Biostat 602 Winter 2017

Lecture Set 6

Point Estimation

Methods of Finding Estimators

Reading: CB 7.1–7.2

Point Estimation

Basic Premise

- Data: $\mathbf{x} = (x_1, \cdots, x_n)$ - realizations of random variables $(X_1, \cdots, X_n)$.
- $X_1, \cdots, X_n$ *i.i.d.* $f_X(x|\theta)$.
- Assume a model $\mathcal{P} = \{f_X(x|\theta) : \theta \in \Omega \subset \mathbb{R}^p\}$ where the functional form of $f_X(x|\theta)$ is known, but θ is unknown.
- Task is to use data $\mathbf{x}$ to make inference on θ

Definition If we use a function of sample $w(X_1, \cdots, X_n)$ as a "guess" of $\tau(\theta)$, where $\tau(\theta)$ is a function of true parameter θ. Then $w(\mathbf{X}) = w(X_1, \cdots, X_n)$ is called a *point estimator* of $\tau(\theta)$. The realization of the estimation, $w(\mathbf{x}) = w(x_1, \cdots, x_n)$ is called the *estimate* of $\tau(\theta)$.

Example 1: Let $X_1, \cdots, X_n$ *i.i.d.* $\mathcal{N}(\theta, 1)$, where $\theta \in \Omega \in \mathbb{R}$.

- Suppose $n = 6$, and $(x_1, \cdots, x_6) = (2.0,\ 2.1,\ 2.9,\ 2.6,\ 1.2,\ 1.8)$.
- Define the estimator $w_1(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. The estimate is 2.1.
- Define the estimator $w_2(X_1, \cdots, X_n) = X_{(1)}$. Its estimate is 1.2.

The estimator is a statistic that is constructed with an objective of making inference about a parameter. Thus, the specific structural form of the function of the parameter $\tau(\theta)$ is crucial in defining an estimator.

Clearly, the class of estimators for a given problem is infinite until we restrict our search to a given class.

We first explore different approaches to obtaining estimators.

Subsequently, we look at methods to evaluate these estimators and search for an optimal one using these criteria.

Method of Moments Estimation

The method of moments is a simple method of estimation that dates back to Karl Pearson, the Father of Statistics, in the late 1800s. It is a method to equate sample moments to population moments and solve the resulting equations for the parameters.

Sample moments	Population moments
$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu_1' = E[X\|\theta] = \mu_1'(\theta)$
$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$	$\mu_2' = E[X^2\|\theta] = \mu_2'(\theta)$
$m_3 = \frac{1}{n}\sum_{i=1}^{n} X_i^3$	$\mu_3' = E[X^3\|\theta] = \mu_3'(\theta)$
$\vdots$	$\vdots$

Point estimator of $\tau(\theta)$ is obtained by solving equations like this.

$$
\begin{aligned}
m_1 &= \mu_1'(\theta) \\
m_2 &= \mu_2'(\theta) \\
\vdots &\quad \vdots \\
m_k &= \mu_k'(\theta)
\end{aligned}
$$

Example 2: Let $X_1, \cdots, X_n$ be *i.i.d.* from $\mathcal{N}(\mu, \sigma^2)$ population. Find method of moments (MoM) estimator for μ, σ^2.

Solution: Note that

$$\mu_1' = E(\mathbf{X}) = \mu = \overline{X}$$

$$\mu_2' = E(\mathbf{X}^2) = [E(\mathbf{X})]^2 + \mathrm{Var}(\mathbf{X}) = \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

The MoM estimators are obtained by setting up the equations

$$\begin{cases} \hat{\mu} = \overline{X} \\ \hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

Solving the two equations above, $\hat{\mu} = \overline{X}$, $\widehat{\sigma^2} = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$, which are the required MoM estimators for μ, σ^2, respectively.

Example 3: Let $X_1, \cdots, X_n$ be *i.i.d.* from $Binomial(k, p)$. Find a MoM estimator for k, p.

Remark: This application is somewhat unusual in the sense that we are interested here in estimating the parameter k which is treated as known in most applications. Examples of such application include (a) estimating the reporting rate of crimes that are typically under-reported such as domestic violence, and (b) estimating detection rate of bugs in a software code.

Solution: The pmf is given by

$$f_X(x|k,p) \;=\; \binom{k}{x} p^x (1-p)^{k-x} \qquad x \in \{0, 1, \cdots, k\}$$

Equating first two sample moments,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_i &= \overline{X} \approx \mu_1' = E(\mathbf{X}) = kp \\
\frac{1}{n}\sum_{i=1}^{n} X_i^2 &\approx \mu_2' = E[\mathbf{X}^2] = (E\mathbf{X})^2 + \mathrm{Var}(\mathbf{X}) = k^2p^2 + kp(1-p)
\end{aligned}$$

Solving these equations,

$$\begin{aligned}
\overline{X} &= \hat{k}\hat{p} \\
\frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \hat{k}^2\hat{p}^2 + \hat{k}\hat{p}(1-\hat{p}) \\
&= \overline{X}^2 + \overline{X}(1-\hat{p}) \\
\hat{p} &= 1 - \frac{(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2)}{\overline{X}} \\
&= \frac{\overline{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}{\overline{X}} \\
\hat{k} &= \frac{\overline{X}}{\hat{p}} = \frac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}
\end{aligned}$$

ARE THESE GOOD ESTIMATORS?

Example 4: Let $X_1, \cdots, X_n$ be i.i.d. Negative Binomial(r, p). Find method of moments estimator for (r, p).

Solution: The moment equations are

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i = E(\mathbf{X}) = \frac{r(1-p)}{p} \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 = E(\mathbf{X}^2) = \left(\frac{r(1-p)}{p}\right)^2 + \frac{r(1-p)}{p^2} \end{aligned}$$

which gives

$$\hat{p} = \frac{m_1}{m_2 - m_1^2} = \frac{\overline{X}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}$$

$$\hat{r} = \frac{m_1 \hat{p}}{1-\hat{p}} = \frac{\overline{X}\hat{p}}{1-\hat{p}}$$

Example 5: Let $X_1, \cdots, X_n$ be *i.i.d.* $Unif(-\theta, \theta)$. What is the MoM estimator for θ?

Remarks

- MoM estimators are used to match sample moments to population moments, the latter of which is typically a function of the model parameters. The estimators for these model parameters are then obtained by solving equations. Thus, to estimate $\tau(\theta)$, one first solves $\overline{X} = \mu(\hat{\theta})$ to obtain MoM estimator $\hat{\theta}$ of θ and then use $\tau(\hat{\theta})$ as the MoM estimator of $\tau(\theta)$. For example,

$$\hat{\theta}_{MoM} = \exp(\overline{X}) \implies \hat{\theta}^{-1}_{MoM} = \exp(-\overline{X}).$$

- It is possible to have multiple moment equations estimating θ. For example, both $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$ estimate the mean λ of a Poisson distribution. The custom in this case is to call the estimator involving lower order moments ($\overline{X}$ in the Poisson case) the MoM estimator.

- The MoM estimator is always calculated in the untransformed scale. For example, in the case of $X_1, \cdots, X_n$ *i.i.d.* from a $Unif(-\theta, \theta)$ population, we know that $|X|_1, \cdots, |X|_n$ is a random sample from $Unif(0, \theta)$. Yet, $\frac{2}{n}\sum_{i=1}^{n} |X|_i$ is not a MoM estimator of θ.

Maximum Likelihood Estimation

Likelihood Function

Definition: Let $X_1, \cdots, X_n \sim \text{i.}i.d.\ f_X(x|\theta)$. The joint distribution of $\mathbf{X} = (X_1, \cdots, X_n)$ is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

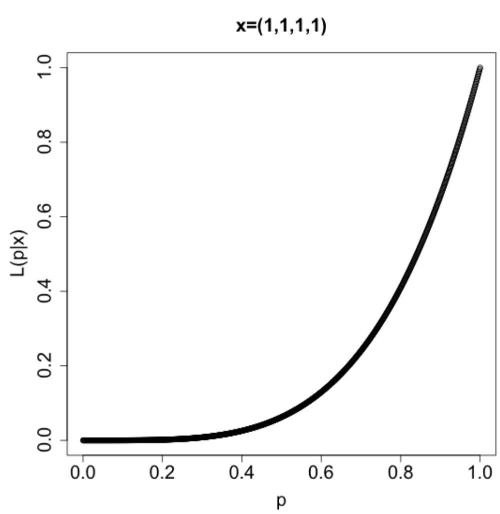
Given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by

$$L(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta)$$

is called the likelihood function.

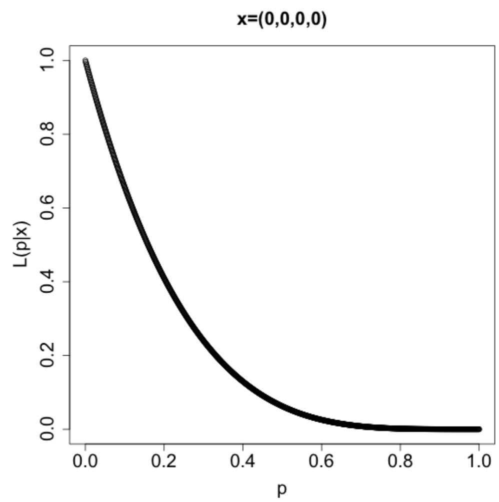
Example 5(a): Let X_1, X_2, X_3, X_4 be $i.i.d.\ Bernoulli(p),\ 0 < p < 1$.

- $\mathbf{x} = (1, 1, 1, 1)^T$
- Intuitively, it is more likely that p is larger than smaller.
- $L(p|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{4} p^{x_i}(1-p)^{1-x_i} = p^4$.



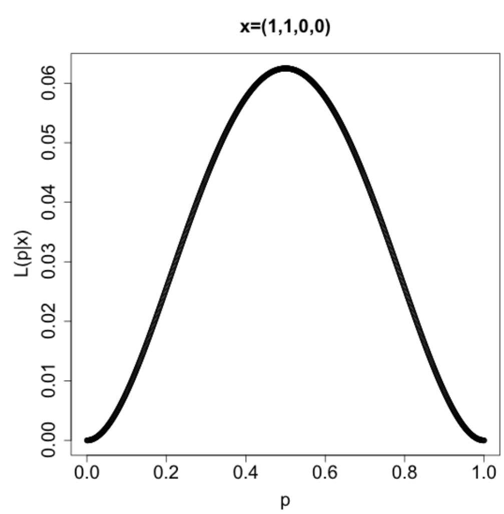
Example 5(b): Let X_1, X_2, X_3, X_4 be *i.i.d. Bernoulli*(p), $0 < p < 1$.

- $\mathbf{x} = (0, 0, 0, 0)^T$
- Intuitively, it is more likely that p is smaller than larger.
- $L(p|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{4} p^{x_i}(1-p)^{1-x_i} = (1-p)^4$.



Example 5(c): Let X_1, X_2, X_3, X_4 be *i.i.d. Bernoulli*(p), $0 < p < 1$.

- $\mathbf{x} = (1, 1, 0, 0)^T$
- Intuitively, it is more likely that p is somewhere in the middle than in the extremes.
- $L(p|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{4} p^{x_i}(1-p)^{1-x_i} = p^2(1-p)^2$.



Maximum Likelihood Estimator

Definition: For a given sample point $\mathbf{x} = (x_1, \cdots, x_n)$, let $\hat{\theta}(\mathbf{x})$ be the value such that $L(\theta|\mathbf{x})$ attains its maximum.

More formally,

$$L(\hat{\theta}(\mathbf{x})|\mathbf{x}) \geq L(\theta|\mathbf{x}) \quad, \forall \theta \in \Omega, \quad \hat{\theta}(\mathbf{x}) \in \Omega.$$

$\hat{\theta}(\mathbf{x})$ is called the *maximum likelihood estimate* of θ based on data $\mathbf{x}$,

$\hat{\theta}(\mathbf{X})$ is the *maximum likelihood estimator (MLE)* of θ.

Example 6: Let $X_1, \cdots, X_n$ be *i.i.d. Exp*(β). Find MLE of β.

Solution: The likelihood function is

$$\begin{aligned} L(\beta|\mathbf{x}) &= f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta) \\ &= \prod_{i=1}^{n} \left[\frac{1}{\beta} e^{-x_i/\beta} \right] = \frac{1}{\beta^n} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\beta} \right) \end{aligned}$$

where $\beta > 0$.

Use the derivative to find potential MLE. Maximizing the likelihood function $L(\beta|\mathbf{x})$ is equivalent to maximize the log-likelihood function

$$\begin{aligned} l(\beta|\mathbf{x}) &= \log L(\beta|\mathbf{x}) = \log \left[\frac{1}{\beta^n} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\beta} \right) \right] \\ &= -\frac{\sum_{i=1}^{n} x_i}{\beta} - n \log \beta \end{aligned}$$

Setting the first derivative of the log-likelihood equal to zero, we get

$$\frac{\partial l}{\partial \beta} = \frac{\sum_{i=1}^{n} x_i}{\beta^2} - \frac{n}{\beta} = 0$$

that simplifies to

$$\sum_{i=1}^{n} x_i = n\beta$$

which yields the solution as

$$\hat{\beta} \;=\; \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

Question: Is $\hat{\beta}$ the maximum likelihood estimator?

Use the double derivative to confirm local maximum.

$$\begin{aligned}
\left.\frac{\partial^2 l}{\partial \beta^2}\right|_{\beta=\overline{x}} &= \left. -2\frac{\sum_{i=1}^{n} x_i}{\beta^3} + \frac{n}{\beta^2}\right|_{\beta=\overline{x}} \\
&= \left.\frac{1}{\beta^2}\left(-\frac{2\sum_{i=1}^{n} x_i}{\beta} + n\right)\right|_{\beta=\overline{x}} \\
&= \frac{1}{\overline{x}^2}\left(-\frac{2n\overline{x}}{\overline{x}} + n\right) \\
&= \frac{1}{\overline{x}^2}(-n) < 0
\end{aligned}$$

Therefore, we can conclude that $\hat{\beta} = \overline{X}$ is unique local maximum on the interval $(0, \infty)$.

Check boundary and confirm global maximum

$\beta \in (0, \infty)$. If $\beta \to \infty$

$$
\begin{aligned}
l(\beta|\mathbf{x}) &= -\frac{\sum_{i=1}^n x_i}{\beta} - n \log \beta \to -\infty \\
L(\beta|\mathbf{x}) &\to 0
\end{aligned}
$$

If $\beta \to 0$, one can also show that $l(\beta|\mathbf{x}) \to -\infty$. This is harder to verify. Visualize this by plotting $l(\beta|\mathbf{x})$ against β.

Since at both ends, L dies off to zero, the local maximum at the interior is indeed the global maximum.

Putting Things Together

1. $\frac{\partial l}{\partial \beta} = 0$ at $\hat{\beta} = \overline{x}$
2. $\frac{\partial^2 l}{\partial \beta^2} < 0$ at $\hat{\beta} = \overline{x}$
3. $L(\beta|\mathbf{x}) \to 0$ (lowest bound) when β approaches the boundary

Therefore $l(\beta|\mathbf{x})$ and $L(\beta|\mathbf{x})$ attains the global maximum when $\hat{\beta} = \overline{x}$

$\hat{\beta}(\mathbf{X}) = \overline{X}$ is the MLE of β.

How do we find MLE?

If the function is differentiable with respect to θ

1. Find candidates that makes first order derivative to be zero
2. Check second-order derivative to check local maximum.
   - For one-dimensional parameter, $\frac{\partial^2 L(\theta)}{\partial \theta^2} < 0$ implies local maximum.
   - For two-dimensional parameter, we need to show
     (a) $\partial^2 L(\theta_1, \theta_2)/\partial \theta_1^2 < 0$ or $\partial^2 L(\theta_1, \theta_2)/\partial \theta_2^2 < 0$.
     (b) Determinant of second-order derivative is positive
3. Check whether boundary gives global maximum.
   - Or clearly justify that boundaries cannot be global maximum.

If the function is NOT differentiable with respect to θ

- Use numerical methods, or
- Directly maximize using inequalities or properties of the function.

Example 7: Let $X_1, \cdots, X_n$ be *i.i.d.* $Uniform(0, \theta)$, where $X_i \in (0, \theta)$ and $\theta > 0$. Find MLE of θ.

Example 8: Suppose n pairs of data $(X_1, Y_1), \cdots, (X_n, Y_n)$ where X_i is generated from an unknown distribution, and Y_i are generated conditionally on X_i.

$$Y_i | X_i \sim \mathcal{N}(\alpha + \beta X_i, \sigma^2)$$

Find the MLE of $(\alpha, \beta, \sigma^2)$.

Solution: The joint distribution of $(X_1, Y_1), \cdots, (X_n, Y_n)$ is

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \prod_{i=1}^{n} f_{\mathbf{Y}}(y_i | x_i) = f_{\mathbf{X}}(\mathbf{x}) \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right]$$

The likelihood function is

$$L(\alpha, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) \;=\; f_{\mathbf{X}}(\mathbf{x}) (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right]$$

The log-likelihood function can be simplied as

$$l(\alpha, \beta, \sigma^2) \;=\; C - \frac{n}{2} \log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial \alpha} \;=\; \frac{2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)}{2\sigma^2} = \frac{n\overline{y} - n\alpha - n\beta\overline{x}}{\sigma^2} = 0$$

$$\hat{\alpha} \;=\; \overline{y} - \hat{\beta}\overline{x}$$

$$\frac{\partial l}{\partial \beta} = \frac{2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i}{2\sigma^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\alpha\overline{x} - \beta \sum_{i=1}^{n} x_i^2}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\overline{x}(\overline{y} - \beta\overline{x}) - \beta \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma} + \frac{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2}{2(\sigma^2)^2} = 0 \\
\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta} x_i)^2
\end{aligned}
$$

Putting Things Together

Therefore, the MLE of $(\alpha, \beta, \sigma^2)$ is

$$
\begin{aligned}
\hat{\alpha}(\mathbf{X}, \mathbf{Y}) &= \overline{Y} - \hat{\beta}\overline{X} \\
\hat{\beta}(\mathbf{X}, \mathbf{Y}) &= \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \\
\hat{\sigma^2}(\mathbf{X}, \mathbf{Y}) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta} X_i)^2
\end{aligned}
$$