STAT 24410

Lec 1A - 9/30

- Midterm in Nov.
- 6 or 7 HW

Example: Genome-Wide Assoc Studies (GWAS)

- Genetic Association

i) cases & controls

ii) disease risk increasing mutations more freq. in cases
↳ $p_m^{cases} > p_m^{controls}$

iii) Single-nucleotide polymorphism (SN(i)P)
↳ A/G → look if $p_A^{cases} \neq p_A^{control}$, if so, close to mutation location
↳ to do genome-wide, need millions of SNPs

iv) Say n cases, m controls
↳

	A	G	count?
cases	n_A	n_G	$2n$
controls	m_A	m_G	$2m$

chromosomes

v) Estimate p_A^{cases}?
↳ $\widehat{p_A^{cases}} = \frac{n_A}{2n}$
↳ why believe this? well here it's the natural estimator
↳ in more complicated scenarios?
↳ it's unbiased ($E[\widehat{p_A}] = p_A$), good accuracy
↳ we'll learn what makes a best estimator for

vi) Testing. H_0: $p_A^{cases} = p_A^{controls}$, H_a: $p_A^{cases} \neq p_A^{controls}$
↳ also need a test statistic & method

↳ $T = \frac{(\hat{p}_A^{cases} - \hat{p}_A^{controls})^2}{(\frac{1}{2n} + \frac{1}{2m})\hat{p}_A(1-\hat{p}_A)}$, $\hat{p}_A = \frac{n_A + m_A}{2n+2m}$ ← freq. A in combined data

+ we claim $T \overset{H_0}{\sim} \chi^2_1$. Why?

a) $n_A \sim$ Bin (sum of Bernoulli); CLT gives approx. $\mathcal{N}$

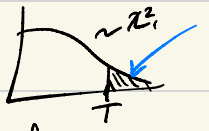
b) $Z \sim \mathcal{N}(0,1) \Rightarrow Z^2 \sim \chi^2_1$

c) this is an approximation, best at the center, where our data is

- is this the "best" test stat.?

↳ criteria to judge: power, type I error

vii) p-value

↳ 

$= 1 - F_{\chi^2_1}(T)$

↳ in fact, $p \overset{H_0}{\sim} Unif(0,1)$ b/c $F_{\chi^2}(T) \sim Unif(0,1)$

↳ only an approx. b/c T was an approx.

viii) 1 million SNPs

↳ now we have $p_1, p_2, \ldots, p_{1m}$

↳ w/ this many p-values, $p < 0.01$ isn't exciting (even under H_0)

- first, suppose we care a/b the smallest p-vals: $p_{(1)} < p_{(2)} < \cdots < p_{(1m)}$

↳ turns out $p_{(1)} \sim Beta(1, n)$, w/ EV $\frac{1}{n+1}$, meaning by chance, we expect to see a $p = \frac{1}{1m+1}$

RVs

- Ω = sample space (e.g. $\mathbb{R}$)
- subsets of Ω = events, elements = sample points

Def: σ-Field. $\mathcal{F}$ is a collection of events s.t.

i) $\Omega \in \mathcal{F}$

ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

iii) $A_n \in \mathcal{F} \Rightarrow \bigcup_n A_n \in \mathcal{F}$

Eg: Borel σ-field.

↳ $\Omega = \mathbb{R}$

↳ $\mathcal{B}$ = smallest σ-field that contains open intervals (a, b)

↳ $[/(a, b)/]$ also in $\mathcal{B}$

Def: Prob. Measure (P). Given $(\Omega, \mathcal{F})$, $P: \mathcal{F} \to [0, 1]$ s.t.

i) $P(\Omega) = 1$

ii) $P(\bigcup_n A_n) = \bigcup_n P(A_n)$ for $A_i \cap A_j = \phi$

Def: RV. on $\sim (\Omega, \mathcal{F}, P)$. X is an RV if $X: \mathcal{F} \to \mathbb{R}$ such that $\forall B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{F}$.

Eg: 2 RVs.

1) $(\Omega_1, \mathcal{F}_1, P_1)$ s.t. $\Omega_1 = \{H, T\}$, $\mathcal{F}_1$ = all subsets, $P_1(\{H\}) = P(\{T\}) = 1/2$

$X_1(\omega) = \begin{cases} 1 & \text{if } H \\ 0 & \text{if } T \end{cases}$

2) $\Omega_2 = [0, 1]$, $\mathcal{F}_2 = \mathcal{B}$, $P_2([b-a]) = b - a$. $X_2(\omega) = \begin{cases} 1 & \text{if } \omega \in [0, 1/2] \\ 0 & \text{otherwise} \end{cases}$

we will only care w/ the dists from here!

- even though these prob. spaces are really diff, the dists are the same
  ↳ $X_1 \overset{D}{=} X_2$

Def. X_1 and X_2 are equal in dist. if $Pr[X_1 \in A] = Pr[X_2 = A] \; \forall A$.

↳ basically an abstraction

Lec 15 - 10/2

- recall $X_1 \overset{D}{=} X_2 \Rightarrow Pr[X_1 \in A] = Pr[X_2 \in A]$

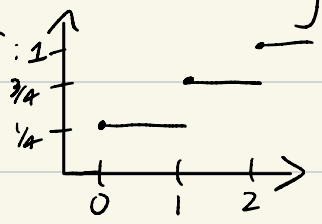
Def: Dist. Fn. X is RV. The CDF of X (denoted by F_X) is a fn. $F_X : \mathbb{R} \to [0,1]$ s.t. $F_X(x) = Pr[X \in (-\infty, x]) = Pr[X \leq x]$

Lemma. $X \overset{D}{=} Y$ iff $F_X(t) = F_Y(t)$ $\forall t \in \mathbb{R}$. No proof.

Eg.

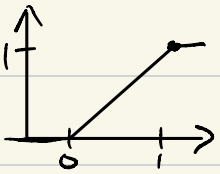
1) Binomial$(2, 1/2)$ (tossing 2 coins)

↳ CDF:



2) Unif$(0,1)$

↳



$F(x) = x$ $\forall x \in [0,1]$

CDF Properties.

1) $x \leq y \Rightarrow F(x) \leq F(y)$
2) right-continuous
3) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$

Questions

a) if $F : \mathbb{R} \to [0,1]$ has properties 1-3, is F a cdf?

b) F is a cdf. How do we simulate RVs w/ this cdf?

Inverse of the Dist. Fn.

+ Case 1: F is continuous & strictly increasing.
  - ↳ $F^{-1}: [0,1] \to \mathbb{R}$ exists; continuous & strictly increasing
- Result 1: say X is RV w/ cdf F
  - ↳ $F(X)$ is a RV
  - ↳ $Pr[F(X) \le u] = Pr[X \le F^{-1}(u)] = F(F^{-1}(u)) = u \to Unif(0,1)!$
    - $\Rightarrow F(X) \sim Unif(0,1)$
- this does not generalize; counterexample = discrete RVs

- Result 2: Let $U \sim Unif(0,1)$. $F^{-1}(U)$ is a RV
  - ↳ $Pr[F^{-1}(U) \le x] = Pr[U \le F(x)] = F(x) \Rightarrow F^{-1}(U) =_D X$

- thus, we can simulate a RV w/ F^{-1} by simulating $Unif(0,1)$, $F^{-1}(U)$
- result 2 generalizes
  - ↳ $F^{-1}(u) = \inf\{x : u \le F(x)\}$ (Q 2)
  - ↳ it can be proved that $F^{-1}(U) =_D X$ (Q 1)
- these results relevant for p-values (by simulation)
- and ordered statistics
  - ↳ $X_1, \dots, X_n$ iid F, $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$

↳ let $U_1, \dots, U_n$ iid $\mathcal{U}(0,1)$

↳ now $F^{-1}(U_1), \dots, F^{-1}(U_n)$ iid F, can show $F^{-1}(U_{(k)}) =_D X_{(k)}$

↳ can learn dist. of $X_{(k)}$ from dist. of $U_{(k)} \sim \text{Beta}(k, n-k+1)$

RVs

Def: Discrete RV. Has a countable support: $S = \{x_1, x_2, \dots\}$, $\Pr[X \in S] = 1$

Notes

a) F_X is piecewise constant if S is the set of isolated points (↗ ε ball, no other points w/in)

b) Dist. is completely determined by $\Pr[X = x_i]$, called the prob. mass function. (↗ pmf)

Eg.

1) Binomial(n, p), $0 < p < 1$, $n \geq 1$

↳ # successes in n trials

↳ $S = \{0, 1, \dots, n\}$

↳ $\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$

- $\text{Bin}(1, p) = \text{Bernoulli}(p)$

↳ if $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, $\sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$

2) Poisson(λ), $\lambda > 0$

↳ $S = \{0, 1, \dots\} = \mathbb{N}$ ↳ memoryless arrivals

↳ $\Pr[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}$

- if $\lim_{n\to\infty} np_n = \lambda$, $\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}$
  ↳ Bin → large, but np_n converges, approx. by poisson

Pf.

$$\binom{n}{k} p_n^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p_n^k (1-p_n)^k = \frac{1}{k!} \cdot \frac{(n)(n-1)\cdots(n-k+1)}{n^k} (np_n)^k (1-p_n)^{-k} \left(1-\frac{np_n}{n}\right)^n$$

in limit... same → 1, λ^k, 1, $e^{-\lambda}$ ($a_n \to a$, $(1+\frac{a_n}{n})^n \to e^a$)

- $X \sim Poi(\lambda_1)$, $Y \sim Poi(\lambda_2) \Rightarrow X+Y \sim Poi(X_1+X_2)$
  ↳ proof = exercise

3) Geometric (in notes)

+ Transformations
- X is discrete w/ support S, $f: \mathbb{R} \to \mathbb{R}$, $f(X) = ?$
  ↳ $f(X)$ is discrete
  ↳ $Pr[f(X)=y] = \sum_{\substack{x \text{ s.t.} \\ f(x)=y}} Pr[X=x]$

Def: Continuous RVs. RV X has a density if $\exists f: \mathbb{R} \to [0,\infty)$ s.t. $Pr[X \in A] = \int_A f(x)\,dx$. ($\forall A$ in Borel)

- $1 = Pr[X \in \mathbb{R}] = \int_{\mathbb{R}} f(x)\,dx$
- density is not unique
  ↳ can modify a countable # of points w/o changing integral

- $F(x) = Pr[X \in (-\infty, x]] = \int_{-\infty}^{x} f(t)\,dt$
- $f(x) = "F'(x)$
  - ↳ remember $\mathcal{U}(0,1)$; not differentiable everywhere

Eg.

1) $\mathcal{U}(a,b)$
   - ↳ $f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & o.w. \end{cases}$, $F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 0 & x > a \end{cases}$

2) $\mathcal{N}(\mu, \sigma^2)$
   - ↳ $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$
   - ↳ standard normal: $Z \sim \mathcal{N}(0,1)$.
     - ↳ $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
     - ↳ $F(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$

3) read a/b gamma, exp, beta

Lec 2A - 10/7

Transformations of Cont. RVs

- say X is a RV w/ density f_X
  - ↳ $h: \mathbb{R} \to \mathbb{R}$, $Y = h(X)$
  - ↳ what is f_Y?

+ Case 1: h strictly increasing & differentiable

$$F_Y(y) = Pr[Y \le y] = Pr[h(X) \le y] = Pr[X \le h^{-1}(y)] = F_X(h^{-1}(y))$$
$$\Rightarrow f_Y(y) = F_X'(h^{-1}(y)) = f(h^{-1}(y)) \frac{\partial h^{-1}(y)}{\partial y}$$

+ General Case: h is 1-to-1, differentiable

$$f_Y(y) = f_X(h^{-1}(y)) \left| \frac{\partial h^{-1}(y)}{\partial y} \right|$$

E.g. $X \sim U(-\frac{\pi}{2}, \frac{\pi}{2})$, $h(x) = \tan(x)$

$$\Rightarrow f_X = \begin{cases} \frac{1}{\pi} & x \in [-\frac{\pi}{2}, \frac{\pi}{2}] \end{cases}$$

↳ note $h: (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$, $h^{-1}(y) = \arctan(y)$, $(h^{-1})'(y) = \frac{1}{1+y^2}$

$$\Rightarrow f_Y = \frac{1}{\pi} \frac{1}{1+y^2}, \quad y \in \mathbb{R}$$

↳ Fact: $Y \sim$ Cauchy

E.g. Many-to-one transformations.

$X \sim N(0,1)$, $Y = X^2$ $(Y \sim \chi^2_1)$

Look into $y > 0$: $F_Y(y) = Pr[Y \le y] = Pr[X^2 \le y]$

$$= Pr[-\sqrt{y} \le X \le \sqrt{y}]$$

$$= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{x^2}{2}} dx$$

Letting $x^2 = t$: $\frac{2}{\sqrt{2\pi}} \int_0^y e^{-t/2} \frac{1}{2\sqrt{t}} dt$

$$= \int_0^y \frac{1}{\sqrt{2\pi}} e^{-t/2} \frac{1}{\sqrt{t}} dt$$

Thus, $f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-y/2} y^{-1/2} : y > 0, \\ 0 \text{ o.w.} \end{cases}$

- "The trick": $\{X^2 \le x\} = \{0 \le x \le \sqrt{x}\} \cup \{-\sqrt{x} \le x \le 0\}$ → disjoint, h is 1-to-1 on each.

Expectation

Intuition: $E[X] \text{ "=" } \sum x \Pr[X \in (x-\varepsilon, x]]$

$$= \sum x \left(F(x) - F(x-\varepsilon)\right)$$

$\varepsilon \to 0$: $\int_{-\infty}^{\infty} x \, dF(x)$

+ Case 1: $X \ge 0$

$E[X] = \int_0^{\infty} x \, dF(x)$ → Riemann-Stiltjes

$$= \begin{cases} \int_0^{\infty} x f(x) \, dx & \text{if } X \text{ has density } f \\ \sum_x x \Pr[X = x] & X \text{ discrete} \end{cases}$$

+ General X

Let $X^+ = \max\{x, 0\}$, and $X^- = \max\{-X, 0\}$. Then $X = X^+ - X^-$.

Def. X has an e.v. if at least one of $E[X^+]$ and $E[X^-]$ is finite. Then $E[X] = E[X^+] - E[X^-]$.

E.g. $X \sim$ Cauchy; $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$, $x \in \mathbb{R}$

$$E x^+ = \int_0^\infty x \cdot \frac{1}{\pi(1+x^2)}\,dx$$

$$= \frac{1}{2\pi}\log(1+x^2)\Big|_0^\infty = +\infty$$

$$E x^- = -\infty$$

} no e.v.!

Properties.

1) X has a density f s.t. $\int_{-\infty}^{\infty} |x| f(x)\,dx < \infty \Rightarrow X$ has a finite expectation:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$$

2) $E[X+Y] = E[X] + E[Y]$

3) $E[cX] = cE[X]$

4) $X \le Y \Rightarrow E[X] \le E[Y]$

5) **Def.** $X \preceq Y$ (Y stochastically dominates X) if $Pr[X > t] \le Pr[Y > t]$.

$X \preceq Y \Rightarrow E[X] \le E[Y]$

6) $E[\mathbb{1}_A(X)] = Pr[X \in A]$

7) $X \perp Y \Rightarrow E[XY] = E[X] \cdot E[Y]$

8) $E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\,dx$

9) $Var(X) = E[(X - E[X])^2]$ (if $E[X] < \infty$)

↳ measure of spread

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

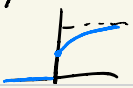
b) $Var(X) = 0 \Leftrightarrow Pr[X = E[X]] = 1$

c) $Var(c \cdot X) = c^2 Var(X)$

d) $X \perp Y \Rightarrow Var(X+Y) = Var(X) + Var(Y)$

10) Mixed Distributions

↳ eg: $Z \sim N(0,1)$, $Z^+ = \max(Z, 0)$

- $Pr[Z=0] = 1/2$:  cdf

- no density; no pmf

Lemma: $X \geq 0$ s.t. F_X is cont. differentiable on $[0,\infty) \setminus S$, where S is a countable set of isolated points $\{x_1, \ldots, x_n\}$.

$$\Rightarrow E[X] = \sum_S x_i Pr[X = x_i] + \int_0^\infty x F_X'(x)\,dx$$

↳ back to Z^+

$$E[Z^+] = 0 \cdot Pr[Z^+ = 0] + \int_0^\infty x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

$$= \frac{1}{\sqrt{2\pi}} - e^{-\frac{x^2}{2}} \Big|_0^\infty = \frac{1}{\sqrt{2\pi}}$$

11) Probability Inequalities

a) Markov

$X \geq 0$, $c > 0 \Rightarrow Pr[X \geq c] \leq \frac{E[X]}{c}$

↳ equality if $Pr[X=0, X=c] = 1$

b) Chebyshev

$c > 0$; $Pr[|X - \mu| \geq c] \leq \frac{Var(X)}{c^2}$

c) Chernoff

$c > 0$, $t \in \mathbb{R}$; $Pr[X \geq c] \leq e^{-tc} E[e^{tx}]$

$$\Pr[X \le c] \le e^{tc} E[e^{-tx}]$$

Proofs

a) $Y = c$ if $X \ge 0$, else 0.

$Y \le X \Rightarrow E[Y] \le E[X] \Rightarrow c \Pr[X \ge c] \le E[X]$

b) Markov for $(X-\mu)^2$

c) Markov for e^{tx}, e^{-tx}

Application 1. Weak Law of Large #s.

Suppose $X_1, \dots, X_n$ iid RVs s.t. $E[X] = \mu < \infty$, $Var(X) = \sigma^2 < \infty$.

Let $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$.

$\Rightarrow \Pr[|\bar{X}_n - \mu| < \varepsilon] \to \infty$ as $n \to \infty$

↳ "converges in probability": $\bar{X}_n \xrightarrow{p} 0$

Pf. $E[\bar{X}_n] = \mu$, $Var(\bar{X}_n) = \frac{\sigma^2}{n}$. By Chebyshev, $\Pr[|\bar{X}_n - \mu| \ge \varepsilon] \le \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$.

Application 2: Large Deviations.

$X_1, \dots, X_n$ iid RVs, finite expectations.

What is $\lim_{n\to\infty} \frac{1}{n} \log\left(\Pr[\bar{X}_n \ge c]\right) (c > \mu)$?

$\Pr[\bar{X}_n \ge c] \to 0$, but how fast?

↳ Chernoff: $\Pr[\sum_i X_i \ge nc] \le e^{-tnc} E[e^{t(\sum_i X_i)}] = e^{-tnc + nk(t)}$ (proof not on midterm)

$k(t) = \log E[e^{tx}]$ = cumulative generating fn.

$\Rightarrow \frac{1}{n} \log \Pr[\bar{X}_n \ge c] \le -[tc - k(t)] \quad \forall t$

$\Rightarrow \le -\sup_t [tc - k(t)] = -I(c)$ → Legendre Transform

(in fact, the $\lim = I(c)$!)

Lec 23 - 10/9

CLT

Def: Convergence in Distribution. X_n converges in dist to X ($X_n \xrightarrow{\mathcal{D}} X$) if $F_{X_n}(t) \to F_X(t)$ $\forall t$.

Def: Moment Gen. Fn. $M: \mathbb{R} \to [0, +\infty]$, $M(t) = E[e^{tX}]$ $\forall t$ (mgf).

Notes:

a) $e^{tX} = \sum_{n=0}^{\infty} \frac{(tX)^n}{n!}$. $E[e^{tX}] = \sum_{n=0}^{\infty} E[X^n] \frac{t^n}{n!}$ → only true if finite sum or Fubini Thm (not important)

↳ so $M^{(n)}(0) = E[X^n]$ = nth moment of X

b) $M_{aX}(t) = M_X(at)$

c) $X \perp Y$, $M_{X+Y}(t) = M_X(t) M_Y(t)$ → powerful; easy to find moments of cumulation of iid RVs

d) $X_1, \dots, X_n$ iid, $M_{\sum X_i}(t) = (M_X(t))^n$

e) $X_n \xrightarrow{\mathcal{D}} X$ iff $M_{X_n}(t) \to M_X(t)$ in a neighborhood of 0

↳ can check convergence of distributions by conv. of fn.s!

CLT Proof.

$X_1, \dots, X_n$ iid $E[X_i] = 0$, $Var(X_i) = \sigma^2$. WTS: $\frac{\sum X_i}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(?, ?)$

↳ $M_{\frac{\sum X_i}{\sqrt{n}}}(t) = \left[M_X\left(\frac{t}{\sqrt{n}}\right)\right]^n$

↳ Taylor: $M_X\left(\frac{t}{\sqrt{n}}\right) = M_X(0) + \frac{t}{\sqrt{n}} M_X'(0) + \frac{t^2}{2n} M_X''(0) + \cdots$

$\approx 1 + 0 + \frac{t^2}{2n}\sigma^2$

$\Rightarrow \left[1 + \frac{t^2\sigma^2}{2n}\right]^n \xrightarrow[n\to\infty]{} e^{\frac{t^2\sigma^2}{2}}$ = mgf for $N(0, \sigma^2)$

Complete proof uses characteristic fn: $E[e^{itx}]$

↳ $|e^{itx}| = 1$; nice integrals, easier math

Random Vectors

$X = (X_1, \dots, X_n) : (\Omega, \mathcal{F}, Pr) \to \mathbb{R}^n$

Def. $X_1, \dots, X_n$ are indep. if $Pr[X_1 \in A_1, \dots, X_n \in A_n] = \prod_{i=1}^{n} Pr[X_i \in A_i]$

Def. cdf. $F_X : \mathbb{R}^n \to [0,1]$, $F_X(x) = Pr[X_1 \le x_1, \dots, X_n \le x_n]$

Def: Discrete RVecs. $\exists\ S \subseteq \mathbb{R}^n$ countable s.t. $Pr[X \in S] = 1$.

Def. Cont. RVecs. X has a density $f : \mathbb{R}^n \to [0, \infty)$ s.t. $Pr[X \in A] = \int_A f(x)\,dx$

Notes

1) $\int_{\mathbb{R}^n} f(x)\,dx = 1$

2) $F(x) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(y_1, \dots, y_n)\,dy_1 \cdots dy_n$

Thm. If f is cont. & differentiable, X has a density $= \frac{\partial^n F}{\partial x_1 \cdots \partial x_n}$

If $X_1, \dots, X_n$ ind, $F_X(x) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$

↳ if F differentiable: $f_X(x) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ → converse also true; can go other way

Marginal Distributions

$X = (X_1, \dots, X_n)$ w/ density f. Let $I \subseteq \{1, 2, \dots, n\} = S_n$ and $X_I = \{X_i\}_{i \in I}$.

Then, $f_{X_I}(x_I) = \int_{\mathbb{R}^{n-|I|}} f(x_{I_1}, \dots, x_{S_n \setminus I})\,dx_{S_n \setminus I}$, $S_n = \{1, 2, \dots, n\}$

Transformations

$X: (\Omega, \mathcal{F}, Pr) \to H \subseteq \mathbb{R}^n$ (H: open region)

$h: H \to G \subseteq \mathbb{R}^n$ (G: open region) s.t. h is 1-to-1 differentiable and $h^{-1}: G \to H$ is differentiable.

↳ X has a density f_X

↳ Let $Y = h(X)$. Y has a density

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)) \left| J(h^{-1}(y)) \right| & y \in G \\ 0 & \text{o.w.} \end{cases}$$

(abs.)

Def: Jacobian. $g: G \to H \subseteq \mathbb{R}^n$. $J(g(y)) = \det\left(\frac{\partial g_i}{\partial y_j}\right)$

E.g.

$X \sim N(0,1) \perp Y \sim \chi^2_n$. $T = \frac{X}{\sqrt{Y}\sqrt{n}} \sim t_n$. Density?

↳ the defn. of the transform is $\mathbb{R}^n \to \mathbb{R}^n$, but this is $\mathbb{R}^2 \to \mathbb{R}^1$...

↳ Plan: $\begin{pmatrix} X \\ Y \end{pmatrix} \to \begin{pmatrix} T \\ Y \end{pmatrix} \xrightarrow[\text{out } Y]{\text{integrate}} T$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$$

$$f_Y(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{y}{2}} y^{\frac{n}{2}-1}, \quad y > 0$$

$\Gamma(\lambda) = \int_0^\infty e^{-x} x^{\lambda-1} dx, \ \lambda > 0$

$\Gamma(1) = 1, \ \Gamma(\frac{1}{2}) = \sqrt{\pi}$

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

$h(x,y) = \left(\frac{\sqrt{n}\,x}{\sqrt{y}}, y\right)$, $h: \mathbb{R} \times (0,\infty) \to \mathbb{R} \times (0,\infty)$

h^{-1}? $\begin{cases} \frac{\sqrt{n}\,x}{\sqrt{y}} = u \\ y = v \end{cases} \to \begin{cases} x = \frac{\sqrt{v}\,u}{\sqrt{n}} \\ y = v \end{cases} \Rightarrow h^{-1}(u,v) = \left(\frac{\sqrt{v}\,u}{\sqrt{n}}, v\right)$

Jacobian: $\begin{vmatrix} \frac{\sqrt{v}}{\sqrt{n}} & \frac{u}{2\sqrt{vn}} \\ 0 & 1 \end{vmatrix} = \frac{\sqrt{v}}{\sqrt{n}} \Rightarrow |J| = \frac{\sqrt{v}}{\sqrt{n}}$

Density of $h(x, y)$ is

$$g(u,v) = f_X\left(\frac{\sqrt{v}u}{\sqrt{n}}\right) f_V(v) \frac{\sqrt{v}}{\sqrt{n}}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{vu^2}{2n}} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{v}{2}} v^{\frac{n}{2}-1} \cdot \frac{\sqrt{v}}{\sqrt{n}}$$

$$\Rightarrow f_T(t) = \int_0^\infty g(u,v)\, dv$$

Lec 3A - 10/14

- recall that RVec transformations are $\begin{pmatrix} X \\ Y \end{pmatrix} \to \begin{pmatrix} X \\ T \end{pmatrix} \to T$

Mixed Distributions

Eg. $X \sim N(0,1)$, $Y \sim$ Bernoulli(p)
- (X, Y) not discrete: $\Pr[(X,Y)=(x,y)]=0$
- no density

Eg. (X, X^2), $X \sim N(0,1)$
- only takes values on the (X, X^2) parabola
- $\Rightarrow \Pr[(X, X^2)=(a,b)\times(c,d)]=0$

Cov & Corr

Def. (X, Y) RVec, $E[X]=\mu_X<\infty$, $E[Y]=\mu_Y<\infty$. The covariance of X and Y is $\mathrm{Cov}(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]$

Def. Correlation of X & Y: $\mathrm{Corr}(X,Y)=\dfrac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}$
- σ_X = std. dev. of X, $\sigma_X<\infty$ (same for Y)

Properties

1) $\mathrm{Cov}(a+bX, c+bY)=bd\cdot\mathrm{Cov}(X,Y)$
2) $\mathrm{Cov}(X,X)=\mathrm{Var}(X)$
3) $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$
4) $X \perp Y \Rightarrow \mathrm{Cov}(X,Y)=0$
5) $\mathrm{Cov}(X,Y)=0 \not\Rightarrow X \perp Y$

- $Z \sim N(0,1)$, S, T random signs | all $\perp$
- $X=SZ$, $Y=TZ$
- $\mathrm{Cov}(X,Y)=0$, but $X \not\perp Y$ — knowing X gives info about Y up to a sign

Conditional Expectation

+ the Plan
  - ↳ (X, Y) RVec
  - ↳ get dist. $F_{X|Y=y} \to E[X|Y=y] =$ fn. of y
    - ↳ so $E[X|Y] = f(y) =$ a RV

Eg. Y discrete: $\Pr[X \in A, Y \in B] = \sum_{y \in B} \Pr[X \in A | Y=y] \Pr[Y=y]$

$= \int_B \Pr[X \in A | Y=y] \, dF_Y(y)$

Def: Conditional Dist of $X|Y=y$. It is $q: \mathcal{F} \times \Omega \to [0,1]$ s.t.

$\Pr[X \in A, Y \in B] = \int_B q(A, y) \, dF_Y(y)$

Def: Corresponding CDF $F_{X|Y=y}(x) = q((-\infty, x], y) = \Pr[X \le x | Y=y]$

Thm.

a) If (X, Y) has density $f(x,y)$, $X|Y=y$ has a density $f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$.

b) If (X, Y) discrete w/ pmf p, $X|Y=y$ is discrete w/ pmf $\frac{\Pr[X, Y]}{\Pr[Y]}$

Recap: $E[X|Y=y]$ is a real #, $E[X|Y]$ is a RV.

Thm. 1 If $X \perp Y$, $E[X|Y] = E[X]$ w/ $\Pr = 1$

Thm. 2: LoTE / Tower Law $E[E[X|Y]] = E[X]$

Thm. 3: Law of Total Variation $\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$

↳ $\mathrm{Var}(Y|X) = E[Y^2|X] - E[Y|X]^2$

E.g. (Mossel's Dice)

Roll a fair die until you get a 6.

↳ Conditional on the event that all rolls are even, what is $E[\#\text{rolls}]$?

X = # rolls until 6; $X \sim \text{Geom}(1/6)$

↳ $\Pr[X=k] = (1-p)^{k-1}p$, $E[X] = \frac{1}{p}$, $\text{Var}(X) = \frac{1-p}{p^2}$

Wrong answer: 3 (as if the die only had 3 faces)

Correct Answer

N = # times you roll 2 or 4 until you get something else (1, 3, 5, 6)

↳ P = value at the last roll

↳ $N \perp P$

$\Rightarrow E[N] = E[N|P] = E[N|P=6]$

$N \sim \text{Geom}(2/3) \Rightarrow E[N] = 3/2 = E[N|P=6]$

↳ soln is 3/2!

↳ why smaller than 3? thought experiment w/ 1M sided die, still only 2 or 4 before 6

↳ most likely to hit 6 on first roll

New Prob: Y = # odd #s observed

$Y|X=k \sim \text{Bin}(k-1, 3/5)$

$E[Y|X=k] = (k-1)(3/5)$, $k = X$, $E[Y|X] = \frac{3}{5}(X-1)$

$E[Y] = E[E[Y|X]] = E[\frac{3}{5}(X-1)] = \frac{3}{5}(6-1) = 3$

Now: $Var(Y)$

↳ $= E[Var(Y|X)] + Var(E[Y|X])$

↳ $Var(Y|X) = ?$ → use binomial → $= (X-1)\left(\frac{3}{5}\right)\left(\frac{2}{5}\right)$

↳ $Var(E[Y|X]) = Var\left(\frac{3}{5}(X-1)\right) = \frac{9}{25}Var(X)$

Rejection Sampling

recall that $F^{-1}(U) \sim F$ but can't always do

Bayesian inference: posterior $= \frac{\text{prior} \times \text{likelihood}}{\text{constant}}$ ← don't always know

Goal: generate samples from dist. w/ density f_X

Assume:

1) $f(x) = \frac{q(x)}{N}$ → known; → normalizing constant
2) I can sample from another dist. $h(x)$
3) $h(x) \geq c \cdot q(x)$, $c > 0$ (same support)

Algorithm

1) Sample Y from h
2) Draw $U \sim U[0,1]$, $U \perp Y$
3) Accept Y and set $X = Y$ if $U \leq \frac{c \cdot q(Y)}{h(Y)}$

Justification

$$Pr\left[x \leq X \leq x+\varepsilon\right] = Pr\left[x \leq Y \leq x+\varepsilon \,\middle|\, U \leq \frac{c \cdot q(Y)}{h(Y)}\right] = \frac{Pr\left[x \leq Y \leq x+\varepsilon,\ U \leq \frac{c \cdot q(Y)}{h(Y)}\right]}{Pr\left[U \leq \frac{c \cdot q(Y)}{h(Y)}\right]}$$

↳ Numerator $\approx Pr\left[x \leq Y \leq x+\varepsilon,\ U \leq \frac{c \cdot q(x)}{h(x)}\right]$

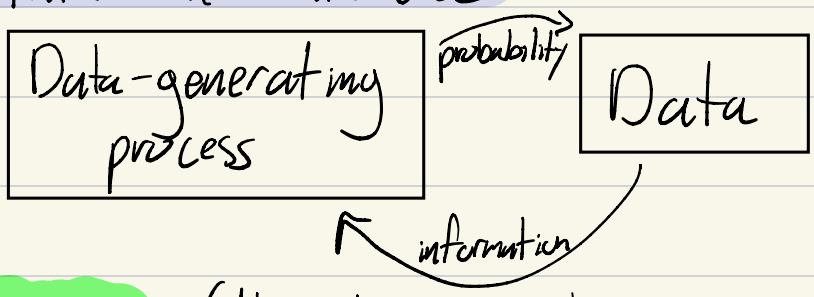
$$= Pr\left[x \leq Y \leq x+\varepsilon\right] Pr\left[U \leq \frac{c \cdot q(x)}{h(x)}\right]$$

$$= \varepsilon \cdot h(x) \cdot \frac{c \cdot q(x)}{h(x)} = \varepsilon \cdot c \cdot q(x)$$

↳ Denominator: $E\left[\mathbb{1}_{\left\{U \leq \frac{c \cdot q(Y)}{h(Y)}\right\}}\right] = E\left[E\left[\mathbb{1} \mid Y\right]\right] = E\left[\frac{c \cdot q(Y)}{h(Y)}\right] = \int \frac{c \cdot q(y)}{h(y)} h(y)\,dy = c \cdot n$

Lec 3B - 10/16

Statistical Inference



Example (the big one)

A firm manufatures batteries & wants to learn their performance from data: $t_1, \ldots, t_n$ = battery durations.

↳ Why do we care?

↳ e.g. T RV models durations; $\Pr[T \geq 10]$?

Model

$T_1, \ldots, T_n$ iid F_T. Goal: find F_T.

↳ We could approx. F_T w/ its empirical dist., as in HW1.
↳ not good for tail probabilities

Dists. Modeling Duration/Lifetime

$T : \Omega : [0, \infty)$ = lifetime/duration of product

Def: Survival Fn. $S : [0,\infty) \to [0,1]$, $S(x) = \Pr[T > x] = 1 - F_T(x)$

$$\Pr[T \leq x + \Delta x \mid T > x] = \frac{\Pr[x < T \leq x + \Delta x]}{\Pr[T > x]} \approx \frac{f(x) \cdot \Delta x}{S(x)}$$

Def: Failure Rate Fn. $h(x) = \frac{f(x)}{S(x)}$, f = density

↳ h can be constant, increasing, decreasing

↳ no aging ↳ aging ↳ reverse aging

Q: If h known, what are f, F_T?

↳ $h(x) = \frac{f(x)}{1-F(x)} = -\frac{(1-F)'(x)}{1-F(x)} = -\frac{d}{dx}\log(1-F(x))$

$\Rightarrow \log(1-F(x)) = -\int_0^x h(t)\,dt + C$

↳ $F(0) \overset{T\geq 0}{=} 0 \Rightarrow C = 0$

$\Rightarrow S(x) = \exp\left(-\int_0^x h(t)\,dt\right)$

$\Rightarrow f(x) = h(x)\exp\left(-\int_0^x h(t)\,dt\right)$.

E.g.

1) $h(x) = \lambda x \quad \forall x$ ($x>0$)

| $\Rightarrow f(x) = \lambda \cdot e^{-\lambda x} \Rightarrow T \sim Exp(\lambda)$

| ↳ Exp. dist. is memoryless: $Pr[X > t+s \mid X > s] = Pr[X > t]$

2) $h(x) = \alpha + \beta x, \quad \alpha, \beta > 0$

| "Gompertz Family"

3) $h(x) = \lambda\beta x^{\beta-1}$

| "Weibull"

Bach 2 Example

Assume $T_1, \ldots, T_n \overset{iid}{\sim} Exp(\lambda)$

↳ λ is our parameter of interest

Questions:

1) How to estimate λ from the data?
2) How good is the estimator?
   ↳ Metric?
3) UQ?

What do we know?

1) $T \sim Exp(\lambda)$

$$E[T] = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx \to \lambda x = y,\ dx = \frac{1}{\lambda} dy$$
$$= \int_0^\infty y \cdot e^{-y} \frac{1}{\lambda} dy = \frac{1}{\lambda} \Gamma(2) = \frac{1}{\lambda}$$
$$E[T^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} dx \to \lambda x = y$$
$$= \int_0^\infty \frac{1}{\lambda} y^2 \cdot e^{-y} \cdot \frac{1}{\lambda} dy = \frac{1}{\lambda^2} \Gamma(3) = \frac{2}{\lambda^2}$$
$$Var[T] = \frac{1}{\lambda^2}$$

2) $\overline{T}_n = \frac{T_1 + \dots + T_n}{n}$

$E[\overline{T}_n] = 1/\lambda$, $Var(\overline{T}_n) = \frac{1}{n\lambda^2}$

↳ LLN: $\overline{T}_n \xrightarrow{P} \frac{1}{\lambda}$

CLT: $\sqrt{n}\left(\overline{T}_n - \frac{1}{\lambda}\right) \xrightarrow{D} N(0, \frac{1}{\lambda^2})$

Estimator for λ

$E[\overline{T}_n] = \frac{1}{\lambda}$

↳ LLN: $\overline{T}_n \sim \frac{1}{\lambda}$ for large n (1st sample moment)

$\Rightarrow \hat{\lambda}_1 = \frac{1}{\overline{T}_n}$ = Method of Moments

Note: what abt 2nd moment?

↳ $E[T^2] = \frac{2}{\lambda^2}$

↳ $\frac{\sum T_i^2}{n} \to \frac{2}{\lambda^2}$, large n

$\Rightarrow \hat{\lambda}_2 = \frac{\sqrt{2n}}{\sqrt{\sum T_i^2}}$

-now we have several estimators — which to use?

Properties of $\hat{\lambda}_1$

1) Consistency.

Thm (Continuous Mapping). $\mathcal{D}$

a) $Z_n \xrightarrow{p} b$, g continuous $\Rightarrow g(Z_n) \xrightarrow{p} g(b)$

b) $Z_n \xrightarrow{\mathcal{D}} Z$, g continuous $\Rightarrow g(Z_n) \xrightarrow{\mathcal{D}} g(Z)$

↳ Thus, $\overline{T}_n \xrightarrow{p} \frac{1}{\lambda} \Rightarrow \frac{1}{\overline{T}_n} \xrightarrow{p} \lambda$

Lemma. If $X_n \xrightarrow{\mathcal{D}} X$, $Y_n \xrightarrow{\mathcal{D}} C$ (constant) $\Rightarrow X_n + Y_n \xrightarrow{\mathcal{D}} X + C$, $X_n Y_n \xrightarrow{\mathcal{D}} C \cdot X$

Thm (Delta Method). X_n s.t. $\sqrt{n}(X_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, g cont. differentiable

$\Rightarrow \sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2 (g'(\theta))^2)$

Pf.

Taylor: $g(X_n) = g(\theta) + g'(\tilde{\theta}_n)(X_n - \theta)$, $\tilde{\theta}_n \in [\theta, X_n]$

$\sqrt{n}[g(X_n) - g(\theta)] = g'(\tilde{\theta}_n)\sqrt{n}(X_n - \theta)$

$g'(\tilde{\theta}_n) \xrightarrow{p} g'(\theta)$; $\sqrt{n}(X_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$ $\Rightarrow$ the product converges

For us, $g(x)=\frac{1}{x}$, $g'(x)=\frac{1}{x^2}$, $\sqrt{n}\left(\bar{T}_n-\frac{1}{\lambda}\right)\xrightarrow{D}\mathcal{N}\left(0,\frac{1}{\lambda^2}\right)$

↳ Delta Method: $\sqrt{n}\left(\frac{1}{\bar{T}_n}-\lambda\right)\xrightarrow{D}\mathcal{N}(0,\lambda^2)$

↳ $\hat{\lambda}_1$ is asymptotically normal

For large n, $\hat{\lambda}_1 \approx \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$

↳ $\frac{\hat{\lambda}_1-\lambda}{\lambda/\sqrt{n}} \approx \mathcal{N}(0,1)$

$$\Rightarrow Pr\left[-1.96 \le \frac{\hat{\lambda}_1-\lambda}{\lambda/\sqrt{n}} \le 1.96\right] \approx 0.95 \Leftrightarrow Pr\left[-1.96+\sqrt{n} \le \frac{\hat{\lambda}_1}{\lambda} \le 1.96+\sqrt{n}\right] \approx 0.95$$

$$Pr\left[\frac{\hat{\lambda}_1}{1.96+\sqrt{n}} \le \lambda \le \frac{\hat{\lambda}_1}{-1.96+\sqrt{n}}\right] \approx 0.95 \rightarrow 95\%\ CI$$

How about small n?

Def. $X \sim Gamma(\alpha,\beta)$ if $f_X(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}\cdot x^{\alpha-1}e^{-\beta x}$, $x>0$

Properties

1) $E[X]=\frac{\alpha}{\beta}$, $Var[X]=\frac{\alpha}{\beta^2}$

2) $X_1,\dots,X_n \overset{iid}{\sim} Gamma(\alpha_i,\beta) \Rightarrow \sum X_i \sim Gamma\left(\sum \alpha_i, \beta\right)$

3) $E\left[\frac{1}{X}\right]=\frac{\beta}{\alpha-1}$, $\alpha>1$

$\sqrt{n}(\hat{\lambda}_1 - \lambda) \xrightarrow{D} \mathcal{N}(0, \lambda^2)$
(asymptotic normality)

Lec 4A - 10/21

Midterm

- cheat sheet (2-sided)
  - ↳ can prnt, handwritten or typed
- + Topics
  - ↳ Ch. 1 - Ch. 3 (everything thru estimation)
- no coding problems, no calculator needed
- practice problems provided
  - ↳ may be different (substantially) from HW, but not harder
  - ↳ e.g. 2 probs similar to HW, 1 novel
  - ↳ mix of computation & proof, but some proof
  - ↳ no explicitly tricky integrals, but e.g. $\int_0^\infty e^{-x}$ is expected (or harder)
- problems: e.g. $T_1, \ldots, T_n$
  a) find MoM estimator for λ, b) use LoLN to show convergence

Estimator Selection (Metrics)

Say $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$, w/ $\hat{\theta}$ an estimator of θ.

Def: Bias. $= E[\hat{\theta}] - \theta$.

↳ unbiased when bias $= 0$.

Def: MSE. $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{bias}^2 + \text{variance}$

Small Sample Properties of $\hat{\lambda}_1$

Note: $Exp(\lambda) = Gamma(1, \lambda) \Rightarrow \sum_{i=1}^{n} T_i \sim Gamma(n, \lambda)$

$$\Rightarrow E[\hat{\lambda}_1] = n E[\textstyle\sum T_i] = \frac{n}{n-1} \cdot \lambda$$

$$E[\hat{\lambda}_1^2] = \frac{n^2}{(n-1)(n-2)} \lambda^2$$

↳ $\Rightarrow Var[\hat{\lambda}_1] = \frac{n^2}{(n-1)^2(n-2)} \cdot \lambda^2$

- Twist: can I construct an unbiased estimator?
  - ↳ $\tilde{\lambda} = \frac{n-1}{n} \cdot \hat{\lambda}_1 \Rightarrow E[\tilde{\lambda}] = \lambda$
    - ↳ any tradeoff?
  - ↳ $Var[\tilde{\lambda}] = \left(\frac{n-1}{n}\right)^2 Var(\hat{\lambda}_1)$
    - ↳ so $\tilde{\lambda}$ has smaller bias & variance → doesn't always happen!

Recap

1) $X_1, \ldots, X_n$ iid
2) $\hat{\lambda} = f(X_1, \ldots, X_n)$
3) Evaluate bias, var.
4) Look @ dist of $\hat{\lambda}$ (asymptotic, small sample)
5) CIs

MoM

↳ $X_1, \ldots, X_n$ iid F

↳ $X \sim F$; $E X^k$ = kth moment

↳ $\frac{\sum X_i^k}{n}$ = kth sample moment

- to estimate k parameters, equate first k moments w/ first k sample moments

E.g.

$X_1, \dots, X_n \overset{iid}{\sim} U[0, \theta]$

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

E.g.

$X_1, \dots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$

↳ $\frac{\alpha}{\beta} = \bar{X}$

↳ 2nd moment: $\frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} = \frac{\sum X_i}{n}$

Note: for consistency, we use LLN and continuous mapping thm.

↳ For the asymptotic dist., we use CLT and delta method

↳ MoM only gives asymptotic guarantees

MLE

$X_1, \dots, X_n \overset{iid}{\sim} F_\theta, \quad \theta \in \mathbb{R}^k$

↳ let $f(x, \theta)$ be the density/pmf (or $f(x|\theta)$, $f_\theta(x)$)

Goal: estimate θ

Def: Likelihood. The likelihood of θ given data $X_1, \dots, X_n$ is the fn. $L(\theta) = L_n(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$.

Def: MLE. $\hat{\theta} = \hat{\theta}_n = \operatorname{argmax}_\theta L_n(\theta)$ → pick the θ that makes the data most likely

Def: Log-likelihood. $\ell(\theta) = \ell_n(\theta) = \log(L_n(\theta))$

E.g.

1) $X_1, \dots, X_n \overset{iid}{\sim} Exp(\lambda)$

$f(x, \theta) = \lambda \exp(-\lambda x), \quad x > 0$

$L_n(\theta) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i), \quad x_i > 0$

$\ell_n(\theta) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i$

$\ell_n'(\theta) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$

↳ $\ell_n'(\theta) = 0 \Rightarrow \lambda = \frac{1}{\bar{x}_n}$ → $\ell''(\lambda) = -\frac{n}{\lambda^2} < 0 \Rightarrow$ global max

2) $X_1, \dots, X_n \overset{iid}{\sim} U[0, \theta]$

$f(x, \theta) = \frac{1}{\theta} : 0 < x < \theta$

$L_n(\theta) = \begin{cases} \frac{1}{\theta^n} & 0 < x_i < \theta \ \forall i \in [n] \\ 0 & \text{o.w.} \end{cases}$

↳ $0 < x_i < \theta \ \forall i \iff 0 < \max_i(x_i) < \theta$

B/c $\frac{1}{\theta^n}$ decreasing, $\hat{\theta} = \max_i(x_i)$ = leftmost val in range

3) (2 params) $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \theta = (\mu, \sigma^2)$

$f(x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)}{2\sigma^2}}$

$L_n(\mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sigma^2} \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$

↳ $\ell_n(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$

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

$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$

Lec 4B - 10/23

Likelihood Estimation

E.g.s

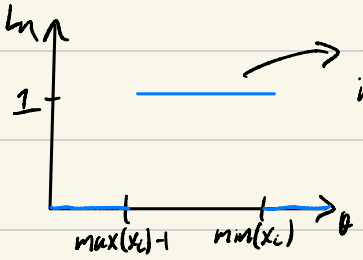
4) Non-Uniqueness

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

$f(x|\theta) = \begin{cases} 1 & \theta < x < \theta+1 \\ 0 & \text{o.w.} \end{cases}$

$L_n(\theta) = \begin{cases} 1 & \theta < x_i < \theta+1 \ \ \forall i \in [n] \\ 0 & \text{o.w.} \end{cases} \iff \theta < \min(x_i) \le \max(x_i) < \theta+1$

$\max(\theta)-1 < \theta < \min(x_i)$



→ any value in this interval is MLE

↳ MoM works, tho

5) MLE DNE

X s.t. $X \sim N(0,1)$ w/ $Pr = \frac{1}{2}$, $X \sim N(\mu, \sigma^2)$ w/ $Pr = \frac{1}{2}$

$f_X(x|\mu, \sigma^2) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right] + \frac{1}{2}\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right]$

$X_1, \dots, X_n \overset{iid}{\sim} f_X$

$L_n(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2)$

↳ Lets look at $\mu = X_1$

↳ $L_n(X_1, \sigma^2) \ge \underbrace{\frac{1}{2^n}\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{↳ } x = x_i \text{ in } N(\mu,\sigma^2)} \cdot \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^{n-1} \cdot e^{-\frac{\sum x_i^2}{2}}}_{\text{↳ take } N(0,1) \text{ as lower bound instead}}$

$\sigma \to 0 \Rightarrow L_n(X_1, \sigma^2) \to \infty$

↳ but $\sigma = 0$ not permissible; no MLE

Why MLE?

+ under mild assumptions
  - consistency
  - asymptotic normality
  - asymptotic variance known
  - efficiency (optimal asymptotic variance)
  - invariance

Theory (not for exams) (θ 1-dim)

Def: Score Fn. $\dot{\ell}_n(\theta | X_n) = \frac{\partial}{\partial\theta} \ell_n(\theta)$

Note: $\ell(\theta) = \sum \log f(x|\theta)$

$\dot{\ell}(\theta) = \sum_{i=1}^{n} \frac{f'}{f}(x|\theta)$ $\quad f' = \frac{\partial}{\partial\theta} f$

Score equation: $\dot{\ell}_n(\theta|x) = 0$ gives MLE

Result 1. $E_{\theta_0}\left[\frac{L_n(\theta)}{L_n(\theta_0)}\right] = 1$

Pf.

$E_{\theta_0}\left[\frac{L_n(\theta)}{L_n(\theta_0)}\right] = \int \frac{L_n(\theta)}{L_n(\theta_0)} f_n(x, \theta_0)\, dx = 1$

↳ **assumption (Common Support):** $\{x : f(x,\theta) > 0\}$ does not depend on θ

↳ e.g. this is why the $U[\theta, \theta+1]$ didn't work!

Result 1 is also $E_{\theta_0}\left[e^{\ell_n(\theta)-\ell_n(\theta_0)}\right]=1$

Result 2. a) $E_{\theta_0}[\dot{\ell}_n(\theta_0)]=0$. b) $-E_{\theta_0}[\ddot{\ell}(\theta_0)]=E[\dot{\ell}(\theta_0)]^2$ (2nd derivative)

Pf.

$E_{\theta_0}\left[e^{\ell_n(\theta)-\ell_n(\theta_0)}\right]=1$

$\hookrightarrow E_{\theta_0}\left[\frac{\partial}{\partial\theta}e^{\ell_n(\theta)-\ell_n(\theta_0)}\right]=0$

$\Rightarrow E_{\theta_0}\left[\dot{\ell}_n(\theta)\cdot e^{\ell_n(\theta)-\ell_n(\theta_0)}\right]=0$

$\hookrightarrow \theta=\theta_0 \Rightarrow$ @

For b) take $\frac{\partial}{\partial\theta}E[\dot{\ell}_n(\theta)]$

Assumption 2: Smoothness of densities. $\frac{\partial}{\partial\theta}E_{\theta_0}[g(x,\theta)]=E_{\theta_0}\left[\frac{\partial}{\partial\theta}g(x,\theta)\right]$

Def: Fisher Information. $I(\theta)=E_\theta[\dot{\ell}^2(\theta)]=E_\theta[-\ddot{\ell}(\theta)]$ $(n=1)$

Notes:

1) $E_\theta[\ddot{\ell}(\theta)]=0 \Rightarrow I(\theta)=Var(\dot{\ell}(\theta))$
2) $Var(\dot{\ell}_n(\theta))=n\cdot I(\theta)$

Thm: Cramer-Rao Inequality. $T(X_n)$ is unbiased for $g(\theta)$. Thus, $Var(T(X_n)) \geq \frac{(g'(\theta))^2}{n\cdot I(\theta)}$.

Note: if $g(\theta)=\theta$, bound is $\frac{1}{n\cdot I(\theta)}$. → high information = low variance

Pf.

$\text{Correlation}^2 \le 1 \Rightarrow \frac{\text{Cov}^2}{\text{Var Var}} \le 1$

$\text{Cov}_\theta^2(T(X_n), \dot{\ell}_n) \le \underbrace{\text{Var}(T(X_n))}_{\text{what we want}} \underbrace{\text{Var}(\dot{\ell}_n)}_{= nI(\theta)}$

$\downarrow$

$\text{Cov}_\theta(T(X_n), \dot{\ell}_n) = E_\theta[T(X_n) \cdot \dot{\ell}]$ → b/c $E[\dot{\ell}_n] = 0$

$= \int T(X_n) \dot{\ell}_n \cdot f_n(\theta, x)\, dx$

↳ Claim: $= \frac{\partial}{\partial\theta} f_n(\theta, x)$

$= \int T(X_n) \frac{\partial}{\partial\theta} f_n(\theta, x)\, dx$

$\overset{?}{=} \frac{\partial}{\partial\theta} \int T(X_n) f_n(\theta, x)\, dx = \frac{\partial}{\partial\theta} g(\theta)$ ← unbiased expectation

MLE Theory

Let $\hat{\theta}_n$ s.t. $0 = \dot{\ell}_n(\hat{\theta}_n) \cong \dot{\ell}(\theta_0) + (\hat{\theta}_n - \theta_0)\ddot{\ell}(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \dddot{\ell}_n(\theta_0)$

$\Rightarrow \sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \dot{\ell}_n(\theta_0) - \frac{1}{n} \ddot{\ell}_n(\theta_0) - \frac{1}{2n}(\hat{\theta} - \theta_0)\dddot{\ell}(\theta_0)$

Notes:

a) $\frac{1}{\sqrt{n}} \dot{\ell}_n(\theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0))$

b) $-\frac{1}{n}\ddot{\ell}(\theta_0) \xrightarrow{P} I(\theta_0)$

c) Last term $\to 0$ as $n \to \infty$

↳ need consistency

$\Rightarrow \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \frac{1}{I(\theta)})$

↳ gives asymptotic normality, asymptotic variance, asymptotic optimality

Thm. $P_{r_{\theta_0}}[\ell_n(\theta_0) > \ell_n(\theta)] \xrightarrow{n\to\infty} 1$ (proof of consistency)

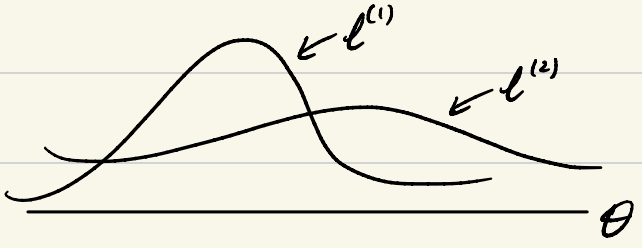
↳ for any fixed θ

- proof requires Jensen

Assumption 3: $\theta_1 \neq \theta_2 \Rightarrow f(x, \theta_1) \neq f(x, \theta_2)$; distinct densities

MLE always consistent, not necessarily unbiased

Lec 5A - 10/28



Which is better to learn about θ?

↳ $\ell^{(1)}$ = peaked likelihood = good ⇒ large derivative

$\ell^{(2)}$ = flat likelihood = bad

$E[\dot{\ell}^2]$ large $= I(\theta)$

$$0 = \dot{\ell}(\theta_0) \cong \dot{\ell}_n(\theta_0) + (\hat{\theta} - \theta_0)\ddot{\ell}_n(\theta_0)$$

$$\Rightarrow \sqrt{n}(\hat{\theta} - \theta_0) = \frac{\frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0)}{-\frac{1}{n}\ddot{\ell}_n(\theta_0)}$$

$$\Rightarrow \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

→ gives asymptotic normality, optimality

E.g. $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(θ)

$$\Rightarrow f(x|\theta) = \theta^x(1-\theta)^{1-x}, \quad x = 0, 1$$

Then $\ell(\theta) = x \log\theta + (1-x)\log(1-\theta)$

$$\dot{\ell}(\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}, \quad \ddot{\ell}(\theta) = \frac{-x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

Score equation: $0 = \dot{\ell}_n(\theta) = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1-\theta}$

$$\Rightarrow \hat{\theta} = \bar{X}_n = \text{sample proportion}$$

$$I(\theta) = E[-\ddot{\ell}(\theta)] = E\left[\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}\right] = \frac{1}{\theta} - \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$$

We know $E[\hat{\theta}] = E[\bar{X}_n] = \theta$ (unbiased)

↳ $\mathrm{Var}(\bar{X}_n) = \frac{1}{n}\mathrm{Var}(X) = \frac{\theta(1-\theta)}{n}$ (smallest var from CR)

$\Rightarrow \hat{\theta}$ = MVUE (min. var. unbiased estimator)

But also, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \theta(1-\theta))$

$$\Rightarrow \hat{\theta} \cong \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right) \text{ for large } n$$

$$\Rightarrow \Pr\left[-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} < z_{\alpha/2}\right] \cong 1 - \alpha \text{ for } z_{\alpha/2} \text{ quantile of } \mathcal{N}(0,1)$$

↳ how do we isolate θ? → want $\Pr[A(x) < \theta < B(x)] = 1 - \alpha$

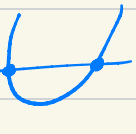
Two Confidence Intervals

1) $\hat{\theta} \pm z_{\alpha/2}\sqrt{n\hat{\theta}(1-\hat{\theta})}$ $\left(\hat{\theta}(1-\hat{\theta}) \cong \theta(1-\theta)\right)$

↳ uses second approximation; the approx. gets worse

2) $Pr\left[\frac{(\hat{\theta}-\theta)^2}{\theta(1-\theta)/n} < z_{\alpha/2}^2\right] = 1-\alpha$

$= Pr\left[(\hat{\theta}-\theta)^2 < z_{\alpha/2}^2 \frac{\theta(1-\theta)}{n}\right]$

↳ gives quadratic: 

Invariance of MLE

Thm. $\hat{\theta}_n$ is MLE of θ and $\tau = g(\theta)$, then $\hat{\tau} = g(\hat{\theta}_n)$ is the MLE of τ.

E.g. $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$.

↳ MLE of μ is $\bar{X}_n$.

$\Rightarrow$ MLE $\mu^2 = (\bar{X}_n)^n$

MLE $e^{\mu} = e^{(\bar{X}_n)}$

Pf.

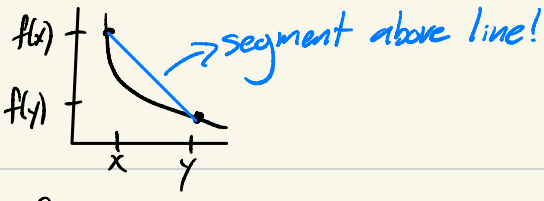
$\hat{\theta}_n = \underset{\text{argmax}}{} L_n(\theta|x)$

Case 1: g bijective.

$\hat{\tau}_n = g(\hat{\theta}) \Rightarrow \hat{\theta}_n = g^{-1}(\hat{\tau})$

$L(\tau) = L(\theta) \leq L(\hat{\theta}_n) = L(\hat{\tau}_n)$

Don't care abt case 2.



Consistency

Def: Convex Fn. f is convex if $\forall x, y,\ 0<\alpha<1$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

↳ if this is "<", we have strict convexity

Note

Z is a RV s.t. $\begin{cases} x & Pr = \alpha \\ y & Pr = 1-\alpha \end{cases} \Rightarrow f(E[Z]) \leq E[f(Z)]$ → holds for all RVs!

Thm: Jensen's Inequality. X RV s.t. $E[|X|]<\infty$, f convex (we also assume differentiable), then $f(E[X]) \leq E[f(X)]$.

Pf.

Tangent has slope $f'(y)$ and is below f.

$\Rightarrow f(x) \geq f(y) + f'(y)(x-y)$ (first-order Taylor)

Let $\mu = E[X]$. Then $f(X) \geq f(\mu) + f'(\mu)(X-\mu)$

$$E[f(X)] \geq f(\mu) = f(E[X])$$

Notes

a) If f strictly convex & $Pr[X=A]<1$ $\forall A$, we get a strict ineq.

b) f convex $\Rightarrow -f$ concave

↳ so if f concave, $f(EX) \geq E f(X)$

c) $f(x)=x^2$ convex $\Rightarrow (E[X])^2 \leq E[X^2]$

$g(x)=\log x$ concave $\Rightarrow \log E[X] \geq E[\log X]$

d) $x_1, \dots, x_n > 0$, $\rho_i \geq 0$ s.t. $\sum \rho_i = 1$

$\Rightarrow \prod_{i=1}^n x_i^{\rho_i} \leq \sum_{i=1}^n \rho_i x_i$

geometric mean — arithmetic mean

Pf: take log, X is RV s.t. $X = x_i$ w/ $Pr = \rho_i$. Then $E \log X \leq \log EX$

e) Holder's Inequality.

$0 \leq X, Y$ RVs, $\frac{1}{p} + \frac{1}{q} = 1$

$\Rightarrow E[XY] \leq (E[X^p])^{1/p} (E[X^q])^{1/q}$

Pf.

Case 1.

$(EX^p)^{1/p} = 1 = (E[Y]^{1/q})^{1/q}$

$XY = (X^p)^{1/p} (Y^q)^{1/q} \overset{d)}{\leq} \frac{1}{p} X^p + \frac{1}{q} Y^q$

$\Rightarrow E[XY] \leq \frac{1}{p} E[X^p] + \frac{1}{q} E[X^q] = 1$

$E[X^p] = 1$ — $E[X^q] = 1$

Case 2.

Use case 1 for $X = \frac{X}{E[X^p]^{1/p}}$, $Y = \frac{Y}{E[Y^q]^{1/q}}$

f) Likelihood Consistency.

$Pr_{\theta_0}[\ell_n(\theta_0) > \ell_n(\theta)] \xrightarrow{n \to \infty} 1$

Pf.

$\frac{1}{n}(\ell_n(\theta) - \ell_n(\theta)) = \frac{1}{n} \sum_{i=1}^n \log \frac{f(x|\theta)}{f(x|\theta_0)} \xrightarrow{Pr_{\theta_0}} E_{\theta_0}\left[\log \frac{f(x|\theta)}{f(x|\theta_0)}\right]$

just log properties

iid RVs

whole quantity is their avg

nontrivial b/c diff density assumption!

not always a single val

$< \log E_{\theta_0}\left[\frac{f(x|\theta)}{f(x|\theta_0)}\right] = \log(1) = 0.$

(Jensen w/ strictly concave log)

$\int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)\, dx = 1$

Lec 5B - 10/30

Multivariate Normal

Def. $X = (X_1, \ldots, X_k)$ has a multivariate normal distribution if $\forall a \in \mathbb{R}^k$, $a^T X$ is normal (any LC is normal).

Notation: $\mu = E[X] = (E[X_1] \ \ldots \ E[X_k])^T$

$\Sigma = Var(X) = E[(X-\mu)(X-\mu)^T] \in \mathbb{R}^{k \times k}$

Notes

a) if Σ is positive-definite $\Rightarrow X$ has density

$$f_X(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \rightarrow |\Sigma| = \det(\Sigma)$$

E.g. $Z \sim \mathcal{N}(0,1)$, $(Z, 1-Z)$ has no density

↳ support is on a line: $x+y=1$; $Cov\begin{pmatrix} Z \\ 1-Z \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ = rank 1

b) (X_1, X_2) bivariate normal & $Cov\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 0 \Rightarrow X_1 \perp X_2$

c) $U \sim \mathcal{N}_k(\mu, \Sigma)$

↳ let $V = a + BU$, $a \in \mathbb{R}^p$, $B \in \mathbb{R}^{p \times k}$

↳ $V \sim \mathcal{N}_p(a + B\mu, B\Sigma B^T)$

Sample Mean & Variance

$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$

↳ $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Notes

a) $E[\bar{X}_n] = \mu$, $Var(S^2) = \sigma^2$

b) $\bar{X} \perp S^2$

Pf. ($n = 2$)

$\bar{X} = \frac{X_1 + X_2}{2}$, $S^2 = \frac{(X_1 - X_2)^2}{2}$

If $X_1 + X_2 \perp X_1 - X_2$, we're done.

↳ note $\begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix}$ is bivariate normal

$Cov(X_1 + X_2, X_1 - X_2) = 0 \Rightarrow (X_1 + X_2) \perp (X_1 - X_2)$ □

c) $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$

$\Rightarrow V = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

$T = \sqrt{n}\frac{\bar{X} - \mu}{S} \sim t_{n-1}$

Confidence Intervals

$X_1, \dots, X_n$ iid F_θ

CIs = probabilistic statements on data of the form

$$Pr_\theta[A(X_1, \dots, X_n) \le \theta \le B(X_1, \dots, X_n)] = \alpha$$

↳ Then $[A(X_1, \dots, X_n), B(X_1, \dots, X_n)]$ is a α- CI

↳ 0.95, 0.99, etc.

Notes

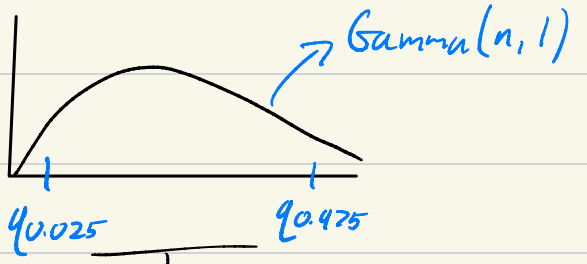
a) ends of the CI are RVs

b) Interpretation: long-term frequency of intervals that cover θ is α.

E.g. $X_1, \ldots, X_n$ iid $Exp(\lambda)$. Recall $X_i \sim Gamma(1, \lambda)$

$\Rightarrow \sum X_i \sim Gamma(n, \lambda)$

Further, $\lambda \sum X_i \sim Gamma(n, 1)$



quantiles := $Pr[Y \le q_\alpha] = \alpha$

$\Rightarrow Pr\left[q_{0.025} \le \lambda \sum X_i \le q_{0.975}\right] = 0.95 \Rightarrow Pr\left[\frac{q_{0.025}}{\sum X_i} \le \lambda \le \frac{q_{0.975}}{\sum X_i}\right] = 0.95$

↳ $\left[\frac{q_{0.025}}{\sum X_i}, \frac{q_{0.975}}{\sum X_i}\right]$ is a 95% CI for λ

+ the key: $\lambda \sum X_i$ has a dist. ⊥ of λ.

Def: Pivot Statistic. $X_1, \ldots, X_n$ iid F_θ. $g(X_1, \ldots, X_n, \theta)$ is called pivot if its dist. does not depend on θ.

If $g(X_1, \ldots, X_n, \theta) \sim F^*$, and L, U quantiles s.t. $Pr[L < g(X_1, \ldots, X_n, \theta) < U]$, we can solve for θ (may not give interval).

Shield #1

E.g.s

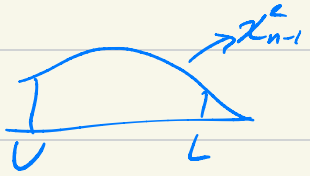
a) $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$

pivot: $g = T = \sqrt{n}\,\frac{\bar{X} - \mu}{S} \sim t_{n-1}$

gives $\bar{X} \pm t_{n-1}^{\alpha/2} \frac{S}{\sqrt{n}}$

b) Normal variance.

pivot g: $V = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ →

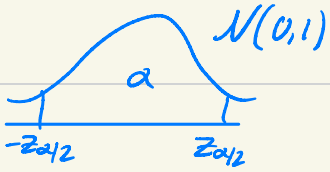


c) Large sample mean.

$X_1, \ldots, X_n$ iid F, $E[X_1] = \mu$, $Var(X_1) = \sigma^2$

$\Rightarrow \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$ (CLT)

$\Rightarrow$ for n large, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1)$



If σ known

$\Rightarrow$ a CI: $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ (approximation)

σ unknown:

$\left[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$ → double approximation

d) MLE. $X_1, \ldots, X_n$ iid F_θ. $\hat{\theta}$ is MLE (valid assumptions)

$\Rightarrow \sqrt{n}(\hat{\theta} - \theta) \approx \mathcal{N}(0, \frac{1}{I(\theta)})$

$\approx \mathcal{N}(0, \frac{1}{I(\hat{\theta})})$ → recall Bernoulli last lec.

e) 2 samples means.

$X_1, \ldots, X_n$ iid $\mathcal{N}(\mu_1, \sigma_1^2) \perp Y_1, \ldots, Y_m$ iid $\mathcal{N}(\mu_2, \sigma_2^2)$

↳ want CI for $\mu_1 + \mu_2$

$\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$

$\Rightarrow (\bar{X} - \bar{Y}) - (\mu_1 - \mu_2) \sim \mathcal{N}(0, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$

★ Ch. 4 on midterm (CIs)

What if we don't have the above scenarios?

↳ Case 1: pivot $g \sim F^*$, F^* unknown

↳ Case 2: T_n is unbiased for θ, but $Var(T_n)$ unknown

Bootstrap

Def: Empirical Dist. F_n Given $X_1, \dots, X_n$ iid F, the EDF $\hat{F}_n$ is the cdf that puts mass $\frac{1}{n}$ at each X_i: $\hat{F}_n(x) = \frac{1}{n} \sum \mathbb{1}_{\{X_i \le x\}}$

Notes

a) $\hat{F}_n$ for fixed x is sample proportion for $Bern(F(x))$

b) $E[\hat{F}_n(x)] = F(x)$

c) $Var(\hat{F}_n(x)) = \frac{1}{n}[F(x)(1-F(x))]$

d) LLN: $\hat{F}_n(x) \xrightarrow{P} F(x)$

e) **Thm (Glivenko-Cantelli):** $\sup_x |\hat{F}_n(x) - F(x)| \longrightarrow 0$ a.s.

Let $\theta = T(F)$

↳ mean $\int x \, dF(x)$

var $\int (x-\mu)^2 \, dF(x)$

median $F^{-1}(1/2)$

Plug in estimator of θ: $\hat{\theta} = T(\hat{F}_n)$ (mean → sample mean)

Lec 6A - 11/4

Midterm

- minimal proofs (similar to HW)
- difficulty on par w/ HW
  ↳ but no copied HW problems

CIs

Pivot Stats: $X_1, \ldots, X_n$ iid F_θ

$g(X_1, \ldots, X_n, \theta) \sim F^*$ is known, ind. of θ

↳ $Pr[L \le g(X_1, \ldots, X_n, \theta) \le R] = 0.95 \rightarrow$ solve for θ

Bootstrap

- relies on empirical distribution fn $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}$

$\theta = T(F)$ mean $\int x \, dF(x)$

$\hat{\theta} = T(\hat{F}_n)$ mean $\int x \, d\hat{F}_n(x) = \bar{X}_n$

E.g. Variance Estimation

$X_1, \ldots, X_n$ iid F, $T_n = g(X_1, \ldots, X_n)$ a statistic, need $Var[T_n]$

Case 0: Can simulate from F

① Take k samples of size n from F.

② For each sample, calculate $g(\cdot)$

↳ $T_{n,1}, T_{n,2}, \ldots, T_{n,k}$

③ Estimate variance w/ $\frac{1}{k}\sum_{j=1}^{k}(T_{n,j}-\bar{T})^2$

General Case

$Var_F[T_n] \approx Var_{\hat{F}_n}[T_n]$

↳ simulate from $\hat{F}_n$ (bootstrap samples ≡ sampling from $X_1, \ldots, X_n$ w/ replacement)

Bootstrap CIs

1. (The Normal Interval) $X_1, \ldots, X_n$ iid F_θ. Let $\hat{\theta} = T_n(X_1, \ldots, X_n)$ be unbiased for θ. $\hat{\sigma}_{boot}$ = bootstrap estimate of sd of θ

The α CI is $[T_n - z_{\alpha/2}\hat{\sigma},\ T_n + z_{\alpha/2}\hat{\sigma}]$

↳ need T_n to be close to normal

2. (Pivot Statistic). $X_1, \ldots, X_n$ iid F_θ, $g(X_1, \ldots, X_n, \theta) \sim F^*$ unknown is our pivot.

We need L, U quantiles for dist. of g

Bootstrap Way:

① Gen bootstrap sample

② Calc. g w/ $\hat{\theta} = \hat{\theta}_{boot}$, $\theta = \hat{\theta}$

e.g. if $g = \hat{\theta} - \theta$, we'd calc. $\hat{\theta}_{boot} - \hat{\theta}$

③ Repeat and get $\hat{L}, \hat{U}$

★ bootstrap not on midterm

Hypothesis Testing

+ "Is this data consistent w/ a certain data-generating mechanism?"

E.g. $X_1, \ldots, X_n$ iid F; is F the cdf of a normal (aka is the data normally distributed)?

Ideas:

1) Plot histogram. Flaw: subjective, other dist.s look similar (t, cauchy)
2) Quantiles (e.g. median, lower/upper quartile)
   - need a comparison; $N(0,1)$ quantiles are known
   - standardize the data: $\frac{X_i - \bar{X}}{s_X}$
   - QQ plot = standard normal vs. sample quantiles, if a line then sample is normal
   - still visual!
3) Quantiles, pt. 2
   ↳ take $\frac{(X_{(\frac{3}{4}n)} - X_{(\frac{1}{4}n)})^2}{\sum (X_i - \bar{X})^2}$ (3rd quartile)

   Generalizes to **Shapiro-Wilks**: $\frac{(\sum a_i X_{(i)})^2}{\sum (X_i - \bar{X})^2}$
4) EDF
   ↳ $\hat{F}_n$ of $\frac{X_i - \bar{X}}{s_X}$, F = cdf of $N(0,1)$
   ↳ **Kolmogorov-Smirnov**: $\sup_x |\hat{F}_n(x) - F(x)|$ (works for any dist.)

Parametric Models

$X_1, \ldots, X_n$ iid F_θ, $\theta \in \Omega$

+ Hypotheses:

Null: $\theta \in \Omega_0 \subset \Omega$

Alternative: $\theta \in \Omega_1 \subset \Omega$

$\left.\begin{array}{l} \Omega_0 \cap \Omega_1 = \emptyset \\ \Omega_0 \cup \Omega_1 = \Omega \end{array}\right\}$

Notes:

a) H_0 = null hypothesis

H_a / H_1 = alternative

b) In our normal example, H_0 = data was normal, H_a = wasn't normal

c) Asymmetry btwn H_0, H_a

↳ data is used to investigate if the data is consistent w/ the null, not alternative

Def. S is the set of all possible values for $(X_1, \ldots, X_n)$. Also, $S = S_0 \cup S_1$, $S_0 \cap S_1 = \emptyset$

↳ S_0 = values for which we do not reject the null (acceptance region)

S_1 = vals for which we reject the null (rejection region)

Normally we define S_1 using a statistic, e.g., $\{x : \sup_x |\hat{F}_n(x) - F(x)|\}$

Def: Test Statistic. T a statistic s.t. $S_1 = \{x : T(x) \in R_1\}$ ($R_1 \subseteq \mathbb{R}$)

E.g. $X_1, \dots, X_n$ iid $N(\mu, 1)$.

$H_0: \mu = \mu_0 \equiv \Omega_0 = \{\mu_0\}$

$H_a: \mu \neq \mu_0 \equiv \Omega_1 = \mathbb{R} \setminus \{\mu_0\}$

$T(x) = |\bar{X}_n - \mu_0|$

$S_1 = \{x : |\bar{X}_n - \mu_0| \geq c\}$ → how to pick c?

Error Types

	$X \in S_0$	$X \in S_1$
$\theta \in \Omega_0$ (H_0)	correct	Type I Error
$\theta \in \Omega_1$ (H_a)	Type II Error	correct

E.g. $X_1, \dots, X_{10}$ iid $\text{Ber}(\theta)$

$H_0: \theta = 1/2$

$H_a: \theta > 1/2$

$T = \sum X_i$; $S_1 = \{\sum X_i \geq 7\}$

+ Type I Error

$$\Pr_{\theta=1/2}[S_1] = \Pr_{\theta=1/2}\left[\sum X_i \geq 7\right] = \sum_{k=7}^{10} \binom{10}{k} (1/2)^k (1/2)^{n-k}$$

$$= 0.17$$

+ Type II Error

↳ for which θ? next time

always easier to calc max < x than max > x!

Lec 6B - 11/6

Error Types

	S_0	S_1
H_0	✓	Type I
H_a	Type II	✓

Def: Power Function. $X_1, \dots, X_n$ iid F_θ. Power fn. is $\pi: \Omega \to [0,1]$ s.t. $\pi(\theta) = Pr_\theta[x \in S_1]$ ($Pr_\theta \Rightarrow$ as if data came from F_θ).

↳ Type I error governed by $\pi(\theta)$, $\theta \in \Omega_0$

Type II error " " $1 - \pi(\theta)$, $\theta \in \Omega_1$

E.g. $X_1, \dots, X_n$ iid $U[0, \theta]$, $\theta > 0$ $(\Omega = \mathbb{R}_+)$

$H_0: 3 \geq \theta \geq 4$

$H_a: \theta < 3$ or $\theta > 4$

MLE of θ is $X_{(n)} \in (0, \theta)$

Rejection region: $S_1 = \{X_{(n)} \geq 4\} \cup \{X_{(n)} \leq 2.9\}$

Power fn:

$$\pi(\theta) = Pr_\theta(S_1) = Pr_\theta(X_{(n)} \geq 4) + Pr_\theta(X_n \leq 2.9)$$

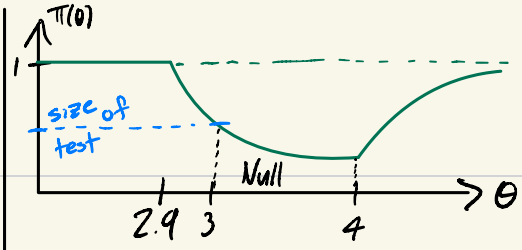
Case $\theta \leq 2.9$: $\pi(\theta) = 0 + 1$

Case $2.9 < \theta < 4$: $\pi(\theta) = 0 + Pr_\theta[X_{(n)} \leq 2.9] = Pr_\theta[X_1 \leq 2.9, \dots, X_n \leq 2.9]$

$$= \left(\frac{2.9}{\theta}\right)^n$$

Case $\theta \geq 4$: $\pi(\theta) = [1 - Pr_\theta(X_{(n)} < 4)] + Pr_\theta[X_{(n)} \leq 2.9]$

$$= 1 - \left(\frac{4}{\theta}\right)^n + \left(\frac{2.9}{\theta}\right)^n$$



→ So the errors aren't clean #s like in elementary statistics

Ideally, $\pi(\theta) = \begin{cases} 0 & \theta \in \Omega_0 \\ 1 & \theta \in \Omega_1 \end{cases}$ or $\pi(\theta) = \begin{cases} \text{low} & \theta \in \Omega_0 \\ \text{high} & \theta \in \Omega_1 \end{cases}$

Solution: make $\pi(\theta)$ low on Ω_0 ≡ find S_1 s.t.
$\pi(\theta) \le \alpha_0 \ \forall \theta \in \Omega_0$ (e.g., $\alpha_0 = 0.1$)

Def: Size of Test. $\sup_{\theta \in \Omega_0} \pi(\theta)$

Def. Test is a level α test if size $= \sup_{\theta \in \Omega_0} \pi(\theta) \le \alpha$.

Note: size is easy to evaluate when $\Omega_0 = \{\theta_0\}$, size $= Pr_{\theta_0}[X \in S_1]$

If $S_1 = \{T = T(x_1, \dots, x_n) \ge c\}$, then a level-$\alpha$ test is s.t. $\sup_{\theta \in \Omega_0} Pr_\theta[T \ge c] \le \alpha$.

E.g. $X_1, \dots, X_n$ iid F_μ, $E[X_i] = \mu$, $Var(X_i) = \sigma^2$ is known.

$H_0: \mu = \mu_0$, $H_a: \mu > \mu_0$

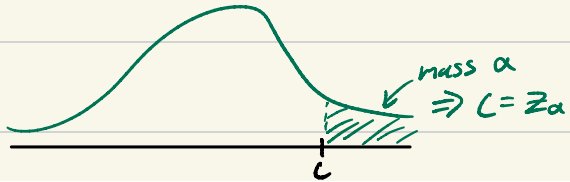
Test Statistic

↳ $S_1 = \{\bar{X}_n - \mu_0 \ge c\}$ → but can't calc probabilities

$S_1 = \left\{\sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \ge c\right\}$ → approx $N(0,1)$ by CLT

$\Rightarrow T = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma} \approx N(0,1)$

$Pr_{\mu_0}[X \in S_1] = Pr_{\mu_0}[T \ge c]$ →



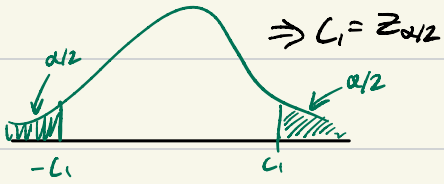
$c = z_\alpha \Rightarrow Pr[Z \sim N(0,1) > z_\alpha] = \alpha$

Now, $S_1 = \left\{x : \bar{X}_n \geq \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right\}$

E.g. Same setting.

$H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$, $T = \sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma}$

$S_1 = \{|T| \geq c_1\}$



$\Rightarrow c_1 = z_{\alpha/2}$ $\quad S_1 = \{|T| \geq z_{\alpha/2}\}$

But neither of these examples say anything a/b the type II error!

Power of the Above Test

$$\Pi(\mu) = Pr_\mu[S_1] = Pr_\mu\left[\left|\sqrt{n}\frac{\bar{X}_n - \mu_0}{\sigma}\right| \geq z_{\alpha/2}\right]$$

$$= Pr_\mu\left[\sqrt{n}\left|\frac{\bar{X}_n - \mu}{\sigma} + \frac{\mu - \mu_0}{\sigma}\right| \geq z_{\alpha/2}\right] \rightarrow z = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \approx N(0,1) \quad (\mu = \mu_0)$$

$$= Pr_\mu\left[\left|Z + \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right| \geq z_{\alpha/2}\right]$$

$$= Pr_\mu\left[Z + \sqrt{n}\frac{\mu - \mu_0}{\sigma} \geq z_{\alpha/2}\right] + Pr\left[Z + \sqrt{n}\frac{\mu - \mu_0}{\sigma} \leq z_{\alpha/2}\right]$$

$$= Pr_\mu\left[Z \geq z_{\alpha/2} - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right)\Big] + Pr_\mu\left[Z \leq -z_{\alpha/2} - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right)\Big]$$

Power depends on

- sample size (n): $\Pi \xrightarrow{n \to \infty} 1$ more samples, more signal $\rightarrow$ more power
- signal ($\mu - \mu_0$): $\Pi \xrightarrow{(\mu - \mu_0) \to \infty} 1$

E.g. Same setting, σ unknown

$T = \sqrt{n}\frac{\bar{X}_n - \mu_0}{s} \approx t_{n-1}$

↳ sample std. dev.

E.g. Variance of Normal

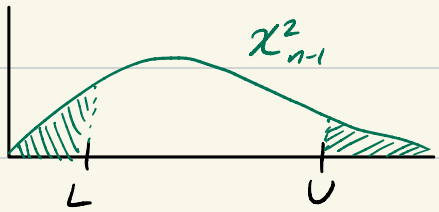
$X_1, \dots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$ $\quad \Omega_0 = \mathbb{R} \times \{\sigma_0\}$

$H_0: \sigma = \sigma_0$, $H_a: \sigma \neq \sigma_0$ $\quad \Omega_1 = \mathbb{R} \times \mathbb{R}_+ \setminus \{\sigma_0\}$

$T = (n-1)\frac{S^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2_{n-1}$ (midterm mention?)

↳ null says $\approx (n-1)$, reject if smaller/greater

↳ $S_1 = \{T < L, \ T > R\}$



Need L, U s.t. $\Pr_{\sigma_0}[S_1] = \alpha$

↳ L, U quantiles of χ^2_{n-1}

Power

$\Pr_\sigma[S] = \Pr_\sigma[T < L, \ T > U]$

Recall $T = (n-1)\frac{S^2}{\sigma_0^2} = \underbrace{(n-1)\frac{S^2}{\sigma^2}}_{\chi^2_{n-1}} \cdot \frac{\sigma^2}{\sigma_0^2}$

$\Rightarrow \Pr_\sigma\left[\chi^2_{n-1} < L\frac{\sigma_0^2}{\sigma^2}, \ \chi^2_{n-1} > U\frac{\sigma_0^2}{\sigma^2}\right]$

Lec 7B – 11/13

Recall hypothesis testing easy when $\pi(\theta)$ constant over Ω_0
↳ if $S_1 = \{T(x) \geq c\}$, $\Rightarrow \mathcal{L}(T(x))$ same for all $\theta \in \Omega_0$

P-Values

Def. p-value is the smallest level α for which we reject the null w/ the observed data.
↳ means we do not need to specify α

- e.g. $p = 0.008$
↳ $\alpha = 0.01 > p \rightarrow$ reject at this level
↳ $\alpha = 0.005 < p \rightarrow$ do not reject

E.g. $X_1, \ldots, X_n$ iid F, $E[X] = \mu$, $Var[X] = \sigma^2$ known

$H_0: \mu = \mu_0$, $H_A: \mu \geq \mu_0$

$T(x) = \sqrt{n}\,\frac{\bar{X}_n - \mu_0}{\sigma} \overset{H_0}{\approx} N(0,1)$

Rejection region: $\{T(x) \geq z_\alpha\}$ → smallest α at which we reject is $T_{obs} = z_\alpha$ = p-value

$\Rightarrow$ if $Z \sim N(0,1)$, $p = Pr[Z \geq T_{obs}]$

E.g. Same as above, but $H_A: \mu \neq \mu_0$

$S_1 = \{|T(x)| \geq z_{\alpha/2}\}$

$\Rightarrow$ p-val is α s.t. $\Pr[Z \geq |T|] = \frac{\alpha}{2}$

$= 2\Pr[Z \geq |T|]$ (under H_0, $\sim U[0,1]$)

Note: $p\text{-val} = \Pr[Z > T_{obs}] = 1 - \Phi(T) \overset{H_0}{\sim} U[0,1]$

↳ helps w/ interpretation: under H_0, p should behave like $\sim U[0,1]$

E.g: Comparing Variances. $X_1, \dots, X_n$ iid $N(\mu_1, \sigma_1^2)$, $Y_1, \dots, Y_m$ iid $N(\mu_2, \sigma_2^2)$

$H_0: \sigma_1^2 = \sigma_2^2$, $H_A: \sigma_1^2 \neq \sigma_2^2$

Notice that $\Omega_0 = \{(\mu_1, \mu_2, \sigma_1, \sigma_2) \in \mathbb{R} : \sigma_1 = \sigma_2\}$ → ugly; we want dist. of T same $\forall \theta \in \Omega_0$!

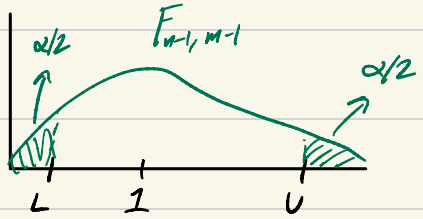
But recall $\frac{(n-1)S_X^2}{\sigma_1^2} \sim \chi^2_{n-1}$, $\frac{(m-1)S_Y^2}{\sigma_2^2} \sim \chi^2_{m-1}$

Note: $U \sim \chi^2_m$, $V \sim \chi^2_n \Rightarrow \frac{U/m}{V/n} \sim F_{m,n}$ ("only find the dist if you are very bored")

↳ use this w/ above to get $\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F_{n-1,m-1}$

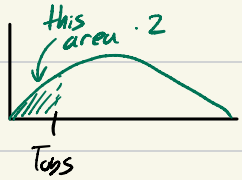
$\Rightarrow T(X) = \frac{S_X^2}{S_Y^2} \overset{H_0}{\sim} F_{n-1,m-1}$

↳ under H_0, ≈ 1.



$\Rightarrow S_1 = \{T > U \text{ or } T < L\}$

$\Rightarrow p$ = take area of tail of $T_{obs} \cdot 2$



Likelihood Ratio Tests

$X_1, \dots, X_n$ iid F_θ, $f(x|\theta)$

Case 1: Simple Hypotheses

$H_0: \theta = \theta_0$, $H_A: \theta = \theta_1$

$LR(X) = \frac{L(\theta_0|x)}{L(\theta_1|x)}$

$\Rightarrow S_1 = \{LR(\theta) \leq c\}$

E.g. H_0: $X_1, \dots, X_n$ iid $\mathcal{N}(1,1)$, H_A: $X_1, \dots, X_n$ iid $\mathcal{N}(2,2)$.

$\theta = (\mu, \sigma^2)$

$$LR(X) = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(x_i-1)^2}{2}\right)}{\left(\frac{1}{2\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(x_i-2)^2}{4}\right)} = (\sqrt{2})^n \exp\left(-\sum_{i=1}^n \left[\frac{x_i^2}{2} - x_i + \frac{1}{2} - \frac{x_i^2}{4} + x_i - 1\right]\right)$$

$$= e^{n/2} \exp\left(-\sum x_i^2/4\right)(\sqrt{2})^n \Rightarrow S_1 = \left\{\sum x_i^2 \geq c\right\}$$

Neyman-Pearson Lemma. H_0, H_A simple hypotheses, $S_1 = \{LR \leq c\}$, α = Type I error, β = Type II. Then, any other test w/ the same α as the LRT has a larger β.

General LRT

$H_0: \theta \in \Omega_0$, $H_A: \theta \in \Omega_1$, $\Omega_0 \cup \Omega_1 = \Omega$

Likelihood Ratio $$\Lambda(X) = \frac{\sup_{\theta \in \Omega_0} L(\theta|x)}{\sup_{\theta \in \Omega} L(\theta|x)}.$$

Thm. $\Omega \in \mathbb{R}^p$ open. Ω_0 is obtained by fixing k coordinates. If MLE assumptions are satisfied,

$$-2 \log \Lambda(X) \overset{H_0}{\sim} \chi^2_k$$

E.g. $X_1, \dots, X_n$ iid $\mathcal{N}(\mu, \sigma^2)$ (2 params $\Rightarrow p = 2$)

$H_0: \mu = \mu_0 \quad \Rightarrow k = 1$

$H_0: \mu = \mu_0, \sigma^2 = 1 \Rightarrow k = 2$

Pf.

For simplicity, $\theta \in \Omega \subset \mathbb{R} \Rightarrow p=1, k=1$, w/ $H_0: \theta = \theta_0$.

Let $\hat{\theta}$ be the MLE.

$$\Rightarrow -2 \log \Lambda(X) = -2\left(\ell(\theta_0) - \ell(\hat{\theta})\right)$$

Taylor: $\ell(\theta_0) \cong \ell(\hat{\theta}) + \underbrace{(\theta_0 - \hat{\theta})\dot{\ell}(\hat{\theta})}_{0} + \frac{1}{2}(\theta_0 - \hat{\theta})^2 \ddot{\ell}(\hat{\theta})$

$$\Rightarrow -2(\ell(\theta_0) - \ell(\hat{\theta})) = \underbrace{(\theta_0 - \hat{\theta})^2}_{\text{normal square}} \underbrace{[-\ddot{\ell}(\hat{\theta})]}_{I(\theta_0)}$$

$\rightarrow$ cancel to get $N(0,1)$

Lec 8A - 11/18

- how to do better on final?
  - ↳ remember midterm mistakes
  - ↳ practice a lot

Tests vs. CIs

E.g. $X_1, \dots, X_n$ iid $N(\mu, 1)$. $H_0: \mu = \mu_0$, $H_A: \mu \neq \mu_0$

$T = \sqrt{n}\left|\frac{\bar{X}_n - \mu_0}{1}\right|$, $S_1 = \left\{\sqrt{n}\left|\frac{\bar{X}_n - \mu_0}{1}\right| \geq z_{\alpha/2}\right\}$ = rejection region @ level α

$$S_0 = \left\{x : \mu_0 - \frac{z_{\alpha/2}}{\sqrt{n}} \leq \bar{X}_n \leq \mu_0 + \frac{z_{\alpha/2}}{\sqrt{n}}\right\}$$

$$= \left\{x : \bar{X}_n - \frac{z_{\alpha/2}}{\sqrt{n}} \leq \mu_0 \leq \bar{X}_n + \frac{z_{\alpha/2}}{\sqrt{n}}\right\}$$

$\Rightarrow$ Given data $X_1, \dots, X_n$, any $\mu_0 \in \bar{X}_n \pm \frac{z_{\alpha/2}}{\sqrt{n}}$ is not rejected by our test!

- ↳ notice that range is a $1-\alpha$ CI

General Case $X_1, \dots, X_n$ iid F_θ.

Let $A(\theta_0)$ denote the acceptance region of $H_0: \theta = \theta_0$, $H_A: \theta \neq \theta_0$ at level α.

Let $S(x) = \{\theta_0 : x \in A(\theta_0)\}$, the set of parameters that are not rejected by our data.

- ↳ then $S(x)$ is a confidence set

Pf. $P_\theta[\theta \in S(x)] = P_\theta[x \in A(\theta)] \geq 1 - \alpha$

Note: the converse is true

Let $S(x)$ be a $1-\alpha$ confidence set. Let $A(\theta_0) = \{x : \theta_0 \in S(x)\}$. Then $A(\theta_0)$ is the acceptance region for a level α test of $H_0: \theta = \theta_0$.

↳ proof is just other way as above

Multiple Testing

Recall the GWAS example: 1M SNPs, 1M tests, 1M p-values

↳ under H_0, $p \sim U[0,1]$.

E.g.

Suppose we have m tests w/ p-values $p_1, \dots, p_m$. Assume $p_1, \dots, p_m$ independent.

↳ if all H_0 true, $\Pr[\exists p_i < \gamma] = \Pr[p_{(1)} \leq \gamma] = 1 - \Pr[p_1 > \gamma, \dots, p_m > \gamma]$

$= 1 - (1-\gamma)^m$

↳ w/ $m = 1000$, $\gamma = 0.001$, $\Pr = 0.63$ — high chance to see small p

Could try to control this at level α: ↗ Sidek Correction

$\Rightarrow 1 - (1-\gamma)^m \leq \alpha \Rightarrow (1-\alpha)^{1/m} \leq 1-\gamma \Rightarrow \gamma \leq 1-(1-\alpha)^{1/m}$

↳ b/c $e^x \approx 1+x$ for small x, $1-(1-\alpha)^{1/m} \approx 1 - e^{-\alpha/m} \approx 1 - (1 - \frac{\alpha}{m})$

$= \frac{\alpha}{m}$

Bonferroni Correction: $\gamma < \frac{\alpha}{m}$

+ What do we control?

	Reject	Non-reject	
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
	$m - R$	R	m

unknown

all letters are non-neg. integers

- if I reject when $p < \gamma$
  - ↳ V = false positives
  - ↳ T = false negatives

Error Rates

a) Family-wide error rates (FWER)
- ↳ $Pr[V > 0]$ ↳ control w/ Sidek, Bonferroni
- ↳ e.g. $\alpha = 0.05$, $m = 10^6 \rightarrow$ Bonferroni: $\gamma < \frac{\alpha}{m} = 5e-8$

quite small! b/c no false positives

b) Per-Family Error Rate (PFER)
- ↳ $E[V]$ (less conservative)

c) False Discovery Rate (FPR)
- ↳ $E[V/R]$

1995 paper

Benjamini & Hochberg. Control FDR at level α.

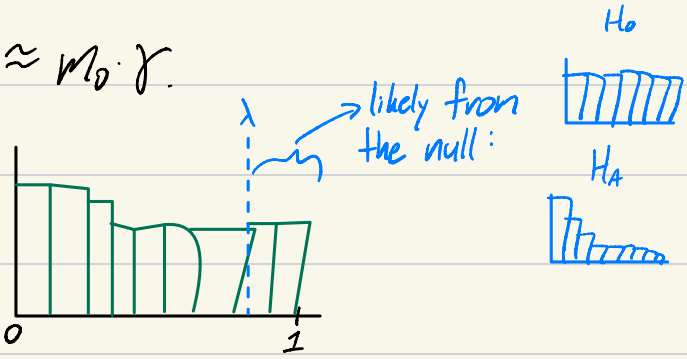
1. Order p-values $p_{(1)} \leq \dots \leq p_{(m)}$
2. Find largest j s.t. $p_{(j)} \leq \frac{\alpha \cdot j}{m}$
3. Reject all H_0 corresponding to $p_{(1)}, \dots, p_{(j)}$.

Note: Estimating FDR.

Given a rejection region $[0, \gamma]$, $E[V] \approx m_0 \cdot \gamma$.

↳ need to know m_0 or $\pi_0 = \frac{m_0}{m}$

Trick: histogram of p-values:



$\Rightarrow \#(p > \lambda) \approx m_0(1-\lambda)$

$\Rightarrow m_0 \approx \frac{\#(p>\lambda)}{(1-\lambda)} \longrightarrow FDR \approx \frac{m_0}{R} = \frac{\#(p>\lambda)\gamma}{R(1-\lambda)}$

- bias-variance tradeoff: large $\lambda \Rightarrow$ pure signal, but low sample size

Pf (BH).

Let $N \subseteq [m]$ be the set of indices for H_0

↳ $|N| = m_0$

↳ Let $\alpha_R = \frac{\alpha \cdot R}{m}$, $R \in [m]$

Then $E[V/R] = E\left[\frac{\sum_{k \in N} \mathbb{1}\{p_k \leq \alpha_R\}}{R}\right]$ → α_R is the threshold, counting qualifying p-vals under H_0

$$= \sum_{k \in N} \sum_{r=1}^{m} \frac{1}{r} Pr[p_k \leq \alpha_R, R = r]$$

Trick: R_k = # discoveries when doing BH at level α w/ $p_k = 0$. → R_k can be R or $R+1$

Then, $\Pr[\bar{p}_k \le \alpha_R, R = r] = \Pr[\bar{p}_k \le \alpha_r, R_k = r]$

↳ by independence, $= \Pr[\bar{p}_k \le \alpha_r] \Pr[R_k = r]$

$$\Rightarrow E[V/R] = \sum_{k \in N} \sum_{r=1}^{n} \frac{1}{r} \alpha_r \Pr[R_k = r]$$

$$= \sum_{k \in N} \frac{\alpha}{m} \underbrace{\sum_{r=1}^{m} \Pr[R_k = r]}_{=1}$$

$$\Rightarrow E[V/R] = \alpha \cdot \frac{m_0}{m} \le \alpha$$

Lec 8B - 11/20

Bayesian Statistics

Def: Bayes Formula. $Pr[A|B] = \frac{Pr[B|A]\,Pr[A]}{Pr[B]}$.

$H_1, \ldots, H_k$ partition of sample space S, D s.t. $Pr[D] > 0$.

$\hookrightarrow Pr[H_i|D] = \frac{Pr[D|H_i]\,Pr[H_i]}{\sum_i Pr[D|H_i]\,Pr[H_i]}$ ← new info

→ how our beliefs are "updated" in light of new information

E.g. CAD = coronary artery disease, PAD = periphery ...

$H_1 = \{CAD+, PAD+\}$, $H_2 = \{CAD+, PAD-\}$, $H_3 = \{CAD-, PAD+\}$,

$H_4 = \{CAD-, PAD-\}$

D = high cholesterol

$Pr[H_i]$ = prior probability, $Pr[D|H_i]$ = likelihood, $Pr[H_i|D]$ = posterior

Def: Third Bayes. $f(\theta|x) = \frac{f(\theta)f(x|\theta)}{h(x)}$

prior = $f(\theta)$ → θ is RV

likelihood = $f(x|\theta)$

posterior = $f(\theta|x)$

marginal dist. of data = $h(x) = \int f(\theta) f(x|\theta)\, d\theta$

Bayesian vs. Frequentist

All starts w/ interp of probability.

↳ coin toss $p = 0.5$

↳ frequentist gets 0.5 from $n \to \infty$ repeats

↳ Bayesians get 0.5 from subjective understanding

Inference

↳ frequentist: θ fixed

↳ Bayesian: θ RV

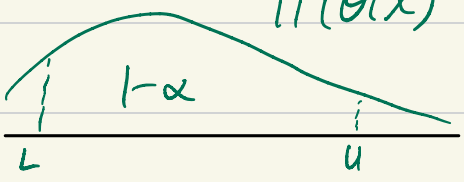
Bayesian Inference

Setup: $X_1, \ldots, X_n$ iid $f(x|\theta)$ → likelihood

↳ prior: $\theta \sim f_\theta(\theta) = \pi(\theta)$

↳ posterior: $\pi(\theta|x) = f(\theta|x) \propto \pi(\theta) f(x|\theta)$ (∝ proportional)

+ How do we use the posterior?



Estimation → this is it!

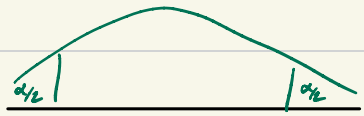
1) posterior mean
2) posterior mode
3) posterior median

- Credible Intervals (equiv of CIs)

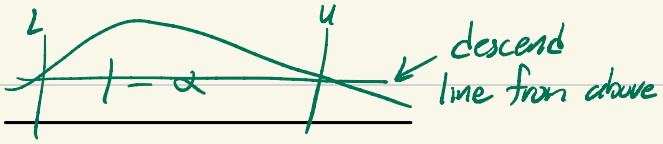
↳ pick L, U s.t. $Pr_{\theta|x}[L \le \theta \le U] = 1-\alpha$

How to pick L, U?

Def: Equal-Tailed Interval. Let $F_{\theta|x}$ be the posterior cdf. Then pick $F^{-1}_{\theta|x}(\alpha/2) \le \theta \le F^{-1}_{\theta|x}(1-\alpha/2)$.



Def: High-Posterior Interval. $I = \{\theta : f(\theta|x) \ge c\}$ s.t. $Pr_{\theta|x}(I) = 1-\alpha$.



+Hyp. Testing

Let $H_0: \theta \in \Omega_0$, $H_A: \theta \in \Omega_1$.

↳take ratio of probabilities: $\dfrac{Pr_{\theta|x}[\theta \in \Omega_0]}{Pr_{\theta|x}[\theta \in \Omega_1]}$

E.g: Coin Tossing.

θ = pr of heads. Let $X_1, \dots, X_n$ iid $Bern(\theta)$.

We need a prior for θ.

↳ let's start w/ $\theta \sim U[0,1]$ → basically no prior information

Likelihood

↳ $X = \sum_i x_i \to f(x|\theta) = \theta^x (1-\theta)^{n-x}$

Posterior

↳ $\pi(\theta|x) \propto f(x|\theta)\pi(\theta) = \theta^x(1-\theta)^{n-x}$, $\theta \in (0,1)$

Def: Beta Dist. $Beta(\alpha, \beta)$ has density $f(x) = \frac{1}{\beta(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, where $\beta(\alpha,\beta) = \frac{1}{\Gamma(\alpha+\beta)}\Gamma(\alpha)\Gamma(\beta)$.

↳ $x \in (0,1)$

Beta Facts

a) $\text{Beta}(1,1) = U[0,1]$

b) W/ $\text{Beta}(\alpha, \alpha)$

↳ if $\alpha < 1$: 

↳ if $\alpha > 1$: 

c) If $X \sim \text{Beta}(\alpha, \beta)$, $E[X] = \frac{\alpha}{\alpha+\beta}$ and $\text{Var}[X] = \frac{\alpha}{\alpha+\beta} \cdot \frac{\beta}{\alpha+\beta} \cdot \frac{1}{\alpha+\beta+1}$

d) If $\alpha, \beta > 1$, the mode is $\frac{\alpha-1}{\alpha+\beta-2}$.

Back to our example: $\pi(\theta|x) \propto \theta^x (1-\theta)^{n-x}$

$\hookrightarrow \theta|X \sim \text{Beta}(X+1, n-X+1)$

Now let's estimate

↳ mean = $\frac{X+1}{n+2}$

↳ mode = $\frac{X}{n}$ → also MLE

Case 2: New Prior

Now let's say $\theta \sim \text{Beta}(\alpha, \beta)$

$\Rightarrow \pi(\theta|X) \propto \underbrace{(\theta^{\alpha-1}(1-\theta)^{\beta-1})}_{\text{↳prior}} \underbrace{(\theta^X (1-\theta)^{n-X})}_{\text{↳likelihood}} = \theta^{\alpha+X-1}(1-\theta)^{\beta+n-X-1}$

↳ $\sim \text{Beta}(\alpha+X, \beta+n-X)$

New posterior depends on the data (X) & prior (α, β).

↳ mean = $\frac{\alpha+X}{\alpha+\beta+n}$ ↳ posterior "sample size" ↗ w/ α = # heads

Also think of weighted mean:

↳ $\frac{\alpha}{\alpha+\beta}\left(\frac{\alpha+\beta}{\alpha+\beta+n}\right) + \frac{X}{n}\left(\frac{n}{\alpha+\beta+n}\right)$ → sample weight

prior mean ↳ prior weight ↳ sample mean

Concrete Example.

$\alpha = \beta = 3$, $n = 10$, $X = 3$

↳ we get mean $= \frac{6}{16} = \frac{3}{8} = 0.375$ > frequentist $\frac{3}{10} = 0.3$

b/c the prior was $\frac{1}{2}$

- but how do we pick priors? ∃ a few camps:

1) use your judgement
2) assume no information
3) pick a convenient prior

↳ Conjugate Family → $\begin{cases} \theta & \text{beta} \\ X & \text{binomial} \\ \theta|X & \text{beta} \end{cases}$ → subjective, but easy to work w/

E.g. Travel to a city. Based on the first bus you see, numbered T, and assuming n busses $\in [N]$, what is n?

+ Frequentist
  - Likelihood: $f(T|N) = \frac{1}{N}$ (discrete uniform)
    - ↳ mean $= \frac{N}{2}$ → MoM ⇒ $\hat{N} = 2T$

+ Bayesian
  - Need prior on $N \in \{1, 2, 3, \ldots\}$ → what to use?

Lec 9A - 12/2

Bayesian Inference

$X_1, \dots, X_n$ iid $f(x|\theta)$ → likelihood

$\theta \sim \pi(\theta)$ → prior

$\pi(\theta|X) \propto \pi(\theta) \prod_i f(x_i|\theta)$ → posterior

↳ prior ↳ likelihood

E.g. (exponential)

$X_1, \dots, X_n$ iid $Exp(\lambda)$

Prior: $\lambda \sim Gamma(k, r)$ (conjugate prior)

Posterior: $\pi(\lambda|X) \propto \pi(\lambda) f(X|\lambda)$

$$= \left(\frac{r^k}{\Gamma(k)} \lambda^{k-1} e^{-r\lambda}\right)\left(\lambda^n e^{-\lambda \sum x_i}\right)$$

$$\propto \lambda^{n+k-1} \exp\left(-\lambda\left(r + \sum x_i\right)\right)$$ → can drop terms w/o params

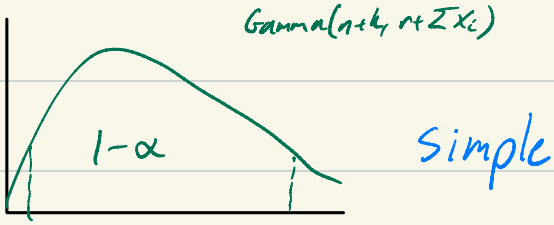
$$\Rightarrow \quad \lambda|X \sim Gamma(n+k,\ r+\sum x_i)$$

Estimation

- posterior mean: $\frac{n+k}{r+\sum x_i}$

↳ recall the MLE is $\frac{1}{\bar{x}}$; same as bayes if $k = r = 0$

Credible Intervals

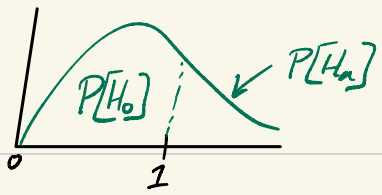


simple

Hyp. Testing

- $H_0: \lambda \in (0,1)$, $H_a: \lambda \geq 1$

- again look @ posterior:



E.g: Normal Mean.

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

Prior: $\mu \sim N(\mu_0, \nu^2)$

↳ can be shown that $\mu|X \sim N\left(\bar{X}\frac{n/\sigma^2}{n/\sigma^2+1/\nu^2} + \mu_0 \cdot \frac{1/\nu^2}{\frac{n}{\sigma^2}+\frac{1}{\nu^2}}, \frac{1}{n/\sigma^2+1/\nu^2}\right)$

(influence of data; influence of prior)

↳ $\frac{n}{\sigma^2}$ = information in the data; n large = more info

↳ $\frac{1}{\nu^2}$ = information in the prior; ν^2 small = lots of info

Selecting the Priors

Most critical & criticized part of Bayesian inference.

Subjective Determination. $\theta \in \Omega$

a) Say Ω is discrete. Then just use past experience

b) If Ω is an interval, two options:

1) discretize (histogram approach)

2) matching a parametric family

E.g. $X \sim Bin(n, p)$, p = pr. of getting A in 24410

From previous years, p has mean 0.7 & var 0.1

If we restrict to Beta priors, $p \sim Beta(\alpha, \beta)$

↳ want $\frac{\alpha}{\alpha+\beta} = 0.7$, $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.1 \Rightarrow \alpha = 0.77, \beta = 0.33$

$\Rightarrow p \sim \text{Beta}(0.77, 0.33)$

Conjugate Priors

Def: A family F of distributions is said to be closed under sampling for a model $f(x|\theta)$ if $\forall f \in F$, the posterior $f(\theta|x) \propto f(\theta) f(x|\theta) \in F$.

E.g.s

Model	Conjugate Prior
Bernoulli	Beta
Normal	Normal
Exponential	Gamma
Poisson	Gamma

- in the previous example problem, we combined subjective & conjugate priors

Noninformative Priors: favor no particular value.

Case 1: $|\Omega| = a$ (finite)
↳ prior = $\frac{1}{a}$ for each value

Case 1b: Ω is countable (e.g. $\mathbb{Z}$)
↳ soln: constant, improper prior (invalid dist.)

Case 2: Ω interval (e.g. $\Omega = [0, 1]$)

↳ Is $U[0,1]$ noninformative?

↳ Reparameterizing to odds:

$\eta = \frac{\theta}{1-\theta} \sim \frac{1}{(1+\eta)^2}$ on $\mathbb{R}_+$ (can be shown)

↳ Unif is not invariant under reparameterization

Case 3: $\Omega = \mathbb{R}$.

↳ soln: flat/constant improper prior (can't integrate to 1)

↳ usable as long as posterior is proper

Jeffrey's Prior

$X_1, \dots, X_n$ iid $f(x|\theta)$ w/ Fisher information $I(\theta)$

Def: Jeffrey's Prior $\pi_J(\theta) \propto I^{1/2}(\theta)$

Thm: Invariance. Let $\theta \sim \pi_J(\theta)$, and $\eta = g(\theta)$. Then η is distributed according to Jeffrey's prior of $g(\theta)$.

Pf: Lec. notes.

E.g. $X \sim Bin(n, \theta)$.

$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

↳ $I(\theta) = \frac{n}{\theta(1-\theta)}$

$\Rightarrow \pi_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} = Beta(1/2, 1/2)$ (proper distribution!)

Note: not all Jeffrey priors are proper.

Lec 9B - 12/4

Statistical Decision Theory

↳ framework for making "decisions"

↳ e.g. choosing estimators: $\theta \in \Omega$ = "truth", $\mathcal{D}$ = collection of decisions

Def: Loss Fn. $L: \Omega \times \mathcal{D} \to [0, \infty)$. $L(\theta, d)$ = loss associated w/ decision $d \in \mathcal{D}$.

E.g.

$\mathcal{D} = \{d_1, d_2\}$ (say d_1 = football game, d_2 = movie)

$\Omega = \{\theta_1, \theta_2\}$ (θ_1 = rain, θ_2 = no rain)

Loss:

θ_1	d_1	1
θ_1	d_2	1/4
θ_2	d_1	0
θ_2	d_2	4/10

Assume a 40% chance of rain.

Expected loss:

↳ choose d_1: $1 \cdot 4/10 + 0 \cdot (1 - 4/10) = 4/10$

↳ choose d_2: $1/4 \cdot 4/10 + 4/10 (1 - 4/10) = 34/100$

So the expected loss favors d_2

Inference as a Decision Problem

Problem: $X_1, \ldots, X_n$ iid $f(x|\theta)$, $\theta \in \Omega$ w/ prior $\pi(\theta)$.

Decision: choice of estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$

↳ our loss fn. is $L: \Omega \times \Omega \to [0, \infty)$

Types of Loss Functions

1) Squared error: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
2) Absolute error: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$
3) Zero-One loss: $L(\theta, \hat{\theta}) = \mathbb{1}_{\{\theta \neq \hat{\theta}\}}$

Def: (frequentist) risk. $R_{\hat{\theta}}(\theta) = R(\theta, \hat{\theta}) = E_{x^n}[L(\theta, \hat{\theta})] = \int L(\theta, \hat{\theta}) f(x^n|\theta) dx^n$

Notes

a) this is a pre-data measure of performance
b) if we use squared error loss, $R(\theta, \hat{\theta}) = MSE(\hat{\theta})$
c) for a given estimator, risk is a fn. of θ

E.g.s

a) X_1, X_2 iid $\begin{cases} \theta-1 & p=1/2 \\ \theta+1 & p=1/2 \end{cases}$

Let's use zero-one loss, w/ estimators

$\hat{\theta}_1 = \frac{X_1 + X_2}{2} \rightarrow R(\theta, \hat{\theta}_1) = Pr_\theta[X_1 = X_2] = 1/2$

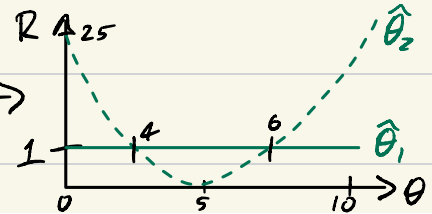
$\hat{\theta}_2 = X_1 + 1 \rightarrow R(\theta, \hat{\theta}_2) = Pr_\theta[X_1 = \theta + 1] = 1/2$

↳ neither risks depend on θ, & risk is equal

b) $X \sim N(\theta, 1)$, $\theta \in (0, 10) = \Omega$ w/ squared error loss

$\hat{\theta}_1 = X \rightarrow R(\theta, \hat{\theta}_1) = E[(X-\theta)^2] = Var[X] = 1$

$\hat{\theta}_2 = 5 \rightarrow R(\theta, \hat{\theta}_2) = E[(5-\theta)^2] = (5-\theta)^2$



Neither estimator always wins
↳ how to choose?

+ Frequentist Approach

Def: Maximum Risk. $\overline{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta})$

Def: Minimax Estimator. $\hat{\theta}$ s.t. $\overline{R}(\hat{\theta}) = \inf_{\tilde{\theta}} \overline{R}(\tilde{\theta})$.

+ Bayesian Approach

Def: Bayes Risk. $r(\hat{\theta}) = \int R(\theta, \hat{\theta}) \pi(\theta) d\theta$.

Def: Bayes Estimator. associated w/ a loss fn. & prior π is $\hat{\theta}$ s.t. $r(\hat{\theta}) = \min_{\tilde{\theta}} r(\tilde{\theta})$.

E.g. $X \sim Bin(n, \theta)$

Prior: $\pi(\theta) = U[0,1]$, loss = square loss

Let $\hat{\theta}_1 = \frac{X}{n}$ and $\hat{\theta}_2 = \frac{\alpha + X}{\alpha + \beta + X}$ (posterior mean for Beta(α, β) prior)

↳ for $\hat{\theta}_1$: $R(\theta, \hat{\theta}_1) = E[(\hat{\theta} - \theta)^2] = Var[\hat{\theta}_1] = \frac{\theta(1-\theta)}{n}$

$\overline{R}(\hat{\theta}) \overset{\theta = 1/2}{=} \frac{1}{4n}$

$r(\hat{\theta}) = \int_0^1 \frac{\theta(1-\theta)}{n} \cdot 1 \, d\theta = \frac{1}{6n}$

↳ for $\hat{\theta}_2$: $R(\theta, \hat{\theta}) = \frac{n\theta(1-\theta) + (\alpha - (\alpha+\beta)\theta)^2}{(\alpha+\beta+n)^2} = \hat{\theta}_2^{\alpha, \beta}$

↳ turns out that $\hat{\theta}_2^{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}} = \frac{n}{4(n+\sqrt{n})^2}$ → does not depend on θ!

$\overline{R}(\hat{\theta}_2^{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}) = \frac{n}{4(n+\sqrt{n})^2}$ → $\hat{\theta}_2^{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}$ wins! maybe b/c of small part of sample space

$r(\hat{\theta}_2^{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}) = \frac{n}{4(n+\sqrt{n})^2}$ → loses here (for large n)

The best Bayes estimator here is $r(\hat{\theta}_2^{1,1}) = \frac{1}{6(n+2)}$ → $\hat{\theta}_2^{1,1}$ = posterior mean

Def: Posterior Risk. $r(\hat{\theta} \mid x^n) = \int L(\theta, \hat{\theta}) \pi(\theta \mid x^n) \, d\theta$

Thm. Let $\hat{\theta} = \hat{\theta}(x^n)$ be the value that minimizes the posterior risk. Then $\hat{\theta}$ is the Bayes estimator.

Pf.

$$r(\hat{\theta}) = \int R(\theta, \hat{\theta}) \pi(\theta) \, d\theta = \int\int L(\theta, \hat{\theta}) f(x \mid \theta) \pi(\theta) \, dx \, d\theta$$

by Fubini: $$= \int\int L(\theta, \hat{\theta}) \pi(\theta \mid x) f(x) \, d\theta \, dx$$

$$= \int r(\hat{\theta} \mid x^n) f(x^n) \, dx^n.$$

Thm. Bayes estimators are:

- square loss → posterior mean
- absolute loss → posterior median
- zero-one loss → posterior mode

Pf. (of ①)

Let X be a RV w/ mean μ.

↳ $E[(X-c)^2] = E[(X-\mu+\mu-c)^2]$

$= E[(X-\mu)^2] + (\mu-c)^2$ → minimized when $c = \mu$

$r(\hat{\theta} \mid x) = \int (\theta - \hat{\theta})^2 \pi(\theta \mid x^n) \, d\theta$

($\hat{\theta} \leftrightarrow x$, $\theta \leftrightarrow c$) ⇒ minimized for $\hat{\theta}$ = posterior mean