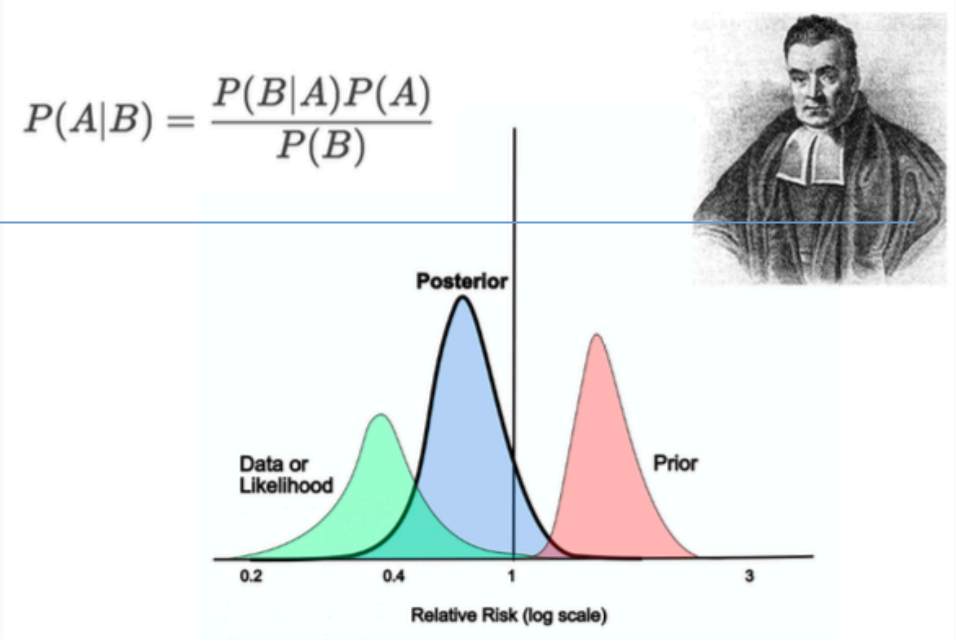
Mathematical Statistics II

STA2212H S LEC9101

Week 5

February 4 2025



Today

1. Recap Jan 28 marginal posterior, hierarchical Bayes
2. Optimality in estimation: efficiency, CRLB MS §6.4
3. Optimality in estimation: decision theory MS §6.2

("finite-sample")

4. HW4 ← corr'd yesterday
5. Office Hour today 3.30 - 4.30

Department Seminar Thursday February 6 11.00 – 12.00
Hydro Building, Room 9014
" Numerical integration in statistical problems "
Alex Stringer, U Waterloo

Choosing priors

- conjugate priors ← exp'l families (analytic calc's easier)
- non-informative priors
  objective
  flat, "ignorance"
- convenience priors * whatever is in software (default) : often conjugate or flat
- minimally/weakly informative priors || ← Zampieri
  - flat
  - 3 normals
- hierarchical priors

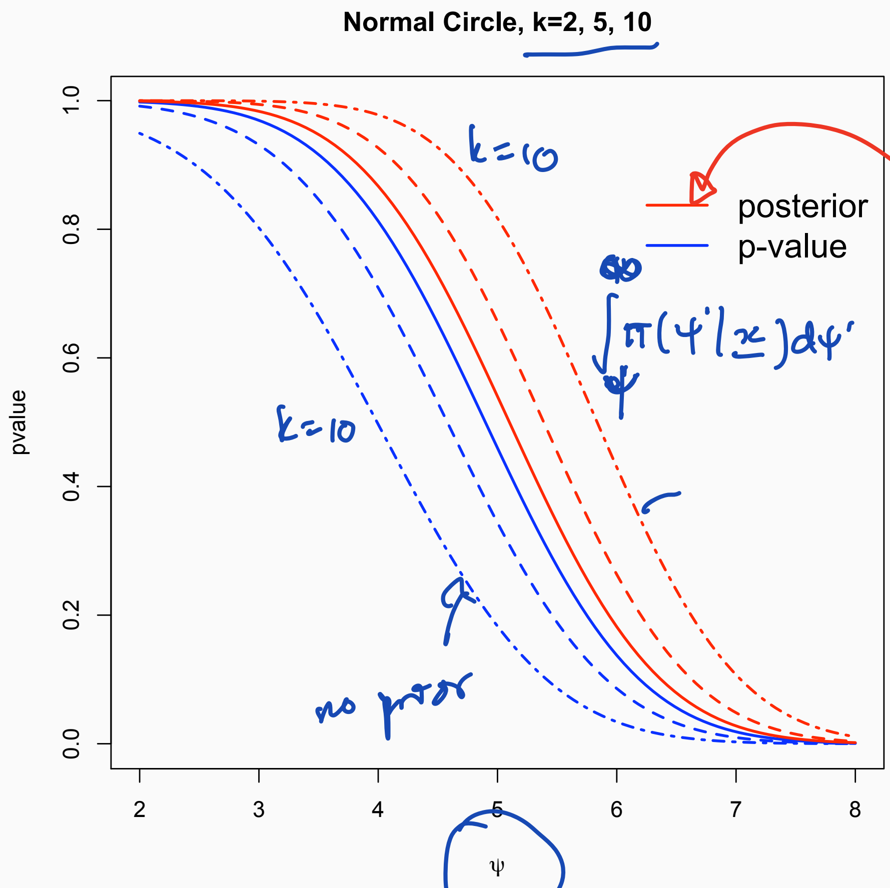
- Bayes estimators
  mode[$\pi(\theta|x)$], $E(\theta|\underset{\sim}{x})$, med[$(\underset{\sim}{\theta}|x)$]

Marginal posterior distributions

- Bayes posterior carries all the information about θ, given $\boldsymbol{x}$ — by definition
- probabilities for any set A computed using the posterior distribution
- $\mathrm{pr}(\boldsymbol{\Theta} \in A \mid \boldsymbol{x}) = \int_A \pi(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta}$
- if $\theta = (\psi, \boldsymbol{\lambda})$, $\quad \underbrace{\pi_{\mathrm{m}}(\psi \mid \boldsymbol{x})}_{\text{marg'l}} = \int \pi(\psi, \boldsymbol{\lambda} \mid \boldsymbol{x}) d\boldsymbol{\lambda} = \dfrac{\int L(\psi, \boldsymbol{\lambda}; \boldsymbol{x}) \pi(\psi, \boldsymbol{\lambda}) d\boldsymbol{\lambda}}{\int L(\psi, \boldsymbol{\lambda}; \boldsymbol{x}) \pi(\psi, \boldsymbol{\lambda}) d\psi d\boldsymbol{\lambda}}$
- or, if $\psi = \psi(\theta)$, $\quad \pi_{\mathrm{m}}(\psi \mid \boldsymbol{x}) = \int_A \pi(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta}, \quad A = \{\boldsymbol{\theta} \in \Theta : \psi(\boldsymbol{\theta}) = \psi\}$
- with marginalization, ‘flat’ priors can have a large influence on the marginal posterior

Example: many normal means

Stein, 1959



$X_i \sim N(\mu_i, 1) \qquad i = 1, \dots, k$

$\pi(\mu_i) \propto 1 \qquad i = 1, \dots, k$

$\pi(\mu_i \mid x_i) = N(x_i, 1)$

$$\pi(\underset{\sim}{\mu} \mid \underset{\sim}{x}) = \prod_{i=1}^{n} \pi(\mu_i \mid x_i)$$

$$\psi = \sum_{i=1}^{k} \mu_i^2$$

$$\pi_m(\psi \mid \underset{\sim}{x}) = \chi_k^{2\prime}\left(\sum_{i=1}^{k} x_i^2\right)$$

exp'd $k + \sum x_i^2$

$\sum X_i^2 \sim \chi_k^{2\prime}(\psi)$

$p(\psi) = \Pr\{\chi_k^{2\prime}(\psi) \le \sum x_i^2\}$

Bayesian hierarchical models

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$
- $\theta_i \mid \mu \sim N(\mu, \sigma^2)$
- $\mu \sim N(\mu_0, \tau^2)$
- $f(x \mid \theta, \mu)$

v_i known

σ^2 known

hyperparameters

$$f(x_i \mid \theta_i) = \frac{1}{\sqrt{2\pi}\sqrt{v_i}} e^{-\frac{1}{2v_i}(x_i-\theta_i)^2}$$

$$\pi(\theta_i \mid \mu) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(\theta_i-\mu)^2}$$

$$\pi(\mu \mid \underline{x}) = \frac{1}{\sqrt{2\pi}\,\tau} e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2}$$

canonical exp. family

$$f(x_i \mid \theta_i) = e^{\left(-\frac{x_i^2}{2v_i} + \frac{\theta_i x_i}{v_i} \underbrace{- \frac{\theta_i^2}{2v_i} - \frac{1}{2}\log v_i - \frac{1}{2}\log\sqrt{2\pi}}_{\text{params. } \ldots}\right)}$$

$$\underbrace{\boxed{-\frac{1}{2v_i}}\,x_i^2}_{s_1} \qquad \underbrace{\frac{\theta_i}{v_i}\,x_i}_{s}$$

$e^{-\frac{1}{\cdots}(\theta_i - \mu)^2}$

Bayesian hierarchical models

- $x_i \mid \theta_i \sim N(\theta_i, v_i)$
- $\theta_i \mid \mu \sim N(\mu, \sigma^2)$
- $\mu \sim N(\mu_0, \tau^2)$
- $\pi(\theta, \mu \mid x)$

hyperparameters

$$e^{-\frac{1}{2}(x_i-\theta_i)^2} \quad e^{-\frac{1}{2}(\theta_i-\mu)^2}$$

$$\pi(\theta_i, x_i, \mu) \propto \underline{f(x_i \mid \theta_i)\,\pi(\theta_i \mid \mu)\,\pi(\mu)}$$

$$\pi(\underset{\sim}{\theta}, \underset{\sim}{x}, \mu) \propto \prod_{i=1}^{n} \cdots$$

before μ: $e^{-\frac{1}{2v_i}(\theta_i - x_i)^2} \quad e^{-\frac{1}{2\sigma^2}(\theta_i-\mu)^2}$

$$\pi(\theta_i \mid x_i, \mu) \propto e^{\theta_i^2 \underbrace{\left(-\frac{1}{2v_i} - \frac{1}{2\sigma^2}\right)} + \theta_i\left(\frac{x_i}{v_i} + \frac{\mu}{\sigma^2}\right) + \cdots}$$

$$N \qquad -\frac{1}{2\,\mathrm{var}} = -\frac{1}{2v_i} - \frac{1}{2\sigma^2}$$

$$E(\theta_i \mid x_i, \mu) = \frac{x_i}{v_i} + \frac{\mu}{\sigma^2}$$

Bayesian hierarchical models

SM 11.4. Eg. 11.25

$\mathrm{var}(\theta_i \mid x_i, \mu)$

$E(\mu \mid x) =$

$\mathrm{var}(\mu \mid x) =$

$E(\theta_i \mid x) =$

$$E(\theta_i \mid x_i, \mu) = \frac{\frac{x_i}{v_i} + \frac{\mu}{\sigma^2}}{\frac{1}{v_i} + \frac{1}{\sigma^2}} = x_i a_i + \mu(1 - a_i)$$

n.b. not centered (?)

grand mean over hosp.

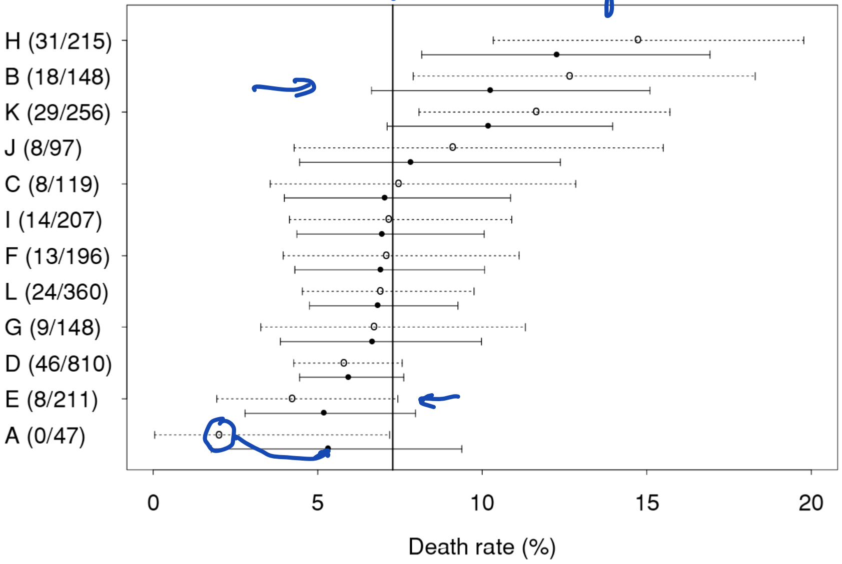


Figure 11.11 Posterior summaries for mortality rates for cardiac surgery data. Posterior means and 0.95 equitailed credible intervals for separate analyses for each hospital are shown by hollow circles and dotted lines, while blobs and solid lines show the corresponding quantities for a hierarchical model. Note the shrinkage of the estimates for the hierarchical model towards the overall posterior mean rate, shown as the solid vertical line; the hierarchical intervals are slightly shorter than those for the simpler model.

no hyperprior

$\mu\left(1-\frac{\sigma^2}{\sigma^2+v_i}\right)$

$\theta_i \sim N(\mu,\sigma^2)$

$\mu \sim N(\mu_0,\tau^2)$

- $$\boxed{\mathrm{E}(\theta_i \mid x) = x_i \frac{\sigma^2}{\sigma^2+v_i} + \mathrm{E}(\mu \mid x)\left(1-\frac{\sigma^2}{\sigma^2+v_i}\right)}$$
- $$\mathrm{E}(\mu \mid x) = \frac{\mu_0/\tau^2 + \sum x_i/(\sigma^2+v_i)}{1/\tau^2 + \sum 1/(\sigma^2+v_i)}$$

$E(x_i) = E[E\{x_i \mid \theta_i, \mu\}]$ (tower formula)

- If $\underbrace{\sigma^2}_{v_i}$ unknown, then need to sample from the posterior, no closed form available

MCMC not algebra

- Figure 11.11 applies similar ideas, plus sampling from the posterior, in logistic regression

Binomial not normal

- recall, in regular models,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\{0, I^{-1}(\theta)\}$$

$I(\theta)$ definition

$$E_\theta\left[\frac{\partial \ell(\theta; X_i)}{\partial \theta}\right]^2$$

- smaller variance means more precise estimation
- Is $I^{-1}(\theta)$ small?

- recall, in regular models,

$I(\theta)$ definition

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\{0, I^{-1}(\theta)\}$$

- smaller variance means more precise estimation
- Is $I^{-1}(\theta)$ small?
- Yes, there's a sense in which it is "as small as possible"

Cramer-Rao lower bound

- recall, in regular models, "almost unbiased" — $I(\theta)$ definition

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} N\{0, I^{-1}(\theta)\}$$

smallest var.

- smaller variance means more precise estimation
- Is $I^{-1}(\theta)$ small?
- Yes, there's a sense in which it is "as small as possible"

- Step 1: suppose $\boldsymbol{X} = X_1, \dots, X_n$ is an i.i.d. sample from a density $f(x;\theta)$
- Let $U = U(\boldsymbol{X}) = \ell'(\theta; \boldsymbol{X})$ — score function
- Let $S = S(\boldsymbol{X})$ be an unbiased estimator of $g(\theta)$ — $\mathrm{E}_\theta\{S(\boldsymbol{X})\} = g(\theta)$
- then $\mathrm{var}_\theta(S) \geq \{\mathrm{Cov}_\theta(S, U)\}^2/\mathrm{Var}_\theta(U)$ $= nI(\theta)$ — proof: Cauchy-Schwarz

lower bound

- Cauchy-Schwartz inequality: Z_1, Z_2, with $\mathrm{E}(Z_1^2) < \infty, \mathrm{E}(Z_2^2) < \infty$, MS Ex 2.7; HW2 STA2112F

$$\{\text{Cov}(Z_1, Z_2)\}^2 \leq \text{var}(Z_1)\text{var}(Z_2)$$

- take $Z_1 = S(\boldsymbol{X})$, an unbiased estimator of $g(\theta)$
- take $Z_2 = U(\boldsymbol{X}) = \Sigma \ell'(\theta; X_i)$

$u_\theta(\underset{\sim}{x})$

score function

$u(\theta; \underset{\sim}{x})$

$= \ell'(\theta; \underset{\sim}{x})$

- then

$$\{\text{Cov}_\theta(S, U)\}^2 \leq \text{var}_\theta(S)\text{var}_\theta(U)$$

-

$$\text{var}_\theta(S) \geq \frac{\text{Cov}_\theta^2(S, U)}{I_n(\theta)}$$

- $$\mathrm{var}_\theta(S) \geq \frac{\mathrm{Cov}^2_\theta(S,U)}{I_n(\theta)}$$
- $\mathrm{Cov}_\theta(S,U)$ $= \int S(\underset{\sim}{x}) \underbrace{\ell'(\theta; \underset{\sim}{x})} f(\underset{\sim}{x};\theta) d\underset{\sim}{x} = \int S(\underset{\sim}{x}) \frac{\partial}{\partial\theta} f(\underset{\sim}{x};\theta)\, d\underset{\sim}{x}$

  $\boxed{E_\theta U = 0}$

  $= \frac{\partial}{\partial\theta} \int S(\underset{\sim}{x}) f(x;\theta) d\underset{\sim}{x}$

  $= g'(\theta)$
- when would we get equality?

$$\boxed{\mathrm{var}_\theta(S) \geq \frac{\{g'(\theta)\}^2}{I_n(\theta)}}$$

If $E_\theta S = \theta$ $\quad g(\theta) = \theta$

$g'(\theta) = 1$

$\mathrm{var}\, S \geq \frac{1}{I_n(\theta)}$

- $$\text{var}_\theta(S) \geq \frac{\text{Cov}^2_\theta(S,U)}{I_n(\theta)}$$

for every value of n

- $\text{Cov}_\theta(S,U)$ $= \dfrac{\{g'(\theta)\}^2}{I_n(\theta)}$

- when would we get equality?

- special case, $g(\theta) = \theta$

only if $\underbrace{u(\theta; \underset{\sim}{X})}_{\text{not strictly a statistic}} = a(\theta)S(\underset{\sim}{X}) + b(\theta) = \ell'(\theta; X)$

$\ell(\theta; \underset{\sim}{X}) = A(\theta)S(\underset{\sim}{X}) + B(\theta) + c(\underset{\sim}{X})$

$f(\underset{\sim}{x}; \theta) = e^{A(\theta)S(\underset{\sim}{X}) + B(\theta) + c(\underset{\sim}{X})}$

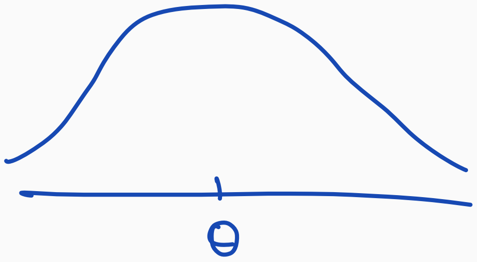
- CRLB attained $\iff$ $U(\theta; X) = A(\theta)S(X) + B(\theta)$ MS: $U_\theta(x)$ (p. 323)
- MS Example 6.12: $X_1, \ldots, X_n$ i.i.d. Poisson(λ) $\lambda^{x_i} e^{-\lambda}/x_i! = f(x_i; \lambda)$
- $\bar{X} = \hat{\lambda}$ has variance $= 1/I(\theta)$ $\hat{\lambda} = \bar{X}$
- For estimating λ^2 three estimators are proposed $\operatorname{var} \hat{\lambda} = \frac{\lambda}{n}$

		$\operatorname{var}(T_j)$	
an unbiased estimator	$T_1 = \boxed{\frac{1}{n}\sum X_i(X_i - 1)}$	$\frac{4\lambda^3}{n} + \frac{2\lambda^2}{n}$	$E\,T_1 = \lambda^2$
the best unbiased estimator	$T_2 = \mathrm{E}(T_1 \mid \bar{X})$	$\frac{4\lambda^3}{n} + \frac{2\lambda^2}{n^2}$	$E\,T_2 = \lambda^2$
the MLE not too bad	$T_3 = \hat{\lambda}^2$	$\frac{4\lambda^3}{n} + \frac{5\lambda^2}{n^2} + \frac{\lambda}{n^3}$	

$E\hat{\lambda}^2 \neq \lambda^2$ [!?]

ntbc $= \lambda^2 + ?/n + \text{rem}_n$

CRLB for λ^2 is $4\lambda^3/n$ (ntbc)



- A more interesting example, logistic density

$$f(x;\theta) = \frac{\exp(x-\theta)}{\{1+\exp(x-\theta)\}^2}$$

$E(X) = \theta$

- CRLB of an unbiased estimator of θ is $3/n$

info. lower bound $I_n^{-1}(\theta)$ is $3/n$

- by previous argument, not attained

in finite samples

- e.g. $\bar{X}$ is unbiased for θ,

$$\underline{\mathrm{var}(\bar{X})} = \frac{\pi^2}{3n} > \frac{3}{n}$$

$= \frac{1}{3}$ ntbc

$E_\theta\{\ell'(\theta;X)^2\}$

$\ell(\theta;x)$

$= \log f(x;\theta)$

What about maximum likelihood estimator?

MS Ch. 6.5

- Suppose $\tilde{\theta}_n$ is ~~a~~ any sequence of estimators with

defined for each n

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

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N\{0, \sigma^2(\theta)\}$$

- Is $\sigma^2(\theta) \geq 1/I(\theta)$?
- Yes, if $\tilde{\theta}_n$ is "regular", and $\sigma^2(\theta)$ continuous in θ technical see MS §6.4, and Thm. 6.6

$$\lim_{n \to \infty} P_{\theta_n}\{\sqrt{n}(\tilde{\theta}_n - \theta_n) \leq y\} = G_\theta(y)$$

for any $\theta_n = \theta + c/\sqrt{n}$

smoothness cond$^{=}$

What about maximum likelihood estimator?

- Suppose $\tilde{\theta}_n$ is a sequence of estimators with

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N\{0, \sigma^2(\theta)\}$$

- Is $\sigma^2(\theta) \geq 1/I(\theta)$?
- Yes, if $\tilde{\theta}_n$ is "regular", and $\sigma^2(\theta)$ continuous in θ

see MS §6.4, and Thm. 6.6

- Is the MLE 'regular'?
- Yes, under the 'usual regularity conditions'
- And, its a.var = lower bound

$$a.var(\hat{\theta}_n) = \frac{1}{nI(\theta)}$$

CRLB

"BAN"

What about maximum likelihood estimator?

- Suppose $\tilde{\theta}_n$ is a sequence of estimators with

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N\{0, \sigma^2(\theta)\}$$

- Is $\sigma^2(\theta) \geq 1/I(\theta)$?
- Yes, if $\tilde{\theta}_n$ is "regular", and $\sigma^2(\theta)$ continuous in θ see MS §6.4, and Thm. 6.6

- Is the MLE 'regular'?
- Yes, under the 'usual regularity conditions'
- And, its a.var = lower bound "BAN"

- there are other regular estimators that are also asymptotically fully efficient
- and might be better in finite samples

- comparison of two consistent estimators

via limiting distributions

- $\sqrt{n}(T_{1n} - \theta) \xrightarrow{d} N\{0, \sigma_1^2(\theta)\}, \quad \sqrt{n}(T_{2n} - \theta) \xrightarrow{d} N\{0, \sigma_2^2(\theta)\}$

- asymptotic relative efficiency of T_{1n}, relative to T_{2n} is $\dfrac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}$

$\sigma_2^2 > \sigma_1^2$

$\Rightarrow \dfrac{\sigma_2^2}{\sigma_1^2} > 1$

ARE T_{1n}, relative to T_2

> 1

T_{1n} is more eff't

- comparison of two consistent estimators — via limiting distributions
- $\sqrt{n}(T_{1n} - \theta) \xrightarrow{d} N\{0, \sigma_1^2(\theta)\}$, $\sqrt{n}(T_{2n} - \theta) \xrightarrow{d} N\{0, \sigma_2^2(\theta)\}$
- asymptotic relative efficiency of T_1, relative to T_2 is $\dfrac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}$ $\quad \dfrac{I^{-1}(\theta)}{\sigma_1^2(\theta)}$
- if T_{2n} is the MLE $\hat{\theta}_n$, then $\sigma_2^2(\theta) = I^{-1}(\theta)$ — as small as possible

ARE of T_1 .. T_2 is $\dfrac{1}{I(\theta)\,\sigma_1^2(\theta)}$

- the MLE is fully efficient

? nfbc? w 10% prob.

- the asymptotic efficiency of T_1 is $1/\sigma_1^2(\theta)I(\theta)$ — relative to the MLE implicit

rel. amount

$\sigma_1^2 \geq I^{-1}(\theta)$

T_1 is less eff. than $\hat{\theta}_{mle}$

- finite-sample approach to optimality in estimation
- start with a loss function $\ell(\hat{\theta}, \theta)$ not lik. fn but a loss fn
- examples: squared error, absolute error, 0-1 loss, Kullback-Liebler

loss incurred by using $\hat{\theta}$ to est. θ, when θ is true value

need a $\hat{\theta}$
want it
'close' to θ

$\ell_1(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ $\quad$ $\ell_p(\hat{\theta}, \theta) = |\hat{\theta} - \theta|^p$ $\quad p \geq 1$

$\ell_2(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$

$\ell_{KL}(\hat{\theta}, \theta) = \int \log \frac{f(x;\theta)}{f(x;\hat{\theta})} f(x;\theta)\,dx$ $\quad$ KL loss

$$\ell_3(\hat{\theta}, \theta) = \begin{cases} 0 & \text{if } \hat{\theta} = \theta \\ 1 & \text{if } \hat{\theta} \neq \theta \end{cases}$$

- finite-sample approach to optimality in estimation
- start with a loss function $L(\hat{\theta}, \theta)$
- examples: squared error, absolute error, 0-1 loss, Kullback-Liebler

- Risk function of $\hat{\theta}$ is expected loss: under $f(x;\theta)$ model for assessing risk

$$R_\theta(\hat{\theta}) = \mathrm{E}_\theta\{L(\hat{\theta}, \theta)\}$$

AoS $R(\hat{\theta}, \theta)$

MSE, MAE, bias/variance trade-off

$$R_{\hat{\theta}}(\theta) = \int L[\hat{\theta}(x), \theta) f(x;\theta) dx$$

Decision theory and Bayes estimators

- finite-sample approach to optimality in estimation
- start with a loss function $L(\hat{\theta}, \theta)$
- examples: squared error, absolute error, 0-1 loss, Kullback-Liebler

- Risk function of $\hat{\theta}$ is expected loss:

$$R_\theta(\hat{\theta}) = \mathrm{E}_\theta\{L(\hat{\theta}, \theta)\}$$

MSE, MAE, bias/variance trade-off

- Risk function depends on θ, and on the form of the estimator

Examples: squared error loss

$\hat{\theta}_1(X) = X$

$\hat{\theta}_2(X) = 3$

$X \sim N(\theta, 1)$

$R_\theta(\hat{\theta})$

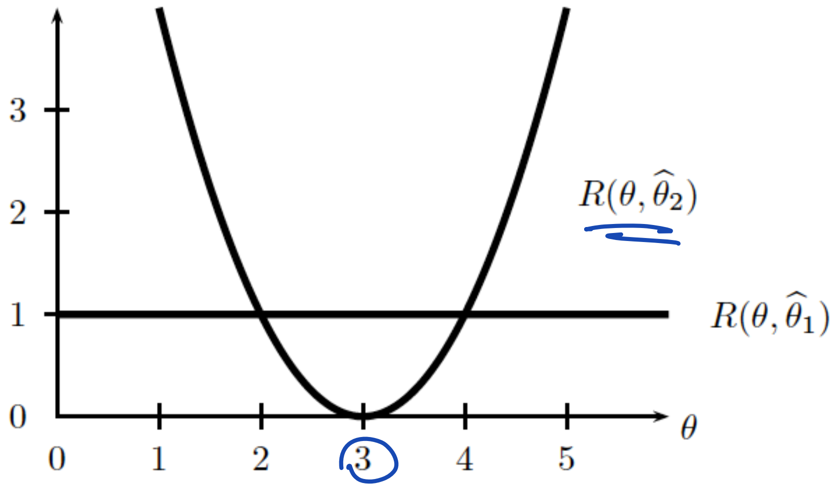
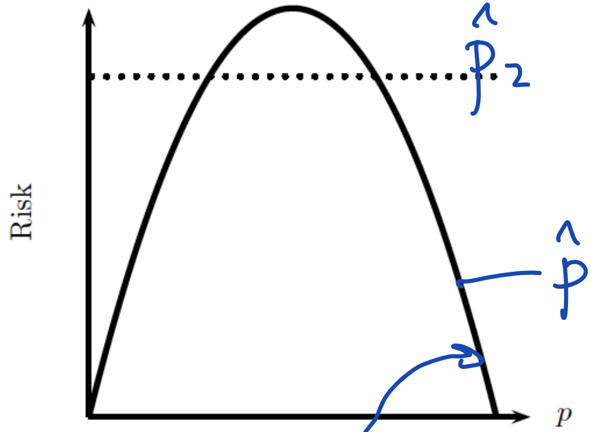


FIGURE 12.1. Comparing two risk functions. Neither risk function dominates the other at all values of θ.

$\hat{p}_2 = \dfrac{X+a}{n+a+b}$

$E(p|X)$ under Beta prior

$a = b = \sqrt{\frac{n}{4}}$



$\hat{p}_2$ risk

$\hat{p}_1 = \bar{X} = X/n$

FIGURE 12.2. Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 12.3. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

$X \sim Binom(n, \cancel{\theta}\, p)$

under L_2

sq'd err

$E_p\{(\frac{X}{n} - p)^2\} = p(1-p)$

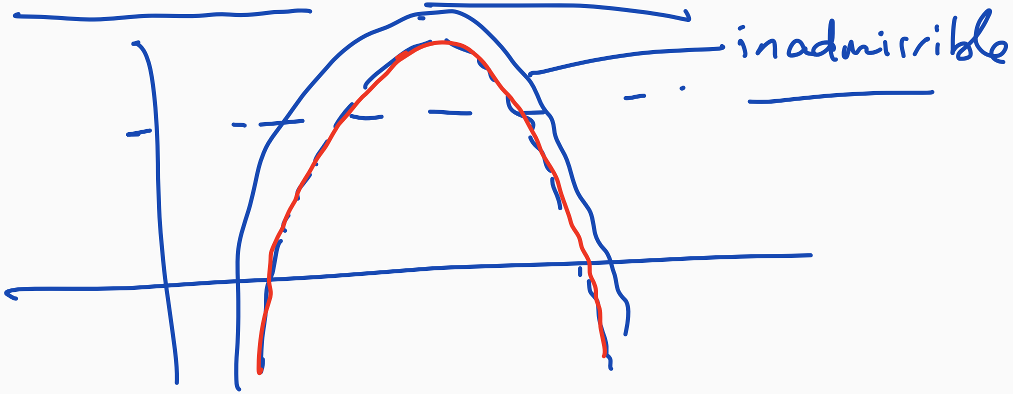
$\alpha = \beta = \sqrt{n/4}$

- an estimator is admissible if no other estimator has a smaller risk function
- For a given loss function L, an estimator $\hat{\theta}$ is inadmissible if there is another estimator $\tilde{\theta}$ with

$$R_\theta(\tilde{\theta}) \leq R_\theta(\hat{\theta}), \quad \text{for all } \theta \in \Theta,$$

and

$$R_{\theta_0}(\tilde{\theta}) < R_{\theta_0}(\hat{\theta}), \quad \text{for some } \theta_0 \in \Theta.$$



Optimality

- an estimator is admissible if no other estimator has a smaller risk function

- For a given loss function L, an estimator $\hat{\theta}$ is inadmissible if there is another estimator $\tilde{\theta}$ with
$$R_\theta(\tilde{\theta}) \le R_\theta(\hat{\theta}), \quad \text{for all } \theta \in \Theta,$$
and
$$R_{\theta_0}(\tilde{\theta}) < R_{\theta_0}(\hat{\theta}), \quad \text{for some } \theta_0 \in \Theta.$$

- MS Ex 6.1; $X \sim \lambda \exp(-\lambda x)$: under squared-error loss, $\hat{\lambda}$ is inadmissible:
Beat by $\tilde{\lambda} = (n-1)\hat{\lambda}/n$
But under a different loss function the MLE has smaller risk than $\tilde{\lambda}$

$L(\hat{\theta}, \theta) = \log(\frac{\theta}{\hat{\theta}}) - 1 - \frac{\theta}{\hat{\theta}}$

- the Bayes risk of an estimator is the average of the risk function, over a prior distribution

$\min_a E(X-a)^2 : \underline{a = EX}$

- $$R_B(\hat{\theta}) = \int R_\theta(\hat{\theta})\pi(\theta)d\theta$$

find a $\hat{\theta}$ to minimize Bayes $\underline{\text{risk}}$

- Optimal Bayes estimators minimize the expected posterior loss:

$$\int L\{\hat{\theta}(x), \theta)\}\pi(\theta \mid x)d\theta$$

- Example: squared-error loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ need to minimize over $\hat{\theta}$

$$\int (\hat{\theta} - \theta)^2\pi(\theta \mid \boldsymbol{x})d\theta$$

$= E_{pos}(a-\theta)^2$

choose a to min $\int (a-\theta)^2 \pi(\theta|x)d\theta$

$\tilde{\theta}_B(\underline{x})$

- solution $\underline{\hat{\theta}(\boldsymbol{x})} = \mathrm{E}(\theta \mid \boldsymbol{x})$ posterior mean

min. Bayes risk over estimator class

Bayes estimators are admissible

- Suppose $\hat{\theta}$ is a Bayes estimator

and is unique

- Suppose we have another estimator $\tilde{\theta}$ with a smaller frequentist risk function:

$$R_\theta(\tilde{\theta}, \theta) \leq R_\theta(\hat{\theta}, \theta)$$

- The Bayes risk of $\tilde{\theta}$ is

$$R_B(\tilde{\theta}) = \int$$

Bayes estimators are admissible

- Suppose $\hat{\theta}$ is a Bayes estimator

and is unique

- Suppose we have another estimator $\tilde{\theta}$ with a smaller frequentist risk function:

$$R_\theta(\tilde{\theta}, \theta) \leq R_\theta(\hat{\theta}, \theta)$$

- The Bayes risk of $\tilde{\theta}$ is

$$R_B(\tilde{\theta}) = \int$$

- instead of minimizing the average (over $\pi(\theta)$) of the risk function we could

$$\min \max R_\theta(\hat{\theta})$$

Definition §6.2

- such estimators are called minimax

Decision theory

- finding the 'best' point estimator $\hat{\theta}$
- best = smallest expected loss
- no asymptotic theory involved
- can find these using a Bayesian argument
- but the justification is not Bayesian
- another non-asymptotic approach to 'best' estimators: UMVU MS 6.3

Bayesian philosophy

F1 Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.

F2 Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.

F3 Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

prop. of method

Bayesian philosophy

B1 Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation. For example, I might say that "the probability that Albert Einstein drank a cup of tea on August 1, 1948" is .35. This does not refer to any limiting frequency. It reflects my strength of belief that the proposition is true.

B2 We can make probability statements about parameters, even though they are fixed constants.

B3 We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

QUESTION 1: Interpreting probability

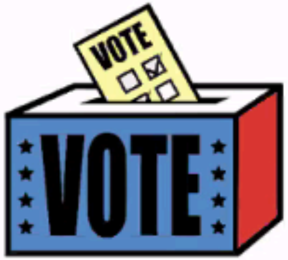
P(Heads) = 0.5 means...

a. If I flip this coin over and over, roughly 50% will be Heads.

b. Heads and Tails are equally plausible.

c. Both a and b make sense.

QUESTION 2: Interpreting probability (again)



P(candidate A wins) = 0.8 means...

a. If we observe this election over & over, candidate A will win roughly 80% of the time.

b. Candidate A is 4 times more likely to win than to lose.

c. The pollster's calculation is wrong.
Candidate A will either win or lose, thus their probability of winning can only be 1 or 0.

QUESTION 3: Bigger picture

I claim that I can predict the outcome of a coin flip.

Mine claims she can distinguish between non-vegan and vegan poutine.

We both succeed in 10 of 10 trials! What do you conclude?

a. My claim is ridiculous. You're still more confident in Mine's claim than in my claim.

b. 10-out-of-10 is 10-out-of-10 no matter the context. Thus the evidence supporting my claim is just as strong as the evidence supporting Mine's claim.

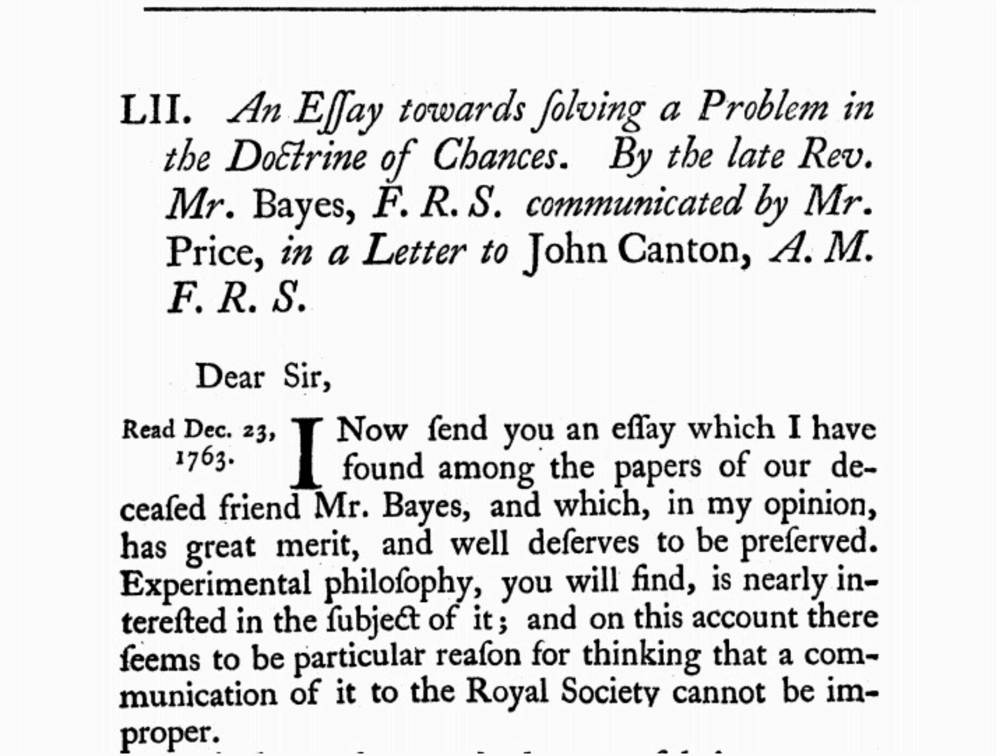
QUESTION 4: Asking questions

You've tested positive for a very rare genetic trait.
If you only get to ask the doctor **one** question, which would it be?

a. P(rare trait | +)
Given the positive test result, what's the probability I actually have the trait?

b. P(+ | rare trait)
If I *don't* have the trait, what's the chance I would have tested positive anyway?

Some history



LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763.

I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

… Some history

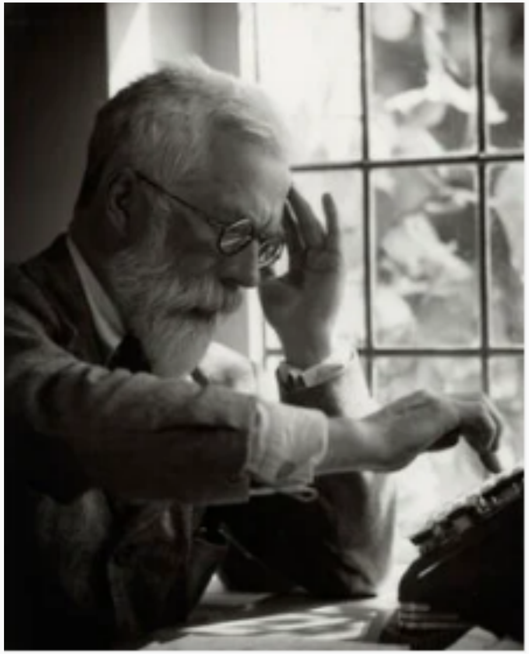
528 *Dr Fisher, Inverse probability*

Inverse Probability. By R. A. Fisher, Sc.D., F.R.S., Gonville and Caius College; Statistical Dept., Rothamsted Experimental Station.

[*Received* 23 July, *read* 28 July 1930.]

I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet that is quite exactly the position in respect of inverse probability. Bayes, who seems to have first attempted to apply the notion of probability, not only to effects in relation to their causes but also to causes in relation to their effects, invented a theory, and evidently doubted its soundness, for he did not publish it during his life. It was posthumously published by Price, who seems to have felt no doubt of its soundness. It and its applications must have made great headway during the next 20 years, for Laplace takes for granted in a highly generalised form what Bayes tentatively wished to postulate in a special case.

Before going over the formal mathematical relationships in



This just in

Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 181–187
https://doi.org/10.1093/jrsssa/qnae044
Advance access publication 17 May 2024

Original Article

Bayesian issues in the 1950s: an episode involving Karl Popper and Jimmie Savage

Stephen M. Stigler

Department of Statistics, University of Chicago, Chicago, IL, USA

Address for correspondence: Stephen M. Stigler, Department of Statistics, University of Chicago, 5747 S. Ellis Ave., Chicago, IL 60637, USA. Email: stigler@uchicago.edu

... This just in

'The trifle in my hand that I wanted to mention is that you may remember, that in the dittoed draft of my book [Savage, 1954], one of the earliest ditto drafts, I attributed to R. A. Fisher the expression of the idea that since the a priori distribution washes out in a large sample, that there ought to be some intrinsic way of analyzing the data in itself without ever postulating a prior distribution at all. I don't remember whether I criticized that argument on the spot, but it's not valid, of course, because the prior distribution does wash out, does so only exponentially, and the rate at which it washes out does depend considerably on what prior distribution it is. Thus for example, since I'm firmly convinced that extrasensory perception does not exist, it would take tremendous amounts of data, of relevant opposing data, to bring me to the opposite point of view. Well, the thing was, we couldn't find this passage anywhere in Fisher and, when I wrote him, he said it was ridiculous, he never could've said any such thing, but Bob Schlaifer has found the reference for me, and it's in Paper 24 of Fisher's collected papers, it's the passage that straddles pages 286 and 287 and I just thought you might like to look at it for yourself.'

Here is the relevant paragraph from Fisher (1934):

$\pi(\theta)$

'As an axiom this supposition [a uniform prior distribution] of Bayes fails, since the truth of an axiom should be manifest to all who clearly apprehend its meaning, and to many writers, including, it would seem, Bayes himself, the truth of the supposed axiom has not been apparent. It has, however, been frequently pointed out that, even if our assumed form for f(x)dx be somewhat inaccurate, our conclusions, if based on a considerable sample of observations, will not greatly be affected; and, indeed, subject to certain restrictions as to the true form of f(x)dx, it may be shown that our errors from this cause will tend to zero as the sample of observations is increased indefinitely. The conclusions drawn will depend more and more entirely on the facts observed, and less and less upon the supposed knowledge a priori introduced into the argument. This property of increasingly large samples has been sometimes put forward as a reason for accepting the postulate of knowledge a priori. It appears, however, more natural to infer from it that it should be possible to draw valid conclusions from the data alone, and without a priori assumptions.—If the justification for any particular form of f(x) is merely that it makes no difference whether the form is right or wrong, we may well ask what the expression is doing in our reasoning at all, and whether, if it were altogether omitted, we could not without its aid draw whatever inferences may, with validity, be inferred from the data. In particular we may question whether the whole difficulty