Contents

1 Point Estimation

Definition: A point estimator is any scalar (or vector) -valued function of the sample. $(x_1, ...x_n) \sim f(x|\theta)$

A point estimator for $\tau(\theta)$ is a statistic $T(x)$ with the purpose of approximating $\tau(\theta)$

1.1 Method of Moments

The k-th moment of a r.v. X is $\mu_k(\theta) = \mathbb{E}_\theta(X^k) = \int\limits_{\mathbb{X}} x^k f(x|\theta)dx$

Given an iid sample $x_1...x_n \overset{\text{iid}}{\sim} f(x|\theta)$ we have sample moments: $\hat{\mu}_k = \frac{1}{n}\sum\limits_{i=1}^{n} x_i^k$

Suppose $\theta \in \Theta \in \mathbb{R}^P$, and that $\mu_k(\theta)$ exists and is fininte for k = 1, ..., p.

Definition: The method of moments estimator of θ is the solution to the system of equations:

$$\mu_1(\theta) = \hat{\mu}_1 \tag{1}$$
$$. \tag{2}$$
$$. \tag{3}$$
$$\mu_p(\theta) = \hat{\mu}_p \tag{4}$$

We call it $\hat{\theta}_{MM}$

Example: $X_i \overset{\text{iid}}{\sim} Beta(\alpha, \beta), \theta = (\alpha, \beta)$

$\mu_1(\theta) = \frac{\alpha}{\alpha+\beta}$

$\mu_2(\theta) = Var_\theta(x_1) + \left(\frac{\alpha}{\alpha+\beta}\right)^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha^2}{(\alpha+\beta)^2}$

$\frac{\alpha}{\alpha+\beta} = \hat{\mu}_1, \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha^2}{(\alpha+\beta)^2} = \hat{\mu}_2$

$\beta = \left(\frac{1-\hat{\mu}_1}{\hat{\mu}_1}\right)\alpha$ from the first equation. Plug this into the second expression and solve:

$\hat{\alpha}_{MM} = \hat{\mu}_1 \left[\frac{\hat{\mu}_1(1-\hat{\mu}_1)}{\hat{\mu}_2-\hat{\mu}_1^2} - 1\right]$

$\implies \hat{\beta}_{MM} = \frac{(1-\hat{\mu}_1)}{\hat{\mu}_1}\hat{\alpha}_{MM}$

Example: $x_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

$\mu_1(\theta) = \hat{\mu}_1, \mu_1(\theta) = \mu, \implies \hat{\mu}_{MM} = \mu$

$\mu_2(\theta) = Var_\theta(x_1) + \mu_1^2 = \sigma^2 + \mu^2 = \hat{\mu}_2$

$\sigma^2_{MM} = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{n-1}{n}s^2$

Note: If μ, σ^2 are the mean/variance of any family, their MM estimators are $\overline{X}, \frac{n-1}{n}s^2$.

For families where parameters are not just the mean and variance, you can find it via two ways: use mean/variance in MM, then calculate parameters, or use parameters in MM then calculate mean/variance. Both yield same results.

Fact: MM estimators are invariant of re-parameterizations.

Let $\eta = \eta(\theta)$ be a 1:1 mapping (intervtible). Then, $\hat{\eta}_{MM} = \eta(\hat{\theta}_{MM})$

Let's say $x_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and we want to estimate $\tau(\theta) = \frac{\mu}{\sigma}$. This isn't 1:1. What can we do? We can do the 'Transformations' method and create a second value τ_2, etc. We can also just plug in the estimators.

Definition: The MM estimator for a parametric function $\tau(\theta)$ is just $\hat{\tau}_{MM}(\theta) = \tau(\hat{(\theta)}_{MM})$

Properties:

- MM equations may have a unique solution, no solution, or many solutions
- Often, MM estimators are used as initial values for another estimation technique (ie. a root finding method)
- Why should it work? Let's say θ^* is the true value of θ. Then Law of Large Numbers says: $\hat{\mu}_k \xrightarrow{P} \mu_k(\theta*)$. Then we are solving $\mu_k(\theta) = \hat{\mu}_k \approx \mu_k(\theta^*)$

Example: $X_i \overset{\text{iid}}{\sim} Bin(m, \theta), i = 1, ..., n$ (m is known)

Find the MM estimator of $\tau(\theta) = ln\frac{\theta}{1-\theta}$

1. Find MM for θ ($\hat{\theta}_{MM}$)
2. Plug in ($\hat{\tau}_{MM}(\theta) = \tau(\hat{\theta}_{MM})$)
3. $\hat{\tau}_{MM}(\theta) = ln\frac{\overline{X}/m}{1-\overline{X}/m}$

2 Maximum Likelihood

$X = (X_1, ..., X_n) \sim f(x|\theta), \theta \in \Theta \subset \mathbb{R}^k$

Notation: $f(x;\theta)$: function of x indexed at θ. Basically, given some set value of the θ.

Likelihood Function: The likelihood function is: $L(\theta;x) = f(x|\theta)$

Notes:

- L is a function of θ for each $x \in \mathbb{X}$ (in sample space)
- Plugging in X for x gives $L(\theta;X)$, a stochastic process (ie. plug in a random X makes this a random function for θ)
- The log-likelihood function is $l(\theta;x) = ln[L(\theta;x)]$
- If $x_i \overset{\text{iid}}{\sim} f(x_i|\theta)$ (f is marginal dist.), then $l(\theta;x) = \sum\limits_{i=1}^{n} ln[f(x_i|\theta)]$ (because $x_i \overset{\text{iid}}{\sim}$, the sum is just a transformation and we can apply LLN, CLT, etc.)

Maximum Likelihood Estimate: If $x \in \mathbb{X}$ is observed, a maximum likelihood estimate of $\theta, \hat{\theta}(x)$, is any value $\theta \in \Theta$ that maximizes $L(\theta|x)$.

$\hat{\theta}(x) = \underset{\theta\in\Theta}{argmax}[L(\theta|x)]$

This is a function of observed data (an estimate, not an estimator).

Maximum Likelihood Estimator: A maximum likelihood estimator (MLE) is $\hat{\theta} = \hat{\theta}(X)$

If an ML estimate exists, then $\hat{\theta}(x) = \underset{\theta\in\Theta}{argmax}[l(\theta;x)]$. This is because $ln(x)$ is a strictly increasing function.

Why does maximum likelihood work? Can we show that $\hat{\theta} \approx \theta_0$ (true parameter)?

Assume $X_1, ..., X_n \overset{\text{iid}}{\sim} f(x_i|\theta)$. $l(\theta|x) = \sum\limits_{i=1}^{n} ln[f(x_i|\theta)]$

$\frac{1}{n}l(\theta|x) = \frac{1}{n}\sum\limits_{i=1}^{n} ln[f(x_i|\theta)] \xrightarrow{P} \mathbb{E}_{\theta_0}(ln[f(x|\theta)]) = \int\limits_{\mathbb{X}} ln[f(x|\theta)]f(x|\theta_0)dx$

It would make sense that $\hat{\theta}(x) = \underset{\theta\in\Theta}{argmax}[l(X|\theta)] \approx \underset{\theta\in\Theta}{argmax}\, \mathbb{E}_{\theta_0}(ln[f(X|\theta)])$ which we hope $= \theta_0$.

Define $D(\theta;\theta_0) = \mathbb{E}_{\theta_0}(ln[f(X|\theta)])$. We will show that $D(\theta_0;\theta_0) - D(\theta;\theta_0) \geq 0 \forall\theta$.

Kullback-Liebler Divergence: Let f_0 and f_1 be any two PDFs/PMFs. The Kullback-Liebler divergence from f_0 to f_1 is $K(f_0, f_1) = -\mathbb{E}_{f_0}[ln\frac{f_1(x)}{f_0(x)}]$

$D(\theta_0;\theta_0) - D(\theta;\theta_0) = \mathbb{E}_{\theta_0}[ln(f(X|\theta_0)) - ln(f(X|\theta))] = -\mathbb{E}_{\theta_0}[ln\frac{f(X|\theta)}{f(X|\theta_0)}]$.

Lemma: For any two PDFs/PMFs $f_0, f_1, K(f_0, f_1) \geq 0$, with equality iff $f_0 \equiv f_1$.

Remeber Jensen's Inequality: When $g(x)$ is convex (happy), $\mathbb{E}[g(x)] \geq g(\mathbb{E}[x])$.

Proof (Discrete Case): Suppose $X \sim f_0$ and set $Z = \frac{f_1(x)}{f_0(x)}$. Let $S_j = \{x : f_j(x) > 0\}$

Since $g(z) = -ln(z)$ is convex and $\mathbb{E}_{f_0}(z) = \sum\limits_{x \in S_0} \frac{f_1(x)}{f_0(x)} f_0(x) = \sum\limits_{x \in S_0} f_1(x) \leq 1$.

By Jensen's Inequality: $K(f_0, f_1) = -\mathbb{E}_{f_0}[ln(Z)] = \mathbb{E}_{f_0}[g(z)] \underset{1}{\geq} g(\mathbb{E}_{f_0}(z)) \underset{2}{\geq} 0$.

This is only 'equal' when g is linear. Since $g(z)$ is not linear, equality in 1 only happens iff $Z = \frac{f_1(x)}{f_0} = c \neq 0, \forall x \in S_0, [S_0 \subset S_1]$. Equality in 2 only happens iff $\sum\limits_{x \in S_0} f_1(x) = 1, [S_1 \subset S_0]$.

Suppose 1 and 2 are equalities. $1 = \sum\limits_{x \in S_1} f_1(x) = \sum\limits_{x \in S_1} c f_0(x) = c \sum\limits_{x \in S_1} f_0(x) = c \sum\limits_{x \in S_0} f_0(x) = c$