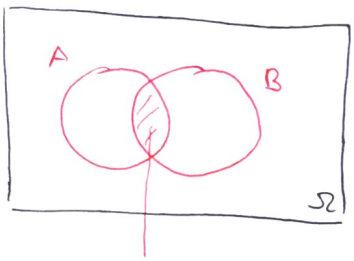
ASTROSTATISTICS

PROBABILITY FUNDAMENTALS



- Sum Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Conditional Probability:

$$P(A \cap B) = P(A|B)\,P(B)$$

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

prob of A given B

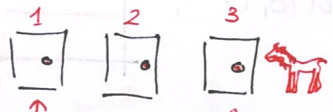
- Law of Total Probability: use partitioning $\bigcup_i \{B_i\} = \Omega$, $B_i \cap B_j \neq \emptyset$ $i \neq j$

$$P(A) = \sum_i P(A|B_i)\,P(B_i)$$

- Bayes theorem:

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

- Example ①: Monty Hall problem



choose door 1 — Monty opens and shows a goat

1	2	3	opened door	stay	change
C	G	G	2/3	C	G
G	C	G	3	G	C
G	G	C	2	G	C

better to change!

earthquake opens door w goat → 2 is ruled out → 50-50 chance

② Extreme MH: $N = 1000$ doors, you pick one, host opens 998 other doors, all w goat

H_i = car is behind i $\quad P(H_i) = 1/1000$

d_k = "data" observation that $N-2$ doors have goats

$$P(H_1|d_k) = \frac{P(d_k|H_1)\,P(H_1)}{P(d_k)} \qquad P(d_k|H_1) = 1/N-1$$

$$P(d_k|H_i) = \begin{cases} 1 & k = i \\ 0 & \text{otherwise} \end{cases}$$

LTP: $P(d_k) = \sum_{i=1}^{N} P(d_k|H_i)\,P(H_i)$

$$P(d_k) = P(d_k|H_1)\,P(H_1) + P(d_k|H_k)\,P(H_k) = \frac{1}{N-1}\cdot\frac{N}{N} + \frac{1}{N} = \frac{1}{N-1}$$

$$P(H_1|d_k) = \frac{\frac{1}{N-1}\cdot\frac{1}{N}}{\frac{1}{N-1}} = \frac{1}{N} \Rightarrow P(H_{k\neq 1}|d_k) = \frac{N-1}{N}$$

- Continuous distributions / random variables

notation: $P_X(X = x) \rightarrow P(x)$

- $P(x,y)$ = joint pdf $= P(x|y)P(y) = P(y|x)P(x) = P(x)P(y)$

↑ if x and y are independent

- Marginals $P(x) = \int P(x,y)\,dy$

use continuous version of LTP $P(x) = \int P(x|y)P(y)\,dy$

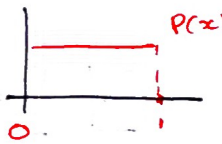
⇒ Baye's Rule: $P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\int P(x|y)P(y)\,dy}$

- Transformations of RVs

Suppose X has $p(x)$, have $y = \Phi(x)$ invertible, so $x = \Phi^{-1}(y)$

$$P(y) = P(x)\left|\frac{dx}{dy}\right| = P(\Phi^{-1}(y))\left|\frac{d\Phi^{-1}(y)}{dy}\right|$$

→ example: $X \sim U(0,1)$



$P(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} = U(x|0,1)$

$y = -\ln x$

$x = e^{-y}$ $\quad dx/dy = -e^{-y} \rightarrow P(y) = U(e^{-y}|0,1)\cdot e^{-y}$

$= \begin{cases} e^{-y} & 0 \le y < \infty \\ 0 & \text{otherwise} \end{cases}$

exponential distribution

- "Delta" method $\quad Y = \Phi(X) \quad \mathrm{Var}(X) = \sigma_X^2 \quad \mathbb{E}(X) = X_0$

want to know $\sigma_Y^2 = \mathrm{Var}(Y)$

- Definitions: $X \sim P(x)$ random variable distributed by P

$$\mathbb{E}[X] = \int x P(x)\,dx$$

$$\mathbb{E}[h(x)] = \int h(x) P(x)\,dx$$

$$\mathrm{Var}[X] = \int (x - \mathbb{E}(X))^2 P(x)\,dx = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- If we have $p(x,y)$ $\quad \mathbb{E}[X] = \iint x\, p(x,y)\, dx\, dy$

$$\mathrm{Cov}(X,Y) = \int (x - \mathbb{E}[X])(y - \mathbb{E}[Y])\, p(x,y)\, dx\, dy$$

→ if have bilinearity: $S = \sum_{i=1}^{n} a_i X_i \quad T = \sum_{j=1}^{k} b_j Y_j$

$$\mathrm{Cov}(S,T) = \mathrm{Cov}\left(\sum a_i X_i, \sum b_j Y_j\right) = \sum_i \sum_j a_i b_j \mathrm{Cov}(X_i, Y_j)$$

~~And~~

- Delta Method want to know $\mathrm{Var}(Y)$ $\quad X \sim p(x) \quad Y = g(X)$

do a Taylor expansion $\quad Y = g(X) = g(X_0) + \left.\frac{dg}{dx}\right|_{x=x_0} (X - X_0) + \dots$

$$\Rightarrow \mathrm{Var}(Y) = \mathbb{E}(Y - \mathbb{E}(Y)) = \left|\left.\frac{dg}{dx}\right|_{x=x_0}\right| \mathrm{Var}(X)$$

accuracy depends on "wiggliness" of $g(x)$ compared to size of variance. e.g. for log not so good

Limit theorems

- LLN law of large numbers $\quad X_1, \dots, X_N$ iid rvs ↑ independent identically distributed random variables $\quad \omega \quad M = \mathbb{E}[X_i]$

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$ running sample mean

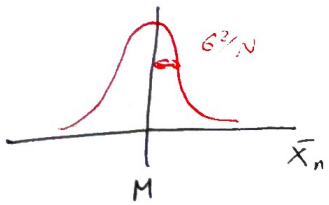
$\Rightarrow N \to \infty \quad \forall \varepsilon > 0 \quad p(|\bar{X}_n - M| > \varepsilon) \to 0$

sample mean converges on the true mean

- CLT central limit theorem $\quad \sigma^2 = \mathrm{Var}(X_i)$

$$P\left(\sqrt{n}\,(\bar{X}_n - M) \le x\sigma\right) \to \Phi(x) = \int_{-\infty}^{x} N(t|0,1)\, dt$$

$\frac{1}{\sqrt{2\pi}} e^{-t^2/2\sigma^2}$ gaussian



ie $\quad P(\bar{X}_n) \to N(\bar{X}_n | M, \sigma^2/N)$

Asymptotic Normality

STATISTICAL INFERENCE

- suppose $\underline{D} = \{X_1, \dots, X_N\}$ $\quad X_i \overset{iid}{\sim} P(x|\theta)$

Want to estimate θ from $\underline{D}$, different ideas:

① Sample Mean $\frac{1}{N}\sum_{i=1}^{N} x_i$

② Sample mean of first $k < N$: $\frac{1}{k}\sum_{i=1}^{k} x_i$

③ $\frac{1}{N-1}\sum_{i=1}^{N} x_i$ doesn't have a name...

④ $\frac{1}{2}[\min(x) + \max(x)]$ Midrange

⑤ median

⑥ bin data → take mode

ESTIMATOR $\hat{M} = f(\underline{X})$ for ESTIMAND $M_\theta = \int x\, P_\theta(x)\, dx$

$\hat{\theta} = f(\underline{X})$ is an estimator for θ

Frequentist properties

$\hat{\theta}$ is unbiased if $E[\hat{\theta}] = \theta$

- Imagine k experiments: $j = 1, \dots, k$ $\quad \underline{X}_j = (X_{1,j}, \dots, X_{N,j})$

get $\underline{X}_1, \dots, \underline{X}_j, \dots, \underline{X}_k$

$\hat{\theta}_1 = \hat{\theta}(\underline{X}_1)$ $\quad$ $\hat{\theta}_j = \hat{\theta}(\underline{X}_j)$

$$\frac{1}{k}\sum_{j=1}^{k} \hat{\theta}_j \longrightarrow \theta \quad \text{as} \quad k \to \infty$$

BIAS $\boxed{B(\hat{\theta}) = E(\hat{\theta}) - \theta}$

Consistency as you gather more data $N \to \infty$

$\hat{\theta} \to \theta$, $\forall \varepsilon > 0 \quad P(|\hat{\theta} - \theta| \geq \varepsilon) \to 0$

Efficiency : $Var(\hat{\theta})$

Mean Square Error $MSE(\hat{\theta}) = E[(\hat{\theta}(x) - \theta)^2]$

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

treat as variables

$\ldots = Var(\hat{\theta}) + B(\hat{\theta})^2$

so for unbiased estimators smallest MSE for most efficient

- MVUE among all unbiased estimators → smallest variance

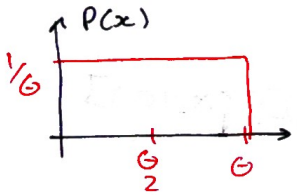
⇒ Bias-Variance Tradeoff not always unbiased best

Examples: ① $X_1, \ldots, X_N \sim N(\mu, \sigma^2)$ $\quad X_1, \ldots, X_N \sim U(a,b)$ $\quad Var = \frac{(b-a)^2}{12}$, $M = \frac{a+b}{2}$

	$\sigma=1, N=10$		
sample mean	σ^2/N	0.10	σ^2/N 0.10
midrange	$\frac{\pi^2\sigma^2}{24\ln(2)}$	0.18	$\frac{6\sigma^2}{(N+2)(N+1)}$ 0.06
	Variance		

for uniform distribution midrange better!

② Uniform distribution



$X_i \sim U(0, \theta)$

→ estimator for θ $\hat{\theta} = \max(\underline{X}) < \theta$ biased

$\hat{\phi} = 2\bar{X} \to E[\bar{X}] = \theta/2 \to 2E[2\bar{X}] = \theta$

maximum likelihood actually $\hat{\theta}$!

draw: $\underline{X} = (0.32, 0.46, 0.97, 2.77, 7.06)$

$\hat{\phi}(\underline{X}) = 4.63$ $\quad \hat{\theta}(\underline{X}) = 7.06$ $\quad \hat{\phi}$ obviously wrong!

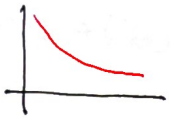
Deriving Estimators

① Method of Moments : match theoretical moments of distribution w/ sample moment

$X_1, \dots, X_N \overset{iid}{\sim} P_\theta(x)$ $\quad$ $\mu_k = \mathbb{E}[X^k] = \int x^k P_\theta(x)\,dx = \text{function}(\theta)$ (theoretical)

$\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} X_i^k$ (average of x^k) $\quad$ $\hat{\mu}_k = \mu_k(\theta) \rightarrow$ solve for $\hat{\theta}$

e.g.: $X_i \sim \exp(\lambda)$



$$p(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = 1/\lambda$, $\bar{x} = \frac{1}{N}\sum x_i \Rightarrow \hat{\lambda} = 1/\bar{x}$

② Method of Least Squares: $\quad \mu = \int x P_\theta(x)\,dx$

$$\hat{\mu}_{LS} = \underset{\mu}{\text{argmin}} \left[\underbrace{\sum_{i=1}^{N} (x_i - \mu)^2}_{SS} \right] \rightarrow \hat{\mu}_{LS} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

③ Maximal Likelihood :

let X_i has pdf $f(x|\theta)$, $\quad X_i \sim f(x|\theta)$

θ = true value, view as function of X

sampling distribution of X

experiment $\rightarrow$ $\bar{x} = (x_1, \dots, x_N)$ $\quad$ $p(\bar{x}|\theta) = \prod_{i=1}^{N} f(x_i|\theta) = L(\theta)$

LIKELIHOOD FNC. of θ ↑ data $\bar{x}$ fixed

Sometimes useful to calculate:

$$\ell(\theta) = \ln(L(\theta)) = \sum_{i=1}^{N} \log[f(x_i|\theta)] \quad \text{LOG LIKELIHOOD}$$

$\Rightarrow$ MLE $\quad \hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} [L(\theta)] = \underset{\theta}{\text{argmax}} [\ell(\theta)]$

need to know how to model the problem

Properties of MLE

if model assumptions are true, ie $L(\theta)$ true

- consistent $\hat{\theta}_{MLE} \to \theta_{true}$ as $N \to \infty$
- asymptotically unbiased $E[\hat{\theta}_{MLE}] \to \theta_{true}$ as $N \to \infty$
- asymptotically normal $(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N(0, I^{-1})$ if θ vector then I matrix

$$\Rightarrow I(\theta) = E\left[\left(\frac{\partial \log[L(\theta)]}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right]$$

EXPECTED FISHER INFORMATION

curvature of likelihood function, 2nd derivative

$$\hat{I} = -\left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\theta = \hat{\theta}_{MLE}}$$

OBSERVED FISHER INFORMATION

$\hat{I} \to I(\theta)$ as $N \to \infty$

- Efficiency: Cramer-Rao Lower Bound (CRLB) $\mathrm{Var}(\hat{\theta}_{MLE}) \geq I^{-1}(\theta)$

asymptotically $N \to \infty$ achieves the CRLB so efficiency I^{-1}

- Functionally invariant $\alpha = g(\theta) \Rightarrow \hat{\alpha}_{MLE} = g(\hat{\theta}_{MLE})$

Multiparameter MLE

$$X_i \overset{iid}{\sim} f(X|\vec{\theta}) \qquad L(\vec{\theta}) = P(\vec{X}|\vec{\theta}) = \prod_{i=1}^{N} f(x_i|\vec{\theta})$$

θ always vector here, all defs the same except fisher info a matrix

$$I_{jk}(\theta) = E\left[-\frac{\partial^2 \log \ell(\theta)}{\partial \theta_j \partial \theta_k}\right]$$

HESSIAN MATRIX

in astrophysics $\hat{I}_{jk} = -\left.\frac{\partial^2 \log \ell(\theta)}{\partial \theta_j \partial \theta_k}\right|_{\theta = \hat{\theta}_{MLE}} \simeq I_{jk}(\theta)$ $N \to \infty$

large sample limit

in freq they only care about this

- $\mathrm{Var}(\hat{\theta}_{MLE,i}) \geq (I^{-1})_{ii}$ by CRLB

$\Rightarrow \mathrm{Cov}(\hat{\theta}_{MLE}) \to I^{-1}$ as $N \to \infty$

Example for Normal RVs

- For $X_i \sim N(\mu, \sigma^2)$ $\quad i = 1, \dots, N$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} N(x_i | \mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\ell(\theta) = \sum_{i=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}(x_i-\mu)^2/\sigma^2$$

overbar, not vector.

① suppose σ^2 is unknown

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = 0 \rightarrow \hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

$$\Rightarrow \frac{\partial^2 \ell}{\partial \mu^2} = -\frac{N}{\sigma^2} \Rightarrow I = -E\left[\frac{\partial^2 \ell}{\partial \mu^2}\right] = \frac{N}{\sigma^2} \Rightarrow I^{-1} = \sigma^2/N$$

So by CRLB $\quad \mathrm{Var}(\hat{\mu}_{MLE}) \geq \sigma^2/N$

Could have calculated more directly

$$\mathrm{Var}(\hat{\mu}_{MLE}) = \mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{N}\sum x_i\right) = \mathrm{Cov}\left(\frac{1}{N}\sum x_i, \frac{1}{N}\sum x_j\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{Cov}(x_i, x_j) = \frac{\sigma^2}{N}$$

saturates the CRLB

$\mathrm{Cov}(x_i, x_j) = \delta_{ij}\sigma^2$ bc of iid assumption

② suppose σ^2 is also unknown $\quad \theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$ multiparameter now!

$$\frac{\partial \ell}{\partial \sigma^2} = \sum_{i=1}^{N} -\frac{1}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{(x_i-\mu)^2}{(\sigma^2)^2} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2} = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{(\sigma^2)^2} \Rightarrow \hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

Want to calculate the Fisher matrix:

$$\frac{\partial^2 \ell}{\partial(\sigma^2)\partial\mu} = \sum_{i=1}^{N} -\frac{(x_i-\mu)}{(\sigma^2)^2} \Rightarrow E\left[-\frac{\partial^2 \ell}{\partial(\sigma^2)\partial\mu}\right] = \sum_{i=1}^{N} \frac{E(x_i-\mu)}{\sigma^4} = 0$$

$$\frac{\partial^2 \ell}{\partial(\sigma^2)^2} = \sum_{i=1}^{N} \frac{1}{2}\frac{1}{\sigma^4} - \frac{(x_i-\mu)^2}{\sigma^6} \Rightarrow E\left[\frac{\partial^2 \ell}{\partial(\sigma^2)^2}\right] = -\frac{N}{2\sigma^4}$$

$$\Rightarrow I = \begin{pmatrix} N/\sigma^2 & 0 \\ 0 & N/2\sigma^4 \end{pmatrix} \Rightarrow I^{-1} = \begin{pmatrix} \sigma^2/N & 0 \\ 0 & 2\sigma^4/N \end{pmatrix}$$

Again, $\mathrm{Var}(\hat{\mu}_{MLE}) = \sigma^2/N$, $\mathrm{Var}(\hat{\sigma^2}_{MLE}) = 2\sigma^4/N$ both saturate the CRLB