36-705

Youtube lecture 28/09/16

04/06/2020

Lecture Notes 7 - Parametric Point Estimation (continued)

- $X_1, \ldots, X_n \sim p(x, \theta)$ $\theta \in \Theta$ - θ can be scalar or vector
  - assume vector here
- $\underline{\theta} = (\theta_1, \ldots, \theta_K)$

2. Method of Moments

$$\mu_1(\theta) = \mathbb{E}[X_i] = \int x \, p(x, \theta) \, dx$$

Sample version :- $m_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$; recall $m_1 \xrightarrow{P} \mu_1(\theta)$ via WLLN

Θ: equate first K sample moments with first K theoretical moments

- Same for 2nd moment

$$\mu_2(\theta) = \mathbb{E}[X_i^2] = \int x^2 p(x, \theta) \, dx$$

Sample version :- $m_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2$; same reasoning

- For k^{th} moment

$$\mu_K(\theta) = \mathbb{E}[X_i^K] = \int x^K p(x, \theta) \, dx$$

Sample version :- $m_K = \frac{1}{n} \sum_{i=1}^{n} X_i^K$, ——— " ———

- We have K terms that depend on θ
- via WLLN, these terms should be 'close' to sample versions
- Equate them to get K equations

Equate $\mu_j(\hat{\theta}) = m_j$ $j = 1, \ldots, K$ and solve

to yield ~~θ^*~~, $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_K)$

(A1): confused by dim.
- Is θ a vector in above form

Example 1

$N(\beta, \sigma^2)$ with $\theta = (\beta, \sigma^2)$.

(*) equate theoretical and sample moments

$\mathbb{E}[X_i] = \beta \qquad \mathbb{E}[X_i^2] = \mathrm{var}(X) - (\mathbb{E}[X_i])^2 = \sigma^2 + \beta^2$

- set:-

$\frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\beta} \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \hat{\sigma}^2 + \hat{\beta}^2$

$\Rightarrow \quad \hat{\beta} = \bar{X} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

LW: only works if you can solve the equations
- when you can solve the equations, you get sample estimators
- Haven't determined whether they are good estimators
- Lots currently an algorithm for generating estimators.

example 2

- Suppose

$X_1, \dots, X_n \sim \mathrm{Binomial}(k, p)$

- where k and p unknown.

we have:-

$kp = \bar{X}_n, \quad \frac{1}{n}\sum_{i=1}^{n} = kp(1-p) + kp^2$

- case where BOTH no. of trials and parameter (success prob. are unknown)
- Flip coin k times; and we only are given X_i
  - no. of heads
    - repeat exp., estimate k and p.

yielding:- typo

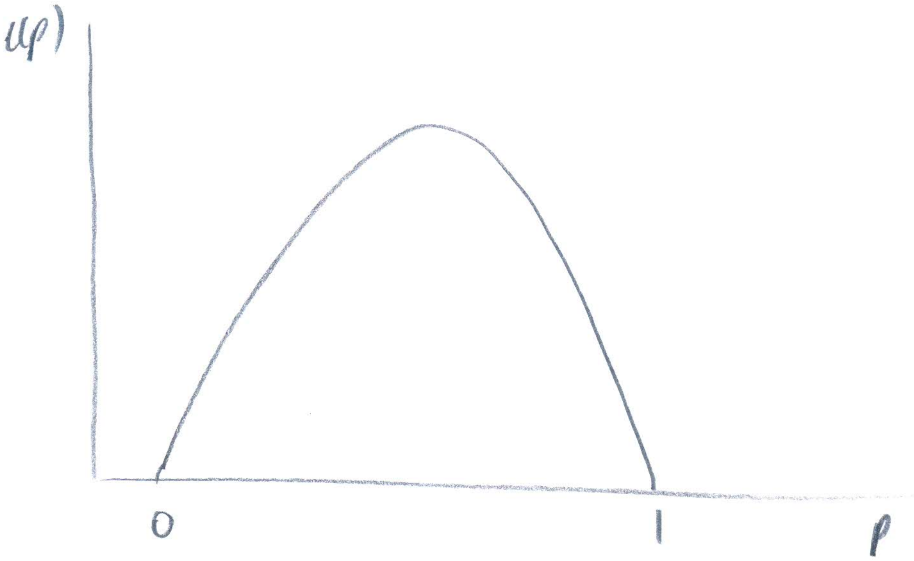
$$\hat{p} = \frac{\bar{X}}{\hat{k}} \qquad \hat{k} = \frac{\bar{X}^2}{\underbrace{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}_{(I)}}$$

LW: Note

(I) can be negative $\Rightarrow$ nonsensical result for estimated coin flips.

- Recall distinction between:-

1) A procedure for generating estimators
2) Evaluating their 'quality'



$\Rightarrow \hat{p} = \bar{X}$

In this case MLE and MOM estimators are the same.

<u>example 4</u>

$X_1, \ldots, X_n \sim N(\mu, 1)$

$$L(\mu, \sigma^2) \propto e^{-\frac{n(\bar{X}-\mu)^2}{2\sigma^2}} e^{-\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{2\sigma^2}}$$

(A1) - verify this; board unclear.

- messy without applying log-trans.

(A2) review calc. ✓

$$\ell(\mu, \sigma^2) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = 0 \qquad \frac{\partial \ell}{\partial \sigma^2} = 0$$

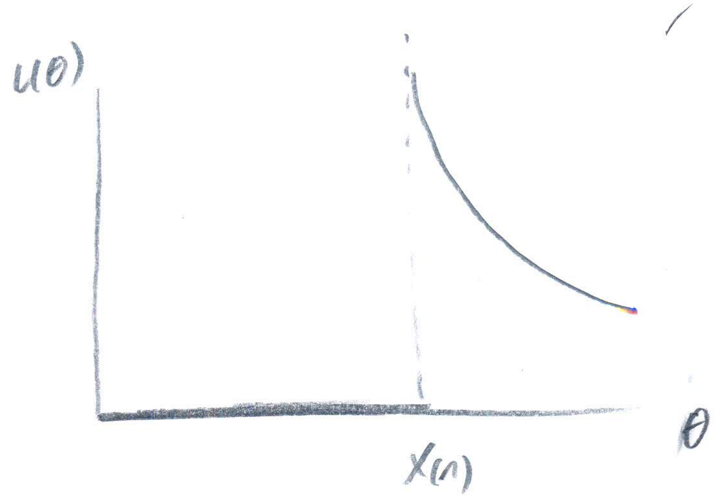
<u>yields</u>: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

<u>example 5</u>

$X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$

$$L(\theta) = \frac{1}{\theta^n} \mathbb{I}(\theta > X_{(n)})$$

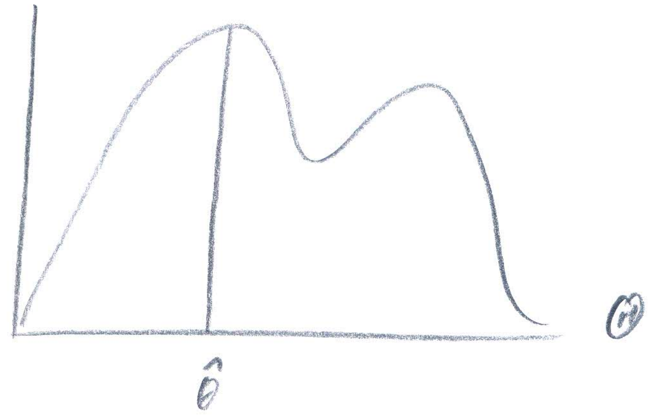
$\hat{\theta} = X_{(n)}$ i.e. $\max(X_1, \ldots, X_n)$



3. Maximum likelihood

- Most common in science journals.
- Systematised by Ronald Fisher

$$L(\theta) = p(X_1, \ldots, X_n; \theta) \overset{iid}{=} \prod_{i=1}^{n} p(X_i; \theta)$$



- $\hat{\theta}$ is the point that maximises $L(\theta)$.
- $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$

(*) $\hat{\theta}$ depends on likelihood function $L(\theta)$, which depends on data $X_1, \ldots, X_n$.

(*) Numerically easier to maximise:-

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

$\log(\cdot)$ - monotone fn

$$\ell(\theta) = \log L(\theta)$$

(*) Certain very desirable properties that ML yields an estimator.

- often solved via:-

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = 0 \quad j = 1, \ldots, k$$

(for sufficiently 'nice' functions)

- classroom settings $\rightarrow$ analytic sol
- Real world $\rightarrow$ numerical methods. (10-725 convex opt.)

Example - Bernoulli

$X_1, \ldots, X_n \sim Ber(p)$

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S (1-p)^{n-S} \qquad S = \sum_{i=1}^{n} X_i$$

$$\ell(p) = S \log(1-p) + (n-S)(1-p)$$

- MoM estimator (for Unif(0, θ))

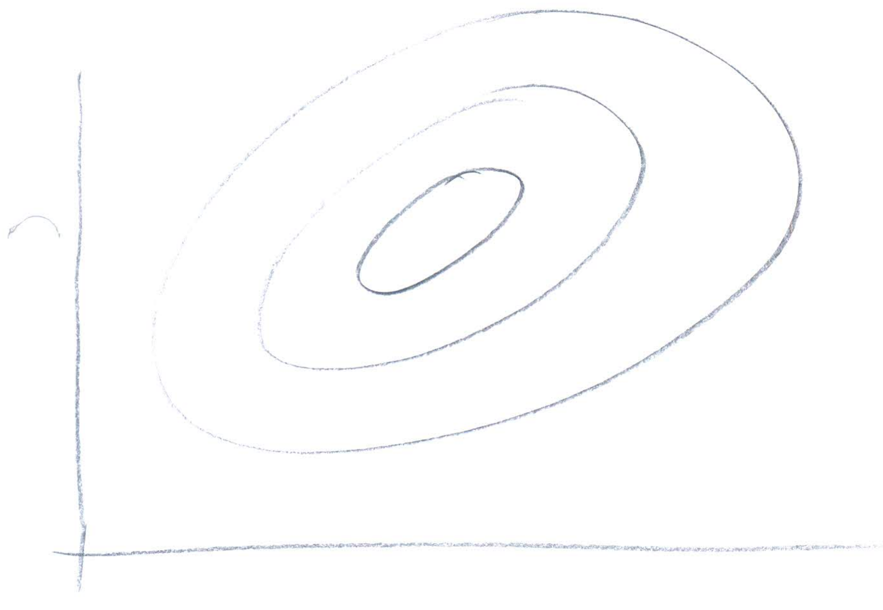
$E[X_i] = \frac{\theta}{2}$ set $\frac{\theta}{2} = \bar{X} \Rightarrow \hat{\theta} = 2\bar{X}$

(*) MLE and MoM yield distinct estimators → how to assess?

(*) An important property of MLE.
- introduce some terminology to disclose property.
- suppose $\theta = (\eta, \xi)$
- $L(\theta)$. $\hat{\theta} = (\hat{\eta}, \hat{\xi})$ (analogous to $\hat{\mu}, \hat{\sigma}$ for normal case) incorrectly,
- focus on likelihood function as function over one parameter 'at a time'
- Profile likelihood :- $L(\eta) = \sup_{\xi} L(\eta, \xi)$ - maximum over the other parameter
(for η)

- Note :- $\hat{\eta} = \arg\max_{\eta} L(\eta) = \arg\max \left(\sup_{\xi} (\eta, \xi) \right)$

(A3) - Fill in details of geometric interpretation of maximisation



(*) overall maximiser of $L(\theta)$ and maximiser of profile likelihood are equivalent

i.e. MLE has a 'good' property called equivariance

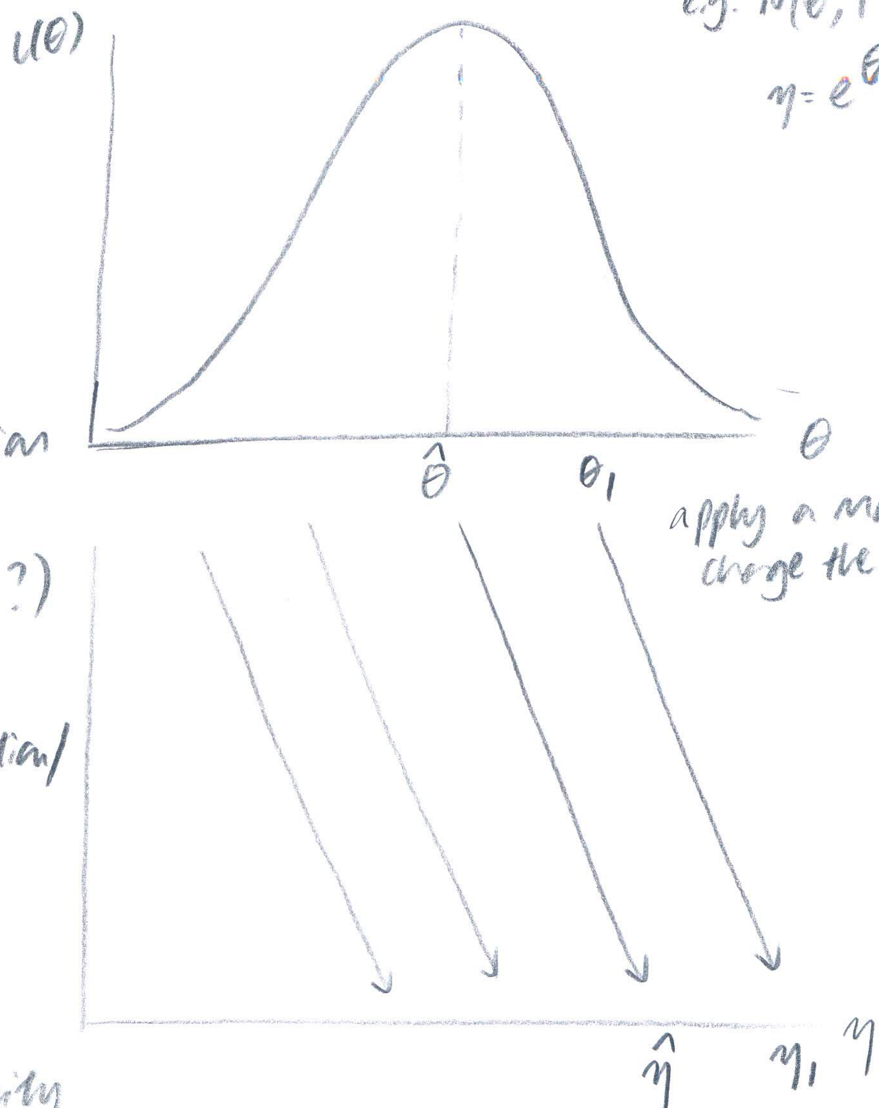
(*) suppose I have an arbitrary function of θ :- AB

$\eta = g(\theta)$

$\hat{\eta} = g(\hat{\theta})$

e.g. $N(\theta, 1)$

$\eta = e^{\theta}$



apply a mapping/ change the scale.

· nothing stopping us from indexing distri with a suitable transformation

- likelihood is just density fn of η (or $\hat{\eta}$?)

- Does not have to be a monotonic transformation/ one-to-one

(AB) need to review for clarity if example doesn't clear it up.

---

example - Bernoulli

$\hat{p} = \bar{X}$

noting $p = \dfrac{e^{\psi}}{1+e^{\psi}}$

$\psi = \log\left(\dfrac{p}{1-p}\right)$

(*) For MLE; via equivariance, $\hat{\psi} = \log\left(\dfrac{\hat{p}}{1-\hat{p}}\right)$

(*) If you took original likelihood, reexpressed it in terms of ψ; we can effect a substitution

(*) useful if we are interested in a function of the parameter we are trying to estimate.

equivariance example

example - normal

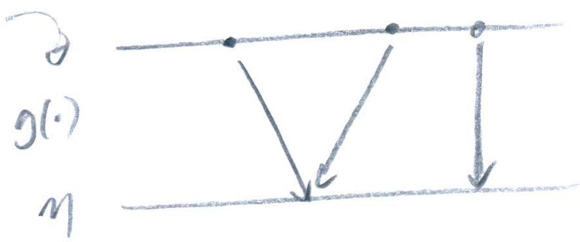
$N(\mu,\sigma^2)$ $\quad \hat{\mu} = \bar{X}$ $\quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$

· via equivariance

$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$ $\quad \widehat{\frac{\mu}{\sigma}} = \frac{\hat{\mu}}{\hat{\sigma}}$ (plug in).

(*) Argument for equivariance when the transformation is not one-to-one

(P4): logic of equivariance, MLE - review



---

4. Bayes Estimator

- this is not Bayesian inference
- currently, view this as another algorithm for generating an estimator
- against frequentist ass. of course so far and ~~as~~

(*) Assume that the parameter θ is a random variable

- define a prior distribution $p(\theta)$
- $p(x_1,\dots,x_n;\theta)$ - joint probability of $x_1,\dots,x_n$, parametrized by θ.

(*) treating θ as a random variable $\Rightarrow$ we can condition on it (as it assumed to be r.v.)

$p(x_1,\dots,x_n|\theta)$ - joint prob, condit. on θ.

- consider:- $\quad p(x_1,\dots,x_n|\theta)p(\theta) = p(x_1,\dots,x_n,\theta)$

Compute the posterior via Bayes theorem

$$p(\theta|x_1,\dots,x_n) = \frac{p(x_1,\dots,x_n|\theta)p(\theta)}{p(x_1,\dots,x_n)} = \frac{p(x_1,\dots,x_n|\theta)p(\theta)}{\int p(x_1,\dots,x_n|\theta)p(\theta)\,d\theta}$$

- Recover standard Bayes theorem over r.v, param semantics.

posterior $\propto$ likelihood $\times$ prior

- Have a density function for θ, can extract a point estimator (but could go further).

Bayes estimator $\hat{\theta} = \mathbb{E}(\theta|x_1,\dots,x_n) = \int \theta\, p(\theta|x_1,\dots,x_n)\,d\theta$

(*) Ultimately end up with $\hat{\theta}$, an estimator i.e. function of the data.

LW: Prefer to see this as purely an algorithm to get an estimator; and not to discuss/endow the semantics of Bayesian formalism with a deeper meaning unworthy.

Example 7

- $X_1,\dots,X_n \sim Bern(\theta)$ $\quad L(\theta) = \theta^S(1-\theta)^{n-S}$ $\quad S = \sum_{i=1}^{n} X_i$

PS: can't see board.

- select prior:-

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{where} \quad \Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,dt$$

- α,β hyperparam; Bayes est in this case is an infinity of possible est.

(*) Beta distribution family

- Posterior:

$$p(\theta|X_1,\dots,X_n) \propto \theta^{S+\alpha-1} \cdot (1-\theta)^{n-S+\beta-1}.$$

Note functional form similarities

(*) Note functional form similarity of prior and posterior

(*) Hence posterior is a beta distribution $Beta(\alpha+S, n-S+\beta)$

$\hat{\theta}_{BAYES} = \mathbb{E}[\theta | X_1, \ldots, X_n]$ (A5) Review

$$= \frac{S+a}{\alpha+\beta+n} \quad \text{②}$$

$$= (1-\lambda)\hat{\theta}_{MLE} + \lambda \bar{\theta}$$

where $\bar{\theta} = \dfrac{\alpha}{\alpha+\beta}$ $\quad \lambda = \dfrac{\alpha+\beta}{\alpha+\beta+n}$ $\quad \hat{\theta}_{MLE} = \dfrac{S}{n}$

(mean of θ wrt prior distn) in this context.

(*) Bayes estimator, is a convex combination of MLE estimator and the mean of prior

(*) Note collapsement of $\hat{\theta}_{BAYES}$ to $\hat{\theta}_{MLE}$ and $\bar{\theta}$

LW: (6.1) Note that at the end of all 3 algorithms, we end up with an estimator:-

$$\hat{\theta} = g(X_1, \ldots, X_n)$$