**36-700: Probability and Mathematical Statistics I** **Fall 2016**

# Lecture 8: September 16

*Lecturer: Siva Balakrishnan*

## 8.1 Review and Outline

Last class we discussed:

- Interpreting moments
- Other notions of convergence: convergence in quadratic mean, convergence in $\ell_1$.
- Central limit theorem.

This class we start discussing statistical estimation formally. We will also preview two other topics: confidence sets and hypothesis testing. This is Chapter 6 of the Wasserman book.

## 8.2 Statistical Estimation

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:

$$X_1, \ldots, X_n \sim F,$$

what can we infer about $F$?

In order to make meaningful inferences about $F$ from a small number of samples 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 possibilities:

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 $\infty$. 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:

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

## 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 $\hat{\theta}$ 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:

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

Similarly, the variance of an estimator is given by:

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

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 of a parameter *consistent* if the estimator converges to the true parameter in probability, i.e. for any $\epsilon$:

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

as $n \to \infty$.

## 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 Asymptotic Normality

Often estimators that we study will have an asymptotically normal distribution. This means that:

$$\frac{\hat{\theta}_n - \theta}{\sqrt{v(\hat{\theta}_n)}}$$

converges in distribution to $N(0, 1)$. We will refer to this property as asymptotic normality.

## 8.6 Confidence Sets

In general, for a parameter $\theta$ we define a $1-\alpha$ confidence set $C_n$ to be any set which has the property that:

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha.$$

We usually refer to $\mathbb{P}_\theta(\theta \in C_n)$ as the coverage of the confidence set $C_n$. The confidence set $C_n$ is a random set (and $\theta$ is a fixed parameter).

One can think about the coverage guarantee in the following way:

You repeat the experiment many times, each time constructing a different confidence interval $C_n$. Then $1-\alpha$ of these different sets will contain the corresponding true parameter. Notice, that the true parameter does not have to be fixed, so in some sense the experiment you conduct can be different each time.

We already saw a way to construct confidence intervals for a Bernoulli parameter using Hoeffding's inequality. More generally, we can always use concentration inequalities to construct confidence intervals. These confidence intervals are often loose and we instead resort to approximate (asymptotic) confidence intervals.

It is often the case that:

$$\frac{\hat{\theta}_n - \theta}{\sqrt{v(\hat{\theta}_n)}}$$

is asymptotically $N(0,1)$. In these cases we have that $\hat{\theta}_n \approx N(\theta, v(\hat{\theta}_n))$. Define, $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. Then we would construct a confidence interval:

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\sqrt{v(\hat{\theta}_n)}, \hat{\theta}_n + z_{\alpha/2}\sqrt{v(\hat{\theta}_n)}).$$

We now need to verify that:

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha,$$

as $n \to \infty$, which is what it means to be an asymptotic confidence interval.

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}(\hat{\theta}_n - z_{\alpha/2}\sqrt{v(\hat{\theta}_n)} \leq \theta \leq \hat{\theta}_n + z_{\alpha/2}\sqrt{v(\hat{\theta}_n)}) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\sqrt{v(\hat{\theta}_n)}} \leq z_{\alpha/2}\right) \\
&\to \mathbb{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha.
\end{aligned}$$

**Example:** Bernoulli confidence sets:

1. We previously constructed confidence sets using Hoeffding's inequality. They took the form:
$$C_n = \left( \hat{p}_n - \sqrt{\frac{\log(2/\alpha)}{2n}}, \hat{p}_n + \sqrt{\frac{\log(2/\alpha)}{2n}} \right).$$

2. If we instead use the normal approximation: we first note that the variance of our estimator is:
$$v(\hat{\theta}_n) = \frac{p(1-p)}{n}.$$
However, we cannot use this variance to create our confidence set, so we instead estimate the variance as:
$$\hat{v}(\hat{\theta}_n) = \frac{\hat{p}_n(1-\hat{p}_n)}{n}.$$
With this we would use the confidence interval:
$$C_n = \left( \hat{p}_n - z_{\alpha/2}\sqrt{\hat{v}(\hat{\theta}_n)}, \hat{p}_n + z_{\alpha/2}\sqrt{\hat{v}(\hat{\theta}_n)} \right).$$
It is easy to verify that this interval is always shorter than the Hoeffding interval but it is only asymptotically correct.

## 8.7 Hypothesis testing

Typically, the way that statistical hypothesis testing proceeds is by defining a so-called **null hypothesis.** We then collect data, and typically the question we ask is whether the data provides enough evidence to *reject* the null hypothesis.

**Example:** Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$, and we want to test if the coin is fair. In this case the null hypothesis would be:
$$H_0 : p = 1/2.$$
We typically also specify a **alternate hypothesis.** In this case, the alternative hypothesis is:
$$H_1 : p \neq 1/2.$$
Typically, hypothesis testing proceeds by defining a **test statistic.** In this case, a natural statistic might be:
$$T = \left| \frac{1}{n}\sum_{i=1}^{n} X_i - p \right|.$$
It might make sense to reject the null hypothesis if $T$ is large. We will be more precise about this later on particularly by defining the different types of errors, and how to set the threshold for $T$.