Stat 2911 Lecture Notes

Class 11, 2017

Uri Keich

Method of moments estimation: definition and examples, MLE: definition (Rice 8.1, 8.2-8.4)

Estimation

Given observations (sample) $x_1, \ldots, x_n$ from a dist. $F(\theta)$, where θ is an unknown parameter, we are interested in estimating θ.

Examples. Bernoulli(p), Poisson(λ), geometric(p), multinomial($m; p_1, \ldots, p_r$), Binomial(m, p).

Assumption: the sample $x_1, \ldots, x_n$ is a particular realization of the iid RVs $X_1, \ldots, X_n \sim F(\theta)$

The Method of Moments

Example. Suppose we are given a sample $x_1, \ldots, x_n$ from a Bernoulli(p) dist.

How would you estimate p?

The method of moments works by noting that

$$\mu_1 = E(\bar{X}_n) = p$$

so rather than estimating p we estimate μ_1!

It's actuall not as Homer-ish as it sounds:

p is a parameter specific to the dist. whereas the mean μ_1 is a universal notion so we can instead concentrate on estimating universal quantities.

For example, how would you estimate μ_1 given a sample $x_1, x_2, \ldots, x_n$?

μ_1 can be estimated by the sample mean $\hat{\mu}_1 = \frac{1}{n}\sum_1^n x_i = \bar{x}$

To generalize this we need to define the moments of the dist.

We can extend our definition of L^1 and L^2 to

$L^k = \{ \underline{X} : \Omega \to \mathbb{R} \text{ is a RV with } \sum_i |x_i|^k p_{\underline{X}}(x_i) < \infty \}$ $(k \in \mathbb{N}_1)$

Note: $\underline{X} \in L^k \iff E|\underline{X}|^k < \infty \iff |\underline{X}|^k \in L^1$

For $\underline{X} \in L^k$ we define its k^{th} moment as

$$\mu_k = E(\underline{X}^k) = \sum_i x_i^k p_{\underline{X}}(x_i) \qquad \text{e.g. } \mu_1 = E(\underline{X})$$

Clearly, the moments are property of the dist. of $\underline{X}$ rather than of $\underline{X}$ itself: the RHS is defined in terms of $p_{\underline{X}}$.

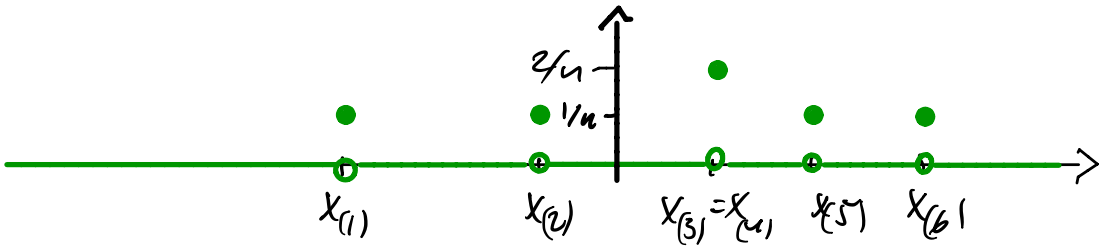
Hence we can just as well refer to the moments of the dist. or the pmf (for discrete RVs).

The moments can be estimated from the sample $x_1, \ldots, x_n$ by the sample moments:

$$\hat{\mu}_k := \frac{1}{n} \sum_i x_i^k$$

Note that the sample moments are the moments of the empirical dist. which is defined by assigning a probability of $1/n$ to each sample point / observation x_i (cumulatively).

The pmf of the empirical dist. associated with the sample $\{x_i\}_1^n$:



$x_{(i)}$ are the ordered sample points (order statistics)

so $\{x_{(i)}\}_1^n = \{x_i\}_1^n$ and $x_{(1)} \le x_{(2)} \le x_{(3)} \cdots \le x_{(n)}$

In the method of moments we express the parameter θ as a function of the moments, which yields an estimate of the parameter by ...

replacing the moments with their estimates: the sample moments.

Examples

1) $X_i \sim$ Bernoulli(p)

$$p = \mu_1 \Rightarrow \hat{p} = \hat{\mu}_1 = \frac{1}{n}\sum_1^n x_i = \bar{x}$$

2) $X_i \sim$ Binom(m, p) m is known but p is not.

$$\mu_1 = E(X_i) = mp \Rightarrow p = \mu_1/m$$

$$\Rightarrow \hat{p} = \quad =$$

Ex. Given a sample $y_1, \dots, y_m$ from a Bernoulli(p) dist. verify that $\hat{p}$ from (1) agrees with $\hat{p}$ from (2) with $x_i = \sum_1^m y_i$

3) $X_i \sim Pois(\lambda)$ $\quad \lambda \geq 0$

$\mu_1 = E(X_i) = \lambda \Rightarrow \hat{\lambda} = \hat{\mu}_1 = \bar{x}$

Recall the Bortkiewicz horse kicks example.

Note: we could have used

$\lambda = \sigma^2 = \mu_2 - \mu_1^2 \Rightarrow \hat{\lambda} = \hat{\mu}_2 - \hat{\mu}_1^2$

Can you guess why that would be an inferior estimator?

It involves the second as well as the first moment, hence its variability would be higher.

4) $X_i \sim$ geometric(p) $\quad p \in [0,1]$

$\mu_1 = E(X_i) = 1/p$

$\Rightarrow p = 1/\mu_1$

$\Rightarrow \hat{p} = 1/\bar{x}$

5) $X_1, \dots, X_n \sim \text{Binom}(m, p)$ where both m and p are unknown

Example: you try to estimate your competitor's production rate and/or failure rate by, say, going through their rejects bin.

Goal: express m and p in terms of moments of lowest possible order.

$\mu_1 = E(X_i) = mp$ $\quad (\mu = \mu_1)$

$\mu_2 - \mu_1^2 = \sigma^2 = mp(1-p)$

$\Rightarrow \sigma^2 = \mu(1-p)$

$\Rightarrow p = 1 - \sigma^2/\mu = \frac{\mu - \sigma^2}{\mu}$ $\qquad m = \mu/p = \frac{\mu^2}{\mu - \sigma^2}$

$\Rightarrow \hat{p} = 1 - \hat{\sigma}^2/\hat{\mu}$ $\qquad \hat{m} = \frac{\hat{\mu}^2}{\hat{\mu} - \hat{\sigma}^2}$

where $\hat{\mu} = \hat{\mu}_1 = \bar{X}$

$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$

$= \frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2$

$\overset{ex}{=} \frac{1}{n}\sum_1^n (X_i - \bar{X})^2$ (sample variance, or variance of the empirical dist.)

Note that the sample variance is often taken as

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_1^n (X_i - \bar{X})^2$$

Can you spot a potential problem with these estimators?

More generally, estimating using the method of moments consists of expressing the moments of the dist. in terms of the unknown parameters $\underline{\theta} = (\theta_1, \dots, \theta_r)$ and then inverting this map to obtain

$$\begin{aligned}\theta_1 &= f_1(\mu_1, \dots, \mu_k)\\ \theta_2 &= f_2(\mu_1, \dots, \mu_k)\\ &\vdots\\ \theta_r &= f_r(\mu_1, \dots, \mu_k)\end{aligned}$$

where $k \geq r$ is minimal and $f_i : \mathbb{R}^k \longrightarrow \mathbb{R}$.

$$\Rightarrow \quad \hat{\theta}_i = f_i(\hat{\mu}_1, \dots, \hat{\mu}_k) \quad \text{where}$$

$$\hat{\mu}_k := \frac{1}{n} \sum_i^n x_i^k$$

Note: for a particular sample $x_1, \dots, x_n$

$$\bar{x} = \sum_1^n x_i \cdot \frac{1}{n} \in \mathbb{R} \quad \text{but}$$

$$\overline{X} = \sum_1^n X_i \cdot \frac{1}{n} \text{ is a RV.}$$

Maximum Likelihood Estimation (MLE)

Given a sample $x_1, \ldots, x_n$ from a dist. with pmf $f_\theta(x)$ we want to estimate θ.

The likelihood function gives the prob. of the observed sample, assuming the sample was generated using f_θ, where θ is known:

$$L(\theta) = P_\theta(X_1 = x_1, \ldots, X_n = x_n)$$

not the prob. of θ!

If the sample is modeled as an iid then

$$L(\theta) = \prod_1^n f_\theta(x_i)$$

Example. $X_i \sim \text{Bernoulli}(\theta) \Rightarrow f_\theta(x) = \begin{cases} 1-\theta & x=0 \\ \theta & x=1 \\ 0 & \text{otherwise} \end{cases}$

In order to analyze $L(\theta)$ it is beneficial to simplify:

$$f_\theta(x) = \theta^x \cdot (1-\theta)^{1-x} \cdot 1_{x \in \{0,1\}}$$

Bernoulli sample $\Rightarrow x_i \in \{0,1\}$

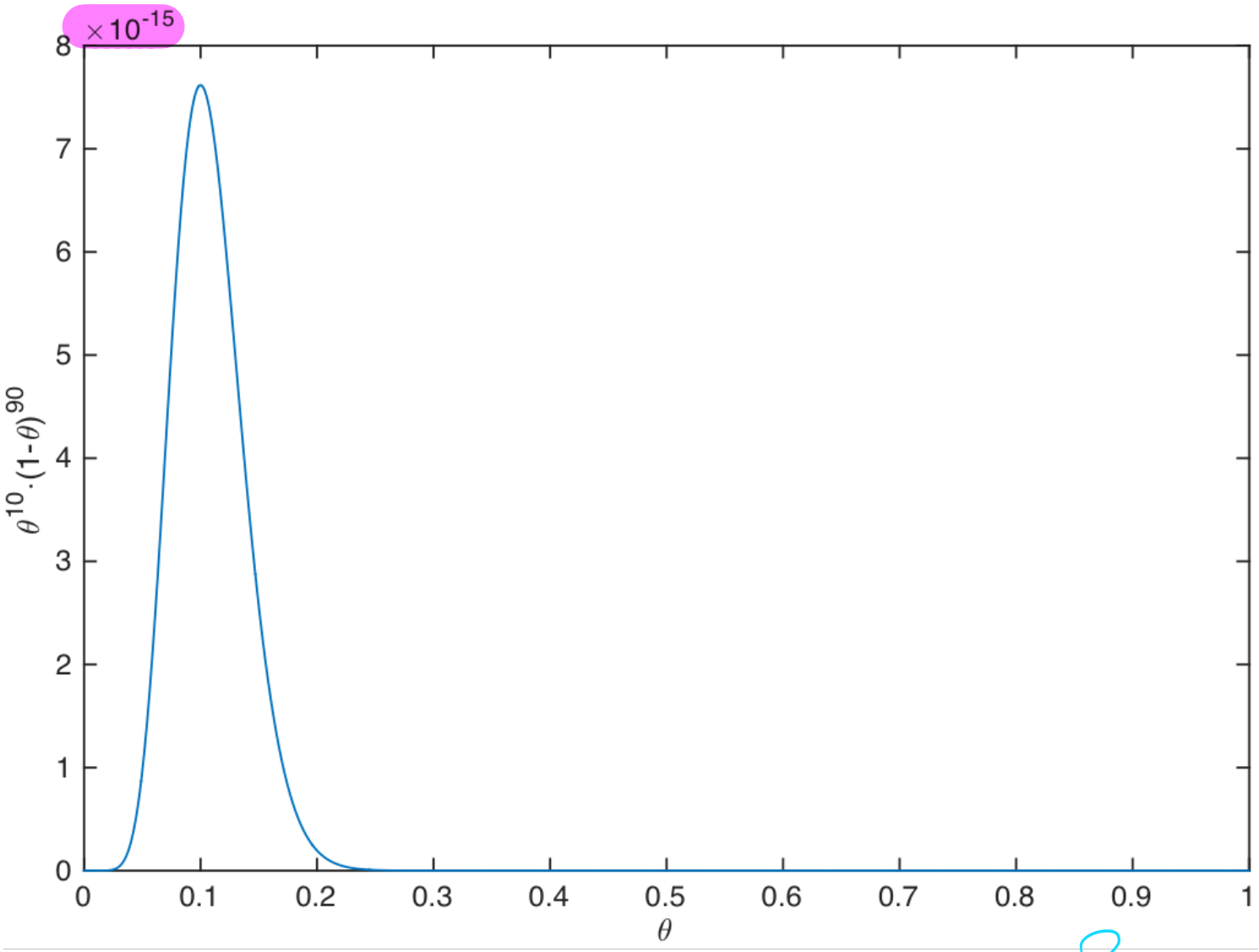
$$\Rightarrow L(\theta) = \prod_1^n \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\sum x_i}\left(1-\theta\right)^{\sum (1-x_i)}$$

$$= \theta^{\sum x_i}\left(1-\theta\right)^{n-\sum x_i}$$

e.g. $n=100$, $\sum_1^n x_i = 10 \Rightarrow L(\theta) = \theta^{10}(1-\theta)^{90}$

$n = 100$, $\sum_1^n x_i = 10$ $\Rightarrow$ $L(\theta) = \theta^{10}(1-\theta)^{90}$



How would you estimate θ?

The MLE is defined as

$$\hat{\theta} = \operatorname{argmax}_\theta L(\theta)$$

Since log is a monotone increasing function

$$\hat{\theta} = \operatorname{argmax}_\theta \ell(\theta),$$

where $\ell(\theta) = \log L(\theta) \overset{\text{iid case}}{=} \sum_1^n \log f_\theta(x_i)$ (more convenient for maximizing!)