36-705

Lecture Notes 7 - Point Estimation - Review

- Basic setup

$X_1, \dots, X_n \sim p(x;\theta)$. Estimate $\theta = (\theta_1, \dots, \theta_K)$ via an estimator from data:-

$$\hat{\theta} = \hat{\theta}_n = w(X_1, \dots, X_n)$$

(i): param - fixed, unknown (frequentist)
estimator - r.v. as a deterministic function of random data $(X_1, \dots, X_n)$

(*) Always found this notational point tricky:-

$$\mathbb{E}_\theta(\hat{\theta}) = \int \cdots \int \hat{\theta}(x_1, \dots, x_n) p(x_1;\theta) p(x_2;\theta) \dots p(x_n;\theta)\, dx_1 \dots dx_n$$

- expectation of an estimator under a probability distribution parametrised by value of parameter at its true value θ.

(ii): notion of sampling distribution

- An estimator $\hat{\theta}_n$ is a random variable.
- It has a distribution (like any other r.v.) with mean and variance.

(iii): consistency of an estimator:-

$\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$ (convergence in probability of an estimator to true parameter value)

i.e. $P(|\hat{\theta}_n - \theta| > \epsilon) \xrightarrow{n \to \infty} 0$

2. MOM

- equate k sample moments with k theoretical moments.
- solve to get $\hat{\theta}_{MOM} = (\hat{\theta}_1, \dots, \hat{\theta}_K)$
- sample moments $m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ · Note $m_k \xrightarrow{P} \mu_k(\theta)$ via WLLN.
- Theoretical moments $\mu_k(\theta) = \mathbb{E}[X_i^k]$
- yielding MoM estimator for k moments

(*) some -

(*) some uncertainty on how to go about applying MoM estimator.
- we need to posit a distn family right? (i.e. statistical model)
- Yes, have to posit a distn family with unknown parameters.
- i.e. assume DGP of $X_1, \ldots, X_n$ is [insert family]
- This is a family of densities/distn's indexed by parameters
- We use the data to estimate the parameters via an estimator
- The procedure for generating estimators are covered in the notes

---

3. Maximum likelihood

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ L(\theta) \qquad L(\theta) = p(X_1, \ldots, X_n; \theta) \overset{IID}{=} \prod_{i=1}^{n} p(X_i; \theta)$$

$$= \underset{\theta}{\text{argmax}}\ \ell(\theta) \qquad \ell(\theta) = \log L(\theta).$$

---

(11) - $L(\mu, \sigma^2)$ was derived earlier.
- use cross-term expansion trick with $\bar{x}$
- constant of prop can be discarded as likelihoods equivalent up to a constant of proport.

(*) Equivariance and profile likelihood

- Profile likelihood
- standard likelihood:-

$$L(\theta) = p(x_1, \ldots, x_n; \theta) \overset{IID}{=} \prod_{i=1}^{n} p(X_i; \theta)$$

$\hat{\theta}_{MLE}$ obtained by solving $\dfrac{\partial \ell(\theta)}{\partial \theta_j} = 0 \quad j = 1, \ldots, K$

- Profile likelihood:-
- Partition parameters:- $\theta = (\eta, \xi)$

$$\text{e.g. } \theta = (\underbrace{\theta_1, \theta_2, \ldots, \theta_m}_{\eta}, \underbrace{\theta_{m+1}, \ldots, \theta_K}_{\xi})$$

Then $L(\theta) = L(\eta, \xi)$

profile likelihood for η is likelihood maximised wrt to the other parameter.

That is:- $L(\eta) = \sup_{\xi} L(\eta, \xi)$

- $\hat{\eta}_{MLE} = \underset{\eta}{\arg\max}\, L(\eta) = \underset{\eta}{\arg\max} \left\{ \sup_{\xi} L(\eta, \xi) \right\}$

- we can therefore find

$$\hat{\theta}_{MLE} = \underset{\theta}{\arg\max}\, L(\theta)$$

OR

$$\hat{\xi}_{MLE} = \underset{\xi}{\arg\max}\, L(\eta, \xi)$$

$$\hat{\eta}_{MLE} = \underset{\eta}{\arg\max}\, L(\eta)$$

(*) equivariance of MLE

- If $\eta = g(\theta)$ (i.e. an arbitrary function of parameter)

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

- suppose g is invertible so $\eta = g(\theta)$ and $\theta = g^{-1}(\eta)$
- define $L^*(\eta) = L(\theta)$ where $\theta = g^{-1}(\eta)$
- so for any η:-

$$L^*(\hat{\eta}) = L(\hat{\theta}) \geq L(\theta) = L^*(\eta)$$

i.e. value of parameter

- why? Because $\hat{\theta}$ maximises likelihood (it is MLE).

0/S 1 - review C&B

- Hence $\hat{\eta} = g(\hat{\theta})$ maximises $L^*(\eta)$
- for non-invertible functions (?) ; this is still true if we define

$$L^*(\eta) = \sup_{\theta : \tau(\theta) = \eta} L(\theta)$$

(i.e. profile likelihood) ?

- 9.14 (Theorem) - Wasserman

- let $\tau = g(\theta)$ be a function of θ
- let $\hat{\theta}_n$ be the MLE of θ.
- Then $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of τ.

PROOF

- let $h = g^{-1}$ denote the inverse of g
- Then $\hat{\theta}_n = h(\hat{\tau}_n)$
- for any τ; $L(\tau) = \prod_{i=1}^{n} f(x_i; h(\tau)) = \prod_{i=1}^{n} f(x_i; \theta) = L(\theta)$

  where $\theta = h(\tau)$

- Hence, for any τ, $L_n(\tau) = L(\theta) \leq L(\hat{\theta}) = L_n(\hat{\tau})$

$\blacksquare$

---

4. Bayes Estimator

- move to Bayesian worldview, only for purposes of generating estimator
- treat θ as r.v.

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

Bayes est.: $\hat{\theta}_{BAYES} = \mathbb{E}[\theta | x^n] = \int \theta p(\theta | x^n) \, d\theta$

Example 7

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

Prior: $\theta \sim Beta(\alpha, \beta)$

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} \, dt$$

Posterior:

$$p(\theta|x^n) \propto p(x^n|\theta)p(\theta)$$

$$\Rightarrow p(\theta|x^n) \propto \theta^S(1-\theta)^{n-S}\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$ (omitting norm constant.)

$$\propto \theta^{S+\alpha-1}(1-\theta)^{n-S+\beta-1}$$ (dropping terms with no θ).

Hence, with appropriate normalisation :-

$$\theta|x^n \sim Beta(S+\alpha, n-S+\beta)$$

Bayes estimator: $\mathbb{E}[\theta|x^n] = \int \theta\, p(\theta|x^n)\, d\theta$

- in our case, we're looking for mean of the Beta distri.
- For $Beta(\alpha,\beta)$, mean is $\frac{\alpha}{\alpha+\beta}$.

Hence $\hat{\theta} = \mathbb{E}[\theta|x^n] = \frac{S+\alpha}{(S+\alpha)+(n-S+\beta)} = \frac{S+\alpha}{\alpha+\beta+n}$

- Note $\hat{\theta} = \frac{S+\alpha}{\alpha+\beta+n} = \lambda\bar{\theta} + (1-\lambda)\hat{\theta}_{MLE}$

$$\lambda = \frac{\alpha+\beta}{\alpha+\beta+n} \qquad \bar{\theta} = \frac{\alpha}{\alpha+\beta}$$

(*): Properties of MoM, MLE → cover in LN9 (from Wasserman) O/S 2

---

example 8

(A1) $X_1,\dots,X_n \sim N(\mu,\sigma^2)$ σ^2 known

Assume conjugate prior on μ

i.e. $p(\mu) \sim N(m,\tau^2)$ m,τ^2 fixed and

- <u>Posterior</u>: $p(\mu|x_1,\ldots,x_n) = \dfrac{p(x_1,\ldots,x_n|\mu)p(\mu)}{\int p(x_1,\ldots,x_n|\mu)p(\mu)\,d\mu}$

- drop norm. constant.

$$\propto p(x_1,\ldots,x_n|\mu)p(\mu)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \frac{1}{\tau\sqrt{2\pi}} \exp\left\{-\frac{1}{2\tau^2}(\mu-m)^2\right\}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \frac{1}{\tau\sqrt{2\pi}} \exp\left\{-\frac{1}{2\tau^2}(\mu-m)^2\right\}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\tau\sqrt{2\pi}}\right) \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} \cdot$$

$$\exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\} \exp\left\{-\frac{1}{2\tau^2}(\mu-m)^2\right\}$$

- We drop the normalisation constants, i.e. any terms not containing μ.

$$\Rightarrow p(\mu|x_1,\ldots,x_n) \propto \exp\left\{\frac{-n}{2\sigma^2}(\bar{x}-\mu)^2\right\} \exp\left\{-\frac{1}{2\tau^2}(\mu-m)^2\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(\frac{n(\bar{x}-\mu)^2}{\sigma^2} + \frac{(\mu-m)^2}{\tau^2}\right)\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(\frac{\tau^2 n(\bar{x}-\mu)^2 + \sigma^2(\mu-m)^2}{\sigma^2\tau^2}\right)\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left(\tau^2 n(\bar{x}-\mu)^2 + \sigma(\mu-m)^2\right)\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left(\tau^2 n\bar{x}^2 - 2\tau^2 n\bar{x}\mu + \tau^2 n\mu^2 + \sigma^2\mu^2 - 2\sigma^2 m\mu + \sigma^2 m^2\right)\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left((\tau^2 n+\sigma^2)\mu - 2(\tau^2 n\bar{x}+\sigma^2 m)\mu + (\tau^2 n\bar{x}^2+\sigma^2 m^2)\right)\right\}$$

· completing the square formula for an arbitrary quadratic:-

$$ax^2+bx+c = a\left(x+\frac{b}{2a}\right)^2 + \left(c-\frac{b^2}{4a}\right)$$

· we complete square in μ, setting:-

$$a=(\tau^2 n+\sigma^2) \quad b=-2(\tau^2 n\bar{x}+\sigma^2 m) \quad c=(\tau^2 n\bar{x}^2+\sigma^2 m^2)$$

yielding:

$$(\tau^2 n+\sigma^2)\mu^2-2(\tau^2 n\bar{x}+\sigma^2 m)\mu+(\tau n\bar{x}^2+\sigma^2 m^2)$$

$$=(\tau^2 n+\sigma^2)\left(\mu-\frac{2(\tau^2 n\bar{x}+\sigma^2 m)}{2(\tau^2 n+\sigma^2)}\right)^2+\left(\tau^2 n\bar{x}^2+\sigma^2 m-\frac{4(\tau^2 n\bar{x}+\sigma^2 m)^2}{4(\tau^2 n+\sigma^2)}\right)$$

$$p(\mu|x_1,\dots,x_n)\propto \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left[(\tau^2 n+\sigma^2)\left(\mu-\frac{\tau^2 n\bar{x}+\sigma^2 m}{\tau^2 n+\sigma^2}\right)^2+\left(\tau^2 n\bar{x}^2+\sigma^2 m-\frac{(\tau^2 n\bar{x}+\sigma^2 m)^2}{(\tau^2 n+\sigma^2)}\right)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2\left(\frac{\sigma^2\tau^2}{\tau^2 n+\sigma^2}\right)}\left(\mu-\frac{\tau^2 n\bar{x}+\sigma^2 m}{\tau^2 n+\sigma^2}\right)^2\right\}\exp\left\{-\frac{1}{2\sigma^2\tau^2}\left(\tau^2 n\bar{x}^2+\sigma^2 m-\frac{(\tau^2 n\bar{x}+\sigma^2 m)^2}{\tau^2 n+\sigma^2}\right)\right\}$$

· Discard right hand exp term (contains no μ terms).

Hence:-

$$p(\mu|x_1,\dots,x_n)\propto \exp\left\{-\frac{1}{2\left(\frac{\sigma^2\tau^2/n}{\tau^2+\sigma^2/n}\right)}\left(\mu-\frac{\tau^2 n\bar{x}+\sigma^2 m}{\tau^2 n+\sigma^2}\right)^2\right\}$$

- Hence the posterior of the mean parameter is Normal, with

$$\mathbb{E}[\mu|x_1,\dots,x_n]=\frac{\tau^2 n}{\tau^2 n+\sigma^2}\bar{x}+\frac{\sigma^2}{\tau^2 n+\sigma^2}m$$ (convex comb of MLE sample mean / prior mean)

$$\mathrm{var}[\mu|x_1,\dots,x_n]=\frac{\sigma^2\tau^2/n}{\tau^2+\sigma^2/n}$$

5. MSE

(clue is in name).

- mean squared error $\mathbb{E}_\theta[(\hat{\theta}-\theta)^2] = \int \cdots \int (\hat{\theta}(x_1, \ldots, x_n) - \theta)^2 p(x_1;\theta) \ldots p(x_n;\theta)\, dx_1 \ldots dx_n$

- Bias $= \mathbb{E}_\theta(\hat{\theta}) - \theta$

- variance $V = Var_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta(\hat{\theta}))^2]$

(n2) - expectation wrt joint distn (that generated the data), not over a distn for θ!

$$\mathbb{E}_\theta[(\hat{\theta}-\theta)^2] = \int (\hat{\theta}(x_1, \ldots, x_n) - \theta)^2 p(x^n, \theta)\, dx^n$$

- IID decomposes joint $p(x^n;\theta)$ into $p(x_1;\theta)p(x_2;\theta) \ldots p(x_n;\theta)$
- $MSE = B^2 + V$

- MSE is a 'metric' (preliminary for evaluating estimators)
- unbiasedness → bias $B = \mathbb{E}_\theta[\hat{\theta}] - \theta = 0 \Rightarrow \mathbb{E}_\theta[\hat{\theta}] = \theta$
- when this occurs then MSE = variance.
- (0151): integrate this with presentation in Bishop and with the various ways of understanding this to get holistic understanding
  - Bishop + interpret. of bias, variance
  - diagram

---

example 10

- note $\hat{S}_n^2 = \frac{n}{n-1} \hat{\sigma}^2_{MLE}$ (correction for biasedness)

- why is

$\mathbb{E}[S_n^2] = \sigma^2 \longrightarrow$ see 1.3.17. (and HW question for proof)

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$

- consider the following MLE :- ⊛ don't forget; this assumes normal Distr.

$$\hat{\mu}_{MLE,MOM} = \bar{X}_n \qquad \hat{\sigma}^2_{MLE} = \frac{1}{n}\hat{\sum_{i=1}}(X_i - \bar{X})^2$$

- $\mathbb{E}[\hat{\mu}_{MLE,MOM}] = \mathbb{E}[\bar{X}_n] = \mu$
- instead of using $\hat{\sigma}^2_{MLE}$ (without bias adjustment); use unbiased sample variance $\hat{S}_n^2$
- $\mathbb{E}[\hat{S}_n^2] = \sigma^2$
- $MSE(\hat{\mu}_{MLE}) = MSE(\bar{X}_n) = B^2 + V = V$ as $B = 0$ (unbiased)

$$= Var_\mu(\bar{X}_n) = \mathbb{E}_\mu[(\bar{X}_n - \mathbb{E}_\mu[\bar{X}_n])^2] = \mathbb{E}[(\bar{X}_n - \mu)^2]$$

$$= \frac{\sigma^2}{n}$$

A

- $MSE(S_n^2) = B^2 + V = V$ as $B = 0$ (unbiased)

$$= Var_{\sigma^2}(S_n^2)$$

$$= \mathbb{E}_{\sigma^2}[(S_n^2 - \mathbb{E}_{\sigma^2}[S_n])^2] = \mathbb{E}[(S_n^2 - \sigma^2)^2]$$

$$= \mathbb{E}[S_n^4 - 2S_n^2\sigma^2 + \sigma^4]$$

$$= \mathbb{E}[S_n^4] - 2\sigma^2\mathbb{E}[S_n^2] + \mathbb{E}[\sigma^4]$$

$$= \mathbb{E}[S_n^4] - 2\sigma^2(\sigma^2) + \sigma^4$$

$$= \mathbb{E}[S_n^4] - \sigma^4 \qquad \text{(also obt. via } Var(S_n^2) = \mathbb{E}[S_n^4] - (\mathbb{E}[S_n^2])^2$$

- Not sure how to compute 4^{th} moment of S_n.
- seems like an appeal to χ^2 distr is used. (or a lot of tedious derivation).
- Going to put this here :-

WIKI:
- Distn of sample variance (itself an r.v.)

WIKI: (also from UNI):-

- For $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$ Note RHS is:- $$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

- $\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$ is an r.v. with χ^2 distri $(n-1)$ degrees of free.

(*) For an r.v. Y with χ^2 distri and m degrees of freedom:-

$$\mathbb{E}[Y] = m \quad Var[Y] = 2m$$

Hence:

$$Var\left[\frac{n-1}{\sigma^2} S_n^2\right] = 2(n-1)$$

- $Var\left[\frac{n-1}{\sigma^2} S_n^2\right] = \left(\frac{n-1}{\sigma^2}\right)^2 Var[S_n^2] = 2(n-1)$

$$\Rightarrow Var_{\sigma^2}(S_n^2) = Var(S_n^2) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{(n-1)}$$

- As required. $\blacksquare$