Properties of Estimators I

Bruce M. Boghosian

Department of Mathematics

Tufts University

Outline

Cumulative distribution functions

- If $f_Y(y)$ is the *probability density function (PDF)* of random variable Y, we define the *cumulative distribution function (CDF)* of Y as

$$F_Y(y) := \text{Prob}(Y < y) = \int_{-\infty}^{y} dz\, f_Y(z).$$

- Note that, by the *Fundamental Theorem of Calculus*, it follows immediately that the PDF is given by

$$f_Y(y) = \frac{d}{dy} F_Y(y)$$

- Sometimes it is easier to figure out the CDF, and then differentiate to obtain the PDF.

Order statistics for $Y_{\max}$

Bruce M. Boghosian

- Suppose that $\vec{y} = \{Y_1, Y_2, \ldots, Y_n\}$ are n i.i.d. random variables with by the continuous PDF $f_Y(y)$. What is the PDF of $Y_{\max} = \max_j Y_j$?
- Note that if $Y_{\max} < y$, then it must be that $Y_j < y$ for $j = 1, 2, \ldots, n$.

$$\begin{aligned}
F_{Y_{\max}}(y) =& \mathrm{Prob}(Y_{\max} < y) \\
=& \mathrm{Prob}(Y_1 < y)\ \mathrm{Prob}(Y_2 < y) \cdots \mathrm{Prob}(Y_n < y) \\
=& F_{Y_1}(y) F_{Y_2}(y) \cdots F_{Y_n}(y) \\
=& [F_Y(y)]^n .
\end{aligned}$$

- It follows that $f_{Y_{\max}}(y) = \frac{d}{dy} F_{Y_{\max}}(y)$, so

$$\boxed{f_{Y_{\max}}(y) = n\,[F_Y(y)]^{n-1}\, f_Y(y).}$$

Order statistics for Y_{min}

- Suppose that $\vec{y} = \{Y_1, Y_2, \ldots, Y_n\}$ are n i.i.d. random variables with by the continuous PDF $f_Y(y)$. What is the PDF of $Y_{\text{min}} = \min Y_j$?
- Note that if $Y_{\text{min}} < y$, then it must be that $Y_j > y$ for $j = 1, 2, \ldots, n$.

$$\begin{aligned}
F_{Y_{\text{min}}}(y) =& \text{Prob}(Y_{\text{min}} < y) = 1 - \text{Prob}(Y_{\text{min}} > y) \\
=& 1 - \text{Prob}(Y_1 > y)\,\text{Prob}(Y_2 > y) \cdots \text{Prob}(Y_n > y) \\
=& 1 - [1 - F_{Y_1}(y)]\,[1 - F_{Y_2}(y)] \cdots [1 - F_{Y_n}(y)] \\
=& 1 - [1 - F_Y(y)]^n .
\end{aligned}$$

- It follows that $f_{Y_{\text{max}}}(y) = \frac{d}{dy}\{1 - [1 - F_{Y_{\text{max}}}(y)]\}$, so

$$\boxed{f_{Y_{\text{min}}}(y) = n\,[1 - F_Y(y)]^{n-1}\, f_Y(y).}$$

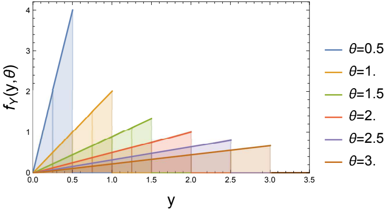
Desirable properties of estimators

- **Unbiasedness:** Suppose that $\vec{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ is a random sample from the continuous PDF $f_Y(y;\theta)$, where θ is an unknown parameter. An estimator $\hat{\theta}(\vec{Y})$ is said to be *unbiased* for θ if $E(\hat{\theta}) = \theta$ for all θ. A similar definition holds for discrete random variables.
- **Efficiency:** Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for parameter θ. If $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$, we say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$.
- **Relative efficiency:** The *relative efficiency* of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is $\frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$.

Unbiasedness Example 1 (continued from last time)

- Consider the one-parameter probability density function

$$f_Y(y;\theta) = \begin{cases} \frac{2y}{\theta^2} & \text{if } 0 \le y \le \theta \\ 0 & \text{otherwise} \end{cases}$$



- Normalization: $\int_0^\theta dy\ \frac{2y}{\theta^2} = 1$
- Mean: $\mu = \int_0^\theta dy\ \frac{2y}{\theta^2} y = \frac{2}{3}\theta$
- Mean square: $E(Y^2) = \int_0^\theta dy\ \frac{2y}{\theta^2} y^2 = \frac{1}{2}\theta^2$
- Variance: $\text{Var}(Y) = \int_0^\theta dy\ \frac{2y}{\theta^2}(y-\mu)^2 = \frac{1}{18}\theta^2$
- Standard deviation: $\sigma_Y = \sqrt{\text{Var}(Y)} = \frac{1}{3\sqrt{2}}\theta$

Example 1: Method of moments

- Let $M_1 = \frac{1}{n}\sum_{j=1}^{n} y_j$ be the sample mean.
- Set the theoretical mean equal to the sample mean: $\frac{2}{3}\theta_e = M_1$
- Hence $\theta_e = \frac{3}{2}M_1$
- MM estimator is then

$$\hat{\theta}_{\rm mm}(\vec{y}) = \frac{3}{2n}\sum_{j=1}^{n} y_j$$

- MM estimator is unbiased

$$E(\hat{\theta}_{\rm mm}(\vec{y})) = \frac{3}{2n}\sum_{j=1}^{n} E(y_j) = \frac{3}{2n}\sum_{j=1}^{n}\frac{2}{3}\theta = \frac{3}{2n}n\left(\frac{2}{3}\theta\right) = \theta.$$

Example 1: Maximum likelihood estimation

- If $\max_j y_j > \theta$, the likelihood will be zero
- So suppose that $\theta > \max_j y_j$
- Likelihood is $L(\vec{y};\theta) = \prod_{j=1}^{n} \left(\frac{2y_i}{\theta^2}\right)$
- This clearly increases as θ decreases, so the MLE estimator is

$$\hat{\theta}_{\text{mle}}(\vec{y}) = \max_j y_j$$

- To calculate $E(\hat{\theta}_{\text{mle}})$, we need $f_{Y_{\max}}(y)$, but we can calculate this using what we know about order statistics.

Example 1: Calculation of CDF

- Note $F_Y(y) = 0$ for $y \leq 0$, and $F_Y(y) = 1$ for $y \geq \theta$
- For $0 < y < \theta$, we have $F_Y(y) = \int_0^y dz\ f_Y(z) = \int_0^y dz\ \frac{2z}{\theta^2} = \frac{y^2}{\theta^2}$
- Hence the CDF is

$$F_Y(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ \frac{y^2}{\theta^2} & \text{if } 0 \leq y \leq 1 \\ 0 & \text{if } y \geq 1 \end{cases}$$

- Hence, from our theorem on order statistics

$$f_{Y_{\max}}(y) = \begin{cases} n\left(\frac{y^2}{\theta^2}\right)^{n-1} \frac{2y}{\theta^2} & \text{if } 0 \leq y \leq \theta \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 2n\frac{y^{2n-1}}{\theta^{2n}} & \text{if } 0 \leq y \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

- Note that this is normalized

$$\int_0^\theta dy\ f_{Y_{\max}}(y) = \frac{2n}{\theta^{2n}} \frac{\theta^{2n}}{2n} = 1.$$

Example 1: Bias of maximum likelihood estimation

- We can now compute the expectation value of the MLE estimator

$$E(\hat{\theta}_{\text{mle}}) = \int_0^\theta dy\ f_{Y_{\max}}(y) y = \int_0^\theta dy\ 2n \frac{y^{2n-1}}{\theta^{2n}} y = \frac{2n}{\theta^{2n}} \frac{\theta^{2n+1}}{2n+1} = \frac{2n}{2n+1}\theta.$$

- The *MLE estimator is biased* since

$$E(\hat{\theta}_{\text{mle}}(\vec{y})) = \frac{2n}{2n+1}\theta \neq \theta$$

- It is *asymptotically unbiased*, since $\lim_{n\to\infty} E(\hat{\theta}_{\text{mle}}(\vec{y})) = \theta$.

Example 1: Construction of unbiased version of MLE estimator

- We can construct an unbiased version of the MLE estimator by defining

$$\hat{\theta}_3(\vec{y}) := \frac{2n+1}{2n}\hat{\theta}_{\text{mle}} := \frac{2n+1}{2n}\max_j y_j.$$

- We can see that $\hat{\theta}_3$ is unbiased since

$$E(\hat{\theta}_3) = \frac{2n+1}{2n}\frac{2n}{2n+1}\theta = \theta,$$

- There is no problem with creating unbiased estimators in this way.
- Note that $\hat{\theta}_3$ is not the MLE estimator, but it is arguably preferable to it.

Example 2: Requirement for a linear estimator

Bruce M. Boghosian

- Suppose $X_1, X_2, \ldots, X_n$ have PDF $f_X(x, \theta)$, with theoretical mean $E(X) = \theta$.
- Hence $E(X_j) = \theta$ for $j = 1, 2, \ldots, n$
- Suppose we construct the estimator $\hat{\theta}(\vec{X}) = \sum_{j=1}^{n} a_j X_j$.
- Then $\hat{\theta}(\vec{X}) = \sum_{j=1}^{n} a_j E(X_j) = \sum_{j=1}^{n} a_j \mu = \left(\sum_{j=1}^{n} a_j\right) \mu$
- So $\hat{\theta}$ is unbiased iff $\sum_{j=1}^{n} a_j = 1$

Example 3: The variance of the normal distribution

Bruce M. Boghosian

- Recall the normal distribution with theoretical mean μ and variance $v = \sigma^2$

$$f_X(x) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-\mu)^2}{2v}\right]$$

- Recall that the MLE and MM estimators were

$$\hat{\mu}(\vec{X}) = \overline{X} := \frac{1}{n}\sum_{j}^{n} X_j$$

$$\hat{v}(\vec{X}) = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2$$

- Clearly, the estimator $\hat{\mu}$ is unbiased from the previous example $\sum_{j=1}^{n}\frac{1}{n} = 1$
- What about the estimator $\hat{v}$?

Example 3: Two useful lemmas

- Lemma 1

$$E(X_j\,\overline{X}) = E\left(X_j \frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\sum_{k=1}^{n} E\left(X_j X_k\right) = \frac{1}{n}\sum_{k\neq j}^{n} E\left(X_j\right) E\left(X_k\right) + \frac{1}{n} E\left(X_j^2\right)$$

$$= \frac{n-1}{n}\mu^2 + \frac{1}{n}(\mu^2 + v) = \mu^2 + \frac{1}{n} v$$

- Lemma 2

$$E(\overline{X}^2) = E\left(\frac{1}{n}\sum_{j=1}^{n} X_j\,\overline{X}\right) = \frac{1}{n}\sum_{j=1}^{n} E\left(X_j\,\overline{X}\right) = \frac{1}{n}\sum_{j=1}^{n}\left(\mu^2 + \frac{1}{n} v\right) = \mu^2 + \frac{1}{n} v$$

Example 3: Expectation of the variance estimator

Bruce M. Boghosian

- Using our lemmas, we have

$$
\begin{aligned}
E(\hat{v}) =& E\left(\frac{1}{n}\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2\right) = E\left(\frac{1}{n}\sum_{j=1}^{n}\left(X_j^2-2X_j\overline{X}+\overline{X}^2\right)\right)\\
=&\frac{1}{n}\sum_{j=1}^{n}\left[E\left(X_j^2\right)-2E\left(X_j\overline{X}\right)+E\left(\overline{X}^2\right)\right]\\
=&\frac{1}{n}\sum_{j=1}^{n}\left[\left(\mu^2+v\right)-2\left(\mu^2+\frac{1}{n}v\right)+\left(\mu^2+\frac{1}{n}v\right)\right]\\
=&\frac{n-1}{n}v
\end{aligned}
$$

- Since $E(\hat{v})\neq v$, the *variance estimator is biased*, but *asymptotically unbiased*.

Example 3: Constructing an unbiased variance estimator

- Reverting to the notation σ^2, instead of v, we have found $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$
- We construct the *sample variance* estimator,

$$\boxed{\hat{S}^2(\vec{X}) := \frac{n}{n-1}\hat{\sigma}^2(\vec{X}) = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2}$$

- There is an associated *sample standard deviation* estimator

$$\boxed{\hat{S}(\vec{X}) := \sqrt{\frac{n}{n-1}}\,\hat{\sigma}(\vec{X}) = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2}}$$

- This is used in interval estimation, instead of the estimated standard deviation.

Example 4: Comparison of efficiency of two estimators of the mean

- Similar assumptions to Example 2, with i.i.d. random variables X_1, X_2, X_3
- Assume $E(X_j) = \mu$ and $\mathrm{Var}(X_j) = \sigma^2$ for $j = 1, \ldots, 3$
- Consider the two estimators,

$$\begin{aligned}\hat{\mu}_1(\vec{X}) &:= \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3\\ \hat{\mu}_2(\vec{X}) &:= \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\end{aligned}$$

- These are both unbiased, from Example 2, so $E(\hat{\mu}_1) = E(\hat{\mu}_2) = \mu$, and

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_1(\vec{X})\right) &:= \frac{1}{16}\sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{16}\sigma^2 = \frac{3}{8}\sigma^2\\ \mathrm{Var}\left(\hat{\mu}_2(\vec{X})\right) &:= \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 = \frac{3}{9}\sigma^2\end{aligned}$$

Example 4: Comparison of efficiency of two estimators of the mean

- We continue our analysis of the two unbiased estimators,

$$\hat{\mu}_1(\vec{X}) := \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3$$
$$\hat{\mu}_2(\vec{X}) := \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$$

- We have found that $\hat{\mu}_2$ is more *efficient* since

$$\mathrm{Var}\left(\hat{\mu}_2(\vec{X})\right) = \frac{3}{9}\sigma^2 < \frac{3}{8}\sigma^2 = \mathrm{Var}\left(\hat{\mu}_1(\vec{X})\right)$$

- The *relative efficiency* of $\hat{\mu}_2$ with respect to $\hat{\mu}_1$ is

$$\frac{\mathrm{Var}\left(\hat{\mu}_1\right)}{\mathrm{Var}\left(\hat{\mu}_2\right)} = \frac{\frac{3}{8}\sigma^2}{\frac{3}{9}\sigma^2} = \frac{9}{8}$$

The Cramér-Rao bound

- Let $f_Y(y;\theta)$ be a continuous PDF with continuous first and second derivatives
- Suppose that $\{y \mid f_Y(y) \neq 0\}$ does not depend on θ
- We are given n samples $\vec{Y} = \{Y_1, Y_2, \ldots, Y_n\}$
- Let $\hat{\theta}(\vec{Y})$ be an unbiased estimator of θ
- Then

$$\mathrm{Var}(\hat{\theta}) \geq \left\{ n\, E\left[\left(\frac{\partial \ln f_Y(Y;\theta)}{\partial \theta}\right)^2\right]\right\}^{-1} = \left\{ -n\, E\left[\frac{\partial^2 \ln f_Y(Y;\theta)}{\partial \theta^2}\right]\right\}^{-1}$$

- This gives us a lower bound on the efficiency of any unbiased estimator.
- The *absolute efficiency* of an unbiased estimator $\hat{\theta}$ is the ratio of the Cramér-Rao lower bound to the variance of $\hat{\theta}$

Example 5: The Bernoulli and binomial distributions

Bruce M. Boghosian

- We may suppose that $p_X(k;p) = p^k(1-p)^{1-k}$ where $k \in \{0,1\}$
- Flip coin n times, and define $X = X_1 + X_2 + \cdots + X_n$ where $X_j \in \{0,1\}$.
- Define the unbiased estimator $\hat{p} = X/n$
- The variance of the result is

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{Var}(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Example 5: Cramér-Rao bound for binomial distribution

Bruce M. Boghosian

- To calculate the Cramér-Rao bound, note

$$\ln p_{X_j}(X_j;p) = X_j \ln p + (1-X_j)\ln(1-p)$$

- Taking derivatives,

$$\frac{\partial \ln p_{X_j}(X_j;p)}{\partial p} = \frac{X_j}{p} - \frac{1-X_j}{1-p}$$

$$\frac{\partial^2 \ln p_{X_j}(X_j;p)}{\partial p^2} = -\frac{X_j}{p^2} - \frac{1-X_j}{(1-p)^2}$$

- Taking the expectation value

$$\left\{-n\,E\left[\frac{\partial^2 \ln p_{X_j}(X_j;p)}{\partial p^2}\right]\right\}^{-1} = \left\{n\left(\frac{p}{p^2}+\frac{1-p}{(1-p)^2}\right)\right\}^{-1} = \frac{p(1-p)}{n}$$

- So $\mathrm{Var}(\hat{p})$ achieves the Cramér-Rao bound. It is maximally efficient.

Summary

Bruce M. Boghosian

- We have reviewed *CDFs* and *order statistics*
- We have reviewed the concept of *unbiasedness* and worked three examples.
- We have studied the concepts of *efficiency* and *relative efficiency* and worked through one example.
- We have learned the statement of the *Cramér-Rao bound*, and the notation of *absolute efficiency*