MATH324 (Statistics) – Lecture Notes

McGill University

Masoud Asgharian

Winter 2019

Contents

1 Lecture 0

2 Lecture 1

3 Lecture 2

3.1 Markov's Inequality

Let X be a random variable and h be a **non-negative** function; ie:

$$h : R \to R^+ \cup \{0\} \quad = \quad [0, \infty)$$

Suppose $E(h(X)) < \infty$,then for some $\lambda > 0$, we have:

$$P(h(X) \geq \lambda) \leq \frac{E[h(X)]}{\lambda} \tag{1}$$

Proof. Suppose X is a continuous random variable:

$$\begin{aligned} E[h(x)] &= \int_x h(x) f_{_X}(x) dx \\ &= \Big(\int_{x:h(x)\geq\lambda} h(x) f_{_X}(x) dx + \int_{x:h(x)<\lambda} h(x) f_{_X}(x) dx \Big) \\ &\geq \int_{x:h(x)\geq\lambda} h(x) f_{_X}(x) dx \qquad \underline{since}\ h \geq 0 \\ &\geq \lambda \int_{x:h(x\geq\lambda} f_{_X}(x) dx = \lambda\ P(h(X) \geq \lambda) \end{aligned}$$

$$\implies P(h(X) \geq \lambda) \leq \frac{E(h(X))}{\lambda}$$

The proof for the discrete case is similar. □

3.2 Tchebyshev's Inequality

Tchebyshev's Inequality is a special case of Markov's Inequality. Consider $h(x) = (x - \mu)^2$, then:

$$\begin{aligned} P(|X - \mu| \geq \lambda) &= P((X - \mu)^2 \geq \lambda^2) \\ &\leq \frac{E[(X - \mu)^2]}{\lambda^2} \qquad\qquad if\ E[(X - \mu)^2] < \infty \end{aligned}$$

Let $\mu = E(X)$, then $E[(X-\mu)^2] = Var(X)$ denoted by σ_x^2. We therefore have:

$$P(|X-\mu_x| \geq \lambda) \leq \frac{\sigma_x^2}{\lambda^2} \qquad where\ \mu_x = E(X) \tag{2}$$

Now consider $\lambda = K\sigma_x$ where K is a known number. Then:

$$P(|X-\mu_x| \geq K\sigma_x) \geq \frac{\sigma_x^2}{K^2\sigma_x^2} = \frac{1}{K^2} \tag{3}$$

This is called **Tchbyshev's Inequality**.

Example 3.1. *Suppose* $K = 3$.

$$P(|X-\mu_x| \geq 3\ \sigma_x) \leq \frac{1}{9}$$

In other words, at least 88% *of the observations are within* 3 *standard deviation from the population mean.*

Going back to the our example:

$$X_i \sim (\mu, 1) \quad , \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We want to study $P(\epsilon \geq \delta) = P(|\bar{X}_n - \mu| \geq \delta)$, first we note that:

$$E(X_i) = \mu \quad ,\ i = 1, 2, \ldots, n$$

Then:

$$\begin{aligned} E(\bar{X}_n) &= E\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \frac{1}{\cancel{n}}(\cancel{n}\mu) \\ &= \mu \end{aligned} \tag{$*$}$$

Thus, using (2) we have:

$$P(|\bar{X}_n - \mu| \geq \delta) \leq \frac{Var(\bar{X}_n)}{\delta^2}$$

Now:

$$\begin{aligned}
Var(\bar{X}_n) &= Var\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \frac{1}{n^2} Var\big(\sum_{i=1}^{n} X_i\big) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + \sum_{1\leq i<j\leq n} \sum_{1\leq i<j\leq n} Cov(X_i, X_j)\Big] \quad \textit{using Thm } 5.12(b) - \textit{page } 271 \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \qquad\qquad \textit{since } \coprod_{1}^{n} X_i \\
&= \frac{1}{n^2} nVar(X) = \frac{Var(X)}{n} \qquad \textit{since } x_i s \textit{ are identically distributed} \\
&= \frac{\delta_X^2}{n} \qquad\qquad (**)
\end{aligned}$$

In our case $X \sim N(\mu, 1)$ so $Var(X) = \delta_X^2 = 1$. Thus $Var(\bar{X}_n) = \frac{1}{n}$

Remark. $X \coprod Y \implies Cov(X, Y) = 0$. *Note that:*

$$X \coprod Y \implies E\big[g_1(X)g_2(Y)\big] = E[g_1(X)].E[g_2(Y)]$$

in particular:

$$X \coprod Y \implies E\big[XY\big] = E[X].E[Y]$$

on the other hand:

$$Cov(X, Y) = E[XY] - E(X)E(Y)$$

thus:

$$X \coprod Y \implies Cov(X, Y) = 0.$$

recall that $X \coprod Y$ means X and Y are independent, i.e. $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ where $f_{X,Y}, f_X and f_Y$ represent respectively the

We therefore have:

$$P(|\bar{X}_n - \mu| \geq \delta) \leq \frac{1}{n\delta^2} \tag{4}$$

Using (4) and the sample size, n, we can find an upper bound for the proportion of deviations which are greater than a given threshold δ.

We can also use (4) for Sample Size Deterministic:

Suppose δ is given and we want $P(|\bar{X}_n - \mu| \geq \delta) \leq \beta$ where β is also given. Then setting $\frac{1}{n\delta^2} = \beta$, we can estimate $n \approx \frac{1}{\beta\delta^2}$.

3.3 Application to Voting

Define $X_i = \begin{cases} 1 & \text{NDP} \\ 0 & \text{otherwise} \end{cases}$. Associated to each eligible voter in Canada we have a binary variable X. Let $p = P(X = 1)$. So p represents the proportion of eligible voters who favor NDP. Of interest is often estimation of p. Suppose we have a sample of size n, $X_1, X_2, \ldots, X_n$.

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample proportion; The counterpart of p which nat be denoted by $\hat{p}$. Note that:

$$\mu_{_X} = E(X) = 1 \times P(X = 1) + 0 \times P(X = 0) = 1 - p + 0 \times (1 - p) = p$$

and:

$$E(X^2) = 1^2 \times P(X = 1) + 0^2 \times P(X = 0) = 1 - p + 0 \times (1 - p) = p$$

From (*) and (**) we find that :

$$E(\hat{p}_n) = E(\bar{X}_n)\mu_{_X} = p$$

and:

$$Var(\hat{p}_n) = E(\bar{X}_n) = \frac{Var(X)}{n} = \frac{\sigma_X^2}{n} = \frac{p(1-p)}{n}$$

Thus using (2), we have:

$$P(|\hat{p}_n - p| \geq \delta) \leq \frac{Var(\hat{p}_n)}{\delta^2} = \frac{p(1-p)}{n\delta^2}$$

Note that the above bound on the probability of derivation depends on p which is *unknown*. We however notice that $p(1-p) \leq \frac{1}{4}$.
Define $\zeta(x) = x(1-x)\ for\ 0 < x < 1$. Then:

$$\begin{aligned}
&\zeta^{'}(x) = 1 - 2x \implies \zeta^{'}(x) = 0 \implies x = \frac{1}{2} \\
&\zeta^{"}(\frac{1}{2}) = -2 \implies x = \frac{1}{2} \qquad \text{which is a } \textbf{maximizer} \\
&\zeta(\frac{1}{2}) = \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4}
\end{aligned}$$

(Note that $\zeta^{"}(x) = -2$ for all $0 < x < 1$)

We therefore find:

$$P(|\hat{p}_n - p| \geq \delta) \leq \frac{1}{4n\delta^2} \tag{5}$$

Using (5) and a given sample size n we can find an upper bound for the probability of derivation by δ and the amount for any given δ.
We can also use (5) for sample size deterministic for a size bound β and derivative δ as follows:

$$\frac{1}{4n\delta^2} = \beta \qquad \implies \qquad n \geq \frac{1}{4\beta\delta^2}$$

This is of course conservative since $p(1-p) \leq \frac{1}{4}$.

4 Lecture 3

4.1 MSE

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 denoted by $\mathbf{MSE(\hat{\Theta}_n)}$.

We note that if $\Theta = E(\hat{\Theta}_n)$, i.e. $\hat{\Theta}_n$ is an unbiased estimation of Θ, then:

$$MSE(\hat{\Theta}_n) = E[|\hat{\Theta}_n - \Theta|^2] = E[(\hat{\Theta}_n - \mu_{n_{\Theta_n}})^2] = Var(\hat{\Theta}_n)$$

Now recall that $Var(X) = 0 \implies P(X = \text{constant}) = 1$ which essentially means random variable X is a constant.

The same comment applies to $\text{MSE}(\hat{\Theta}_n)$. 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 at least the above comment tells 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:

$$\begin{aligned}
\text{MSE}(\hat{\Theta}_n) &= E[(\hat{\Theta}_n - \Theta)^2] \\
&= E\Big[\Big((\hat{\Theta}_n - E(\hat{\Theta}_n)) + \big(E(\hat{\Theta}_n) - \Theta\big)\Big)^2\Big] \\
&= E\big[\big(\hat{\Theta}_n - E(\hat{\Theta}_n)\big)^2\big] + \big(E(\hat{\Theta}_n) - \Theta\big)^2 + 2 \cdot E\big[\big(\hat{\Theta}_n - E(\hat{\Theta}_n)\big)\big] \cdot \big(E(\hat{\Theta}_n) - \Theta\big) \\
&= E\big[(\hat{\Theta}_n - E(\hat{\Theta}_n))^2\big] + E\big[\overbrace{(E(\hat{\Theta}_n) - \Theta)^2}^{\text{not a r.v.}}\big] + 2 \cdot E\Big[\overbrace{\big(E(\hat{\Theta}_n) - \Theta\big)}^{\text{not a r.v.}} \cdot \big(\hat{\Theta}_n - E(\hat{\Theta}_n)\big)\Big] \\
&= Var(\hat{\Theta}_n) + \big[\overbrace{E(\hat{\Theta}_n) - \Theta}^{\text{Bias}(\hat{\Theta}_n)}\big]^2 + 2 \cdot Bias(\hat{\Theta}_n) \cdot \overbrace{E[(\hat{\Theta}_n - E(\hat{\Theta}_n))]}^{E(\hat{\Theta}_n) - E(\hat{\Theta}_n) = 0} \\
&= Var(\hat{\Theta}_n) + Bias^2(\hat{\Theta}_n)
\end{aligned}$$

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.

4.2 Unbiased Estimators

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

Definition. *An estimator $\hat{\Theta}_n$ of an **ESTIMANAL** Θ is said to be **unbiased** if $E(\hat{\Theta}_n) = \Theta$, for all possible values of Θ.*

Example 4.1. $X_i \overset{iid}{\sim} N(\mu, \sigma^2) \qquad i = 1, 2, \ldots, n$

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

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

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

$$\begin{aligned}
Var(\bar{X}_n) &= Var\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + 2\cdot\sum\sum_{1\le i<j\le n}\overbrace{Cov(X_i, X_j)}^{0}\Big] && \text{Theorem 5.12(b) - page 271} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) && \coprod_{i=1}^{n} X_i \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^{\cancel{2}}}\cdot\cancel{n}\sigma^2 = \frac{\sigma^2}{n} && \text{identically distributed} \\
&\implies MSE(\bar{X}_n) = Var(\bar{X}_n) + \overbrace{Bias^2(\bar{X}_n)}^{0} = Var(\bar{X}_n) = \frac{\sigma^2}{n}
\end{aligned}$$

An inspection of the above calculation shows that for unbiased μ 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:

$$Cov(X_i, X_j) = 0 \qquad \text{where } i \neq j$$

Suppose $X_1, \ldots, X_n$ have the same mean value μ. Then:

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

Suppose further that $X_1, \ldots, X_n$ have the same variance σ^2 and $Cov(X_i, X_j) = 0,\ i \neq j$.

Then:

$$\begin{aligned}
Var(\bar{X}_n) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{1\leq i<j\leq n} Cov(X_i, X_j)\Big] && \text{Theorem 5.12(b) - Page 271} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) && \text{Orthogonality: i.e. } Cov(X_i, X_j) = 0 \ \ if\ i \neq j \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n\cancel{^2}}\cancel{n}\sigma^2 = \frac{\sigma^2}{n} && \text{having the same variance} \\
&\implies MSE(\bar{X}_n) = Var(\bar{X}_n) = \frac{\sigma^2}{n}
\end{aligned}$$

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

4.3 Stein's Paradox

We will learn later that if $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ 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 a nice optimal properties are not preserved in higher dimensions. In fact if:

$$X_i \overset{iid}{\sim} N(\mu_{_X}, 1), \quad Y_i \overset{iid}{\sim} N(\mu_{_Y}, 1) \text{ and } Z_i \overset{iid}{\sim} (\mu_{_Z}, 1)$$

then, we can find the 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 **inadmissible estimator** of $\begin{pmatrix} \mu_{_X} \\ \mu_{_Y} \\ \mu_{_Z} \end{pmatrix}$.

4.4 Admissibility

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

$$MSE(\tilde{\Theta}) \leq MSE(\hat{\Theta}) \qquad \text{for all possible values of } \Theta$$

and this inequality is strict for some values of Θ.

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

5 Lecture 4

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 $\tilde{X_{n,C}}$