Bayesian inference | example | Kaiser ICU case study ①

random sample of $n = 112$ medical records of heart attack patients, $s = 4$ of them had an unplanned transfer to the Intensive Care Unit (ICU)

Goal: population proportion of ICU transfers for such patients $\triangleq \theta = ?$ (unknown)

Estimate θ from the dataset $\underset{\sim}{y} = (y_1, \dots, y_n)$,

$$y_i = \begin{cases} 1 & \text{if ICU} \\ 0 & \text{not} \end{cases}$$

and construct a 95% interval for θ using the Bayesian approach to inference

For comparison: 95% frequentist interval from two

points of view: ($\boxed{1}$ Neyman, Central Limit Theorem) ②

$$\hat{\theta} = \bar{y} = \frac{s}{n} = \frac{4}{112} \doteq 0.0357$$

$$\widehat{SE}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \doteq \sqrt{\frac{(.036)(1-.036)}{112}} \doteq 0.0175$$

approximate (large-sample) 95% confidence interval (CI) $A: \hat{\theta} \pm 1.96\,\widehat{SE}(\hat{\theta})$

$\boxed{\text{Neyman}}$ $\leftarrow \Phi^{-1}(1-\frac{.05}{2})$

$= (.001, .070)$

$\boxed{2}$ (Fisher, maximum likelihood)

$$\hat{\theta}_{MLE} = \bar{y} = \frac{s}{n} = \frac{4}{112} \doteq .0357$$

$$\widehat{SE}(\hat{\theta}_{MLE}) = \sqrt{\hat{I}^{-1}(\hat{\theta}_{MLE})}$$

likelihood function

$$\ell(\theta|y) = \theta^{s}(1-\theta)^{n-s}$$

log likelihood function $\ell\ell(\theta|y) = s\log\theta + (n-s)\log(1-\theta)$

$$\frac{d}{d\theta} \ell(\theta|y) = \frac{s}{\theta} - \frac{n-s}{1-\theta} = \frac{s-n\theta}{\theta(1-\theta)}$$

$= 0$ iff $\theta = \hat{\theta}_{MLE} = \frac{s}{n}$ $\boxed{\text{calculating the Fisher information}}$

$$\frac{d^2}{d\theta^2} \ell(\theta|y) = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}, \text{ so}$$

$$-\frac{d^2}{d\theta^2} \ell(\theta|y) \Big|_{\theta=\frac{s}{n}} = \hat{I}(\hat{\theta})$$

(after simplification) $= \dfrac{n}{\hat{\theta}_{MLE}(1-\hat{\theta}_{MLE})} =$

---

so $\hat{SE}(\hat{\theta}_{MLE}) = \sqrt{\dfrac{\hat{\theta}_{MLE}(1-\hat{\theta}_{MLE})}{n}}$ ← same as (Neyman)

$\doteq 0.0175$ $\boxed{\text{approximate (large-sample)}}$

$\boxed{\text{Fisher}}$

95% confidence interval: $\hat{\theta} \pm 1.96\, \hat{SE}(\hat{\theta})$

so $\begin{pmatrix}\text{Neyman approx.} \\ \text{95\% CI}\end{pmatrix} = \begin{pmatrix}\text{Fisher} \\ \text{approx.} \\ \text{95\% CI}\end{pmatrix}$ ④

$= (.001, .070)$

Bayesian approach

reminder of Bayes's theorem for (T/F) true/false propositions

U = unknown (T/F)

D = data (T/F)

problem context provides $P(U)$, $P(D)$, $P(D|U)$;

what is $P(U|D)$?

definition

$$P(U|D) = \begin{cases} \dfrac{P(U, D)}{P(D)} & \text{if } P(D) > 0 \\ \text{undefined} & \text{else} \end{cases}$$

(and)

so therefore

$$P(U, D) = P(D) \cdot P(U|D)$$

definition $\quad P(D|U) = \begin{cases} \dfrac{P(D,U)}{P(U)} & \text{if } P(U)>0 \\ 0 & \text{else} \end{cases}$ (5)

so therefore

$$P(D,U) = P(U)\cdot P(D|U)$$

but $P(U,D) = P(D,U)$; therefore

$$P(D)\cdot P(U|D) = P(U)\cdot P(D|U)$$

and

$$\boxed{P(U|D) = \frac{P(U)\cdot P(D|U)}{P(D)}} \quad (*)$$

Bayes's theorem for T/F propositions

In the Kaiser Fed case study

$U \leftrightarrow 0<\theta<1$ (unknown)

$D \leftrightarrow \underset{\sim}{Y} = (Y_1, \dots, Y_n)$

turns out, as we saw in AMS 131, that

The analogue of Bayes's theorem (6) when $\mathcal{U}$ and $\mathcal{D}$ are not T/F propositions is obtained by simply translating the ingredients in (*) above into unconditional and conditional probability distributions:

$$p(\theta \mid \underset{\sim}{y}) = \frac{\overbrace{p(\theta)}^{\text{prior distribution for } \theta} \cdot \overbrace{p(\underset{\sim}{y} \mid \theta)}^{\text{sampling dist. for } y \text{ given } \theta}}{\underbrace{p(\underset{\sim}{y})}_{\text{sampling distribution for } y}}$$

$p(\theta \mid \underset{\sim}{y})$: posterior distribution for θ given $\underset{\sim}{y}$

once the data vector $\underset{\sim}{y}$ is known the left hand side LHS

$p(\theta \mid \underset{\sim}{y})$ is a function of θ for fixed $\underset{\sim}{y}$

therefore the right-hand side (RHS) also has to be regarded as a function of θ for fixed $\underset{\sim}{y}$

$$\frac{p(\theta)\cdot p(\underset{\sim}{y}\mid\theta)}{p(\underset{\sim}{y})}$$

two consequences:

(a) $p(\underset{\sim}{y})$ is a constant (in θ), playing the role of $\frac{1}{\text{(normalizing constant)}}$, stuck into the equation to make the LHS integrate to 1; thus

$$p(\theta\mid y) \propto p(\theta)\, p(y\mid\theta)$$

is proportional to

(b) Bayesians think of the sampling distribution $p(\underset{\sim}{y} \mid \theta)$ of $\underset{\sim}{y}$ given θ as a function of θ for fixed $\underset{\sim}{y}$, <u>exactly as</u> Fisher did when he defined <u>maximum likelihood</u>:

$$\ell(\theta \mid \underset{\sim}{y}) = c \cdot p(\underset{\sim}{y} \mid \theta)$$

arbitrary constant $c > 0$

Thus $p(\theta \mid \underset{\sim}{y}) \propto p(\theta) \cdot \ell(\theta \mid \underset{\sim}{y})$

posterior distribution for θ given $\underset{\sim}{y}$ $\propto$ (prior distribution for θ) $\cdot$ (likelihood function)

Now we already know in the Kaiser Ken case study that

$$\ell(\theta \mid \underline{z}) = \theta^{s}(1-\theta)^{n-s}, \text{ so}$$

$$p(\theta \mid \underline{z}) \propto p(\theta) \cdot \theta^{s}(1-\theta)^{n-s}$$

What shall we use for the prior distribution $p(\theta)$?

$p(\theta)$ quantifies all information (if any) about θ <u>external</u> to the data set $\underline{z}$

Suppose (as was the case in the Kaiser example) that little was known about θ external

to Z other than the obvious fact that $0 < \theta < 1$. Then we want $p(\theta)$ to be a <u>low-information</u> (otherwise known as <u>diffuse</u>) prior distribution: it should be essentially flat across most or all of $(0, 1)$.

One possibility:

$$p(\theta) = \begin{cases} 1 & \text{for } 0 < \theta < 1 \\ 0 & \text{else} \end{cases}$$

This is the <u>Uniform</u> distribution on $(0, 1)$. $\theta \sim \text{Uniform}(0, 1)$

More generally, it would be nice (II) to find a <u>family</u> of distributions on $(0,1)$ that made calculating the posterior distribution easy.

Q:	Is there such a family?
A:	yes: it's called the family of <u>Beta</u> distributions $\text{Beta}(\alpha,\beta)$, <u>indexed</u> by the two quantities $\begin{pmatrix}\alpha>0\\ \beta>0\end{pmatrix}$:
its density is	$p(\theta) = \underbrace{c}_{\text{normalizing constant}} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$
Note:	$\text{Beta}(\underset{\alpha}{1}, \underset{\beta}{1}) = \text{Uniform}(0,1)$

Why is this family so useful?

Recall that

$$p(\theta|z) \propto p(\theta) \cdot \theta^{s}(1-\theta)^{n-s}$$

with

$p(\theta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$ the posterior dist.

becomes $p(\theta|z) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{s}(1-\theta)^{n-s}$

$$= \theta^{(\alpha+s)-1}(1-\theta)^{(\beta+n-s)-1}$$

In other words, if the prior is Beta(α, β) then the posterior is Beta$(\alpha+s, \beta+n-s)$

When the prior & posterior are different members of the same family of probability distributions,

The prior is said to be <u>conjugate</u> to the likelihood function.

Thus we have our first conjugate updating rule:

$$\left(s = \sum_{i=1}^{n} y_i \right)$$

$$\left\{ \begin{array}{l} \theta \sim \text{Beta}(\alpha, \beta) \\ (y_i \mid \theta) \overset{IID}{\sim} \text{Bernoulli}(\theta) \end{array} \right\} \rightarrow (\theta \mid \underset{\sim}{y}) \sim \text{Beta}(\alpha + s, \beta + n - s)$$

$\boxed{\text{R demo:}}$ plot Uniform$(0,1)$ = Beta$(1,1)$ prior and Beta$(\alpha+s, \beta+n-s)$ = Beta$(1+4, 1+112-4)$ = Beta$(5, 109)$ posterior on same scale.

Use qbeta(.025, 5, 109) and (14)

qbeta(.975, 5, 109) to compute

95% Bayesian interval & compare

with frequentist 95% CI (.001, .070)

qbeta(.025, 5, 109) ≐ .0145

qbeta(.975, 5, 109) ≐ .0882

Bayes: (.0145, .0882) symmetric

Neyman, Fisher: (.001, .070) symmetric