36-705 03/06/2020

Youtube lecture 26/09/1

[redacted]

- Likelihood function

- A means of generating estimators
- Bayesian inference
- sufficiency

Define: $X^n = (X_1, \dots, X_n)$; joint density $p(x^n; \underline{\theta}) = p(x_1, \dots, x_n; \theta)$

[redacted]

$\theta \in \Theta$

(*) The likelihood function

$L : \Theta \to [0, \infty)$ (maps parameter space to some numbers)

$L(\theta) \equiv L(\theta; x^n) = p(x^n; \theta)$

UW: Treat x^n as fixed, view likelihood as a function of parameters θ

(*) If data is IID then likelihood is:-

$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$ IID case only

(*) If θ is fixed then $p(x^n; \theta)$ is a probability density for x

(*) If $x^{(n)}$ is fixed and view $L(\theta)$, it is no longer a probability, but a function of θ.

(*) log-likelihood function

$\ell(\theta) = \log L(\theta)$ (TO) - Review example 1

UW: Probability density fn and likelihood fns live on different space

- Likelihood functions are an equivalence class of functions (defined up to a constant of proport.)

Example 2

- $X_1, \dots, X_n \sim N(\mu, 1)$

$L(\mu) = \prod_i \frac{1}{(2\pi)^{1/2}} \exp^{-\frac{1}{2}(x_i - \mu)^2} = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$

- Add and subtract $\bar{x}$; throw away non μ terms yielding $L(\mu) \propto \exp\{-\frac{n}{2}(\bar{x}-\mu)^2\}$

(A1): Review the proport. calculation

(*) This is the likelihood function, a function of μ, not x.

LW: Also looks like normal p.d. centred at $\bar{x}$; in this case then they are the same; but in general, likelihood and p.d. are distinct

example 3

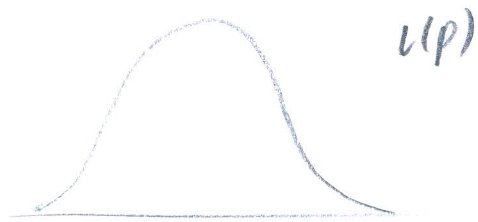
- Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

LW: Reversing roles; fix X at obs. values $X = x$ ... and request prob. of data as function of p.

(*) $L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \propto p^X (1-p)^{n-X}$

where $X = \sum_{i=1}^{n} x_i$ (sum of coin flips)

LW: sample space - sequences of 0s and 1s
- Parameter space - [0, 1]
- Discrete probability function but
- continuous likelihood function

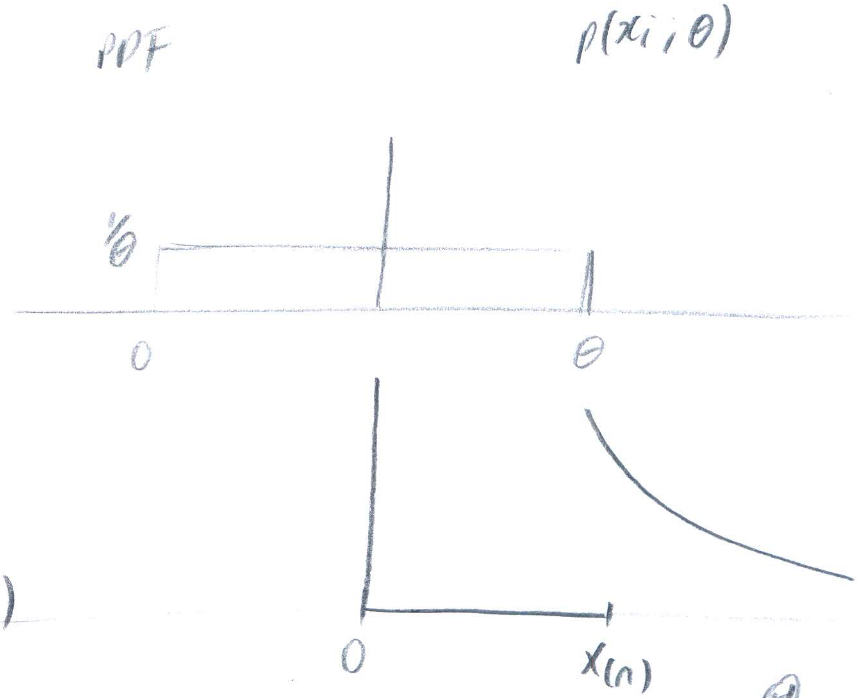


example 5

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

LW: Good way to think about likelihood function.

For every possible θ, what is probability of getting a particular dataset (for this) a particular θ)

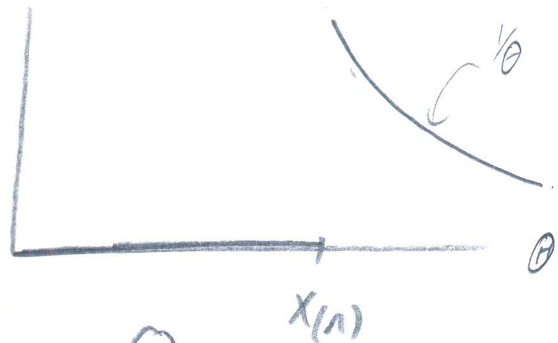


$L(\theta) = \prod_{i=1}^{n} p(X_i;\theta)$

$X_{(n)}$ - maximum observation.

$$L(\theta) = \begin{cases} 0 & \theta < X_{(n)} \\ \left(\frac{1}{\theta}\right)^n & \theta > X_{(n)} \end{cases}$$

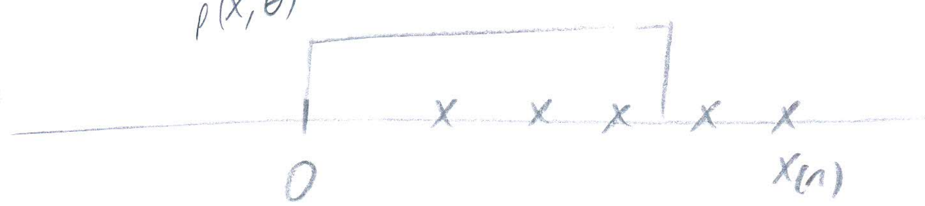
$L(\theta) =$



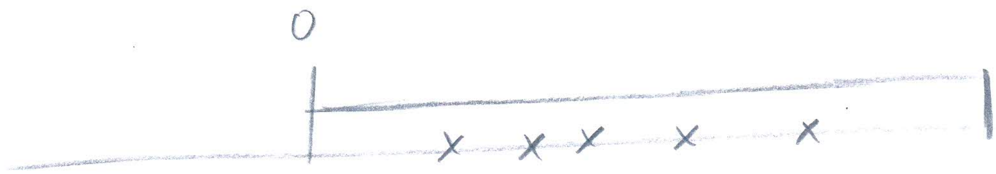
(?) Found this tricky; don't understand intuition (A2)

LW: If $\theta \in [0, X_{(n)}] \rightarrow$ could not have gotten dataset

$\theta \in [X_{(n)}, \infty) \rightarrow$ non-zero probability of getting dataset (at that value of θ)



$\theta < X_{(n)}$



$\theta > X_{(n)}$

- okay; a little clearer intuitively

(*) Note distinction between PDF and likelihood.

- notation: - likelihood also written $L_{X_1, \ldots, X_n}(\theta)$ to remind reader of dataset

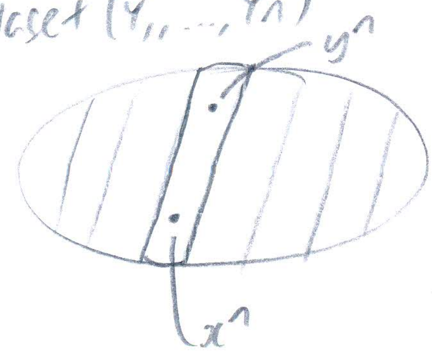
(*) Two likelihoods are the same if they are proportional → gives partition.

$(X_1, \ldots, X_n) \sim (Y_1, \ldots, Y_n)$ if $L_{X_1 \ldots X_n}(\theta) \propto L_{Y_1 \ldots Y_n}(\theta)$

- 2 datasets are in the same partition of a partitioned sample space if the likelihood of dataset $(X_1, \ldots, X_n)$ is proportional to likelihood of dataset $(Y_1, \ldots, Y_n)$

any constant not depend. on θ

LW:



- Take set of all possible datasets
- create partitions so that datasets are on the same partition if likelihood our 2 datasets are the same shape.

- creates an equivalence class/partition.

(*) This is the minimal sufficient partition

- sufficiency and minimal sufficiency related via likelihood fn

(*) Asking if a statistic is sufficient

↳ can I compute the likelihood function from this statistic?

- YES ⟶ it is a sufficient statistic

e.g. $\bar{X}$ is a sufficient statistic as I can compute likelihood (for Normal)

entire dataset is — " — — " —

(*) minimal sufficiency ⟶ that partition induced by statistic is in a $_\wedge$ sense minimal reduction of data

AND

⟶ in a sense, smallest amount of info needed to construct likelihood

(*) informally, likelihood function itself is minimal sufficient statistic

LW: likelihood induces a partition, telling you two datasets have the same likelihood; that partition is minimal sufficient

(*) (on) The minimal sufficient statistic has all information you need to compute the likelihood function

- whether likelihood is suitable for $_\wedge$ estimating the parameters ~~to be estimated~~ depends - different question

LW: many likelihoods need to be computed numerically (in practice)

example 6

- $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n \sim N(\underline{\mu}, \Sigma)$ — symm. P.D.

likelihood:

$$L(\underline{\mu}, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_i (\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu})\right)$$

- maximising over μ and Σ is a little nasty.

- consider $l(\mu, \Sigma)$

$$l(\underline{\mu}, \underline{\Sigma}) = -\frac{n}{2}\log|\underline{\Sigma}| - \frac{1}{2}\underbrace{\sum_{i=1}^{n}(\underline{x}_i - \underline{\mu})^T \underline{\Sigma}^{-1}(\underline{x}_i - \underline{\mu})}$$

A4 - Review calc.

$\sum_i (\underline{x}_i - \underline{\mu})^T \underline{\Sigma}^{-1}(\underline{x}_i - \underline{\mu})$ - relaxing vector not.

- Add/subtract $\bar{X}$
- cross prod. 0
- $tr(A) = \Sigma$ diag. for sq.
- $tr(a) = a$
- $tr(ABC) = tr(CAB) = tr(BCA)$

$$= \underbrace{\sum_i (x_i - \bar{X})^T \Sigma^{-1}(x_i - \bar{X})}_{\text{quad. form i.e. scalar}} + n(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

$$= \sum_i tr\{(x_i - \bar{x})^T \Sigma^{-1}(x_i - \bar{x})\} + n(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

$$= \sum_i tr\{\Sigma^{-1}(x_i - \bar{X})(x_i - \bar{x})^T\} + n(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

$$= tr\sum_i\{\Sigma^{-1}(x_i - \bar{X})(x_i - \bar{X})^T\} + n(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

- Σ^{-1} does not contain index i

$$= tr\,\Sigma^{-1}\sum_i (x_i - \bar{X})(x_i - \bar{X})^T + n(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

- $S_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^T$

$$= n\,tr\,\Sigma^{-1}\frac{1}{n}\sum_i (x_i - \bar{X})(x_i - \bar{X})^T + n(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

A5 - Find and correct typo.

$$= n\,tr\,\Sigma^{-1}S_n + n(\bar{X} - \mu)^T \Sigma^{-1}(\bar{X} - \mu)$$

Hence,

$$l(\underline{\mu}, \underline{\Sigma}) = -\frac{n}{2}\log|\underline{\Sigma}| - \frac{n}{2}tr(\underline{\Sigma}^{-1}\underline{S}) - \frac{n}{2}(\underline{\bar{X}} - \underline{\mu})^T \underline{\Sigma}^{-1}(\underline{\bar{X}} - \underline{\mu})$$

- maximising this:-

set $\underline{\mu} = \underline{\bar{X}}$ and $\underline{\Sigma} = \underline{S}$

- HW: why focus on maximisation?
- HW: cover section 2. Likelihood, Sufficiency & Likelihood Principle after we review Lecture Notes 7 (later)

Lecture Notes 7 - Point Estimation

W A6: clarify if r.v is vector or scalar.

1. Introduction

- $X_1, \dots, X_n \sim p(x; \underline{\theta})$. want to estimate $\theta = (\theta_1, \dots, \theta_K)$ given data

(*) Take a function of data i.e. a statistic (but not quite)

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

- $\hat{\theta}$ - indicates estimate of param
- $\hat{\theta}_n$ - reminder of sample size

E1

(*) Distinction between parameter and estimator.

θ - Parameter - fixed unknown number (no randomness)

$\hat{\theta}$ - Estimator - r.v. - has distrib, mean, variance.

(*) Most common estimator construction principles - Parametric point estimation.

1. MOM - method of Moments
2. Maximum likelihood (MLE)
3. Bayesian estimators.

Context

1. - Invented in 1900s → computationally v. simple → resurgence in MLearning
2. - Most common point estimator for parametric models.
   - under certain conditions; it is in a sense, optimal
3. - Bayesian estimators (not Bayesian inference, nets, rule, theorem).

- For some ML models; MLE difficult; MOM easier

Q: How do we know which to use?

- 2 questions
  - i) Method for constructing estimators.
  - ii) Evaluating estimators according to some criteria

(*) Evaluating estimators:-

1. Consistency → if I gave you more data, will your estimator converge in probability to ...

2. Bias-variance
3. Mean squared error
4. *minimax theory $\longrightarrow$ a way of formulating and evaluating estimators wrt optimality
5. Robustness $\longrightarrow$ estimator may be optimal under certain conditions; what if those conditions are relaxed?
- presence of outliers etc. $\rightarrow$ fash. in high-dim statistics

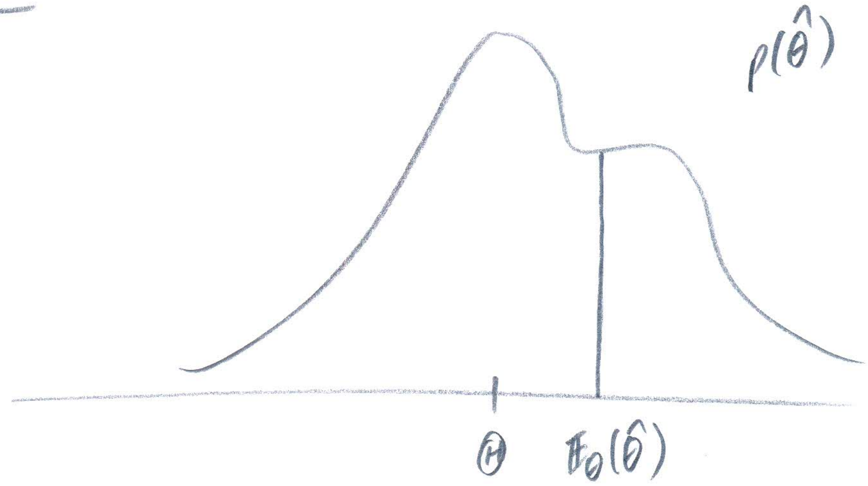
LW: minimax theory
- covered in cursory detail $\rightarrow$ see 36-702

(*) Terminology

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

- remarks: subscript $\theta \not\Rightarrow \theta$ is random - it is a fixed unknown no.
  - informs us that we are putting in a particular value of θ when we compute $\hat{\theta}(x_1, \ldots, x_n)$; when true value is θ.
  - $\hat{\theta}$ is random

2. Bias :- $\mathbb{E}_\theta(\hat{\theta}) - \theta$



3. Sampling distribution
- distribution of $\hat{\theta}_n$

4. Standard error
- standard deviation in special context when r.v. is an estimator

$\sqrt{var(\hat{\theta}_n)} = se(\hat{\theta}_n)$

5. $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$ (large sample/asympt. property)