Lecture 5. Sufficient Statistics. Intro to Estimation

Sufficient statistics

produce smaller well-behaved model

imposing assump.

Parametric family: $X_1, \ldots, X_n \sim$ $i.i.d. F(\cdot, \theta)$. θ is the only unknown parameter

↑ don't know

eg1) $x_1, \ldots, x_n \sim iid \quad N(\mu, \sigma^2) \quad (\theta = \mu, \sigma^2)$

eg2) $x_1, \ldots x_n \sim iid \quad x_i = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$ so $\theta = p$

$X = (x_1, \ldots, x_n)$

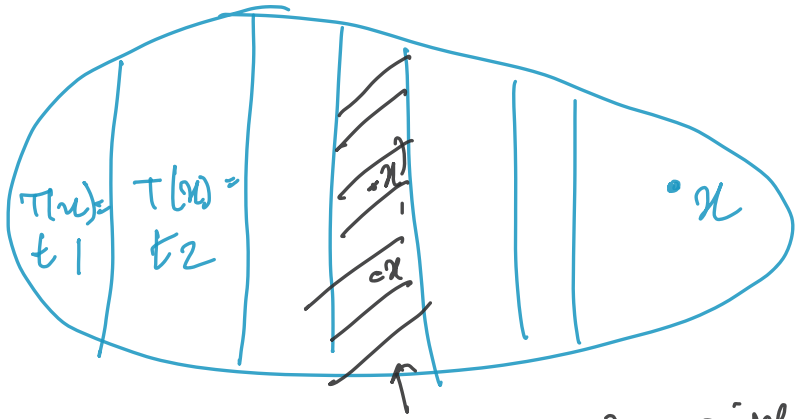
Definition 1.

Statistic $T(X)$ is sufficient for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

$X \mid T(X)$ doesn't depend on θ

Just reporting statistics / summary

one slice - diff values of t



Suppose $T(x) = t$ is given.
What's the distribution of x in slice — doesn't depend on θ.

→ how x is distributed depends on θ

Sufficient statistics

* Joint distbn of $T(x)$ comes from X

Example Let $X = (X_1, ..., X_n)$ be a random sample from $N(\mu, \sigma^2)$. Suppose that σ^2 is known.
$T(X) = \overline{X}_n$ is sufficient for μ.

$X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ (σ^2: known)

$\theta = \mu$

pdf is a product of pdf

1) Figure out distbn of $X | X_n$

Joint distbn of X: $f_X(x_1, .. x_n) = \prod_i f_{x_i}(x_i)$ = Pdf of gaussian

$= \prod \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right\}$

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

$T(X) = \bar{X}_n \sim N(\mu, \sigma^2/n)$ — now figure pdf & suff. stat.

$$f_{T(n)}(t) = \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}} \exp\left\{-\frac{n}{2\sigma^2}(t-\mu)^2\right\}$$

would is ratio of them

$$f_{X|T(X)}(x_1, \ldots, x_n | \bar{x}_n) = \frac{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right\}}{\sqrt{2\pi}\frac{\sigma}{\sqrt{n}} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\}}$$

constant

$$= \text{const} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{2\mu}{2\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2} + \frac{n}{2\sigma^2}\left[\bar{x}^2 - 2\mu\bar{x} + \mu^2\right]\right\}$$

cancel out

& thus doesn't depend on μ.

Factorization Theorem

Theorem 1 (Factorization Theorem).

Let $f(x|\theta)$ be the pdf of X. Then $T(X)$ is a sufficient statistic if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that $f(x|\theta) = g(T(x)|\theta)h(x)$.

dep. on θ — only on data

Eg 1: $X_1, \dots, X_n \sim$ iid $N(\mu, \sigma^2)$ $\theta = (\mu, \sigma^2)$

(pdf of a gaussian)

$$f_X(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right\} =$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{2\mu}{2\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2}\right\}}_{g(T;\theta)} \cdot \underbrace{1}_{h(x)}$$

parameter

data

Suff. st $T(x) = \left(\sum x_i^2, \sum x_i\right)$

eg say: $(s^2, \bar{X}_n)$ ↙ — 1 to 1 correspondence — s^2 derived before in class/ex ⟷ $s^2 = \frac{1}{n-1}\left(\sum x_i^2 - n(\bar{x})^2\right)$

Eg2) Bernoulli(p)

$X = (X_1, \ldots, X_n) \sim$ iid $\begin{bmatrix} 1 & p \\ 0 & 1-p \end{bmatrix}$

Q — what is ~~joint~~ stat. suff.

Step 1:

* write down joint distn

$$f_X(x_1, \ldots x_n) = \prod_{i=1}^{n} f_{x_i}(x_i)$$

$$= \prod p^{x_i} (1-p)^{1-x_i}$$

$$= p^{\sum x_i} (1-p)^{\sum (1-x_i)}$$

Step 2: Suff'ent statistics

$$= p^{\sum x_i} (1-p)^{n - \sum x_i}$$

$T(x) = \sum x_i$

$g(T, \theta) \cdot h(x) = 1$

↑ $f_{x_i}(x_i) = p^{x_i}(1-p)^{1-x_i}$

$x_i = 1 : p^1 (1-p)^0$

$p^1 \cdot 1 = p.$

$x_i = 0 : p^0 (1-p)^1$

$1 \cdot 1$

data shows so is suff. st.

Factorization Theorem: examples

drew from unit interval

Ex3) $X_1, \dots, X_n \sim iid\ \mathcal{U}[\theta, 1+\theta]$

* what is pdf for one obs $f_{x_i}(x_i) = \begin{cases} 1 & x_i \in [\theta, 1+\theta] \\ 0 & \text{otherwise} \end{cases}$

* then joint distn:

$= \mathbb{1}_{[\theta, 1+\theta]}(x_i)$ — indicator fn.

(pdf) $f_x(x) = \prod_i f_{x_i}(x_i) = \begin{cases} 1 & \theta \le \min_i x_i \le \max_i x_i \le 1+\theta \\ 0 & \end{cases}$

points in between not informative,

sufficient statistic $T(x) = (\min_i x_i, \max_i x_i)$

whose pdf is $g(T, \theta)$ & $h(x) = 1$

Factorization Theorem: examples

Estimators. Unbiased estimators

calculated from data

unbiased: if on avg correct

We say that $\hat{\theta} = T(X)$ is *unbiased* for θ if $\mathbb{E}_\theta[T(X)] = \theta$ for all possible values of θ where $\mathbb{E}_\theta$ is the expectation when θ is the true parameter value.

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

$\mathbb{E}\,T(X) = \mathbb{E}\,\hat{\theta} = \theta$

$\text{Bias} = \mathbb{E}\,\hat{\theta} - \theta$

Eg 1) $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ ← unknown parameter

is a discrete that takes values $x \in \{0, 1, 2, 3, \ldots\}$

and probability $P(X_i = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ } all prob. are +ve & sum to 1

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}\cdot e^{\lambda} = 1$$

$$E[X_i] = \sum_{k=0}^{\infty} k \cdot e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k}{(k-1)!}$$

$$= e^{-\lambda} \cdot \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

$$= e^{-\lambda} \cdot \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \qquad j = k-1$$

$$= e^{-\lambda} \cdot \lambda e^{\lambda} = \underline{\lambda}$$

$\hat{\lambda} = \overline{X_n}$ ↑ unbiased

$\hat{\lambda}_1 = X_5$ (or X_i) ↑ unbiased est.

can also do weighted avg.

$\mathrm{Var}(X_i) = \lambda$ also unbiased. ↑

Thus, is a property.

Ex2) $X_i \sim$ iid Bernoulli(p)

$EX_i = p$

$\hat{p} = \frac{\#\{X_i = 1\}}{n} = \frac{1}{n}\sum X_i$

↑ unbiased.

Unbiasness

yes, many cases.

Is it easy to get an unbiased estimator? An example.

eg) $x_1, \dots, x_n \sim$ iid Bernoulli (p)

change parameter to be $\theta = \frac{1}{p}$ (not just p as before)

If I put $\hat{\theta} = \frac{1}{\hat{p}}$ will not unbiased bec

$$E\hat{\theta} = E\frac{1}{\hat{p}} \neq \frac{1}{Ep}$$

face a non-linear relation.

Q) Why not love UB. est.?

Suppose $\hat{\theta} = T(x_1, \dots, x_n)$

$$E(\hat{\theta}) = \sum_{(x_1,\dots x_n) \in \{0,1\}^n} T(x_1, \dots, x_n) \cdot p^{\sum x_i} (1-p)^{n-\sum x_i}$$

↑ polynomial

unbiased $\Rightarrow$ $= \frac{1}{p}$

$\forall p$ ↑ impossible to be equal to $1/p$.

$\frac{1}{p}$ cannot be written as polynomial

Thus, no unbiased est.

Non-linear trans- introduce bias —— and can correct

by introducing Bootstrap bias corr

Bootstrap bias correction

Setup

Parameter of interest $\theta = g(\mu)$, $\mu = \mathbb{E}[X_i]$

Natural estimate $\hat{\theta} = g(\bar{X}_n)$ is biased if g is non-linear.

Bootstrap bias-correction

(1) For each $b = 1, ..., B$ generate $\{X^*_{ib}\}$ from $\{X_1, ..., X_n\}$ with replacement;

(2) Calculate $\overline{X}^*_b = \frac{1}{n}\sum_{i=1}^n X^*_{ib}$;

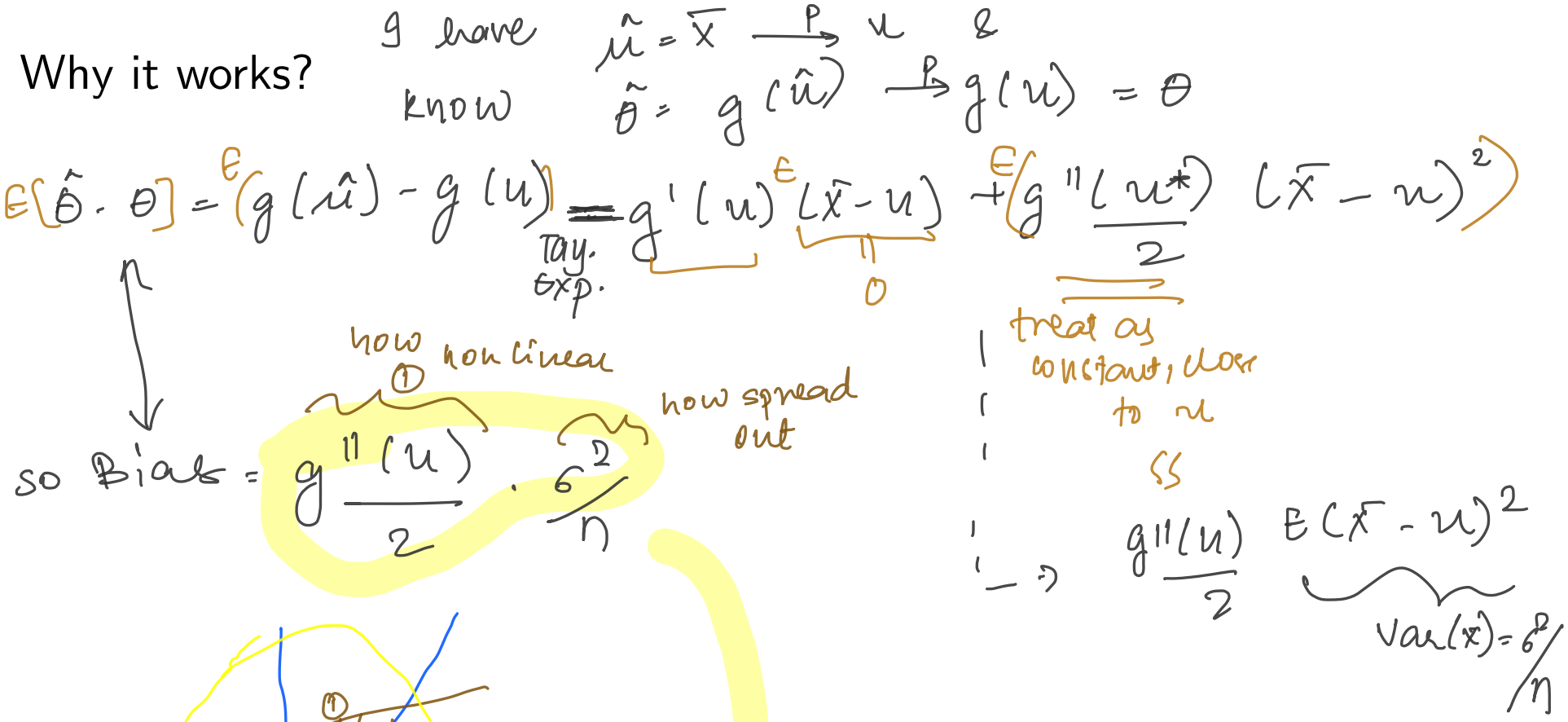
(3) Estimate $\theta^*_b = g(\bar{X}^*_b)$;

(4) $\text{Bias}^* = \frac{1}{B}\sum_{b=1}^B \theta^*_b - \hat{\theta} \approx \text{Bias}$.

(5) Use $\tilde{\theta} = \hat{\theta} - \text{Bias}^*$ as your estimate.

Bootstrap bias correction

Why it works?



my
And Bias* $\simeq \frac{g''(\hat{u})}{2} \cdot \frac{(\sigma^*)^2}{n}$

Many unbiased est.

Estimators. Efficiency

how close the realizes are

$$\underline{\mathrm{MSE}}(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$$

from the truth.

Theorem 2. But don't know, so we have

$$MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + \mathbb{V}ar(\hat{\theta}).$$

$$E[(\hat{\theta} - \theta)^2] \pm E\hat{\theta} = E[(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2)$$

random

$$= E[(\hat{\theta} - E\hat{\theta})^2 + (E\hat{\theta} - \theta)^2 + 2(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta)]$$

$$Var(\hat{\theta}) + Bias^2 + E[(\hat{\theta} - E\hat{\theta}) \cdot (E\hat{\theta} - \theta)]$$

$$= 0$$

θ : pop constant, exists

Trade-off between bias and variance

* $\hat{\theta}$ is some unbiased est. of θ — i.e $E\hat{\theta} = \theta$

Introduce new est. $\tilde{\theta} = \underbrace{(1-c)\tilde{\theta} + c \cdot 0}$ c is small constant

shrink it to close 0.

Mean of $\tilde{\theta}$:

$$E[\tilde{\theta}) = (1-c)E\hat{\theta} + c \cdot 0 = (1-c)\theta$$

$$\text{so bias}(\tilde{\theta}) = E\tilde{\theta} - \tilde{\theta} = -c\theta$$

Now, var of $\tilde{\theta}$

$$\text{Var}(\tilde{\theta}) = \text{Var}((1-c)\hat{\theta}) = (1-c)^2 \text{Var}(\hat{\theta})$$

$$MSE(\tilde{\theta}) = (1-c)^2 \underbrace{\text{Var}(\hat{\theta})}_{\approx \sigma^2} + c^2\theta^2 = g(c)$$ trying to minimize it

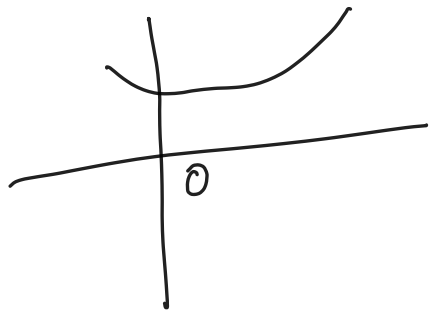
θ

Trade-off between bias and variance

$$g'(c) = -2\sigma^2(1-c) + 2\theta^2 c$$

$g'(0) = -2\sigma^2(1-0) < 0$ is -ve so 0 not an optimal point.

so need more shrinkage.



Connection between efficiency and sufficient statistics

Theorem 3 (Rao-Blackwell).

Assume that $T(X)$ is a sufficient statistic for parameter θ and $\hat{\theta} = \delta(X)$ is an estimator for θ. Define $\hat{\theta}_2 = \mathbb{E}[\delta(X)|T(X)]$ is an estimator for θ as well and $MSE(\hat{\theta}_2) \leq MSE(\hat{\theta})$. In addition, if $\hat{\theta}$ is unbiased, then $\hat{\theta}_2$ is unbiased as well.