8/31/22 (Week 1)

Topic: Review (Ch 1-3)

X is RV w/ sample space $\mathcal{X}$

Random variables (RVs), X say, are defined by their distribution function (CDF).

$$F_X(x) = F(x) = P_r(X \le x)$$

Ex) Discrete

$$X = \begin{cases} -1, & \text{w.p. } 0.2 \\ 0, & \text{w.p. } 0.3 \\ 1, & \text{w.p. } 0.5 \end{cases}$$

$$\mathcal{S} = \{-1, 0, 1\}$$

Continuous

$$X \sim Exp(\tfrac{1}{5})$$

$$\mathcal{S} = (0, \infty)$$

Mixture

Insurance policy reimburse up to some benefit level, C, with some deductible, d.

X = policy holders loss $\sim Exp(\frac{1}{5})$

Y = payout from insurance co.

$$= \begin{cases} 0, & x < d \\ x-d, & d \le x < C+d \\ C, & x \ge C+d \end{cases}$$

$$\mathcal{S}_y = \{0, C\} \cup (0, C) = [0, C]$$

$Pr(Y=a) = 0$ if $a \in [d\ \ c+d)$ so $\longrightarrow$ $f_Y(y) = \frac{1}{5} e^{-\frac{(y+d)}{5}} I\{0 < y < C\}$

The CDF is a probability measure/law so,

Def
1. $P_r(\mathcal{X}) = 1$
2. If $A \subset \mathcal{X}$ then $P_r(A) \ge 0$
3. If mutually disjoint $A_1, A_2, \ldots$

$$P_r\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P_r(A_i)$$

Def: Conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Law of Total Probability

For events $B_1, \ldots, B_n$ s.t. $\bigcup_{i=1}^{n} B_i = \mathcal{X}$

If and $B_i \cap B_j = \emptyset$, for $i \neq j$

and $Pr(B_i) > 0$ for all i

then For any event $A \in \mathcal{X}$

$$Pr(A) = \sum_{i=1}^{n} Pr(A|B_i) \cdot P(B_i)$$

All RVs have a CDF. Many RVs have a density function as well.

$$f_X(x) = f(x) = F_X'(x) = \frac{d}{dx} F_X(x)$$

Def: Likelihood is the density function but viewed as a function of the parameters.

different from textbook

$$f(x;\theta) = f(\theta|x) = L(\theta|x)$$

is a function of θ.

Read as "the likelihood for θ, given $X = x$."

Applying the Law of Total Prob. to jointly distributed RVs, (X, Y) yields:

$$f_Y(y) = \int_{\mathcal{X}} f_{Y|X=x}(y|X=x) \cdot f_X(x)\, dx$$

(in the case where both X, Y are continuous)

9/2/22 (week 1)

Bayes' Rule/Law

– combines law of tot. prob w/ def. of conditional prob.

For $A, B_1, \dots, B_n$ where B_i are disjoint w/ all B_j, $i \neq j$, $\bigcup_{i=1}^{n} B_i = \mathcal{X}$ and $P(B_i) > 0$ for all i,

we have that

$$\Pr(B_j \mid A) = \frac{\Pr(A \mid B_j)\Pr(B_j)}{\sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i)}$$

def. of cond.

$$\frac{\Pr(B_j \cap A)}{\Pr(A)} \xrightarrow[\text{law of tot.}]{} \frac{\Pr(A \cap B_j)}{\sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i)}$$

def. of cond

"(Reverse) Conditioning"

Jointly Distributed RVs

Ex)
(X, Y)
(X_1, X_2)
$(X_1, X_2, \dots, X_n)$

Q: what is the sample space for jointly distributed RVs, say, X w/ sample space $\mathcal{X}$ and Y w/ samp. space $\mathcal{Y}$?

Distribution Notation:

	Both discrete	Both continuous
joint	$p_{XY}(x,y) = Pr(X=x, Y=y)$ $F(x,y) = Pr(X \le x, Y \le y)$	$Pr\left(\binom{X}{Y} \in A\right) = \iint_A f(x,y)\,dy\,dx$ $F(x,y) = Pr(X \le x, Y \le y)$
Marginal	$p_X(x) = \sum_{y \in \mathcal{Y}} p(x,y)$	$f_X(x) = F_X'(x) = \int_{\mathcal{Y}} f(x,y)\,dy$ where $F_X(x) = Pr(X \le x)$ $= \lim_{y \to \infty} F(x,y)$ $= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u,y)\,dy\,du$
conditional	$p_{X\|Y}(x\|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \begin{cases} \frac{f_{XY}(x,y)}{f_Y(y)}, & \text{if } 0 < f_Y(y) < \infty \\ 0, & \text{otherwise} \end{cases}$

Special EX

X discrete, Y continuous

marginal

$$Pr(X=x) = p_X(x) \qquad F_Y(y) = \sum_{x \in \mathcal{X}} Pr(Y \le y, X \le x)$$

$$f_Y(y) = F_Y'(y)$$

conditional

$$Pr(X=x \mid Y=y) = \frac{f_{Y|X}(y|x)\, Pr(X=x)}{f_Y(y)}$$

In general, knowing the marginal distb'n of X and of Y is <u>NOT</u> enough information for us to determine the joint distb'n of (X, Y)....

unless....

X and Y are independent ← $X \perp\!\!\!\perp Y$ (abbreviation)

Def: Independent RVs

For RVs $(X_1, \dots, X_n)$ w/ joint distb'n fnctn

$$F(x_1, \dots, x_n),$$

we say $(X_1, \dots, X_n)$ are independent RVs if

$$F(x_1, \dots, x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot \ldots \cdot F_{X_n}(x_n).$$

(It can be shown that this is equivalent to saying that the joint pmf or joint density factors.)

Indicator fnctn

$$\mathbb{I}\{0 < x < 1\} = \begin{cases} 1, & \text{if } 0 < x < 1 \\ 0, & \text{o/w} \end{cases}$$

$$E(\mathbb{I}\{0 < X < 1\}) = P_r(0 < X < 1) = 1 \cdot P_r(0 < X < 1) + 0 \cdot P_r(X \notin (0,1))$$

Next week...

- expectation, variance, covariance
- conditional expectation & variance
- moment generating functions
- methods of estimation

Legend

Notation
Examples, questions
Definitions
Proofs &/or theorems
Looking ahead/planning/topics

Prof. Suzy notes to self

9/7/22 (week 2)

Topic : Review (Ch 4)

Def: Moment Generating Function (MGF)

of a discrete RV X is:

$$M(t) = \sum_{x \in \mathcal{X}} e^{tx} p_X(x)$$

of a continuous RV X is:

$$M(t) = \int_{\mathcal{X}} e^{tx} f_X(x) dx$$

The MGF of RV does not always exist (ex. Cauchy) but when it does, it uniquely determines the RV. (The Characteristic function, like the CDF, always exists but is a complex function.)

Def: The moments of a RV X are

$$E(X^r) \text{ for } r = 1, 2, \ldots$$

The r^{th} derivative of MGF, $M(t)$, evaluated at $t=0$, is the r^{th} moment of X;

Ie. $M^{(r)}(0) = E(X^r)$,

provided $M(t)$ exists in an open interval containing zero.

The first and second moments of a RV determine its expectation & variance.

Expected Values	Discrete	Continuous
$E(X) =$	$\sum_{x \in \mathcal{X}} x \, p_X(x)$	$\int_{\mathcal{X}} x \, f_X(x) \, dx$
$E[g(X)] =$	$\sum_{x \in \mathcal{X}} g(x) \, p_X(x)$	$\int_{\mathcal{X}} g(x) \, f_X(x) \, dx$
$E[Y \mid X = x] =$	$\sum_{y \in \mathcal{Y}} y \, p_{Y\mid X}(y \mid x)$	$\int_{\mathcal{Y}} y \, f_{Y\mid X}(y \mid x) \, dy$
$E[g(Y) \mid X = x] =$	$\sum_{y \in \mathcal{Y}} g(y) \, p_{Y\mid X}(y \mid x)$	$\int_{\mathcal{Y}} g(y) \, f_{Y\mid X}(y \mid x) \, dy$

"The expected value is the sum of the <u>possibilities</u> of a RV times their <u>probabilities</u>."

Note: $E[g(X)] \neq g[E(X)]$

Ex) $X = \begin{cases} 1, & \text{w.p. } 1/2 \\ 2, & \text{w.p. } 1/2 \end{cases}$; $g(x) = \frac{1}{x}$

$$E(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2}$$

$$E[g(X)] = \frac{1}{1} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$$

$$g(E(X)) = g\left(\tfrac{3}{2}\right) = \frac{2}{3}$$

However, we do have the following result.

<u>Jensen's Inequality</u>

For any convex function, g, and any RV X,

$$g(E[X]) \leq E[g(X)]$$

provided $E[g(X)]$ and $g(E[X])$ exist and are finite.

concave ~~convex~~

Incorrect! These a swapped!

See pg 21

Expectation is a linear operator:

$$E\left[\sum_{i=1}^{n}(a_i + b_i X_i)\right] = E\left[(a_1 + b_1 X_1) + (a_2 + b_2 X_2) + \dots + (a_n + b_n X_n)\right]$$

$$= E[a_1 + b_1 X_1] + E[a_2 + b_2 X_2] + \dots + E[a_n + b_n X_n]$$

$$\vdots$$

$$= \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i E[X_i]$$

Markov's Inequality

If X is a positive RV for which $E(X)$ exists,

$$Pr(X \geq t) \leq \frac{E(X)}{t}, \quad \text{for any } t \in \mathbb{R}.$$

"mean slasher"

Ex) (of Markov's inequality)

Chebyshev's Inequality

"variance difference"

If X is a RV whose first and second moments exist, then f any $t > 0$:

$$Pr(|X - E(X)| > t) = Pr((X - E(X))^2 > t)$$

$$\leq \frac{E[(X - E(X))^2]}{t^2}$$

$$= Var(X)/t^2$$

Law of Iterated (Total) Expectation

For RVs X and Y,

$E[Y|X]$ is a RV

because it is a function of X, which is not fixed. It always holds that

$$E[E(Y|X)] = E[Y].$$

(For a proof, see pg. 149.)

Note: $E[Y|X=x]$ is a function of x and is thus NOT a RV since $X=x$ is fixed.

Variance

If RV X has $E(X) < \infty$ then

$$Var(X) = E\{[X - E(X)]^2\}$$

$$\vdots$$

$$= E(X^2) - [E(X)]^2$$

Variance is a non-linear operator:

$$Var\left(\sum_{i=1}^{n} a_i + b_i X_i\right) = Var\big((a_1 + b_1X_1) + (a_2 + b_2X_2) + \dots + (a_n + b_nX_n)\big)$$

$$= Var\left(\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i X_i\right)$$

$$= Var\left(\sum_{i=1}^{n} b_i X_i\right)$$

$$= \dots ? \longrightarrow \text{see next pages!}$$

If we are only interested in one RV then:

$$Var(a + bX) = b^2 Var(X)$$

"Variance is the average (squared) distance between the possibilities of a RV and its expectation."

E-V-E Formula (Iterated Variance)

For RVs X and Y, we have that

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$$

(Proof on pg 151)

Covariance

If X, Y are jointly distb'td RVs whose expectations exist,

$$Cov(X,Y) = E[(X-E(X))(Y-E(Y))]$$

$$\vdots$$

$$= E(XY) - E(X)E(Y)$$

Furthermore,

$$Cov(X,X) = Var(X)$$

Properties of variance + covariance:

Let $U = a + \sum_{i=1}^{n} b_i X_i$, $V = c + \sum_{j=1}^{m} d_j Y_j$

for RVs $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$

$$Cov(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{m} b_i d_j Cov(X_i, Y_j)$$

In particular,

$$\begin{aligned}
Var(U) &= Var\left(a + \sum_{i=1}^{n} b_i X_i\right) \\
&= Var\left(\sum_{i=1}^{n} b_i X_i\right) \\
&= Cov\left(\sum_{i=1}^{n} b_i X_i, \sum_{i=1}^{n} b_i X_i\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j Cov(X_i, X_j)
\end{aligned}$$

Ex) If $X_1, \dots, X_n$ are independent (and identically) distributed,

(IID)

what is $Var\left(\sum_{i=1}^{n} X_i\right)$?

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i).$$

9/9/22 (Week 2)

Topic: Estimation Part I (Ch 4+8)

Setting: $X_1, \ldots, X_n$ IID $f(x;\theta)$ is marginal density
$f_X(x)$

Q) What's the difference btwn a statistic and an estimate?

Both however are fnctns of the random sample.
more general
targets a particular parameter

Deriving an estimator ("Recipes")

Method 1: Method of moments

Consider the first few moments of the population distb'n

$$\left.\begin{aligned} \mu_1 &= E[X^1] \\ \mu_2 &= E[X^2] \\ \mu_3 &= E[X^3] \\ &\vdots \end{aligned}\right\}$$

Create a system of equations that can be solved for the parameter(s) θ

Then, substitute the sample estimates of these moments into solution for θ above.

So
$$\left[\begin{aligned} \hat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i^1 \\ \hat{\mu}_2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ \hat{\mu}_3 &= \frac{1}{n}\sum_{i=1}^{n} x_i^3 \\ &\vdots \end{aligned}\right.$$

take the place of $\mu_1, \mu_2, \ldots$ and the result is the estimator!

Method 2: Maximize the Likelihood

Q) If $(x_1, x_2, \dots, x_n)$ are an $\boxed{\text{IID}}$ sample from a population w/ distb'n $F_X(x)$ and density $f_X(x)$, then what is the joint density of $(X_1, X_2, \dots, X_n)$?

$$(X_1, X_2, \dots, X_n) \sim \prod_{i=1}^{n} f_{X_i}(x) = f_X^n(x) = f^n(x;\theta)$$

If we think of this joint distribution as a function of the <u>parameter(s)</u> for fixed (observed) data $(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)$ then we are referring to the likelihood of the parameter(s), given the data.

both can be vectors!

likelihood: $lik(\theta) = f(x;\theta)$

log-likelihood: $\ell(\theta) = \log[lik(\theta)]$

Once we have a likelihood for θ; often we can maximize this function (w.r.t. θ). The maximum (global) is often a useful estimate for θ.

~ Paradigm Shift! ~

Method 3: Use Bayes' Theorem

Treat the parameter, θ, as a RV. Come up w/ an initial guess for the distribution of $\Theta \sim f_{\Theta}(\theta)$.

Typically a prior is denoted as $\theta \sim p(\theta)$ or $\theta \sim \pi(\theta)$

Given a likelihood function for θ, conditioned upon the observed data, $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$, use Bayes' theorem to find the conditional distribution $f(\theta | x)$.

Typically, this posterior density is denoted as

$$\theta | \underset{\sim}{x} \sim \pi(\theta | \underset{\sim}{x})$$

Altogether we have

likelihood : $f(\underset{\sim}{x};\theta)$

prior : $\pi(\theta)$

prior distn for θ

likelihood for θ given data $\underset{\sim}{x}=(x_1,x_2,\dots,x_n)$

posterior : $$\pi(\theta|\underset{\sim}{x}) = \frac{\pi(\theta)f(\underset{\sim}{x};\theta)}{\underbrace{\int_{\Theta}\pi(\theta)f(\underset{\sim}{x};\theta)d\theta}_{\text{"normalizing" constant}}}$$

Q) What is Θ?

Θ is the parameter space

Often, we can ignore the "normalizing" constant and specify the posterior up to proportionality:

$$\pi(\theta|\underset{\sim}{x}) \propto \pi(\theta)f(\underset{\sim}{x};\theta)$$

"is proportional to"

Note: The entire distn of the posterior is a distribution function estimate for θ!

We can derive point estimates for θ by considering different qualities of the posterior.

For example: posterior mean
posterior mode

Q) Are these the only ways to derive an estimator?
no! there are infanite numb of ways to derive an estimator

Q) How do we know if an estimator is useful?
This is what we'll discuss next!

<u>Setting</u>:
Given an sample $(x_1, x_2, \ldots, x_n)$ of RVs that follow a distribution depending on unknown parameter θ, denote $\hat{\theta}_n = \hat{\theta}_n(\underset{\sim}{x})$ as an estimator for θ

Note: $\hat{\theta}_n(\underset{\sim}{X})$ is a RV ; $\hat{\theta}_n(\underset{\sim}{x})$ is a fixed constant.

Desireable characteristics for estimators:

- consistency

$\hat{\theta}_n$ is <u>consistent</u> for θ if, for all $\varepsilon > 0$

$$\lim_{n \to \infty} \Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

Q) What type of convergence is this?
this is an example of 'limit' in probability

Ex) (weak) LLN: sample moments $\xrightarrow{\Pr}$ pop. moments

Note: Continuous functions presene consistency

- unbiased

$\hat{\theta}_n$ is <u>unbiased</u> if $E[\hat{\theta}_n] = \theta$,

ie. the center of its sampling distb'n is θ.

Evaluating an estimator

Def: Mean Square Error

If we are targeting parameter θ w/ an estimator $\hat{\theta}_n$, then

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

trick is to $\pm E(\hat{\theta}_n)$

$$\vdots$$

$$= \underbrace{(E(\hat{\theta}_n) - \theta)^2}_{\text{bias}^2} + \underbrace{Var(\hat{\theta}_n)}_{\text{variance}}$$

Strategies to show consistency:

If $\hat{\theta}_n$ is unbiased – subsititute $E(\hat{\theta}_n)$ in for θ then apply a limiting $\leq$

If $\hat{\theta}_n$ (potentially) biased – then we have to usually work w/ the CDF of $\hat{\theta}_n$

$$Pr(|\hat{\theta}_n - \theta| > \varepsilon) = \underbrace{Pr(\hat{\theta}_n > \theta + \varepsilon)}_{\text{evaluate separately}} + \underbrace{Pr(\hat{\theta}_n < \theta - \varepsilon)}_{\text{evaluate separately}}$$

But, Sometimes you have to get more creative!

Eg. Jensen's $\leq$ could be used to prove biasedness.
(the strict version)

Topic - Detour for errata

Jensen's Inequality

Recall

For any ~~convex~~ concave up function, g, and any RV X,

$$g(E[X]) \le E[g(X)]$$

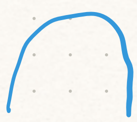
provided $E[g(X)]$ and $g(E[X])$ exist and are finite.

Correction: My heuristic for remembering concave/convex doesn't work!

Instead

looks like a cave!

convex = concave up
$f'' > 0$



concave down
$f'' < 0$

Q) When is the inequality strict?

When the concavity is strict (no plateaus)

Now back to properties of estimators...

Ex) Suppose $X_1, \dots, X_n$ are IID from $U(0, \theta)$.
Consider the following estimates and determine if they are consistent.

Consistent	Unbiased	Estimate
yes	yes	$\hat{\theta}_1 = 2\bar{X}$
no	yes	$\hat{\theta}_2 = 2X_1$
yes	no	$\hat{\theta}_3 = X_{(n)}$
no	no	$\hat{\theta}_4 = 1/X_1^2$

n^{th} order statistic, ie. largest observation

only use the first observation! (not necessarily the minimum!)

Note:

X_1 has density

$$f_{X_1}(x) = \frac{1}{\theta} \mathbb{I}\{0 < x \leq \theta\}$$

and CDF

$$F_{X_1}(x) = P_r(X_1 \leq x) = \frac{x}{\theta} \mathbb{I}\{0 < x \leq \theta\}$$

$\hat{\theta}_1 = 2\bar{X}$

9-14-22

Unbiased?

$$E[2\bar{X}] = 2E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = 2 \times \left(\frac{1}{n}\right) \times E[X_1 + X_2 + \dots + X_n]$$

$$= \frac{2}{n} \times (E(X_1) + E(X_2) + \dots + E(X_n))$$

$$= \frac{2}{n} \cdot n\,E(X_1) = 2\left(\frac{\theta}{2}\right) = \theta \quad \checkmark$$

is unbiased

Consistent?

$\bar{X}$ is consistent for $E[X_1] = \frac{\theta}{2}$. Why?

A: b/c sample moments are consistent for population moments (LLN)

$g(x) = 2x$ is a continuous function

$\therefore \quad g(\bar{X}) = 2\bar{X}$ is consistent for $2E[X_1] = \theta$

$\checkmark$ is consistent

$\boxed{\hat{\theta}_2 = 2X_1}$

Unbiased?

$$E(\hat{\theta}_2) = E(2X_1) = 2E(X_1) = 2\cdot\frac{\theta}{2} = \theta \quad \checkmark$$ is unbiased

Consistent?

$$\begin{aligned} Pr(|\hat{\theta}_2-\theta|>\varepsilon) &= Pr(|\hat{\theta}_2 - E(\hat{\theta}_2)|>\varepsilon) \\ &= Pr(|2X_1 - E(2X_1)|>\varepsilon) \\ &\le \frac{Var(2X_1)}{\varepsilon^2} \\ &= \frac{4\,Var(X_1)}{\varepsilon^2} \leftarrow \text{not a fnctn of } n \text{ so not helpful.} \end{aligned}$$

Chebyshev: $P(|X-E(X)|>\varepsilon) \le \frac{Var(X)}{\varepsilon^2}$

Let's try the CDF approach...

$$\begin{aligned} Pr(|\hat{\theta}_2-\theta|>\varepsilon) &= Pr(2X_1>\theta+\varepsilon) + Pr(2X_1<\theta-\varepsilon) \\ &= Pr\left(X_1 > \frac{\theta+\varepsilon}{2}\right) + Pr\left(X_1 < \frac{\theta-\varepsilon}{2}\right) \\ &= 1 - Pr\left(X_1 \le \frac{\theta+\varepsilon}{2}\right) + \left(\frac{\theta-\varepsilon}{2}\right)\Big/\theta \\ &= 1 - \frac{\left(\frac{\theta+\varepsilon}{2}\right)}{\theta} + \frac{\left(\frac{\theta-\varepsilon}{2}\right)}{\theta} \\ &= 1 - \frac{\theta+\varepsilon}{2\theta} + \frac{\theta-\varepsilon}{2\theta} \\ &= \frac{2\theta - \theta + \varepsilon + \theta - \varepsilon}{2\theta} \\ &= 1 \end{aligned}$$

✗ not consistent

Note: Estimator is function of X_1 only... so we really didn't need to do all that work!

$\hat{\theta}_3 = X_{(n)}$

(scratch work)

CDF for $\hat{\theta}_3$: $\Pr(\hat{\theta}_3 \le x) = \Pr(X_{(n)} \le x)$

by def on $X_{(n)}$ $= \Pr(X_1 \le x, X_2 \le x, \ldots, X_n \le x)$

by indep $= \Pr(X_1 \le x) \cdot \Pr(X_2 \le x) \cdots \Pr(X_n \le x)$

by identical $= [\Pr(X_1 \le x)]^n$

$= \left(\frac{x}{\theta}\right)^n \mathbb{I}\{0 < x < \theta\}$

density for $\hat{\theta}_3$:

$f_{\hat{\theta}_3}(x) = n \cdot x^{n-1} \cdot \frac{1}{\theta} \mathbb{I}\{0 < x \le \theta\}$

Unbiased?

$$E[\hat{\theta}_3] = E[X_{(n)}] = \int_0^\theta \frac{n}{\theta} \cdot x \cdot x^{n-1}\, dx = \frac{n}{\theta}\int_0^\theta x^n dx$$

$$= \frac{n}{\theta}\left(\frac{x^{n+1}}{n+1}\Big|_{x=0}^{\theta}\right)$$

$$= \frac{n}{\theta}\left[\frac{\theta^{n+1}}{n+1} - 0\right] = \frac{n\theta^{n+1}}{(n+1)\theta}$$ X

NOT unbiased

Consistent?

$$\Pr(|\hat{\theta}_3 - \theta| > \varepsilon) = \Pr(X_{(n)} > \theta + \varepsilon) + \Pr(X_{(n)} < \theta - \varepsilon)$$

$$= \int_{\theta+\varepsilon}^{\theta} f_{X_{(n)}}(x)dx + \int_0^{\theta-\varepsilon} f_{X_{(n)}}(x)dx$$

$(\varepsilon > 0)$

$$= 0 + \int_0^{\theta-\varepsilon} \frac{n}{\theta} x^{n-1} dx$$

(can assume $\varepsilon < \theta$)

$$= \frac{n}{\theta}\int_0^{\theta-\varepsilon} x^{n-1} dx$$

$$= \frac{n}{\theta}\left[\frac{x^n}{n}\Big|_{x=0}^{\theta-\varepsilon}\right]$$

$$= \frac{1}{\theta}\left[(\theta-\varepsilon)^n - 0\right]$$

$\lim_{n\to\infty} \frac{(\theta-\varepsilon)^n}{\theta} = 0$ since $\theta \in (0,1)$ ✓ is consistent

$\hat{\theta}_4 = 1/X_1^2$

Unbiased?

$$E\left[\frac{1}{X_1^2}\right] = \int_0^\theta \frac{1}{X_1^2} f_{X_1}(x_1)\,dx_1 = \int_0^\theta \frac{1}{X_1^2}\cdot\frac{1}{\theta}\,dx_1 = \frac{1}{\theta}\int_0^\theta \frac{1}{X_1^2}\,dx_1$$

$$= \frac{1}{\theta}\left[\frac{-1}{X_1}\Big|_{X_1=0}^{\theta}\right]$$ undefined X

not unbiased b/c expectation doesn't exist!

Consistent?

Again, estimator is a function of X_1 only. So what happens as $n \to \infty$?

Nothing. The estimator doesn't change w/ the sample size.

X not consistent

■

9-16-22

Ex) Stakeholder analysis of using a consistent estimator

θ_1 = dosage that max benefit/min harm
θ_2 = change in B cell counts after using medication
} possible parameters

$\hat{\theta}_1$ = 10 mg/kg
$\hat{\theta}_2$ = "X" change in fluoresence intensity
} possible estimates

Suppose $\hat{\theta}_n$ = 10mg/kg is a consistent estimator for θ = dosage that max benefit + min harm

Choice/Decision: Decide whether or not to use a drug to treat Systemic Lupus Erythematous within the first few years of diagnosis. Here is an example of a pilot study currently ongoing.

Stakeholder	Potential results	
	Harm	**Benefit**
Medical practitioners perscribe $\hat{\theta}_n$ dose to patient	possibly not all patients are represented in the population for which we have a sample	for the majority of the population this estimated dosage will be the best dosage
Medication users take $\hat{\theta}_n$ dosage		

- Example harms: cost of money, time, effort; negative impact to reputations; can be tangible or intangible with immediate or delayed effects
- Example benefits: earning or gaining money; removal of a harm; saved time or effort; improved reputation; demonstration of expertise.

Source: Tractenberg, R. E. (2019). Teaching and Learning about ethical practice: The case analysis. https://doi.org/10.31235/OSF.IO/58UMW

9-19-22
(Week 4)

Topic: Estimation Part II (Ch. 8)

Large Sample Theory for MLEs

Setting: $X_1, \dots, X_n \overset{IID}{\sim} f(x;\theta)$

$lik(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$

$\ell(\theta) = \sum_{i=1}^{n} \ln(f(x_i;\theta))$

$\hat{\theta}_n$ = value of θ that maximizes $lik(\theta)$ — MLE specifically

θ_0 = true, unknown value of θ

$n \to \infty$

Def: The score is the gradient (first derivative) of the likelihood fnctn.

$$S(\theta) = \frac{\partial}{\partial\theta}\ell(\theta)$$

rate of change in (log) likelihood

Note: $\hat{\theta}_n$ (the MLE for θ, given $\underset{\sim}{X}_{obs}$) is a "zero" of $S(\theta)$
ie. $S(\hat{\theta}_n) = 0$

Thm: If $f(x;\theta)$ is "smooth enough", then the MLE is consistent.

Note: The expected value of $S(\theta)$ is 0 at $\theta = \theta_0$.
b/c ...

mistake starts here!

$$E[S(\theta)] = E\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] = \cancel{\int \left[\frac{\partial}{\partial\theta}\ell(\theta)\right] f(x;\theta)\,dx}$$

$$= \cancel{\int \left[\frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial\theta}\right] f(x;\theta)\,dx}$$

$$= \cancel{\int \frac{\partial}{\partial\theta} f(x;\theta)\,dx \underset{*}{=} \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \frac{\partial}{\partial\theta}1} = 0$$

Fixed version of $E[s(\theta)] = 0$:

$$E[s(\theta)] = E\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] = \int\cdots\int\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] f(x_1,\ldots,x_n;\theta_0)\,dx_1\ldots dx_n$$

at $\theta=\theta_0$

$$= \underbrace{\int\cdots\int}_{n \text{ times}} \frac{\frac{\partial}{\partial\theta} f(x_1,\ldots,x_n;\theta)}{f(x_1,\ldots,x_n;\theta)} f(x_1,\ldots,x_n;\theta_0)\,dx_1\cdots dx_n$$

$$= \int\cdots\int \frac{\partial}{\partial\theta} f(x_1,\ldots,x_n;\theta_0)\,dx_1\ldots dx_n$$

* if we can interchange $\frac{\partial}{\partial\theta}$ & $\int$

$$= \frac{\partial}{\partial\theta}\int\cdots\int f(x_1,\ldots,x_n;\theta_0)\,dx_1\ldots,dx_n$$

$$= \frac{\partial}{\partial\theta}(1)$$

$$= 0$$

■

Related to

* Calc Thm: Leibniz Integral Rule (special case)

$$\frac{d}{dx}\left(\int_a^b f(x,u)\,du\right) = \int_a^b\left[\frac{\partial}{\partial x} f(x,u)\right]du$$

Def: The Fisher Information is the variance of the score.

$$I_n(\theta) = E\left\{\left[\frac{\partial}{\partial\theta} \ell(\theta)\right]^2\right\}$$

2^{nd} moment of score fnctn

Thm: Information Identity

If $f(x;\theta)$ is "smooth enough", then

$$I_n(\theta) = E\left\{\left[\partial/\partial\theta\, \ell(\theta)\right]^2\right\} = -E\left[\frac{\partial^2}{\partial\theta^2} \ell(\theta)\right]$$

Thm: Asymptotic Normality of MLEs

If $f(x;\theta)$ is "smooth enough", then

$$\sqrt{n I_n(\theta_0)}\,(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,1).$$

Q) What does it mean for an estimate to be "optimal"?

Def: Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are estimators of θ that have the same bias. Ie. $E[\hat{\theta}_1] - \theta = E[\hat{\theta}_2] - \theta$. The effeciency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \text{Var}(\hat{\theta}_2) / \text{Var}(\hat{\theta}_1).$$

Note: If we are comparing asy. variance of an estimator, we call this the asymptotic relative efficiency.

Thm: Cramér-Rao Inequality

Suppose $X_1, \ldots, X_n$ are IID $f(x;\theta)$, where $f(x;\theta)$ is "smooth enough". Let $T = T(\underset{\sim}{X})$ be an unbiased estimate of θ. Then $\text{Var}(T) \geq \dfrac{1}{n I_n(\theta)}$.

Cramér-Rao Lower Band

Note: An unbiased estimate w/ variance equal to the CR-LB is said to be effecient.

Note: As $n \to \infty$, the MLE is asymptotically effecient.

Q) Is asymptotic unbiasedness the same thing as consistent? Why/why not?

Notation

9-21-22

Note: Textbook uses $f(X_i|\theta)$

$f(X_i;\theta)$ density for X_i

If $X_1, \dots, X_n$ are IID then the likelihood is:

$$lik(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

and the log-likelihood is:

$$\ell(\theta) = \sum_{i=1}^{n} \log(f(X_i;\theta))$$

The score function is the gradient of the log-likelihood:

$$\frac{\partial}{\partial\theta}\ell(\theta)$$

The score function has mean zero and varience equal to the Fisher Information

$$I_n(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ell(\theta)\right]^2\right\}$$

Info about θ contained in $(X_1, \dots, X_n)$

Your textbook considers the score for a single RV, X:

$$\frac{\partial}{\partial\theta}\log f(X;\theta)$$

where the Fisher Info is thus

$$I(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right\}$$

is the info about θ contained in X alone.

Consider the log density $\log(f(X;\theta))$:

Q) What is the 1st (population) moment?

$$E[\underbrace{\log(f(X;\theta))}_{g(X)}] = \int \underbrace{[\log(f(x;\theta))]}_{g(X)} f(x;\theta_0)\,dx$$

$\uparrow$ density for X!

Q) What is the 1st sample moment?

$$\frac{1}{n}\sum_{i=1}^{n} \log(f(x_i;\theta)) = \frac{1}{n}\ell(\theta)$$

Now consider the gradient of the log density $\frac{\partial}{\partial\theta}\log(f(X;\theta))$:

Q) What is the 1st (population) moment?

$$E\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = \int \frac{\partial}{\partial\theta}\log f(x;\theta)\, f(x;\theta_0)\,dx$$

$$= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta_0)\,dx \overset{*}{=} \int \frac{\partial}{\partial\theta} f(x;\theta_0)\,dx = \frac{\partial}{\partial\theta}\int f(x;\theta_0)\,dx = \frac{\partial}{\partial\theta}(1) = 0$$

Q) What is the (population) variance?

$$\mathrm{Var}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = E\left\{\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]^2\right\} - \left\{E\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]\right\}^2$$

$$= I(\theta)$$

* at $\theta = \theta_0$

Warm-up group work: 9-23-22

5 mins

Identify strategies, stuck points, approachs you tried to solve assigned HW 8 problem

	#1	#2	#3
Sec 1:	Seth Sherry Miles	Koji Annie Amy	Brian Patty Guy Tillie
Sec 2:	Mwangangi Tinashe Zack	Ben H Atesh Jonathan	Joey Jason Rodas
	Alex Jorge Ben S	Sarah Hellman Ian	Nancy Noha Gertrud

Review + consider:

What strategies/approaches were most useful?

Sufficiency

Setting: $X_1, \dots, X_n$ IID $f(x;\theta)$

$$\text{lik}(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

$$\ell(\theta) = \sum_{i=1}^{n} \ln(f(x_i;\theta))$$

$\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$ is an estimator for θ.

not just MLE!

Q) Is there an estimator that contains as much information about θ as the entire sample, $x_1, \dots, x_n$?

Def: A statistic $T = T(X_1, \dots, X_n)$ is **sufficient** for parameter θ if

$$(X_1, \dots, X_n) \mid T = t$$

follows a distribution that does not depend on θ.

Thm: Factorization Theorem

Statistic $T(X_1, \dots, X_n)$ is sufficient for θ iff

$$f(x_1, \dots, x_n; \theta) = g[T(x_1, \dots, x_n); \theta] \cdot h(x_1, \dots, x_n)$$

likelihood

must involve all of the observed data!

Exponential Family

The family of probability distb'n functions that have sufficient statistics of the same dimension as the parameter space is called the exponential family.

1-Parameter Exponential family:

$$f(x;\theta) = \exp\{c(\theta)T(x) + d(\theta) + S(x)\}$$

for all $x \in A$ where $A \perp\!\!\!\perp \theta$

K-parameter Exponential family:

$$f(x;\theta) = \exp\left\{\sum_{j=1}^{K} c_j(\theta)T_j(x) + d(\theta) + S(x)\right\}$$

for all $x \in A$ where $A \perp\!\!\!\perp \theta$

Note: If T is sufficient for θ, then the MLE is a function of T.

We can see this is the case b/c...

$T(X_1,\ldots,X_n)$ sufficient means

$$\text{lik}(\theta) = \underbrace{f(x_1,\ldots,x_n;\theta)}_{\substack{\text{maximize} \\ \text{wrt } \theta}} = g\underbrace{[T(x_1,\ldots,x_n),\theta]}_{\substack{\text{maximize} \\ \text{wrt } \theta}} \cdot \underbrace{h(x_1,\ldots,x_n)}_{\perp\!\!\!\perp \theta}$$

Thm: Rao-Blackwell Theorem

Let $\hat{\theta}$ be an estimator for θ s.t. $E(\hat{\theta}^2) < \infty$.
If T is sufficient for θ and if $\tilde{\theta} = E[\hat{\theta} \mid T]$,
then, for all θ,

MSE $\longrightarrow$ $E[(\tilde{\theta}-\theta)^2] \leq E[(\hat{\theta}-\theta)^2]$.

Furthermore, the inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

Note: If an estimator is not a function of a sufficient statistic, and if a sufficient statistic exists, then the estimator can be improved!

Group Work:

9-26-22

Dissecting Proofs

Example: Information Identity

Define $I(\theta) = E\left\{\left[\frac{\partial}{\partial\theta} \log f(X;\theta)\right]^2\right\}$.

If $f(\cdot)$ is "smooth enough", the we have

$$E\left\{\left[\frac{\partial}{\partial\theta} \log f(X;\theta)\right]^2\right\} = -E\left[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta)\right].$$

1. Confusing steps?

combining identities in a useful way

how does $\frac{\partial}{\partial\theta} \int \left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right] f(x;\theta)\,dx = \int \left[\frac{\partial^2}{\partial\theta^2} \log f(x;\theta)\right] f(x;\theta)\,dx + \int \left[\frac{\partial}{\partial\theta} \log f(x;\theta)\right]^2 f(x;\theta)\,dx$

2. Useful techniques?

the fact that $\int f(x;\theta)\,dx = 1$; swapping $\frac{\partial}{\partial\theta}$ and $\int \text{—} dx$

and rearrange $\frac{\partial}{\partial\theta} \log f(x;\theta) = \frac{\partial/\partial\theta\, f(x;\theta)}{f(x;\theta)}$

3. Narrative?

use property of density fnctns + swap diff. & integ. + take 2nd deriv applying calc rules + rearrange identities = result ∎

Example: Working Thru Steps of Cramér-Rao

(For me, these were the most confusing steps in this proof.)

$\boxed{\text{pg } 301}$

$$E[ZT] = E\left\{\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log(f(X_i;\theta))\right] T(X_1,\ldots,X_n)\right\} =$$

$$\underbrace{\int\cdots\int}_{n \text{ times}} t(x_1,\ldots,x_n)\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log(f(x_i;\theta))\right] f(x_1,\ldots,x_n;\theta)\, dx_1 \cdots dx_n$$

$$=$$

$$\int\cdots\int t(x_1,\ldots,x_n)\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log(f(x_i;\theta))\right]\left[\prod_{j=1}^{n} f(x_j;\theta)\, dx_j\right]$$

---

and note $\displaystyle\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i;\theta)}{f(x_i;\theta)} \prod_{j=1}^{n} f(x_j;\theta) =$

$$\frac{\frac{\partial}{\partial\theta} f(x_1;\theta)}{f(x_1;\theta)}\left(f(x_1;\theta)\, f(x_2;\theta) \cdots f(x_n;\theta)\right)$$

$$+ \frac{\frac{\partial}{\partial\theta} f(x_2;\theta)}{f(x_2;\theta)}\left(f(x_1;\theta), \cdots f(x_n;\theta)\right)$$

$$+ \cdots + \frac{\frac{\partial}{\partial\theta} f(x_n;\theta)}{f(x_n;\theta)}\left(f(x_1,\theta) \cdots f(x_n;\theta)\right)$$

$\boxed{\text{pg } 301}$

$$E[Z] = E\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log(f(X_i;\theta))\right]$$

$$= E\left[\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(X_i;\theta)}{f(X_i,\theta)}\right]$$

$$= \sum_{i=1}^{n} E\left[\frac{\partial/\partial\theta\, f(X_i;\theta)}{f(X_i;\theta)}\right]$$

$$= \sum_{i=1}^{n} \left\{ \int \left[\frac{\partial/\partial\theta\, f(x_i;\theta)}{f(x_i;\theta)}\right] f(x_i;\theta_0)\, dx_i \right\}$$

at $\theta=\theta_0$

$$= \sum_{i=1}^{n} \left\{ \int \frac{\partial}{\partial\theta} f(x_i;\theta_0)\, dx_i \right]$$

$$= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial\theta} \int f(x_i;\theta_0)\, dx_i \right\}$$

$$= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial\theta} (1) \right\}$$

$$= \frac{\partial}{\partial\theta} f(x_1;\theta)\left[f(x_2;\theta)\cdots f(x_n;\theta)\right]$$

$$+ \frac{\partial}{\partial\theta} f(x_2;\theta)\left[f(x_1,\theta) f(x_3,\theta)\cdots f(x_n,\theta)\right]$$

$$+ \cdots$$

$$+ \frac{\partial}{\partial\theta} f(x_n;\theta)\left[f(x_1;\theta) f(x_2;\theta)\cdots f(x_{n-1};\theta)\right)$$

$$= \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i,\theta).$$

Hence

$$E[ZT] =$$

$$\int\cdots\int t(x_1,\ldots,x_n)\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log(f(x_i;\theta))\right]\left[\prod_{j=1}^{n} f(x_j;\theta)\,dx_j\right]$$

$$= \int\cdots\int t(x_1,\ldots,x_n)\,\frac{\partial}{\partial\theta}\prod_{i=1}^{n} f(x_i;\theta)\,dx_i$$

$$= \frac{\partial}{\partial\theta}\int\cdots\int t(x_1,\ldots,x_n)\prod_{i=1}^{n} f(x_i;\theta)\,dx_1\cdots dx_n$$

$$= \frac{\partial}{\partial\theta} E[T(X_1,\ldots,X_n)]$$

$\blacksquare$

HW 9 #1

Bayesian Estimation/Prediction

θ_0 = prob. that bball player successfully makes a shot

prior: $\pi(\theta) \sim U[0,1]$

obs. data: 2 successful shots in a row

assume: outcomes (of shots) are independent

(a) what is the posterior density of θ?

(b) What would you estimate is the probability that this player makes a third shot?

What is the (probability) model for the data?

Let $x = \begin{cases} 0, & \text{miss} \\ 1, & \text{score} \end{cases}$ $\quad X \sim \text{Bern}(\theta_0)$

$$Pr(X = x) = \theta^x (1-\theta)^{1-x}$$

Now we can evaluate the likelihood for θ given the observed outcomes (data):

$x_1 = 1, x_2 = 1$

$$Pr(X_1 = 1, X_2 = 1) = Pr(X_1 = 1) \cdot Pr(X_2 = 1)$$
$$= \theta^1 (1-\theta)^{1-1} \cdot \theta^1 (1-\theta)^{1-1}$$
$$= \theta^2$$

What is the prior density on θ?

$$f_\Theta(\theta) = 1 \cdot \mathbb{I}\{0 \le \theta \le 1\} = \pi(\theta)$$

Now we can evaluate the posterior, conditioned upon the observed data:

$$\pi(\theta \mid x_1=1, x_2=1) = \frac{\pi(\theta) \cdot f(x_1=1, x_2=1; \theta)}{\int_0^1 \pi(\theta) f(x_1=1, x_2=1; \theta)\, d\theta}$$

$$= \frac{1 \cdot \mathbb{I}\{0 \le \theta \le 1\} \cdot \theta^2}{\int_0^1 1 \cdot \theta^2\, d\theta}$$

$$\vdots$$

$$= \frac{\theta^2}{\theta^3/3 \,\big|_{\theta=0}^{1}} \mathbb{I}\{0 \le \theta \le 1\} = \cdots = 3\theta^2, \quad \text{for } 0 \le \theta \le 1$$

Finally, we can check our answer by verifying that $\underline{\int \pi(\theta \mid \underset{\sim}{x})\, d\theta = 1}$:

$$\int_0^1 3\theta^2\, d\theta = \ldots = 1$$

Part (b) is a question about how to use the posterior to estimate the true value of θ.

$$E(\theta \mid x_1=1, x_2=1) = \int_0^1 \theta \cdot \pi(\theta \mid x_1=1, x_2=1)\, d\theta$$

$$= \int_0^1 3\theta^3\, d\theta$$

Group Work Results

9-28-22

for Dissecting Proofs Worksheet

Cramér-Rao Inequality

Most confusing steps: $E[Z]=0$

$Cov(Z,T) = E[ZT]$

See Example: Working Thru Steps of Cramér-Rao above!

Tricks + techniques: chain rule

Leibniz rule for diff + int.

properties of score +

definition of Fisher info.

Story:

Rao-Blackwell Thm

Most confusing steps: "$Var(\hat{\theta}|T) = 0$ only if..."

understanding what is meant by $\tilde{\theta}$.

how does comparing MSE's come down to comparing variances?

Note: $E(\hat{\theta}) = E[E(\hat{\theta}|T)]$ by law of iterated expectat.

so $\hat{\theta}$ and $\tilde{\theta} = E[\hat{\theta}|T]$ have the same bias!

Also note: If $\hat{\theta}$ is a function of T, then $\tilde{\theta} = E[\hat{\theta}|T] = E[\hat{\theta}(T)|T]$ is <u>not</u> <u>random</u>!

Tricks & techniques: law of iterated expectation and E-V-E property of conditional variance

Story:

Factorization Thm

Most confusing steps: $\dfrac{P_r(\underset{\sim}{X}=\underset{\sim}{x}, T=t)}{P_r(T=t)} = \dfrac{h(\underset{\sim}{x})}{\sum_{T(\underset{\sim}{x})=t} h(\underset{\sim}{x})}$

how to get $g(t;\theta)\sum_{T(\underset{\sim}{x})=t} h(\underset{\sim}{x})$?

Suppose $X_1, \dots X_n$ are continuous over sample space $\mathcal{X}$. Then

$$P_r(\underset{\sim}{X}=\underset{\sim}{x}, T=t) = P_r(X_1=x_1, X_2=x_2, \dots, X_n=x_n, T(X_1,\dots,X_n)=t)$$

$$= \int_A \dots \int f(x_1, x_2, \dots, x_n; \theta)\, dx_1 \cdots dx_2$$

where A is $\{\underset{\sim}{x} \in \mathcal{X} : T(x_1,\dots,x_n)=t\}$

$$\text{(by assumption)} = \int_A \dots \int g(T(x_1,\dots,x_n)) \cdot h(x_1,\dots,x_n)\, dx_1 \cdots dx_n$$

Tricks & techniques:

expand joint density terms.

manipulate sums.

assume A, deduce B. then assume B, deduce A.

Story:

Topic: Estimation Part III
(ch 8)

Confidence Intervals

– quantify the uncertainty inherent to point estimation using properties of random sampling from an assumed model

indirect assessment of uncertainty

For IID data

$$(X_1, \dots, X_n) \sim \underbrace{\prod_{i=1}^{n} f(x_i; \theta_0)}_{\text{assumed model}}$$

parameter fixed, unknown always a constant

Recall

$\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$ is a point estimate for θ_0

is random, has a sampling distb'n

but

$\hat{\theta}_n = \hat{\theta}(x_1, \dots, x_n)$ is the point estimate evaluated for observed data.

is fixed, data has been observed

Similarly,

A confidence interval for θ_0 is a random interval ... until the data is observed.

random b/c it is a fnctn of $X_1, \dots, X_n$

Process:

Use the sampling distb'n of $\hat{\theta}_n$ (in particular the sampling variance of $\hat{\theta}_n$) to identify a lower bound (LB) and upper bound (UB) on the most plausible values for θ_0.

Interpretation:

Although we say we are $(1-\alpha)\times 100\%$ confident that the true value of θ (ie. θ_0) lies w/in [LB, UB], what we mean is something a bit more involved...

Based on the assumed model for the data, the probability that the random interval $[LB(\hat{\theta}_n), UB(\hat{\theta}_n)]$ contains the value of θ that generated the data, θ_0, is $(1-\alpha)$.

Tips & techniques:

Often, it is useful to plot the density (or mass) function for the sampling distb'n of $\hat{\theta}_n$ to identify which distb'n quantiles to use in the CI.

Example of exact and approximate CIs

HW 8 #2b

$X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Exp}(\tau)$

Note: This version is consistent w/ the parameterization in your textbook (the "rate" parameterization vs. "scale")

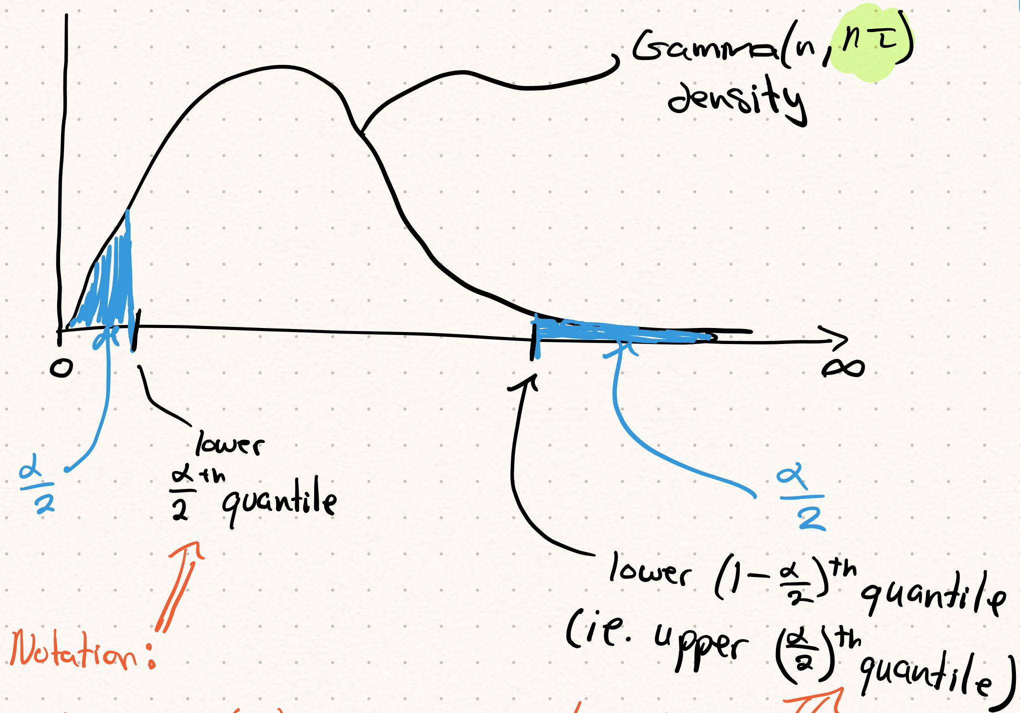
$$\text{lik}(\theta) = \prod_{i=1}^{n} f(x_i; \tau) = \prod_{i=1}^{n} \left[\tau e^{-\tau x_i} \mathbb{I}\{x_i \geq 0\} \right] = \tau^n e^{-\tau \sum_{i=1}^{n} x_i} \mathbb{I}\{x_{(1)} \geq 0\}$$

$\hat{\theta}_{MLE} = \bar{X}$ ⟵ To do: use this sampling distb'n to find a $(1-\alpha)100\%$ CI for τ.

Given : $\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \tau)$

Derive : $\bar{X} \sim \text{Gamma}(n, n\tau)$

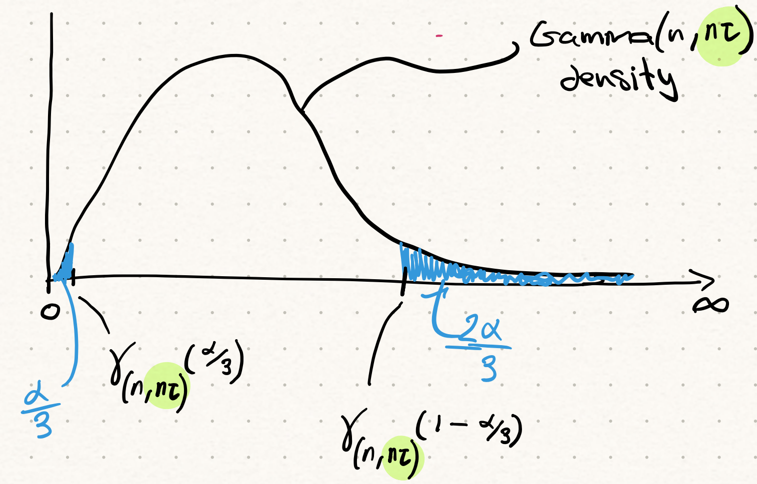
All changes are highlighted in green. View the notes for 9-28-22 to see the other version



Notation: $\gamma_{(n, n\tau)}\left(\frac{\alpha}{2}\right)$

Notation: $\gamma_{(n, n\tau)}\left(1 - \frac{\alpha}{2}\right)$

Note, we could asymmetrically choose the quantiles, e.g.



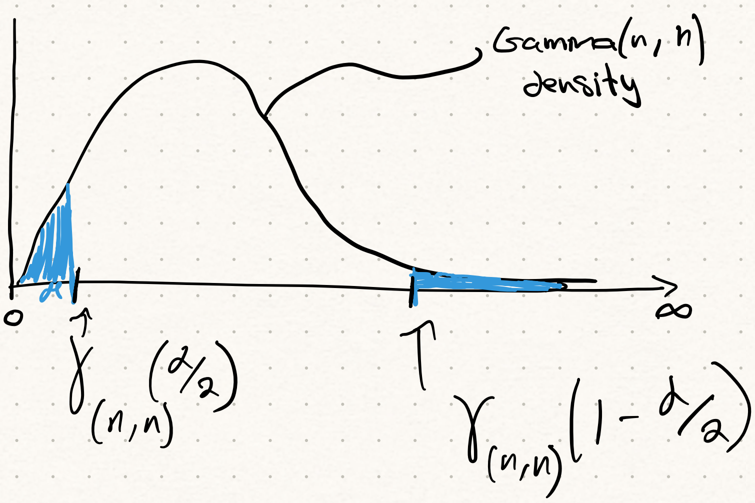
But, in either case, since τ is unknown, we can't find these exact quantiles. Instead, we'll try to find a way to express this idea in term of quantiles from a distribution w/ no unknown parameters

Using properties of the Gamma distb'n we note that

$$\underline{\tau \bar{X}} \sim \text{Gamma}(n, n)$$

This is called a "pivot" b/c the distb'n does NOT depend on any unknowns.

Hence



and these quantiles don't depend on any unknowns!

Eg. In R: $\gamma_{(n,n)}(\alpha/2)$ is found w/ the code

"qgamma($\frac{\alpha}{2}$, shape = n, rate = n, lower.tail = T)"

So we have, by definition of quantiles

$$\Pr\left(\gamma_{(n,n)}(\alpha/2) \le \tau\bar{X} \le \gamma_{(n,n)}(1-\alpha/2)\right)$$

$$= \Pr\left(\frac{\gamma_{(n,n)}(\alpha/2)}{\bar{X}} \le \tau \le \frac{\gamma_{(n,n)}(1-\alpha/2)}{\bar{X}}\right)$$

$$= 1-\alpha$$

Note: If we invert this, we get the same answer as before (w/ the scale parameterization)

Hence

$$\left[\frac{\gamma_{(n,n)}(\alpha/2)}{\bar{X}}, \frac{\gamma_{(n,n)}(1-\alpha/2)}{\bar{X}}\right]$$

is a $(1-\alpha)100\%$ CI for τ. ∎

HW 8 #3b

$X_1, \dots, X_n \overset{IID}{\sim} Exp(\tau)$

Note: This version is consistent w/ the parameterization in your textbook

$$lik(\theta) = \prod_{i=1}^n f(x_i;\tau) = \prod_{i=1}^n \left[\tau e^{-\tau x_i} \mathbb{I}\{x_i \geq 0\}\right] = \tau^n e^{-\tau \sum_{i=1}^n x_i} \mathbb{I}\{x_{(1)} \geq 0\}$$

$\hat{\theta}_{MLE} = \bar{X}$

To do: use the CLT to find an approx. $(1-\alpha)100\%$ CI for τ.

CLT:

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - E[X_1]}{\sqrt{\frac{Var(X_1)}{n}}} \xrightarrow{n \to \infty} N(0,1)$$

for IID sample $X_1, \dots, X_n$

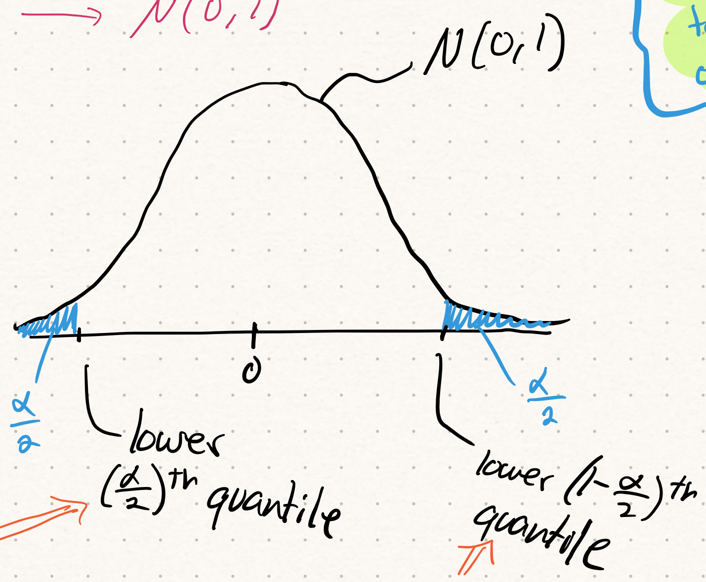
$E[X_1] = \frac{1}{\tau}$, $Var(X_1) = \frac{1}{\tau^2}$

All changes are highlighted in green. View the notes for 9-28-22 to see the other version

Thus we have

$$\frac{\hat{\tau}_{MLE} - \frac{1}{\tau}}{\left(\frac{1}{\tau^2 n}\right)^{1/2}} \xrightarrow{n \to \infty} N(0,1)$$

is a pivot!



Notation:

$z_{(\frac{\alpha}{2})}$ $z_{(1-\frac{\alpha}{2})}$

By definition of quantile:

$$\Pr\left(z_{(\frac{\alpha}{2})} \le \frac{\hat{\tau}_{MLE} - \frac{1}{\tau}}{(\tau^2 n)^{-1/2}} \le z_{(1-\frac{\alpha}{2})} \right)$$

$$= \Pr\left(z_{(\frac{\alpha}{2})} \le \tau\sqrt{n}\left(\bar{X} - \frac{1}{\tau}\right) \le z_{(1-\frac{\alpha}{2})} \right)$$

$$= \Pr\left(z_{(\frac{\alpha}{2})} \le \tau\sqrt{n}\bar{X} - \sqrt{n} \le z_{(1-\frac{\alpha}{2})} \right)$$

$$= \Pr\left(z_{(\frac{\alpha}{2})} + \sqrt{n} \le \sqrt{n}\,\tau\,\bar{X} \le z_{(1-\frac{\alpha}{2})} + \sqrt{n} \right)$$

$$= \Pr\left(\frac{z_{(\frac{\alpha}{2})} + \sqrt{n}}{\bar{X}\sqrt{n}} \le \tau \le \frac{z_{(1-\frac{\alpha}{2})} + \sqrt{n}}{\bar{X}\sqrt{n}} \right)$$

$$= \Pr\left(\frac{\frac{z_{(\frac{\alpha}{2})}}{\sqrt{n}} + 1}{\bar{X}} \le \tau \le \frac{\frac{z_{(1-\frac{\alpha}{2})}}{\sqrt{n}} + 1}{\bar{X}} \right)$$

$$= 1 - \alpha$$

Note: If we invert this, we get the same answer as before (w/ the scale parameterization

Hence $\left[\frac{z_{(\frac{\alpha}{2})}/\sqrt{n} + 1}{\bar{X}}, \frac{z_{(1-\frac{\alpha}{2})}/\sqrt{n} + 1}{\bar{X}} \right]$

is a $(1-\alpha)100\%$ approx. CI for τ.

∎

9-30-22

Bayesian Credible Intervals

Credible Intervals – quantify our personal feelings of uncertainty about the value of a parameter that generated the observed data based on an assumed model

direct assessment of uncertainty

If $X_1, \ldots, X_n$ are IID

$$(X_1, \ldots, X_n) \sim \prod_{i=1}^{n} f(x_i; \theta)$$

$$\theta \sim \pi(\theta)$$

parameter described as a RV

both parts form the assumed model

The observed data $(x_1, \ldots, x_n)$ are realized values from the joint distb'n $\prod_{i=1}^{n} f(x_i; \theta_0)$.

fixed, unknown value of θ that "produced" the observed data

The goal of Bayesian inference is to use the data to describe plausible values for θ_0 though a posterior distb'n

$$\pi(\theta | x_1, \ldots, x_n)$$

Data is fixed, NOT random

A credible interval for θ is a random interval, always.

random b/c it is a function of a RV w/ density $\pi(\theta | x_1, \ldots, x_n)$

Process:

Use the posterior distb'n of θ (given the observed data) to identify a lower bound (LB) and upper bound (UB) on the most plausible values for θ_0.

We choose LB and UB based directly upon quantiles of the posterior.

Interpretation:

We say a W% credible interval [LB, UB], contains θ_0 w/ probability W.

Although this is easier to interpret than a confidence interval, what's harder to communicate is the rationale behind the posterior distb'n.

Example derivation of a Bayesian credible interval

HW 10 #2

To do: use Beta prior to derive posterior distb'n for θ and then find a credible interval for θ

100 items randomly sampled
3 defects found
} Data

θ_0 = proportion of total defective items in the population

$lik(\theta) = \binom{100}{3}\theta^3(1-\theta)^{100-3}$ if we let $X = \begin{cases} 0, & \text{not defective} \\ 1, & \text{defective} \end{cases}$

where $X \sim Bern(\theta)$.

Given
$\pi(\theta) \sim Beta(a, b)$ means $\pi(\theta) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$
is the probability distb'n we are going to use to express our uncertainty about θ_0.

Detour for Notes on $\Gamma(\cdot)$ function

For positive integer a: $\Gamma(a) = (a-1)!$

$\Gamma(a+1) = a\Gamma(a)$

For any a besides negative integers or zero: $\Gamma(a) = \dfrac{\Gamma(a+1)}{a}$

In general, $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\,dt$.

With $lik(\theta)$ and $\pi(\theta)$ we can now find the posterior density:

$$\pi(\theta \mid x_{obs}) = \frac{lik(\theta)\,\pi(\theta)}{\int_\Theta lik(\theta)\,\pi(\theta)\,d\theta}$$

y = # of successes out of 100 trials

$$\pi(\theta \mid y=3) = \frac{\cancel{\binom{100}{3}}\theta^{3}(1-\theta)^{100-3}\cdot\frac{\cancel{\Gamma(a+b)}}{\cancel{\Gamma(a)\Gamma(b)}}\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \cancel{\binom{100}{3}}\theta^{3}(1-\theta)^{100-3}\cdot\frac{\cancel{\Gamma(a+b)}}{\cancel{\Gamma(a)\Gamma(b)}}\theta^{a-1}(1-\theta)^{b-1}\,d\theta}$$

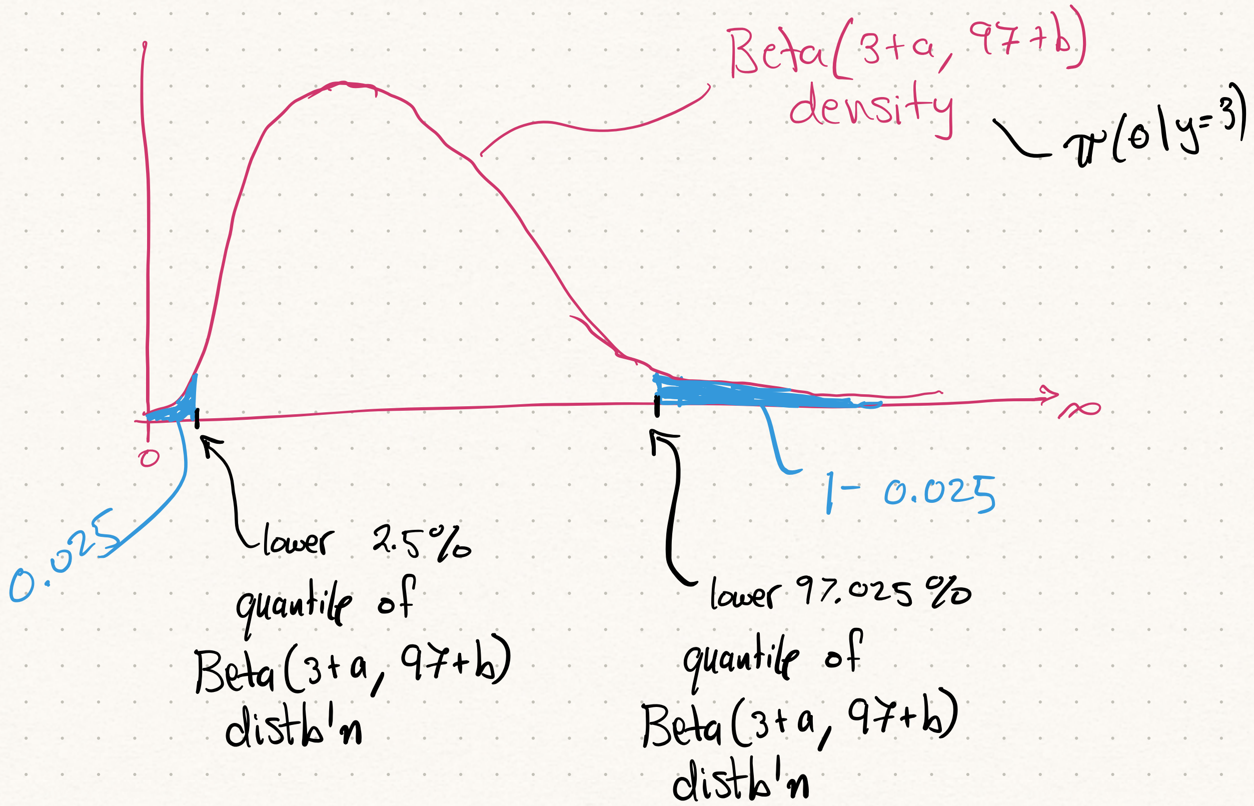
$$= \frac{\theta^{3+a-1}(1-\theta)^{100-3+b-1}}{\int_0^1 \theta^{3+a-1}(1-\theta)^{100-3+b-1}\,d\theta}$$

looks like $\text{Beta}(3+a, 97+b)$

$$\pi(\theta \mid y=3) \sim \text{Beta}(3+a, 97+b)$$

So $\theta \mid x=3 \sim \text{Beta}(3+a, 97+b)$ is the posterior distribution for θ, given the observed data.

For given values of a and b, we can find any quantiles we may want!



In R: qbeta(0.025, shape1 = 3+a, shape2 = 97+b, lower.tail = T)

What we're doing is using the shape of the posterior density to find an interval that describes the most typical values for θ_0. Such credible intervals may also be called highest posterior density regions (hpd for short).

For $W = 95\%$, say,

If $a = b = 1$ then a 95% credible interval for θ_0 is

$$[0.013, 0.842],$$ but

if $a = 0.5$, $b = 5$ then a 95% credible interval for θ_0 is

$$[0.0001, 0.4096].$$

Group Work:

Create a mind-map relating as many theorems from Ch. 8 as you can.

- MLE is consistent
- Identity for Fisher Info
- Asymptotic normality of MLE
- Cramér-Rao lower bound
- Factorization thm for sufficient stats
- MLE is a function of sufficient stat
- Rao-Blackwell Theorem for estimation w/ sufficient statistics