Lecture 5
Point estimators.

1 Estimators. Properties of estimators.

An estimator is a function of the data. If we have a parametric family with parameter θ, then an estimator of θ is usually denoted by $\hat{\theta}$.

1.1 Unbiasness

Let X be our data. Let $\hat{\theta} = T(X)$ be an estimator where T is some function.

We say that $\hat{\theta}$ is *unbiased* for θ if $E_\theta[T(X)] = \theta$ for all possible values of θ where E_θ denotes the expectation when θ is the true parameter value. Thus, the concept of unbiasness means that we are on average right. The *bias* of $\hat{\theta}$ is defined by $\text{Bias}(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$. Thus, $\hat{\theta}$ is unbiased if and only if its bias equals 0. Thus, sample average and sample variance are unbiased estimators of population mean and population variance correspondingly.

There are some cases when unbiased estimators do not exist. As an example, let $X_1, ..., X_n$ be a random sample from a Bernoulli(p) distribution. Suppose that our parameter of interest $\theta = 1/p$. Let $\hat{\theta} = T(X)$ be some estimator. Then $E[\hat{\theta}] = \sum_{(x_1,...,x_n)\in\{0,1\}^n} T(x_1, ..., x_n) P\{(X_1, ..., X_n) = (x_1, ...x_n)\}$. We know that for any $(x_1, ..., x_n) \in \{0,1\}^n$, $P\{(X_1, ..., X_n) = (x_1, ...x_n)\} = p^{\sum x_i}(1-p)^{\sum(1-x_i)}$ which is a polynomial of degree n in p. Therefore, $E[\hat{\theta}]$ is a polynomial of degree at most n in p. However, $1/p$ is not a polynomial at all. Hence, there are no unbiased estimators in this case.

The example above is very typical in the sense that parameter p has an unbiased estimator $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_p$, but the parameter of interest is a non-linear function of p. Notice that $E\frac{1}{\xi} \neq \frac{1}{E\xi}$, and the bias appears from the non-linear transformation. This bias can be partially corrected by bootstrap.

1.1.1 Bootstrap bias correction

Another task for which the bootstrap is used is bias-correction. Suppose, $EZ = \mu$ and, we're interested in a non-linear function of μ, say $\theta = g(\mu)$. Here Z may be a random variable coming from transformations of observed: $Z_i = h(X_i)$. We do have an unbiased estimate of μ, say $\hat{\mu} = \overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. We may try to use this in order to estimate $\theta : \hat{\theta} = g(\bar{Z})$. Estimator $\hat{\theta}$ is reasonable but is biased unless $g()$ is linear. The bias is $\text{Bias} = E\hat{\theta} - g(\mu)$. We can estimate the bias using the bootstrap:

1. For each $b = 1, ..., B$ generate a bootstrap sample, $\{Z^*_{ib}\}$ from set $\{Z_1, ..., Z_n\}$ with replacement;

2. Calculate $\overline{Z}_b^* = \frac{1}{n}\sum_{i=1}^n Z_{ib}^*$;

3. Estimate $\theta_b^* = g(\bar{z}_b^*)$;

4. $\text{Bias}^* = \frac{1}{B}\sum_{b=1}^B \theta_b^* - \hat{\theta} \approx \text{Bias}$.

5. Use $\tilde{\theta} = \hat{\theta} - \text{Bias}^*$ as your estimate.

Why does it work? Let's denote $G_1(\mu) = \frac{dg(\mu)}{d\mu}$ and $G_2(\mu) = \frac{d^2 g(\mu)}{d\mu^2}$. Notice that if CLT works we have $\sqrt{n}(\bar{Z} - \mu) \Rightarrow N(0, \sigma_z^2)$, where $\sigma_z^2 = Var(Z_i)$; or $\bar{Z} - \mu = O_p(1/\sqrt{n})$. Then

$$\widehat{\theta} - \theta = g(\overline{Z}) - g(\mu) = G_1(\mu)(\bar{Z} - \mu) + \frac{1}{2}G_2(\mu)(\bar{Z} - \mu)^2 + o_p(\frac{1}{n}),$$

$$Bias = E(\widehat{\theta} - \theta) = \frac{1}{2}G_2(\mu)E(\bar{Z} - \mu)^2 = \frac{1}{2}G_2(\mu)\frac{\sigma_z^2}{n} + o(\frac{1}{n}),$$

and similarly

$$Bias^* = \frac{1}{2}G_2(\bar{z})\frac{s_z^2}{n} + o_p(\frac{1}{n}).$$

As a result,

$$Bias^* - Bias = o_p(\frac{1}{n}).$$

This procedure eliminates the leading term in bias ($O(1/n)$), but not the whole of the bias. The remaining bias is of order $o(1/n)$. Notice that in principle there was an asymptotic approach to eliminate bias as well (as we did get the formula for the leading term $\frac{1}{2}G_2(\mu)\frac{\sigma_z^2}{n} + o(\frac{1}{n})$). One in principle could have approximated it by $\frac{1}{2}G_2(\overline{Z})\frac{s_z^2}{n}$, but the bootstrap does this automatically.

Example. Assume we wish to estimate the skewness of a distribution $\theta = \frac{E(X-EX)^3}{[Var(X)]^{3/2}}$. A natural estimate is

$$\hat{\theta} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3}{s^3}.$$

This is a non-linear function of $\bar{Z} = (\bar{X}, \frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n X_i^3)$. As such it will most likely have bias, which we can correct with the bootstrap.

1.1.2 Efficiency: MSE

Another concept that evaluates the performance of estimators is the MSE (Mean Squared Error). By definition, $\text{MSE}(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2]$. Last time we showed a useful decomposition for MSE:

$$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + V(\hat{\theta}).$$

Estimators with smaller MSE are considered to be better, meaning more *efficient*. Quite often there is a trade-off between the bias of the estimator and its variance. Thus, we may prefer a slightly biased estimator to an unbiased one if the former has much smaller variance in comparison to the latter one.

Example Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$. Let $\hat{\sigma}_1^2 = s^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$ and $\hat{\sigma}_2^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/n$ be two estimators of σ^2. We know that $E[\hat{\sigma}_1^2] = \sigma^2$. So $E[\hat{\sigma}_2^2] = ((n-1)/n)E[\hat{\sigma}_1^2] =$

$((n-1)/n)\sigma^2$, and $\text{Bias}(\hat{\sigma}_2^2) = \sigma^2/n$. We also know that $(n-1)\hat{\sigma}_1^2/\sigma^2 \sim \chi^2(n-1)$. What is $V(\chi^2(n-1))$? Let $\xi_1, ..., \xi_{n-1}$ be a random sample from $N(0,1)$. Then $\xi = \xi_1^2 + ... + \xi_{n-1}^2 \sim \chi^2(n-1)$. By linearity of expectation, $E[\xi] = (n-1)$. By independence,

$$\begin{aligned} E[\xi^2] &= E[(\xi_1^2 + ... + \xi_{n-1}^2)^2] \\ &= \sum_{i=1}^{n-1} E[\xi_i^4] + 2 \sum_{1 \le i < j \le n-1} E[\xi_i^2 \xi_j^2] \\ &= 3(n-1) + 2 \sum_{1 \le i < j \le n-1} E[\xi_i^2] E[\xi_j^2] \\ &= 3(n-1) + (n-1)(n-2) \\ &= (n-1)(n+1), \end{aligned}$$

since $E[\xi_i^4] = 3$. So

$$V(\xi) = E[\xi^2] - (E[\xi])^2 = (n-1)(n+1) - (n-1)^2 = 2(n-1).$$

Thus, $V(\hat{\sigma}_1^2) = V(\sigma^2 \xi/(n-1)) = 2\sigma^4/(n-1)$ and $V(\hat{\sigma}_2^2) = ((n-1)/n)^2 V(\hat{\sigma}_1^2) = 2\sigma^4(n-1)/n^2$. Finally, $\text{MSE}(\hat{\sigma}_1^2) = 2\sigma^4/(n-1)$ and

$$\text{MSE}(\hat{\sigma}_2^2) = \sigma^4/n^2 + 2\sigma^4(n-1)/n^2 = (2n-1)\sigma^2/n^2.$$

So, $\text{MSE}(\hat{\sigma}_1^2) < \text{MSE}(\hat{\sigma}_2^2)$ if and only if $2/(n-1) < (2n-1)/n^2$, which is equivalent to $3n < 1$. So, for any $n \ge 1$, $\text{MSE}(\hat{\sigma}_1^2) > \text{MSE}(\hat{\sigma}_2^2)$ in spite of the fact that $\hat{\sigma}_1^2$ is unbiased.

In general, the idea of minimizing MSE is not in agreement with unbiasedness: one may get better efficiency if we allow for some bias. Here is "Stein's shrinkage" idea. Assume that the parameter set Θ is bounded and $\hat{\theta} = T(X)$ is an unbiased estimator of θ: $ET(X) = \theta$. Take any fixed point $\theta^* \in \Theta$ and shrink the initial estimator towards it:

$$\hat{\theta}_1 = (1-c)T(X) + c\theta^*.$$

Here c characterize the amount of shrinkage. The new estimator is somewhat biased $Bias(\hat{\theta}_1) = c(\theta^* - \theta)$ but is less dispersed $Var(\hat{\theta}_1) = (1-c)^2 Var(\hat{\Theta})$. So, we have

$$MSE(\hat{\theta}_1) = c^2(\theta^* - \theta)^2 + (1-c)^2 Var(\hat{\theta}).$$

One may calculate the derivative of MSE with respect to c at $c = 0$ and find that it is negative, and thus some small positive amount of shrinkage $c > 0$ will improve the efficiency of the initial estimator.

1.2 Asymptotic properties.

1.2.1 Consistency

Imagine a thought experiment in which the number of observations n increases without bound, i.e. $n \to \infty$. Suppose that for each n, we have an estimator $\hat{\theta}_n$.

We say that $\hat{\theta}_n$ is *consistent* for θ if $\hat{\theta}_n \to_p \theta$.

Example Let $X_1, ..., X_n$ be a random sample from some distribution with mean μ and variance σ^2. Let $\hat{\mu} = \hat{\mu}_n = \overline{X}_n$ be our estimator of μ and $s^2 = s_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$ be our estimator of σ^2. By the Law of large numbers, we know that $\hat{\mu} \to_p \mu$ as $n \to \infty$. In addition,

$$\begin{aligned} s^2 &= \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1) \\ &= \sum_{i=1}^n (X_i - \mu)^2/(n-1) - (n/(n-1))(\overline{X}_n - \mu)^2 \\ &= (n/(n-1))(\sum_{i=1}^n (X_i - \mu)^2/n) - (n/(n-1))(\overline{X}_n - \mu)^2 \end{aligned}$$

By the Law of Large Numbers, $\sum_{i=1}^n (X_i - \mu)^2/n \to_p E[(X_i - \mu)^2 = \sigma^2$ and $\overline{X}_n - \mu = \sum_{i=1}^n (X_i - \mu)/n \to_p E[X_i - \mu] = 0$. By using the Continuous Mapping Theorem, $(\overline{X}_n - \mu)^2 \to_p 0$. In addition, $n/(n-1) \to_p 1$. So, by the Slutsky theorem, $s^2 \to_p \sigma^2$. So $\hat{\mu}$ and s^2 are consistent for μ and σ^2 correspondingly.

1.2.2 Asymptotic Normality

We say that $\hat{\theta}$ is *asymptotically normal* if there are sequences $\{a_n\}_{n=1}^\infty$ and $\{r_n\}_{n=1}^\infty$ and constant σ^2 such that $r_n(\hat{\theta} - a_n) \Rightarrow N(0, \sigma^2)$. Then r_n is called the *rate of convergence*, a_n - the *asymptotic mean*, and σ^2 - the *asymptotic variance*. In many cases, one can choose $a_n = \theta$ and $r_n = \sqrt{n}$. We will use the concept of asymptotic normality for confidence set construction later on. For now, let us consider an example.

Example Let $X_1, ..., X_n$ be a random sample from some distribution with mean μ and variance σ^2. Let $\hat{\mu}$ and s^2 be the sample mean and the sample variance correspondingly. Then, by the Central limit theorem, $\sqrt{n}(\hat{\mu} - \mu) \Rightarrow N(0, \sigma^2)$. As for s^2,

$$\sqrt{n}(s^2 - \sigma^2) = (n/(n-1))[\sum_{i=1}^n ((X_i - \mu)^2 - \sigma^2)/\sqrt{n} - (\sqrt{n}(\overline{X}_n - \mu)/n^{1/4})^2] + (\sqrt{n}/(n-1))\sigma^2$$

By the Central limit theorem, $\sum_{i=1}^n ((X_i - \mu)^2 - \sigma^2)/\sqrt{n} \Rightarrow N(0, \tau^2)$ with $\tau^2 = E[((X_i - \mu)^2 - \sigma^2)^2]$. Note that $\tau^2 = \mu_4 - 2\sigma^2 E[(X_i - \mu)^2] + \sigma^4 = \mu_4 - \sigma^4$ with $\mu_4 = E[(X_i - \mu)^4]$. By Slutsky theorem, $\sqrt{n}(\overline{X}_n - \mu)/n^{1/4} \to_p 0$. In addition, $(\sqrt{n}/(n-1))\sigma^2 \to_p 0$. So, by the Slutsky theorem again, $\sqrt{n}(s^2 - \sigma^2) \Rightarrow N(0, \tau^2)$.