→ 3rd Lecture: Tuesday Jan 16th, 2018

• **MSE**:

To study estimation error we started by studying $P(|\hat{\theta}_n - \theta| > \delta)$, deviation above a given threshold δ, by bounding this probability. One may take a different approach by studying average Euclidean distance, i.e. $E[|\hat{\theta}_n - \theta|^2]$, which is denoted by $MSE(\hat{\theta}_n)$.

We note that if $\theta = E(\hat{\theta}_n)$, i.e. $\hat{\theta}_n$ is an unbiased estimator of θ, then $MSE(\hat{\theta}_n) = E[|\hat{\theta}_n - \theta|^2] = E[(\hat{\theta}_n - \mu_{\hat{\theta}_n})^2] = Var(\hat{\theta}_n)$.

Now recall that $Var(X) = 0 \Rightarrow P(X = \text{constant}) = 1$ which essentially means r.x. is a constant. The same comment applies to $MSE(\hat{\theta}_n)$. We want to find the closest estimator $\hat{\theta}_n$ to θ which means that we want to minimize $E[(\hat{\theta}_n - \theta)^2]$ over all possible estimators, ideally atleast

The above comment tell us that in real applications we cannot expect to find an estimator whose MSE is equal to zero. Let's try to understand the MSE a bit more

$$
\begin{aligned}
MSE(\hat{\theta}_n) &= E[(\hat{\theta}_n - \theta)^2] \\
&= E\left[\left\{(\hat{\theta}_n - E(\hat{\theta}_n)) + (E(\hat{\theta}_n) - \theta)\right\}^2\right] \\
&= E\left[\{\hat{\theta}_n - E(\hat{\theta}_n)\}^2 + \{E(\hat{\theta}_n) - \theta\}^2 + 2\{E(\hat{\theta}_n) - \theta\}\{\hat{\theta}_n - E(\hat{\theta}_n)\}\right] \\
&= E\left[\{\hat{\theta}_n - E(\hat{\theta}_n)\}^2\right] + E\left[\overbrace{\{E(\hat{\theta}_n) - \theta\}}^{\text{NOT a r.v.}}{}^2\right] \\
&\quad + 2E\left[\underbrace{\{E(\hat{\theta}_n) - \theta\}}_{\text{NOT a r.v.}}\{\hat{\theta}_n - E(\hat{\theta}_n)\}\right]
\end{aligned}
$$

$$= Var(\hat{\theta}_n) + [\underbrace{E(\hat{\theta}_n) - \theta}_{Bias(\hat{\theta}_n)}]^2 + 2\,Bias(\hat{\theta}_n)\,\underbrace{E[(\hat{\theta}_n - E(\hat{\theta}_n))]}_{\substack{0 \\ E(\hat{\theta}_n) - E(\hat{\theta}_n) = 0}}$$

$$= Var(\hat{\theta}_n) + Bias^2(\hat{\theta}_n)$$

Roughly speaking, bias measures how far off the target we hit on the average while variance measures how much fluctuation our estimator may show from one sample to another.

- Unbiased Estimators:

In almost all real applications, the class of possible estimators for an estimand is huge and the best estimator, i.e. the one that minimizes MSE no matter what the value of the estimand is, almost never exists. Thus we try to reduce the class of potential estimators by imposing a plausible restriction, for example $Bias(\hat{\theta}_n) = 0$

- Def. An estimator $\hat{\theta}_n$ of an estimand θ is said to be unbiased if $E(\hat{\theta}_n) = \theta$, for all possible values of θ

Example: $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$

suppose both μ and σ^2 are unknown. Consider $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{E(X_i)}^{\mu} = \frac{1}{\cancel{n}} \times \cancel{n}\mu = \mu.$$

Thus $\bar{X}_n$ is an unbiased estimator of μ. As for the $MSE(\bar{X}_n)$, we need to find $Var(\bar{X}_n)$.

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$

Thm 5.12(b) page 271
$$= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\mathrm{Var}(X_i) + 2\underset{1\le i<j\le n}{\sum\sum}\underbrace{\mathrm{cov}(X_i, X_j)}_{0}\right\}$$

$\coprod_{i=1}^{n} X_i$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)$$

identically distributed
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^{\cancel{2}}}\times\cancel{n}\sigma^2 = \frac{\sigma^2}{n}$$

$$MSE(\bar{X}_n) = \mathrm{Var}(\bar{X}_n) + \underbrace{\mathrm{Bias}^2(\bar{X}_n)}_{0} = \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

An inspection of the above calculation shows that for unbiasedness we only require a common mean μ while for calculating the variance we would only require a common variance σ^2 and orthogonality, i.e.

$$\mathrm{cov}(X_i, X_j) = 0 \text{ if } i \neq j.$$

– suppose $X_1, \ldots, X_n$ have the same mean value μ. Then
$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{\cancel{n}}\times\cancel{n}\mu = \mu.$$

– suppose further that $X_1, \ldots, X_n$ have the same variance σ^2 and $\mathrm{cov}(X_i, X_j) = 0$, $i \neq j$. Then
$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)$$

Thm 5.12(b) page 271
$$= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\mathrm{Var}(X_i) + 2\underset{1\le i<j\le n}{\sum\sum}\mathrm{cov}(X_i, X_j)\right\}$$

orthogonality, i.e. $\mathrm{cov}(X_i, X_j) = 0$, $i \neq j$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)$$

having the same variance
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^{\cancel{2}}}\times\cancel{n}\sigma^2 = \frac{\sigma^2}{n}$$

Thus $MSE(\bar{X}_n) = Var(\bar{X}_n) = \frac{\sigma^2}{n}$

if $X_1, \dots, X_n$ have the same mean value and variance and they are orthogonal.

- Remark (Stein's Paradox)

We will learn latter that if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i=1,\dots,n$

then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has many optimal properties. A paradox due to Charles Stein, however, shows that such nice optimal properties are not preserved in higher dimensions. In fact if

$$X_i \overset{iid}{\sim} N(\mu_X, 1), \; Y_i \overset{iid}{\sim} N(\mu_Y, 1) \text{ and } Z_i \overset{iid}{\sim} N(\mu_Z, 1),$$

then we can find biased estimators of $\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix}$ which are closer to $\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix}$ than $\begin{pmatrix}\bar{X}_n\\ \bar{Y}_n\\ \bar{Z}_n\end{pmatrix}$ for any $\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix}$

We may then say that $\begin{pmatrix}\bar{X}_n\\ \bar{Y}_n\\ \bar{Z}_n\end{pmatrix}$ is an <u>inadmissible</u> estimator of $\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix}$.

- Admissibility:

An estimator $\hat{\theta}$ is called admissible if there is no estimator $\tilde{\theta}$ such that

$$MSE(\tilde{\theta}) \le MSE(\hat{\theta}) \text{ for all possible value of } \theta$$

and the inequality is strict for some values of θ.

What this example tells us is that by allowing a little of bias we may be able to reduce variance considerably and hence find an estimator which closer to the target than the most natural unbiased estimator. Note that this phenomenon happens only when the dimension is at least 3. (12)

– We now want to restrict the class of estimators even further. Suppose $X_1, \ldots, X_n$ have the same mean μ and variance σ^2 and they are orthogonal, i.e $Cov(X_i, X_j) = 0$, $i \neq j$. Consider $\tilde{X}_{n,\underset{\sim}{c}} = \sum_{i=1}^{n} c_i X_i$ and

$$\mathcal{C} = \{\tilde{X}_{n,\underset{\sim}{c}} : \underset{\sim}{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n, \ \sum_{i=1}^{n} c_i = 1\}.$$

Note that

$$E(\tilde{X}_{n,\underset{\sim}{c}}) = E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E(X_i)$$

$$= \sum_{i=1}^{n} c_i \mu = \mu \underbrace{\sum_{i=1}^{n} c_i}_{1} = 1 \times \mu = \mu$$

Thus $\tilde{X}_{n,\underset{\sim}{c}}$ is an unbiased estimator of μ for any $\underset{\sim}{c} \in \mathbb{R}^n$ as long as $\sum_{i=1}^{n} c_i = 1$. Then $\mathcal{C}$ is the class of all unbiased linear estimators of μ. We want to find the best estimator within $\mathcal{C}$, i.e.

$$\min_{\underset{\sim}{c} \in \mathbb{R}^n} MSE(\tilde{X}_{n,\underset{\sim}{c}}) \quad \text{s.t.} \quad \sum_{i=1}^{n} c_i = 1 \qquad (\dagger)$$

First we note that $MSE(\tilde{X}_{n,\underset{\sim}{c}}) = Var(\tilde{X}_{n,\underset{\sim}{c}})$ since $\tilde{X}_{n,\underset{\sim}{c}}$ is an unbiased estimator of μ when $\sum_{i=1}^{n} c_i = 1$. On the other hand

$$Var(\tilde{X}_{n,\underset{\sim}{c}}) = Var\left(\sum_{i=1}^{n} c_i X_i\right)$$

Thm 5.12 page 271

$$= \sum_{i=1}^{n} c_i^2 Var(X_i) + 2 \sum\sum_{1 \le i < j \le n} Cov(c_i X_i, c_j X_j)$$

$$= \sum_{i=1}^{n} c_i^2 \sigma^2 + \cancel{2 \sum\sum_{1 \le i < j \le n} c_i c_j \underbrace{Cov(X_i, X_j)}_{0}}_{0}$$

$$= \sigma^2 \sum_{i=1}^{n} c_i^2$$

Thus (†) is equivalent to

$$\min_{\underset{\sim}{c} \in \mathbb{R}^n} \sigma^2 \sum_{i=1}^{n} c_i^2 \quad \text{s. to} \quad \sum_{i=1}^{n} c_i = 1 \qquad (\ddagger)$$

using Lagrange Theorem (‡) is equivalent to

$$\min_{\underset{\sim}{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n} \{ \underbrace{\sigma^2 \sum_{i=1}^{n} c_i + \lambda \left(\sum_{i=1}^{n} c_i - 1 \right)}_{\varphi_\lambda(\underset{\sim}{c})} \}.$$

$$\frac{\partial \varphi_\lambda(\underset{\sim}{c})}{\partial c_i} = 2\sigma^2 c_i + \lambda \, , \quad i = 1, 2, \ldots, n$$

$$\frac{\partial}{\partial \lambda} \varphi_\lambda(\underset{\sim}{c}) = \sum_{i=1}^{n} c_i - 1$$

$$\begin{cases} \dfrac{\partial}{\partial c_i} \varphi_\lambda(\underset{\sim}{c}) = 2\sigma^2 c_i + \lambda = 0 \, , \quad i = 1, 2, \ldots, n \\[2ex] \dfrac{\partial}{\partial \lambda} \varphi_\lambda(\underset{\sim}{c}) = 0 \Rightarrow \sum_{i=1}^{n} c_i = 1 \end{cases}$$

Thus $c_i = -\dfrac{\lambda}{2\sigma^2}$, $i = 1, 2, \ldots, n$ and using the last equation $\displaystyle\sum_{i=1}^{n} -\frac{\lambda}{2\sigma^2} = 1 \Rightarrow \lambda = -\frac{2\sigma^2}{n}$ and

therefore $c_i = -\dfrac{\lambda}{2\sigma^2} = -\dfrac{\left(-\frac{2\sigma^2}{n}\right)}{2\sigma^2} = \dfrac{1}{n}$, $i = 1, \ldots, n$

We can further find

$$\mathcal{H} = \left[\frac{\partial^2}{\partial c_i \partial c_j} \varphi_\lambda(\underset{\sim}{c}) \right]_{i,j=1,\ldots,n}$$

and

Show that

$$\underset{\sim}{x}^T \mathcal{H} \underset{\sim}{x} \geq 0 \quad \text{for any } \underset{\sim}{x} \in \mathbb{R}^n$$
$$= 0 \quad \text{iff } \underset{\sim}{x} = 0$$

This then guarantees that $\underset{\sim}{c}^* = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ is indeed a minimizer; in fact the unique minimizer. To summarize:

$$\tilde{X}_{n,\underset{\sim}{c}^*} = \sum_{i=1}^{n} \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n .$$ Thus $\bar{X}_n$ is the best unbiased linear estimator.

– Estimating Variance

So far we confined ourselves to estimation of the population mean. Now suppose we are interested in estimating variance from $X_1, \ldots, X_n$ where X_i's have the same mean value μ, the same variance σ^2 and they are orthogonal, i.e. $\mathrm{cov}(X_i, X_j) = 0$, $i \neq j$. A natural estimator of

$$\sigma^2 = \mathrm{Var}(X) = E[(X-\mu)^2]$$

is its sample counterpart, i.e.

$$S^2_{n,*} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Now first question is if $S^2_{n,*}$ is an unbiased estimator of σ^2, i.e. $E(S^2_{n,*}) = \sigma^2$

$$(X_i - \mu)^2 = [(X_i - \bar{X}_n) + (\bar{X}_n - \mu)]^2$$

$$= (X_i - \bar{X}_n)^2 + (\bar{X}_n - \mu)^2 + 2(\bar{X}_n - \mu)(X_i - \bar{X}_n)$$

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2 + 2(\bar{X}_n - \mu) \underbrace{\sum_{i=1}^{n} (X_i - \bar{X}_n)}_{0}$$

$$= \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2 \quad \text{I}$$

Taking expectation we find

$$E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] = E\left[n S^2_{n,*}\right] + E\left[n(\bar{X}_n-\mu)^2\right] \quad \text{(II)}$$

$$\text{RHS} = \sum_{i=1}^{n} \underbrace{E(X_i-\mu)^2}_{\sigma^2} = n\sigma^2$$

Note that $E(\bar{X}_n-\mu)=0$, i.e. $E(\bar{X}_n)=\mu$. Thus

$$E\left[n(\bar{X}_n-\mu)^2\right] = n\, E\left[(\bar{X}_n-\mu)^2\right] = n\,\mathrm{Var}(\bar{X}_n).$$

On the other hand $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$. We therefore have

$E\left[n(\bar{X}_n-\mu)^2\right] = n\cdot \mathrm{Var}(\bar{X}_n) = n\cdot\frac{\sigma^2}{n} = \sigma^2$ and hence from (II)

$$n\sigma^2 = E(n S^2_{n,*}) + \sigma^2$$

$$\Rightarrow E(S^2_{n,*}) = \left(\frac{n-1}{n}\right)\sigma^2 = \left(1-\frac{1}{n}\right)\sigma^2$$

meaning that $S^2_{n,*}$ is NOT an unbiased estimator of σ^2. Multiplying both sides of the last equation by the reciprocal of $(1-\frac{1}{n})$ we find $E\left(\frac{n}{n-1}S^2_{n,*}\right)=\sigma^2$. Note, however, that

$$\frac{n}{n-1}S^2_{n,*} = \frac{\cancel{n}}{n-1}\cdot\frac{1}{\cancel{n}}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$$

Thus $\boxed{S^2_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}$ is an unbiased estimator of σ^2.

Why "$n-1$"? "$n-1$" is the dimension of $\overbrace{\underbrace{\{X_i-\bar{X}_n;\ i=1,\dots,n\}}_{V}}^{\text{span}}$

$n-1=\dim(\text{span } V)$. Note however

$\dim(\text{span } W) = n$ where $W=\{X_i-\mu,\ i=1,\dots,n\}$.

We discuss these issues further in Chapter 11 where learn regression.

– Two sample problems:

So far we only considered sampling from one population. We may have samples from two or more populations and may want to make inference about differences between the populations. Suppose, for example, we want to study the difference between the average salaries of men and women,

Men	Women
X_1	Y_1
$\vdots$	$\vdots$
X_m	Y_n

where X_i's have the common mean μ_X and Y_j's the common mean μ_Y. We want to estimate $\mu_X - \mu_Y$. The natural estimate is $\bar{X}_m - \bar{Y}_n$. Show that

$$E[\bar{X}_m - \bar{Y}_n] = \mu_X - \mu_Y$$

hence $\bar{X}_m - \bar{Y}_n$ is an unbiased estimator of $\mu_X - \mu_Y$.

Assume further that Xs and Ys are independent, Xs have common variance σ_X^2, Ys have common variance σ_Y^2, $\mathrm{Cov}(X_i, X_j) = 0$, $i \neq j$, $\mathrm{Cov}(Y_i, Y_j) = 0$, $i \neq j$.

Find $\mathrm{Var}(\bar{X}_m - \bar{Y}_n)$. Hint: Use Thm 5.12.

The difference between two proportions can be treated similarly. Note that proportions are essentially means of binary variables.