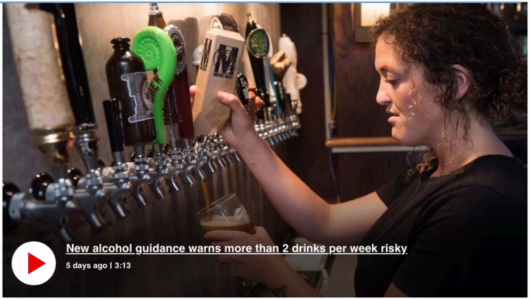
Mathematical Statistics II

STA2212H S LEC0101

Week 3

January 24 2023



Drinking less is better

We now know that even a small amount of alcohol can be damaging to health.

Science is evolving, and the recommendations about alcohol use need to change.

Research shows that no amount or kind of alcohol is good for your health.
It doesn't matter what kind of alcohol it is—wine, beer, cider or spirits.

Drinking alcohol, even a small amount, is damaging to everyone, regardless of age, sex, gender, ethnicity, tolerance for alcohol or lifestyle.

That's why if you drink, it's better to drink less.

Alcohol consumption per week

Drinking alcohol has negative consequences. The more alcohol you drink per week, the more the consequences add up.

0 drinks per week Not drinking has benefits, such as better health, and better sleep.	No risk	0
1 to 2 standard drinks per week You will likely avoid alcohol-related consequences for yourself and others.	Low risk	1 2
3 to 6 standard drinks per week Your risk of developing several different types of cancer, including breast and colon cancer, increases.	Moderate risk	3 4 5 6
7 or more standard drinks per week Your risk of heart disease or stroke increases. **Each additional standard drink** Radically increases the risk of these alcohol-related consequences.	Increasingly high risk	7 8 + ++

During pregnancy, none is the only safe option.

A standard drink means:

Beer
341 ml (12 oz) of beer
5% alcohol

or

Cooler, cider, ready-to-drink
341 ml (12 oz) of drinks
5% alcohol

or

Wine
142 ml (5 oz) of wine
12% alcohol

or

Spirits
(whisky, vodka, gin, etc.)
43 ml (1.5 oz) of spirits
40% alcohol

Aim to drink less

Drinking less benefits you and others. It reduces your risk of injury and violence, and many health problems that can shorten life.

Here is a good way to do it

Count how many drinks you have in a week.

Set a weekly drinking target. If you're going to drink, **make sure you don't exceed 2 drinks on any day.**

Good to know

You can reduce your drinking in steps! Every drink counts: any reduction in alcohol use has benefits.

It's time to pick a new target

What will your weekly drinking target be?

0 1 2 3 4 5 6

Tips to help you stay on target

- Stick to the limits you've set for yourself.
- Drink slowly.
- Drink lots of water.
- For every drink of alcohol, have one non-alcoholic drink.
- Choose alcohol-free or low-alcohol beverages.
- Eat before and while you're drinking.
- Have alcohol-free weeks or do alcohol-free activities.

Canadian Centre on Substance Use and Addiction

The Canadian Centre on Substance Use and Addiction was commissioned by Health Canada to produce Canada's Guidance on Alcohol and Health.
This document is a summary for the public of the new guidance. For more information, please visit www.ccsa.ca.

link

Today

Start Recording

1. Recap
2. Bayesian Inference
3. Optimality in Estimation MS 6
4. H3: comments on HW1, 2; Examples ...

Upcoming seminars of interest

- January 30 3.30 – 4.30 Chiara Sabatti Details
  “Human populations and gene mapping”

- January 30 6.00 – 7.00 pm Vera Liao Details
  “Introduction to Explainable AI”

Recap

- K-L divergence $K(f : f_0) = \int \log \frac{f_0(x)}{f(x)} f_0(x) dx$

re consistency of mle

- empirical c.d.f. is nonparametric MLE of $F(x)$

- profile likelihood and log-likelihood functions $\ell_{prof}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$
- constrained MLE $\hat{\lambda}_\psi = \arg\sup_\lambda \ell(\psi, \lambda)$

parameter of interest; nuisance par

- asymptotically $\ell_{prof}(\psi)$ can be used for inference as $\ell(\theta)$

finite sample properties might be poor

- Bayesian inference: prior, likelihood, posterior
- Bayesian philosophy: parameters modelled as random variables
- Bayesian estimation: posterior mean, median, mode

Example: Bivariate normal

Table 3.1 *Scores from two tests taken by 22 students,* `mechanics` *and* `vectors`.

	1	2	3	4	5	6	7	8	9	10	11
`mechanics`	7	44	49	59	34	46	0	32	49	52	44
`vectors`	51	69	41	70	42	40	40	45	57	64	61

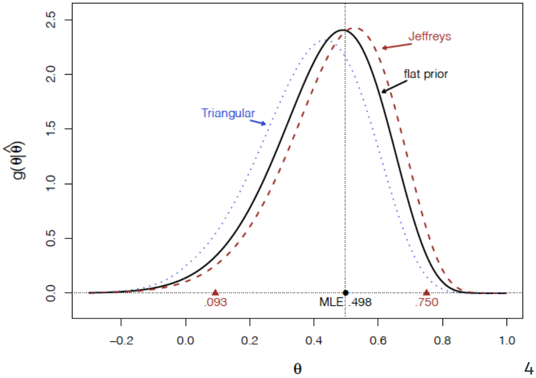
	12	13	14	15	16	17	18	19	20	21	22
`mechanics`	36	42	5	22	18	41	48	31	42	46	63
`vectors`	59	60	30	58	51	63	38	42	69	49	63

Table 3.1 shows the scores on two tests, `mechanics` and `vectors`, achieved by $n = 22$ students. The sample correlation coefficient between the two scores is $\hat{\theta} = 0.498$,

$$\hat{\theta} = \sum_{i=1}^{22}(m_i - \bar{m})(v_i - \bar{v}) \Big/ \left[\sum_{i=1}^{22}(m_i - \bar{m})^2 \sum_{i=1}^{22}(v_i - \bar{v})^2\right]^{1/2}, \quad (3.10)$$

with m and v short for `mechanics` and `vectors`, $\bar{m}$ and $\bar{v}$ their averages. We wish to assign a Bayesian measure of posterior accuracy to the true correlation coefficient θ, "true" meaning the correlation for the hypothetical population of all students, of which we observed only 22.

If we assume that the joint (m, v) distribution is bivariate normal (as



Example: Binomial

SM Ex.11.11

Table 11.2 Mortality rates r/m from cardiac surgery in 12 hospitals (Spiegelhalter *et al.*, 1996b, p. 15). Shown are the numbers of deaths r out of m operations.

A	0/47	B	18/148	C	8/119	D	46/810	E	8/211	F	13/196
G	9/148	H	31/215	I	14/207	J	8/97	K	29/256	L	24/360

provided the mode lies inside the parameter space. Here $\tilde{J}(\theta)$ is the second derivative matrix of $-\tilde{\ell}(\theta)$. This expansion corresponds to a posterior multivariate normal

prior for hospital A $Beta(1, 1)$

posterior mean

Example: Binomial

SM Ex.11.11

580 11 · *Bayesian Models*

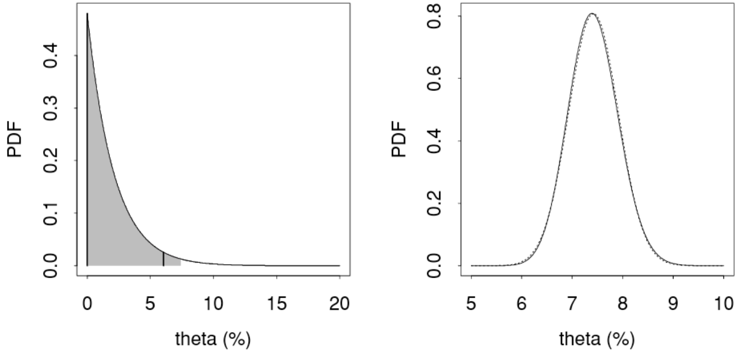


Figure 11.1 Cardiac surgery data. Left panel: posterior density for θ_A, showing boundaries of 0.95 highest posterior credible interval (vertical lines) and region between posterior 0.025 and 0.975 quantiles of $\pi(\theta_A \mid y)$ (shaded). Right panel: exact posterior beta density for overall mortality rate θ (solid) and normal approximation (dots).

hospital A

all hospitals

208 failures, 2814 observations

Choosing priors

- conjugate priors
- non-informative priors — flat, “ignorance”
- convenience priors
- minimally/weakly informative priors
- hierarchical priors

Exponential families and conjugate priors

$$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}; \qquad \pi(\theta;\alpha,\beta) = K(\alpha,\beta)\exp\{\alpha c(\theta) - \beta d(\theta)\}$$

Exponential families and conjugate priors

$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\}$; $\qquad \pi(\theta;\alpha,\beta) = K(\alpha,\beta)\exp\{\alpha c(\theta) - \beta d(\theta)\}$

Example: $f(x;\theta) = \theta(1-\theta)^x, x = 0, 1, \ldots; 0 < \theta < 1$

$f(x;\theta) = \exp\{c(\theta)T(x) - d(\theta) + S(x)\};$ $\pi(\theta;\alpha,\beta) = K(\alpha,\beta)\exp\{\alpha c(\theta) - \beta d(\theta)\}$

Example: $f(x;\theta) = \theta(1-\theta)^x, x = 0, 1, \ldots; 0 < \theta < 1$

Example: $f(x;\mu) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x-\mu)^2\}$

$f(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}}\exp\{-\frac{x^2}{2\sigma^2}\}$

- if parameter space is closed (interval), e.g. $\Theta = [a, b]$, then $\pi(\theta) \sim U(a, b)$ represents 'indifference'

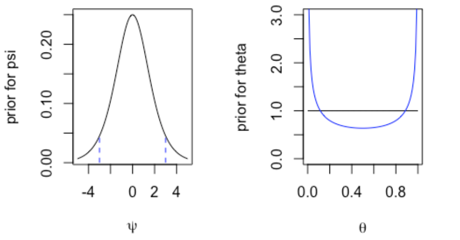
- example: Beta (1,1) prior for Bernoulli probability — last week

- example 5.34: $X \sim N(\mu, 1), \pi(\mu) \propto 1$

Flat priors

- if parameter space is closed (interval), e.g. $\Theta = [a, b]$, then $\pi(\theta) \sim U(a, b)$ represents 'indifference'
- example: Beta (1,1) prior for Bernoulli probability — last week
- example 5.34: $X \sim N(\mu, 1), \pi(\mu) \propto 1$
- improper priors can lead to proper posteriors — ntbc
- priors flat in one parameterization are not flat in another

- Example: $X \sim Bin(n, \theta), 0 < \theta < 1; \theta \sim U(0, 1)$
- log-odds ratio $\psi = \psi(\theta) = \log\{\theta/(1 - \theta)\}$
- $\pi(\psi) = \dfrac{e^{\psi}}{(1 + e^{\psi})^2}, -\infty < \psi < \infty$
- prior probability $-3 < \psi < 3 \approx 0.9$
- an invariant prior: $\pi(\theta) \propto I^{1/2}(\theta)$



Jeffreys' prior

- $\pi(\theta) \propto I^{1/2}(\theta)$

- Example: $X \sim Bin(n, \theta) \qquad I(\theta) = n/\{\theta(1-\theta)\}, \quad 0 < \theta < 1$

- Example 5.35: $X \sim Poisson(\lambda), \qquad I(\lambda) = 1/\lambda, \quad \lambda > 0$ posterior proper?

- Jeffreys' prior for multiparameter $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$ not recommended even by Jeffreys

- Example: $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2) \qquad I(\mu, \sigma^2) =$

Marginalization

- Bayes posterior carries all the information about θ, given $\boldsymbol{x}$ by definition
- probabilities for any set A computed using the posterior distribution
- $\mathrm{pr}(\boldsymbol{\Theta} \in A \mid \boldsymbol{x}) =$
- if $\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda})$, ...
- or, if $\psi = \psi(\boldsymbol{\theta})$
- in this context, 'flat' priors can have a large influence on the marginal posterior

Not all likelihood functions are regular

Example: $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

… Not all likelihood functions are regular

MS Exercise 5.1

$X_1, \ldots, X_n$ i.i.d. $f(x;\boldsymbol{\theta}) = a(\theta_1, \theta_2)h(x), \quad \theta_1 \leq x \leq \theta_2$

Optimality of estimators

- recall, in regular models,

$I(\theta)$?

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

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

- recall, in regular models,

$I(\theta)$?

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

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

- recall, in regular models,

$I(\theta)$?

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

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

- Step 1: suppose $\boldsymbol{X} = X_1, \ldots, X_n$ is an i.i.d. sample from a density $f(x; \theta)$
- and suppose that $\mathrm{E}_\theta\{S(\boldsymbol{X})\} = g(\theta)$
- then $\mathrm{var}(S) \geq \{\mathrm{Cov}_\theta(S, U)\}^2 / \mathrm{Var}_\theta(U)$

proof: Cauchy-Schwarz

- recall, in regular models, $I(\theta)$?

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

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

- Step 1: suppose $\boldsymbol{X} = X_1, \dots, X_n$ is an i.i.d. sample from a density $f(x;\theta)$
- and suppose that $\mathrm{E}_\theta\{S(\boldsymbol{X})\} = g(\theta)$
- then $\mathrm{var}(S) \geq \{\mathrm{Cov}_\theta(S,U)\}^2/\mathrm{Var}_\theta(U)$ proof: Cauchy-Schwarz

- Cauchy-Schwartz inequality: for random variables X, Y, with $\mathrm{E}(X^2) < \infty, \mathrm{E}(Y^2) < \infty$,

$$\{\mathrm{Cov}(X, Y)\}^2 \leq \mathrm{var}(X)\mathrm{var}(Y)$$

- now suppose $X_1, \dots, X_n$ i.i.d. with density $f(x; \theta)$
- and suppose $S(\boldsymbol{X})$ is unbiased for $g(\theta)$
- and recall $U(\boldsymbol{X}) = \Sigma\ell'(\theta; X_i)$ score function
- then

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

- $$\{\mathrm{Cov}_\theta(S,U)\}^2 \le \mathrm{var}_\theta(S)\mathrm{var}_\theta(U)$$
- special case $g(\theta) = \theta$

- $$\{\mathrm{Cov}_\theta(S,U)\}^2 \leq \mathrm{var}_\theta(S)\mathrm{var}_\theta(U)$$
- special case $g(\theta)=\theta$
- when would we get equality?

Example: Poisson

Unbiased estimator of λ^2: $S_1(\boldsymbol{X}) = (1/n)\Sigma X_i(X_i - 1)$ ntbc

Maximum likelihood estimator of λ^2: $S_2(\boldsymbol{X}) = \{(1/n)\Sigma X_i\}^2$

$$\begin{aligned} \mathrm{var}(S_1) &= \frac{4\lambda^3}{n} + \frac{2\lambda^2}{n} \\ \mathrm{var}(S_2) &= \frac{4\lambda^3}{n} + \frac{5\lambda^2}{n^2} + \frac{\lambda}{n^3} \end{aligned}$$

Cramer-Rao lower bound: $\{g'(\lambda)\}^2/nI(\lambda) = (2\lambda)^2/(n/\lambda) = 4\lambda^3/n$

Note: CRLB cannot be attained even by an unbiased estimator

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

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"

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_1, relative to T_2 is $\dfrac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}$

Asymptotic efficiency

- 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_1, relative to T_2 is $\dfrac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}$

- if T_{1n} is the MLE $\hat{\theta}_n$, then $\sigma_1^2(\theta) = I^{-1}(\theta)$

  as small as possible

- the MLE is fully efficient

- the asymptotic, relative to MLE efficiency of T_2 is $\sigma_2^2(\theta)I(\theta)$

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, K-L divergence

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

- 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

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, K-L divergence

- 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

AoS 12.2, 12.3; MS Ex.6.1

12.2 Comparing Risk Functions 195

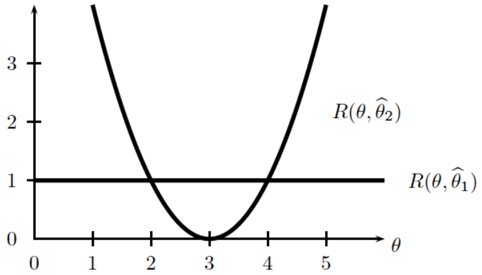


FIGURE 12.1. Comparing two risk functions. Neither risk function dominates the other at all values of θ.

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

Examples: squared error loss

$X \sim Binom(n, \theta)$

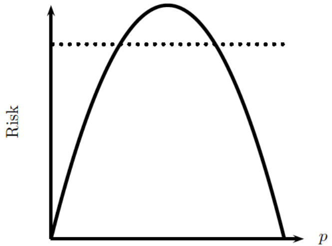


FIGURE 12.2. Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 12.3. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

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

Optimality

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

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

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

- the Bayes risk of an estimator is the average of the risk function, over a prior distribution
- $$R_B(\hat{\theta}) = \int R_\theta(\hat{\theta})\pi(\theta)d\theta$$
- 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 $\hat{\theta}$
$$\int (\hat{\theta} - \theta)^2\pi(\theta \mid \boldsymbol{x})d\theta$$
- solution $\hat{\theta}(\boldsymbol{x}) = \mathrm{E}(\theta \mid \boldsymbol{x})$ posterior mean

- 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

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

- 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

- parameter $\theta = (\theta_1, \ldots, \theta_p)$
- model $f(x^n \mid \theta), \quad x^n = (x_1, \ldots, x_n)$
- joint posterior

$$\pi(\theta \mid x^n) \propto f(x^n \mid \theta)\pi(\theta), \quad \theta \in \mathbb{R}^p$$

Multi-parameter models

- parameter $\theta = (\theta_1, \dots, \theta_p)$
- model $f(x^n \mid \theta)$, $\quad x^n = (x_1, \dots, x_n)$
- joint posterior

$$\pi(\theta \mid x^n) \propto f(x^n \mid \theta)\pi(\theta), \quad \theta \in \mathbb{R}^p$$

- marginal posterior — for θ_1

$$\pi_m(\theta_1 \mid x^n) = \int \pi(\theta \mid x^n) d\theta_2 \dots d\theta_p$$

- marginal posterior — for $\psi(\theta)$

$$\pi_m(\psi \mid x^n) = \int_{\{\theta:\psi(\theta)=\psi\}} \pi(\theta \mid x^n) d\theta$$

Bayesian inference: Multi-parameter models

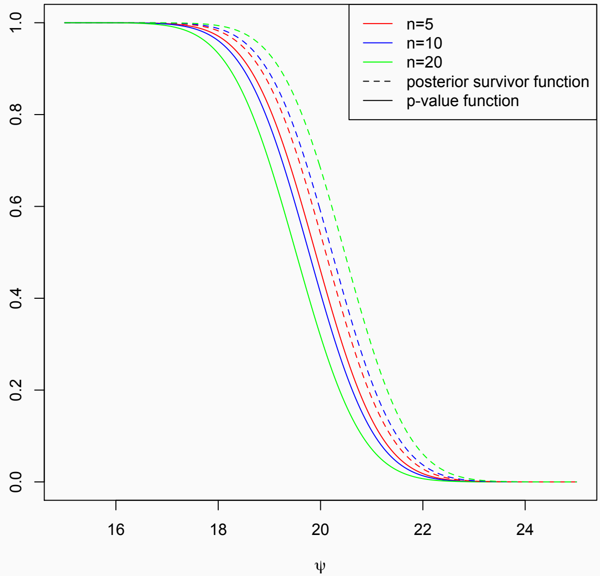
- model: $x_i \sim N(\mu_i, 1), i = 1, \ldots, n$
- prior: $\pi(\mu)d\mu \propto d\mu$
- posterior $\pi(\mu \mid x^n) \propto \prod_{i=1}^n \pi(\mu_i \mid x_i) = \prod_{i=1}^n \phi(x_i, 1/n)$

Bayesian inference: Multi-parameter models

- model: $x_i \sim N(\mu_i, 1), i = 1, \ldots, n$
- prior: $\pi(\mu)d\mu \propto d\mu$
- posterior $\pi(\mu \mid x^n) \propto \prod_{i=1}^n \pi(\mu_i \mid x_i) = \prod_{i=1}^n \phi(x_i, 1/n)$
- $\psi = \sum_{i=1}^n \mu_i^2$ squared length of mean vector

$$\pi(\psi \mid x^n) = \int_A \pi(\mu \mid x^n) d\mu$$

- $\mu_i \mid x_i \sim N(x_i, 1) \implies \sum \mu_i^2 \mid x^n \sim \chi_n^2(\sum x_i^2)$



n=5
n=10
n=20
posterior survivor function
p-value function
0.0
0.2
0.4
0.6
0.8
1.0
16
18
20
22
24
ψ

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.

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

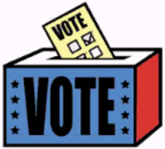
P(Heads) = 0.5 means...

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

B b. Heads and Tails are equally plausible.

P 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?

Bayesian hierarchical models

SM 11.4, Eg. 11.25

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

Bayesian hierarchical models

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

Bayesian hierarchical models

SM 11.4. Eg. 11.25

$$E(\mu \mid x) =$$

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

$$E(\theta_i \mid x) =$$

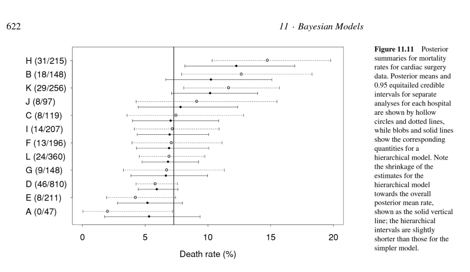


Figure 11.11 Posterior summaries for mortality rates for cardiac surgery data. Posterior means and 0.95 equitailed credible intervals for separate analyses for each hospital are shown by hollow circles and dotted lines, while blobs and solid lines show the corresponding quantities for a hierarchical 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.

Bayesian hierarchical models

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

- If σ^2 unknown, then need to sample from the posterior, no closed form available
- Figure 11.11 applies similar ideas, plus sampling from the posterior, in logistic regression