself converges to the true parameter σ^2. I'm too lazy to show it here, but the math is clearly explained here: Variance of Sample Variance.

Note that under the assumption that the data is normally distributed, we get equation (11) as our mean MLE estimator and equation (14) as our variance MLE estimator.

3 Interval Estimation/Confidence Intervals

One of the aims in probability and statistics is quantifying uncertainty. Regardless of however good an point estimator is, it is nonetheless a single quantity which we cannot model probabilistically. Sure, we might be able to compute certain statistics, but simply computing the mean or variance of an estimator doesn't really tell us with what probability we've captured the right value. To do this from a frequentist perspective, we introduce the confidence interval.

In essence, for a random sample from a probability distribution, we can estimate the true parameter θ by using two estimators, $\hat{\Theta}_\ell$ for the lower estimator, and $\hat{\Theta}_h$ for the higher estimator. These are calculated such that $P(\hat{\Theta}_\ell\le\theta\le\hat{\Theta}_h)\ge 1-\alpha$, where $1-\alpha$ is the confidence level. 0.95 is a common value, but that seems more like inertia in some fields that don't really do stats, and merely use the tools. In truth, it's pretty arbitrary.

As a note, the randomness does not come from θ, as it is modeled as a deterministic quantity. Instead, it comes from the two estimators $\hat{\Theta}_\ell$ and $\hat{\Theta}_h$. The interval itself is the random quantity in question. To interpret the interval, it is fundamentally tied to multiple trials. Because the random quantity is not the parameter itself, but rather the interval itself, we are instead saying that if we constructed an interval repeatedly by sampling from a population, we would expect about $(1-\alpha)100\%$ of the intervals to contain the true parameter.

To actually find this confidence interval, there are some things to consider. We can rewrite the probability as $P(|\hat{\Theta}-\theta| \leq t) \geq \alpha$, where we let $\hat{\Theta}_\ell = \hat{\Theta} - t$, and similarly for the larger estimator. In the cases where the distribution of $\hat{\Theta}-\theta$ is unknown, we are shit out of luck, which can happen when our estimator doesn't converge to some distribution or if the distribution depends on the unknown θ itself. However, in many cases, $\hat{\Theta}-\theta$ is both asymptotically unbiased and also asymptotically normal, or in other words the random variable:

$$\frac{\hat{\Theta}-\theta}{\sqrt{\mathrm{Var}(\hat{\Theta})}} \tag{22}$$

approaches the standard normal as the number of samples N increases, for every value of θ. The estimator is a random variable, and as $N \to \infty$ approaches a normal distribution. Further, by saying it is asymptotically unbiased, the mean approaches θ in the limit, meaning that the new random variable is properly standardized.

Then, if we can calculate $\mathrm{Var}(\hat{\Theta})$, either exactly or approximately, we can move forward with constructing confidence intervals easily.

3.1 Interval Estimation with Unknown Mean and variance

Technically in most real life setting we don't know the underlying mean or variance, but we often assume iid properties. As such, say that we wanted to estimate the mean with the sample mean. As we established before, we need to compute the sample variance as well. As we don't have the underlying variance, this will also need to be estimated (approximation), which we will do with our unbiased variance estimator. As a refresher, these are:

$$\hat{\Theta}_\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$
$$\hat{\Theta}_{\sigma^2} = \frac{N}{N-1}\sum_{i=1}^{N}(X_i - \hat{\Theta}_\mu)^2$$

Then as the sample mean is asymptotically normal by the CLT, we normalize it as

$$\frac{\hat{\Theta}-\theta}{\sqrt{\sigma^2 n^{-1}}}$$

As the variance of the sample mean $\frac{\sigma^2}{n}$. which ultimately gives us:

$$P\left(|\hat{\Theta}_\mu - \theta| \leq t\right) \geq 1-\alpha \tag{23}$$

$$\to P\left(-t \leq \hat{\Theta}_\mu - \theta \leq t\right) \geq 1-\alpha \tag{24}$$

$$\to P\left(-t \leq \frac{\hat{\Theta}_\mu - \theta}{\sqrt{\hat{\Theta}_{\sigma^2}/n}} \leq t\right) \geq 1-\alpha \tag{25}$$

Here, having standardized the estimator for the mean, we can go to our standard normal table for any particular setting of $1-\alpha$. It's easier to calculate the standard score if we try to look at the complement of this event, so let's rename the standardized normal as Z. Then by De Morgan's rules, we get that

$$P(Z < -t) + P(Z > t) \leq \alpha \tag{26}$$

By symmetry of the standard normal, these probabilities are exactly equal, which means that the property we need to satisfy is simple $P(Z > t) \leq \frac{\alpha}{2}$. We can once again take the complement of this event, giving us finally that:

$$P(Z < t) \geq 1 - \frac{\alpha}{2} \tag{27}$$

as the property we want to satisfy. Now we can look at the standard normal table (given some confidence interval $1-\alpha$ we want), and look at the value $1-\frac{\alpha}{2}$ as the cumulative probability we want instead.

After finding the proper value of t, we can then compute the proper bounds of our interval:

$$\to P\left(-\hat{\Theta}_\mu - t\sqrt{\frac{\hat{\Theta}_{\sigma^2}}{n}} \leq -\theta \leq -\hat{\Theta}_\mu + t\sqrt{\frac{\hat{\Theta}_{\sigma^2}}{n}}\right) \geq 1-\alpha \tag{28}$$

$$\to P\left(\hat{\Theta}_\mu - t\sqrt{\frac{\hat{\Theta}_{\sigma^2}}{n}} \leq \theta \leq \hat{\Theta}_\mu + t\sqrt{\frac{\hat{\Theta}_{\sigma^2}}{n}}\right) \geq 1-\alpha \tag{29}$$

which gives us the closed interval of lower and higher estimation bounds for our confidence interval. Essentially, for this setup, as long as the lower bound and upper bound are less and more than the calculated range, the confidence interval is at least $(1-\alpha)100\%$.

3.2 Student's t-distribution

The slight issue is that there are two assumptions when constructing this interval; firstly, we are assuming that the sample mean is actually normal, which, irrespective of if the variance is known or not, is always just an approximation. Secondly, as the variance of the underlying distribution is unknown, we are estimating it, which means the estimated variance is not exactly the variance of even the sample mean.

The first condition is just an unfortunate reality, but in any case, we at least have that confidence bounds get more accurate with increasing sample size. The Edgeworth Series and bootstrapping can get more accurate estimates, but they are much more involved. For the second condition, for small samples, the variance estimator can be somewhat inaccurate, which motivates the **Student's t-distribution with $n-1$ degrees of freedom**.

This distribution is based on the assumption of normality of the data, but can still be a good approximation if the the underlying distribution is symmetric and has finite variance. If the distribution is sub-Gaussian (ie the tail decays at at least an exponential rate), then it is even more accurate.

Thankfully, there is also a t-distribution table that we can look up, and essentially we calculate the value of t based on the $n-1$ t-distribution CDF table instead of the standard normal table, with $n-1$ being determined by the number of samples.

Once the number of samples is large enough (by convention usually around $n \geq 50$), we can resume using the standard normal tables (as we are "close enough" to the population variance), and we call this the **z-score**. Using the t-distribution (when we have a small number of samples) gives us the **t-score**.

I would like to come back and do more research on the derivation of the t-distribution. Tt is fundamentally based on the Cochran's theorem and the Chi-square distribution with $n-1$ degrees of freedom, which comes about from the sum of n Gaussian random variables (giving $n-1$ degrees of freedom, hence the name). Wikipedia has some nice details on all of these topics.

3.3 Small Extension

The estimator for the variance need not be the exact unbiased formulation we have already explored. Depending on the underlying distribution, we might be able to compute different unbiased estimators. An example in *Bertsekas and Tsitsiklis* is that of a Bernoulli random variable, which takes unknown mean θ and variance $\theta(1-\theta)$. We estimate θ using the sample mean $\hat{\Theta}$, which is unbiased, and the estimate the variance as $\hat{\Theta}(1-\hat{\Theta})$. Then for this estimate, we have:

$$\begin{aligned}\mathbb{E}[\hat{\Theta}(1-\hat{\Theta})] &= \mathbb{E}[\hat{\Theta}] - \mathbb{E}[\hat{\Theta}^2] && (30)\\ &= \theta - (\mathrm{Var}(\hat{\Theta}) + \theta^2) = \theta(1-\theta) - \mathrm{Var}(\hat{\Theta}) && (31)\\ &= \theta(1-\theta) - \mathrm{Var}\left(\frac{X_1 + \ldots + X_N}{N}\right) && (32)\\ &= \theta(1-\theta) - \frac{\theta(1-\theta)}{N} && (33)\end{aligned}$$

This estimator is not unbiased, but as $N \to \infty$, the term $\frac{\theta(1-\theta)}{N} \to 0$, and so the estimator is asymptotically unbiased. We could also be conservative, and note that the variance is maximized at $\theta = \frac{1}{2}$, meaning the variance is always less than $\frac{1}{4}$, and simply use this loose upper bound as the variance estimator.

4 Conclusion

In the next set of notes, I will go on to discuss hypothesis testing, which follows naturally from confidence intervals.

References

[Neal 2019] Neal, Brady: *On the Bias-Variance Tradeoff: Textbooks Need an Update.* 2019. – URL https://arxiv.org/abs/1912.08286