36-700: Probability and Mathematical Statistics **Spring 2019**

Lecture 8: Point Estimation

Lecturer: Jing Lei

8.1 Review and Outline

Last class we discussed:

- WLLN
- CLT
- Delta Method

This class we start discussing statistical estimation formally.

8.2 Statistical Models

The central preoccupation of statistics (and machine learning) is to understand/estimate things about some underlying population on the basis of samples. Formally, the typical setup is given:

$$X_1, \dots, X_n \sim F,$$

what can we infer about F?

In order to make meaningful inferences about F from a finite number of sample points we typically restrict F in some natural way. In this case, we will denote by $\mathcal{F}$ the set of possible distributions F. This is called the **statistical model**. Broadly, there are two categories:

1. **Parametric model:** In a parametric model, the set of possible distributions $\mathcal{F}$ can be described by a finite number of parameters. Here are a few examples:

   (a) A Gaussian model: This is a simple two parameter model. Here we suppose that:

$$\mathcal{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

  (b) The Bernoulli model: This is a one parameter model where:

$$\mathcal{F} = \{P(X = 1) = p, P(X = 0) = 1 - p, 0 \leq p \leq 1\}.$$

2. **Non-parametric model:** A non-parametric model is one which where $\mathcal{F}$ cannot be parameterized by a finite number of parameters. Here are a few popular examples:

   (a) Estimating the CDF: Here the model consists of any valid CDF, i.e. a function that is between 0 and 1, is monotonically increasing, right-continuous and equal to 0 at $-\infty$ and 1 at ∞. We are given samples $X_1, \ldots, X_n \sim F$ and the goal is to estimate F.

   (b) Density estimation: In density estimation, we are given samples $X_1, \ldots, X_n \sim f_X$, where f_X is an unknown density that we would like to estimate. It turns out that the class of all possible densities is too big for this problem to be well posed so we need to assume some smoothness on the density. A typical assumption is that the model is given by:

$$\mathcal{F} = \left\{ f : \int (f''(x))^2 dx \leq C < \infty, \int f(x) dx = 1, f(x) \geq 0 \right\}.$$

8.3 Point Estimation

Point estimation in statistics refers to calculating a single "best guess" of the value of an unknown quantity of interest. The quantity of interest could be a parameter or for instance a density function.

Typically, we will use θ to denote the unknown parameter of interest, and $\hat{\theta}$ or $\hat{\theta}_n$ to denote a point estimator. A point estimator is a function of the data $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n),$$

so that $\hat{\theta}_n$ is a random variable.

The bias of an estimator is written as:

$$b(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta.$$

Similarly, the variance of an estimator is given by:

$$v(\hat{\theta}_n) = \text{Var}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2.$$

In classical statistics, often the starting point was to identify *unbiased* estimators, and then find unbiased estimators with small (or minimal variance). In modern statistics, we often use biased estimators because the reduction in variance often justifies the bias.

We call an estimator $\hat{\theta}_n$ of a parameter θ *consistent* if $\hat{\theta}_n \xrightarrow{P} \theta$, i.e. for any ϵ:

$$\mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) \to 0,$$

as $n \to \infty$.

Remark: In this framework, θ is typically a function of F (i.e., $\theta = \theta(F)$). The $\mathbb{E}(\cdot)$ and $\mathbb{P}(\cdot)$ used above refer to the randomness in the data $X_1, ..., X_n$ iid from F. In the following, we may use $\mathbb{E}_\theta(\cdot)$ and $\mathbb{P}_\theta(\cdot)$ to emphasize that the underlying distribution is the one associated with θ.

8.4 The Bias-Variance decomposition

One way to compute the quality of an estimator is via its mean squared error:

$$\text{MSE} = \mathbb{E}_\theta(\theta - \hat{\theta}_n)^2.$$

The MSE can be decomposed as the sum of the squared bias and variance, i.e.:

$$\begin{aligned}\text{MSE} &= \mathbb{E}_\theta(\theta - \hat{\theta}_n)^2 \\ &= \mathbb{E}_\theta(\theta - \mathbb{E}_\theta(\hat{\theta}_n) + \mathbb{E}_\theta(\hat{\theta}_n) - \hat{\theta}_n)^2 \\ &= b(\hat{\theta}_n)^2 + v(\hat{\theta}_n).\end{aligned}$$

A simple consequence of this decomposition is that: if $b(\hat{\theta}_n) \to 0$, and $v(\hat{\theta}_n) \to 0$ then the estimator $\hat{\theta}_n$ is consistent. This is because if both bias and variance tend to 0 then we have convergence in quadratic mean which in turn implies convergence in probability.

Example: Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$, and our estimator:

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

What is the bias of this estimator? What is its variance? Is the estimator consistent?

8.5 Finding estimators

Point estimation is aligned with the general statistical pursuit of trying to understand an underlying population from a sample. Particularly, under the hypothesis that the sample was generated from some parametric statistical model, a natural way to understand the underlying population is by estimating the parameters of the statistical model.

One can of course wonder what happens if the model is wrong? Or where do models really come from? In some rare cases, we actually know enough about our data to hypothesise a reasonable model. Most often however, when we specify a model, we do so hoping that it can provide a useful approximation to the data generation mechanism. The George Box quote is worth remembering in this context: "all models are wrong, but some are useful."

8.5.1 Method of moments

We saw this idea before we will see a couple of other examples today. Broadly, the idea is that we can estimate the sample moments in a straightforward way:

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^n X_i, \\ m_2 &= \frac{1}{n}\sum_{i=1}^n X_i^2, \\ &\vdots \\ m_k &= \frac{1}{n}\sum_{i=1}^n X_i^k. \end{aligned}$$

On the other hand the population moments are some functions of the unknown parameters. We can denote the population moments as: $\mu_1(\theta_1, \ldots, \theta_k), \ldots, \mu_k(\theta_1, \ldots, \theta_k)$.

The method-of-moments prescribes estimating the parameters: $\theta_1, \ldots, \theta_k$ by solving the system of equations:

$$\begin{aligned} m_1 &= \mu_1(\theta_1, \ldots, \theta_k) \\ &\vdots \\ m_k &= \mu_k(\theta_1, \ldots, \theta_k). \end{aligned}$$

We already saw an application of this idea to estimating the mean and variance of a Gaussian:

Example 1: If $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, we would solve:

$$\frac{1}{n}\sum_{i=1}^n X_i = \theta, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = \theta^2 + \sigma^2,$$

to obtain the estimators:

$$\begin{aligned}\hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2,\end{aligned}$$

which of course are natural estimators of these quantities. You can see that the method-of-moments estimators can be biased since in this example the estimator for the variance has non-zero bias.

Example 2: Suppose $X_1, \ldots, X_n \sim \text{Bin}(k, p)$, i.e. each X_i is the sum of k $\text{Ber}(p)$ RVs, and that both k and p are unknown. Casella and Berger point out that this model has been used to model crimes, where usually neither the total number of crimes k nor the reporting rate p are known.

Let us first compute the first two moments of the binomial:

$$\begin{aligned}\mu_1 &= kp \\ \mu_2 &= kp(1-p) + k^2p^2.\end{aligned}$$

Denoting,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

we solve the system of equations:

$$\begin{aligned}\bar{X} &= kp \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= kp(1-p) + k^2p^2.\end{aligned}$$

This yields the estimators:

$$\begin{aligned}\hat{p} &= \frac{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}{\bar{X}}, \\ \hat{k} &= \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.\end{aligned}$$

It is worth pointing out that in this case there are no a-priori "obvious" estimators for the parameters of interest. Further, the method-of-moments estimators in this case can be negative (while the true parameters cannot be). However, when the method-of-moments estimator is negative it is because the sample mean is dominated by the sample variance - which is a sign of excessive variability.

8.5.2 Maximum Likelihood Estimation

The most popular technique to derive estimators is via the principle of maximum likelihood. Suppose that $X_1, \ldots, X_n \sim f_\theta$ where f_θ denotes either the pmf or pdf of the population.

The *likelihood* function is defined by:

$$L(\theta | X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i).$$

A maximum likelihood estimator of θ based on the sample $\{X_1, \ldots, X_n\}$ is any value of the parameter at which the function $L(\theta | X_1, \ldots, X_n)$ is maximized (as a function of θ with $X_1, \ldots, X_n$ held fixed).

Unlike the method-of-moments estimator we can see that the MLE will always obey natural parameter restrictions, i.e. the range of the MLE estimator is the same as the range of the parameter. Intuitively, the MLE is just the value of the parameter which is most likely to have generated the sample and thus is a natural choice as a point estimate. It also has many optimality properties in certain settings: indeed, much of the early theoretical development in statistics revolved around showing the optimality of MLE in various senses.

There are some common difficulties with using the MLE: the first is computational, how do we find the global maximum of the likelihood function and how do we verify/certify that it is in fact the global maximum? The next two are related to numerical sensitivity: (1) we can often only approximately maximize the likelihood (using a numerical procedure like gradient descent) so we need that the likelihood is a well-behaved function of the parameters. (2) Similarly, it is also natural to hope that the estimator is a well-behaved function of the data, i.e., we would hope that a slightly different sample would not result in a vastly different MLE.

The typical way to compute the MLE (suppose that we have k unknown parameters) is to either analytically or numerically solve the system of equations:

$$\frac{\partial}{\partial \theta_i} L(\theta | X_1, \ldots, X_n) = 0 \qquad i = 1, \ldots, k.$$

Example 1: Suppose $X_1, \ldots, X_n \sim N(\theta, 1)$, then the likelihood function is given as:

$$L(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-(X_i - \theta)^2/2).$$

A frequently useful simplification is to observe that if θ maximizes $L(\theta|X_1, \ldots, X_n)$ then θ also maximizes $\log L(\theta|X_1, \ldots, X_n)$. We can also of course ignore constants.

So we have that:

$$\hat{\theta} = \arg\max_{\theta} \log L(\theta|X_1, \ldots, X_n) = \arg\min_{\theta} \sum_{i=1}^{n} (X_i - \theta)^2.$$

So that we obtain:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

We need to further verify that the second derivative of the log-likelihood is negative: in this case the second derivative is $-2n$.

Example 2: Suppose that $X_1, \ldots, X_n \sim \text{Ber}(p)$, then the log-likelihood is given by

$$\begin{aligned} \log L(p|X_1, \ldots, X_n) &\propto \sum_{i=1}^{n} X_i \log p + (1 - X_i) \log(1 - p) \\ &= n\bar{X} \log p + n(1 - \bar{X}) \log(1 - p), \end{aligned}$$

which is maximized at

$$\hat{p} = \bar{X}.$$

8.5.3 Invariance of the MLE

Roughly, the invariance property of the MLE states: suppose that the MLE for a parameter θ is given by $\hat{\theta}$ then the MLE for a parameter $\tau(\theta)$ is $\tau(\hat{\theta})$.

Let us denote the likelihood as a function of the transformed parameter by $L^*(\eta|X_1, \ldots, X_n)$. For invariance we need to consider two cases:

1. τ is an invertible transformation: In this case there is a mapping from the likelihood of a point η to a point $\theta = \tau^{-1}(\eta)$, i.e.

$$L^*(\eta|X_1, \ldots, X_n) = L(\tau^{-1}(\eta)|X_1, \ldots, X_n),$$

and we have that the MLE for η:

$$\begin{aligned}\hat{\eta} =&\arg\sup_{\eta} L^*(\eta|X_1,\ldots,X_n) = \arg\sup_{\eta} L(\tau^{-1}(\eta)|X_1,\ldots,X_n) \\ =&\tau(\arg\sup_{\theta} L(\theta|X_1,\ldots,X_n)) = \tau(\hat{\theta}).\end{aligned}$$

2. When τ is not invertible: In this case, it is possible that several parameters $\theta_1, \theta_2, \ldots$ map to the same value η, so we need some care in defining the induced likelihood.

   Particularly, if we define:

   $$L^*(\eta|X_1,\ldots,X_n) = \sup_{\theta:\tau(\theta)=\eta} L(\theta|X_1,\ldots,X_n),$$

   then one can verify that the MLE is invariant.

We will discuss other properties of the MLE in future lectures.

8.6 Bayes Estimators

The third general method to derive estimators is a bit different philosophically from the first two. At a high-level we need to understand the Bayesian approach (we will of course stay clear of any philosophical questions): in our approaches so far (so-called "frequentist" approaches) we assumed that there was a fixed but unknown true parameter, and that we observed samples drawn i.i.d from the population (whose density/mass function/distribution was indexed by the unknown parameter).

In the Bayesian approach we consider the parameter θ to be random. We have a *prior* belief of the distribution of the parameter, which we update after seeing the samples $X_1, \ldots, X_n$. We update our belief using Bayes rule and the updated distribution is known as the *posterior* distribution.

We denote the prior distribution by $\pi(\theta)$, and the posterior distribution as $\pi(\theta|X_1,\ldots,X_n)$. Using Bayes' rule we have that:

$$\pi(\theta|X_1,\ldots,X_n) = \frac{\pi(\theta)L(\theta|X_1,\ldots,X_n)}{\int_\theta \pi(\theta)L(\theta|X_1,\ldots,X_n)}.$$

The posterior distribution is a distribution over the possible parameter values. In this lecture we are focusing on point estimation so one common candidate is the posterior mean (i.e. the expected value of the posterior distribution).

Ignoring any philosophical questions, one can view this methodology as a way to generate candidate point estimators, by specifying reasonable prior distributions. In practice however this calculation is hard to do analytically, so we often end up specifying priors out of convenience rather than any real prior belief. Particularly, a convenient choice of prior distribution is one for which the posterior distribution belongs to the same family as the prior distribution: such priors are called *conjugate priors.*

Example 1: Binomial Bayes estimator: Suppose $X_1, \dots, X_n \sim \text{Ber}(p)$. We will first need to define the Beta distribution: a RV has a Beta distribution with parameters α and β if its density on $[0, 1]$ is

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

For us it will be sufficient to ignore the normalizing part and just remember that the Beta density:

$$f(p) \propto p^{\alpha-1}(1-p)^{\beta-1}.$$

The mean of the Beta distribution is: $\alpha/(\alpha+\beta)$.

Let us denote

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

There are two candidate priors we will consider:

1. The flat/uninformative prior: $\pi(p) = 1$ for $0 \le p \le 1$. In this case, the posterior density is:

$$\begin{aligned} f(p|X_1, \dots, X_n) &\propto p^{n\bar{X}}(1-p)^{n(1-\bar{X})} \\ &= p^{(n\bar{X}+1)-1}(1-p)^{(n(1-\bar{X})+1)-1}. \end{aligned}$$

   This is just a $\text{Beta}(n\bar{X}+1, n(1-\bar{X})+1)$ distribution. So our estimate (the posterior mean) would be:

$$\begin{aligned} \hat{p} &= \frac{n\bar{X}+1}{n+2} \\ &= \frac{n}{n+2}\bar{X} + \frac{2}{n+2}\frac{1}{2} \\ &= w\bar{X} + (1-w)\frac{1}{2}, \end{aligned}$$

   which can be viewed as a convex combination of the MLE and the prior mean 1/2.

2. The other common prior is the one that is conjugate to the bernoulli likelihood, i.e. the Beta prior. A similar calculation as the one above will show that if use $\pi(p) \sim \text{Beta}(\alpha, \beta)$, then the posterior distribution will also be a Beta distribution:

$$f(p|X_1, \dots, X_n) \sim \text{Beta}(\alpha + n\bar{X}, \beta + n(1-\bar{X})),$$

and our Bayes estimator would be:

$$\hat{p} = \frac{\alpha + n\bar{X}}{\alpha + \beta + n}.$$

Example 2: Gaussian Bayes estimator: Here we suppose that our prior belief is that the parameter has distribution $N(\mu, \tau^2)$ and we observe $X_1, \ldots, X_n$ drawn from $N(\theta, \sigma^2)$. We assume that σ^2, μ, τ^2 are all known.

A fairly involved calculation in this case (see for instance Problem 1 of Chapter 11 of Wasserman) will show that the posterior distribution is also a Gaussian with parameters:

$$\begin{aligned} \hat{\mu} &= \frac{n\tau^2}{\sigma^2 + n\tau^2} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) + \frac{\sigma^2}{\sigma^2 + n\tau^2} (\mu) \\ \hat{\sigma}^2 &= \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2}, \end{aligned}$$

which suggests the point estimate:

$$\hat{\mu} = \frac{n\tau^2}{\sigma^2 + n\tau^2} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) + \frac{\sigma^2}{\sigma^2 + n\tau^2} (\mu).$$

Which can once again be seen as a convex combination of our prior belief and the MLE, i.e.:

$$\hat{\mu} = w \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) + (1 - w)\mu,$$

for some $0 \leq w \leq 1$.