4. Bayesian Inference

Bayesian Inference

So far we have followed the frequentist approach:

- we have treated unknown parameters as a fixed constants, and
- we have imagined repeated sampling from our model in order to evaluate properties of estimators, interpret confidence intervals, calculate p-values, etc.

We now take a different approach: in Bayesian inference, *unknown parameters* are treated as *random variables*.

In subjective Bayesian inference, probability is a measure of the strength of belief.

Before any data are available, there is uncertainty about the parameter θ. Suppose uncertainty about θ is expressed as a "prior" pdf (of pmf) for θ.

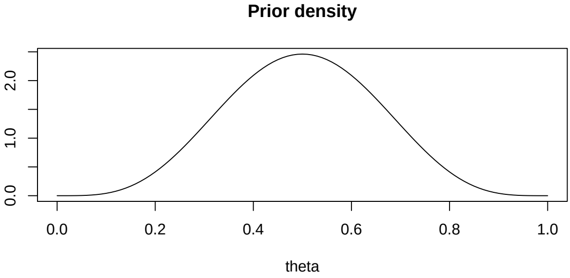
Then, once data are available, we can use Bayes' theorem to combine our prior beliefs with the data to obtain an updated "posterior" assessment of our beliefs about θ.

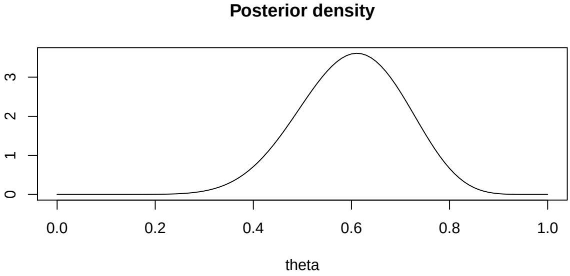
Example

Suppose we have a coin which we think might be a bit biased.

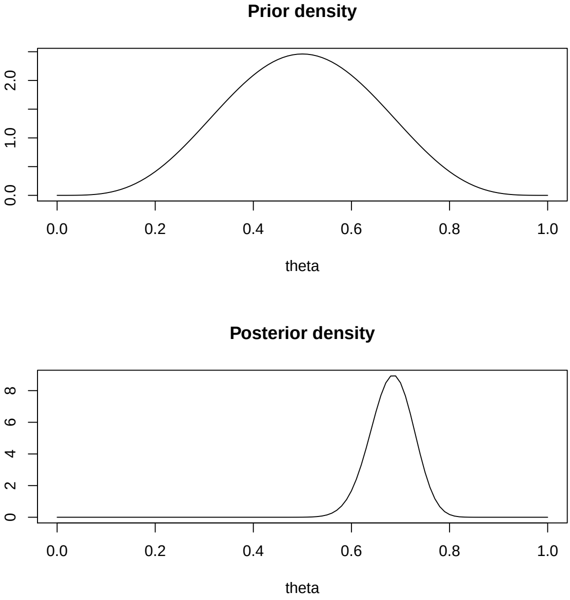
Let θ be the probability of getting a head when we flip it.

Prior: Beta(5, 5). Data: 7 heads from 10 flips.





Prior: Beta(5, 5). Data: 70 heads from 100 flips.



4.1 Introduction

Suppose that, as usual, we have a probability model $f(\underset{\sim}{x} \mid \theta)$ for data $\underset{\sim}{x}$. ← likelihood

In this section we write $f(\underset{\sim}{x} \mid \theta)$ (rather than $f(\underset{\sim}{x}; \theta)$) to indicate that $\underline{x}$ is conditional on θ, we have a conditional distribution/density.

Suppose also, before observing $\underset{\sim}{x}$, we summarise our beliefs about θ in a prior density $\pi(\theta)$.

That is, we treat θ as a random variable.

Once we have observed $\underset{\sim}{x}$, our updated beliefs about θ are contained in the conditional density of θ given $\underset{\sim}{x}$, which is called the <u>posterior density</u> $\pi(\theta \mid \underset{\sim}{x})$.

Theorem (Bayes' theorem - continuous version)

For continuous random variables Y and Z, the conditional density $f(z|y)$ of Z given Y satisfies

$$f(z|y) = \frac{f(y|z)f(z)}{f(y)} \qquad (*).$$

Proof By definition of conditional density,

$$f(z|y) = \frac{f(y,z)}{f(y)} \quad ① \qquad \text{and} \qquad f(y|z) = \frac{f(y,z)}{f(z)} \quad ②.$$

From ② $f(y,z) = f(y|z)f(z)$ and substituting into ① gives $(*)$. $\square$

Note: marginal pdf of Y is

$$f(y) = \int_{-\infty}^{\infty} f(y, z)\, dz = \int_{-\infty}^{\infty} f(y \mid z) f(z)\, dz \qquad (**).$$

(similar expression for $f(z)$).

With $\underline{x}$ and θ in place of y and z we have

$$\pi(\theta \mid \underline{x}) = \frac{f(\underline{x} \mid \theta)\pi(\theta)}{f(\underline{x})} \qquad \leftarrow \text{like } (*)$$

where $$f(\underline{x}) = \int_{\text{all } \theta} f(\underline{x} \mid \theta)\pi(\theta)\, d\theta \qquad \leftarrow \text{like} (**).$$

As usual for conditional densities, we treat $\pi(\theta | \underset{\sim}{x})$ as a function of θ, with data $\underset{\sim}{x}$ fixed.

Since $\underset{\sim}{x}$ is fixed, $f(\underset{\sim}{x})$ is just a constant, and so

$$\boxed{\begin{array}{c} \pi(\theta | \underset{\sim}{x}) \propto f(\underset{\sim}{x} | \theta) \times \pi(\theta) \\ \\ \text{posterior} \propto \text{likelihood} \times \text{prior} \end{array}}$$

Example Conditionally on θ, suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$.

$$P(X_i = 1 \mid \theta) = \theta, \quad P(X_i = 0 \mid \theta) = 1 - \theta$$

i.e. $f(x_i \mid \theta) = \theta^{x_i} (1-\theta)^{1-x_i}, \qquad x_i = 0, 1.$

So likelihood $f(\underline{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$

$= \theta^{r} (1-\theta)^{n-r}$ where $r = \sum_{1}^{n} x_i$

A natural prior here is a Beta(a,b) pdf:

$$\pi(\theta) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}, \qquad 0 < \theta < 1.$$

Here $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,d\theta$ beta function

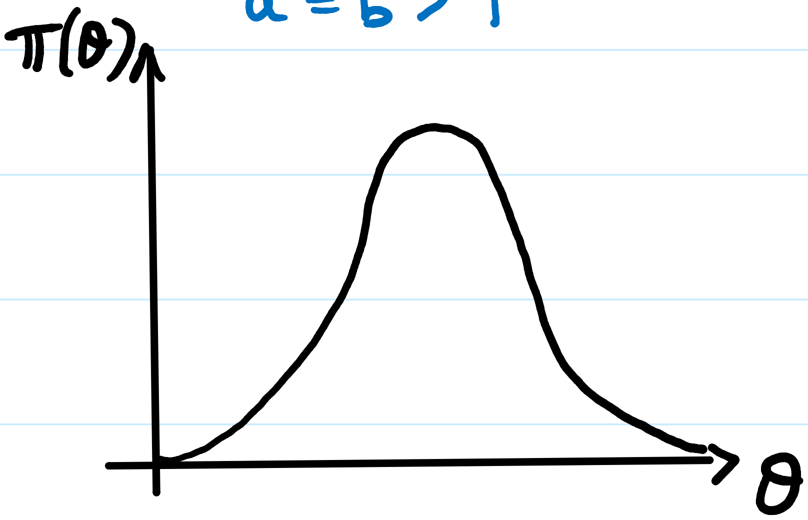
$$= \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$$

and $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\,du$

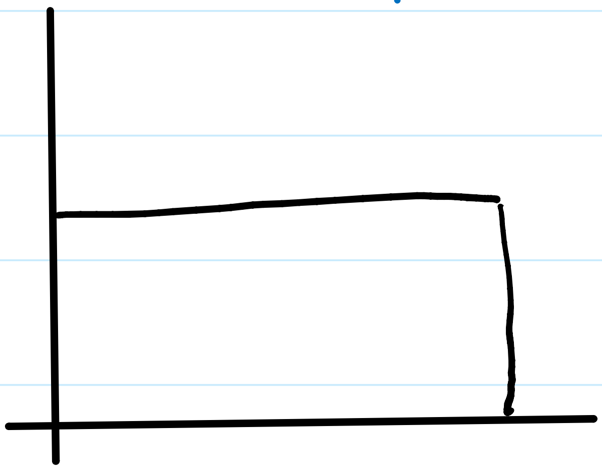
$\Gamma(a+1) = a\,\Gamma(a)$ for $a > 0$

$\Gamma(n) = (n-1)!$ for n positive integer.

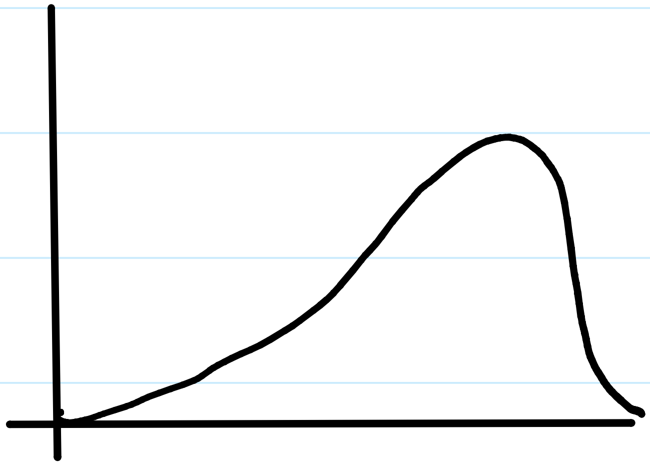
$a = b > 1$



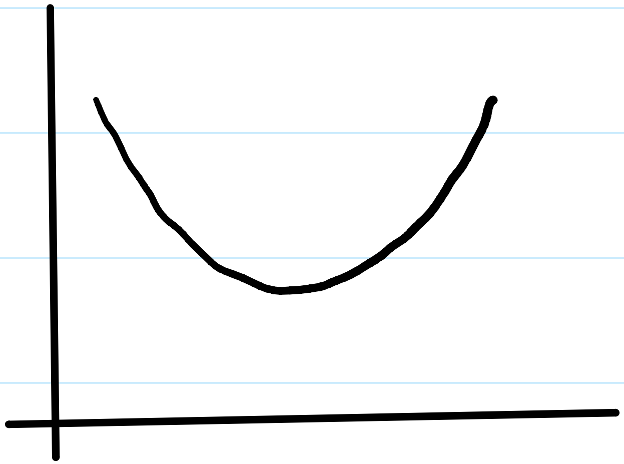
$a = b = 1$



$a > b > 1$



$a = b < 1$



We are assuming a, b known, and $a > 0$, $b > 0$.

chosen to reflect our prior beliefs

Now posterior $\propto$ likelihood $\times$ prior, so

$$\pi(\theta \mid \underset{\sim}{x}) \propto \theta^{r}(1-\theta)^{n-r} \times \theta^{a-1}(1-\theta)^{b-1}$$

$$= \theta^{r+a-1}(1-\theta)^{n-r+b-1} \qquad (3)$$

The RHS of (3) depends on θ exactly as for a Beta$(r+a, n-r+b)$ density.

Hence the constant of proportionality in ③ must be

$$\frac{1}{B(r+a,\, n-r+b)},$$ and the posterior distribution

is a Beta $(r+a,\, n-r+b)$.

So pdf $$\pi(\theta \mid \underline{x}) = \frac{1}{B(r+a,\, n-r+b)} \theta^{r+a-1} (1-\theta)^{n-r+b-1},$$

$$0 < \theta < 1.$$

<u>Note</u>: <u>no need</u> to do any integration.

Example Conditional on θ, suppose $X_1 \ldots X_n \overset{iid}{\sim}$ Poisson(θ).

Suppose prior for θ is a Gamma(α, β) pdf:

$$\pi(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \qquad \theta > 0$$

where $\alpha > 0$, $\beta > 0$ known

posterior $\propto$ likelihood $\times$ prior

$$\pi(\theta | \underline{x}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\theta} \theta^{x_i}}{x_i!} \right) \times \theta^{\alpha-1} e^{-\beta\theta}$$

$$\propto \theta^{r+\alpha-1} e^{-(n+\beta)\theta} \qquad \text{where } r = \sum x_i .$$

So the posterior distribution is a Gamma,

$\pi(\theta \mid \underline{x})$ is a $\mathrm{Gamma}(r+\alpha, n+\beta)$ pdf

[because $\pi(\theta \mid \underline{x})$ depends on θ as for a

$\mathrm{Gamma}(r+\alpha, n+\beta)$].

Example (MRSA)

[Example from `www.scholarpedia.org`.]

Let θ denote the number of MRSA infections per 10,000 bed-days in a hospital.

Suppose we observe $y = 20$ infections in 40,000 bed-days, i.e. in $10{,}000N$ bed-days where $N = 4$.

- A simple estimate of θ is $y/N = 5$ infections per 10,000 bed-days.
- The MLE of θ is also $\widehat{\theta} = 5$ if we assume that y is an observation from a Poisson distribution with mean θN, so

$$f(y \mid \theta) = (\theta N)^y e^{-\theta N}/y!\,.$$

However, other evidence about θ may exist.

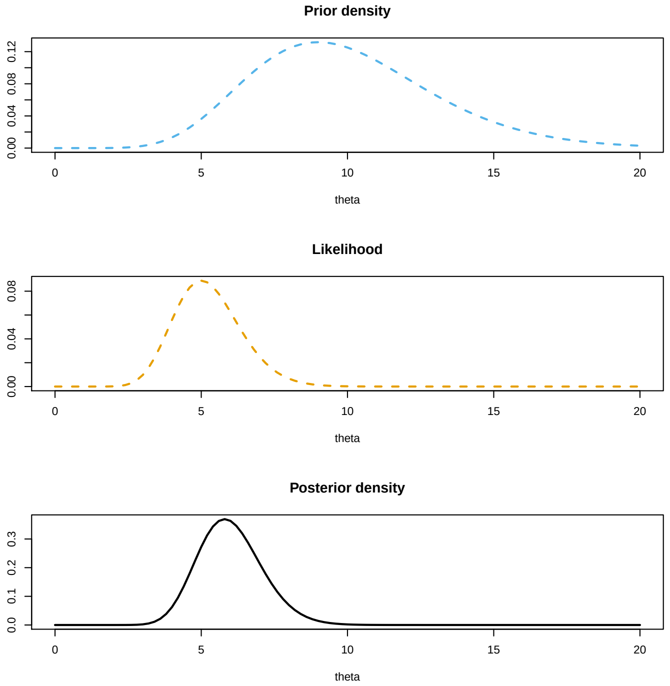
Suppose this other information, on its own, suggests plausible values of θ of about 10 per 10,000, with 95% of the support for θ lying between 5 and 17.

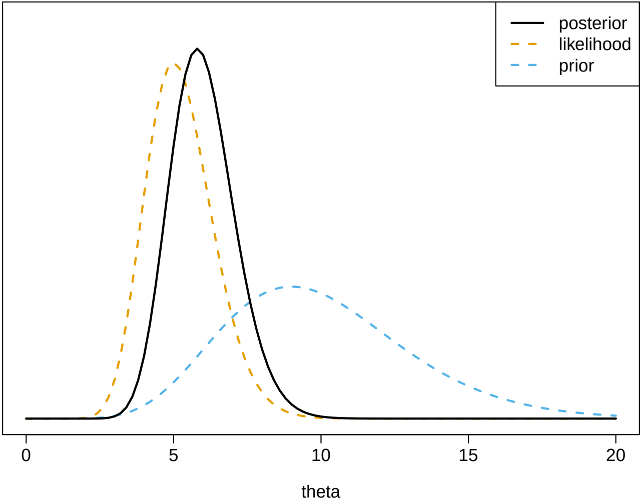
We can use a prior distribution to describe this. A Gamma pdf is convenient here:

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \quad \text{for } \theta > 0.$$

Taking $\alpha = 10$, $\beta = 1$ gives approximately the properties above.

- The posterior combines the evidence from the data (i.e. the likelihood) and the other (i.e. prior) evidence. We can think of the posterior as a compromise between the likelihood and the prior.
- Calculated on board in lectures: the posterior is another Gamma.


Prior density
0.00
0.04
0.08
0.12
0
5
10
15
20
theta
Likelihood
0.00
0.04
0.08
0
5
10
15
20
theta
Posterior density
0.0
0.1
0.2
0.3
0
5
10
15
20
theta


posterior
likelihood
prior
0
5
10
15
20
theta

4.2 Inference

All information about θ is contained in the posterior density $\pi(\theta|\underline{x})$.

Posterior summaries

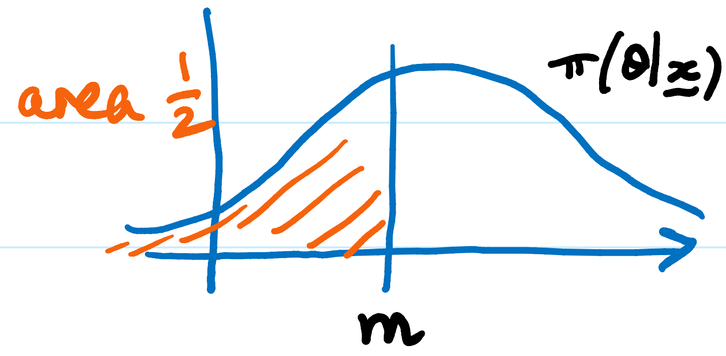
Sometimes summaries of $\pi(\theta|\underline{x})$ are useful, e.g.

(i) the posterior mode (value of θ at which $\pi(\theta|\underline{x})$ is max)

(ii) the posterior mean $E(\theta|\underline{x})$

expectation over θ
($\underline{x}$ is <u>fixed</u>)

(iii) posterior median, m such that $\int_{-\infty}^{m} \pi(\theta | \underline{x})\, d\theta = \frac{1}{2}$



(iv) $\text{var}(\theta | \underline{x})$

(v) other quantiles of $\pi(\theta | \underline{x})$.

Example Conditional on θ, suppose $X \sim \text{Binomial}(n, \theta)$.

We write this as: $X \mid \theta \sim \text{Binomial}(n, \theta)$.

Prior $\theta \sim U(0,1)$.

posterior $\propto$ likelihood $\times$ prior

$$\pi(\theta \mid \underline{x}) \propto \binom{n}{x} \theta^x (1-\theta)^{n-x} \times 1$$

$$\propto \theta^x (1-\theta)^{n-x}$$

So $\theta \mid \underset{\sim}{x} \sim \text{Beta}(x+1, n-x+1)$.

Posterior mean

$$E(\theta|\underline{x}) = \int_0^1 \theta \, \pi(\theta|\underline{x}) \, d\theta$$

$$= \frac{1}{B(x+1, n-x+1)} \int_0^1 \theta^{x+1} (1-\theta)^{n-x} \, d\theta$$

$$= \frac{1}{B(x+1, n-x+1)} \cdot B(x+2, n-x+1)$$

$$= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \cdot \frac{\Gamma(x+2)\Gamma(n-x+1)}{\Gamma(n+3)}$$

$$= \frac{x+1}{n+2}$$ using $\Gamma(a+1) = a\Gamma(a)$ twice

So even when all trials are successes ($x=n$), this point estimate is $\dfrac{n+1}{n+2} < 1$ (seems sensible especially if n small).

Posterior mode is $\dfrac{x}{n}$ (same as MLE).

For large n, i.e. when the likelihood contribution dominates that from the prior, posterior mean and mode will be close.

Interval estimation

Frequentist $\longrightarrow$ confidence interval

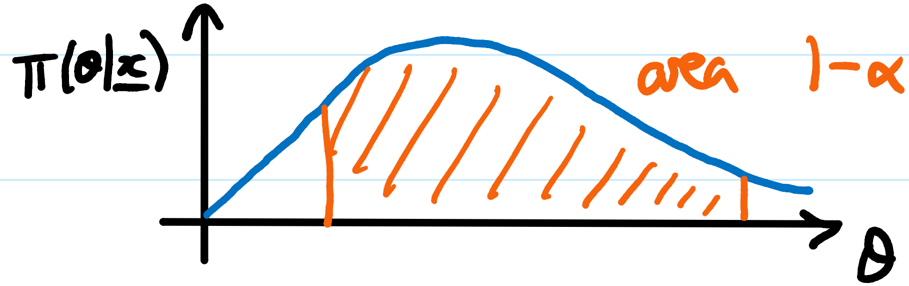
Bayesian $\longrightarrow$ credible interval

Let Θ be the parameter space.

Definition A $100(1-\alpha)\%$ (posterior) credible set for θ is a subset C of Θ such that

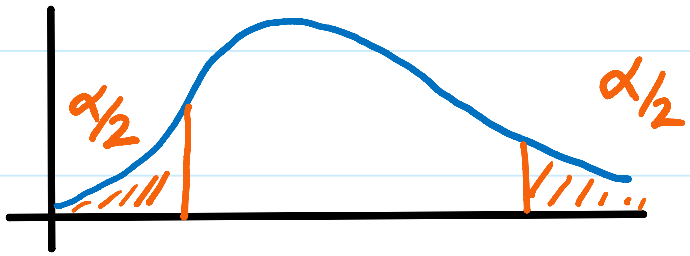
$$\int_C \pi(\theta \mid \underline{x})\, d\theta = 1 - \alpha .$$

Note this is just saying $P(\theta \in C \mid \underset{\sim}{x}) = 1 - \alpha$



A **credible interval** is when set C is an interval, $C = (\theta_1, \theta_2)$ say.

The interval (θ_1, θ_2) is called **equal-tailed** if
$P(\theta \leq \theta_1 \mid \underset{\sim}{x}) = P(\theta \geq \theta_2 \mid \underset{\sim}{x})$



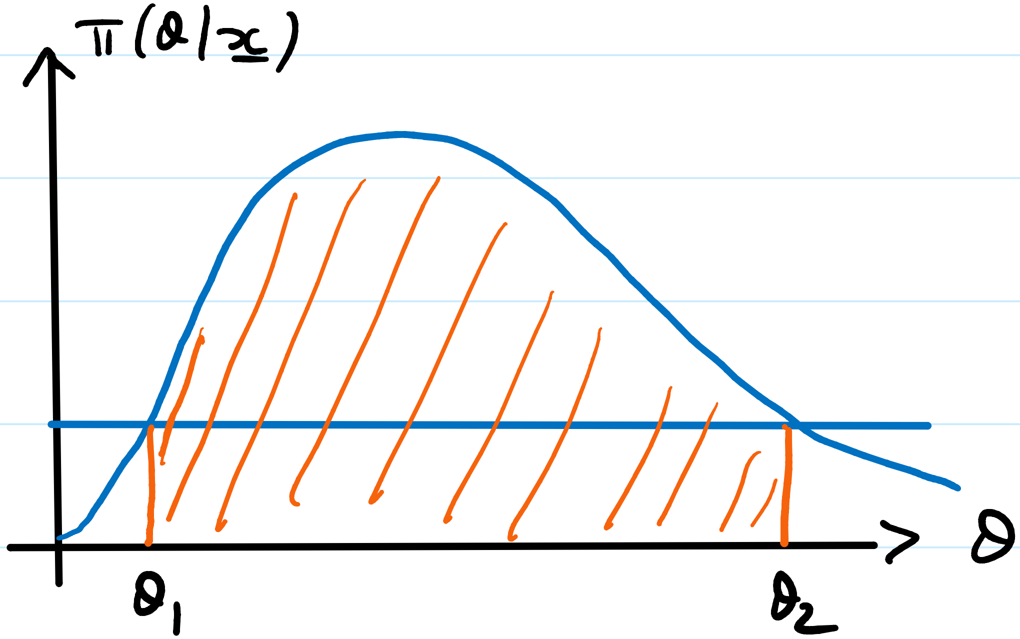
In words: "the probability that θ lies in C, given the observed data $\underline{x}$, is $1-\alpha$"

Very simple!

This is <u>not</u> true of a confidence interval.

Definition We call C a highest posterior density (HPD) credible set if $\pi(\theta | \underline{x}) \geq \pi(\theta' | \underline{x})$ for all $\theta \in C$ and all $\theta' \notin C$.

E.g. (θ_1, θ_2) here:



An HPD interval has minimal width among all $1-\alpha$ credible intervals.

Multi-parameter models

θ may be a vector. If so, everything above still applies, all integrals over θ mean multiple integrals over all components of θ.

e.g. $\theta = (\psi, \lambda)$, so posterior $\pi(\psi, \lambda \mid \underline{x})$.

All info about ψ is contained in the marginal posterior for ψ, which is $\pi(\psi \mid \underline{x}) = \int \pi(\psi, \lambda \mid \underline{x})\, d\lambda$

integrate over all λ to find marginal distribution

Prediction

Let X_{n+1} represent a future observation.

Assume, conditional on θ, that X_{n+1} has density $f(x_{n+1} | \theta)$ independent of $X_1 \ldots X_n$.

The density of X_{n+1} given $\underset{\sim}{x}$, called the posterior predictive density, is a conditional density, found by the usual rules of probability:

$$f(x_{n+1} | \underset{\sim}{x}) = \int f(x_{n+1}, \theta | \underset{\sim}{x}) \, d\theta$$

integrate over all θ to find marginal density

$\underset{\sim}{x} = (x_1, \ldots, x_n)$ here

$$= \int \underbrace{f(x_{n+1} \mid \theta, \underline{x})}_{f(x_{n+1} \mid \theta) \text{ by the independence above}} \pi(\theta \mid \underline{x})\, d\theta$$

$$f(u, v \mid w) = f(u \mid v, w)\, f(v \mid w)$$

$$= \int f(x_{n+1} \mid \theta)\, \pi(\theta \mid \underline{x})\, d\theta.$$

4.3 Prior information

How do we choose a prior $\pi(\theta)$?

(i) If substantial prior knowledge exists, we could ask a subject-area expert.

(ii) If we have little prior knowledge we might want a prior that expresses "prior ignorance"

is this possible? maybe $\theta \sim U(0,1)$ for a prior probability value

(iii) We might want to choose a "conjugate" prior for ease of calculation (by hand)

	prior	lik		posterior
e.g.	Beta +	Bernoulli	$\longrightarrow$	Beta
	Gamma +	Poisson	$\longrightarrow$	Gamma
	$\vdots$			

Note (iii) can overlap with (i) and (ii).

Example Conditional on θ, let $X_1 \dots X_n$ be independent $N(\theta, \sigma^2)$ where σ^2 known.

Let prior be $\theta \sim N(\mu_0, \sigma_0^2)$ where μ_0, σ_0^2 known.

Then $\pi(\theta \mid \underline{x}) \propto f(\underline{x} \mid \theta)\, \pi(\theta)$

$$\propto \exp\left[-\frac{1}{2} \sum \frac{(x_i - \theta)^2}{\sigma^2}\right] \exp\left[-\frac{1}{2} \frac{(\theta - \mu_0)^2}{\sigma_0^2}\right]$$

Now complete the square:

$$\frac{(\theta-\mu_0)^2}{\sigma_0^2} + \sum \frac{(x_i-\theta)^2}{\sigma^2} = \theta^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{x}}{\sigma^2}\right) + \text{constant}$$

$$= \frac{1}{\sigma_1^2}(\theta-\mu_1)^2 + \text{constant}$$

after completing the square

where

$$\mu_1 = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}} \qquad (1)$$

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \qquad (2)$$

Hence $\pi(\theta | \underline{x}) \propto \exp\left(-\frac{1}{2\sigma_1^2}(\theta - \mu_1)^2\right)$

a $N(\mu_1, \sigma_1^2)$ pdf

So $\theta | \underline{x} \sim N(\mu_1, \sigma_1^2)$.

weight $\frac{1}{\sigma_0^2}$

① says: posterior mean μ_1 = weighted av. of prior mean μ_0 and sample mean $\bar{x}$

weight $\frac{n}{\sigma^2}$

The precision of a random variable is $\frac{1}{\text{variance}}$.

② says: posterior precision = prior precision + data precision.

Improper priors

If $\sigma_0^2 \to \infty$ above then $\pi(\theta \mid \underline{x})$ is approx $N\left(\bar{x}, \frac{\sigma^2}{n}\right)$.
i.e. the likelihood contribution dominates the prior contribution as $\sigma_0^2 \to \infty$.

This corresponds to prior $\pi(\theta) \propto c$, a constant,
i.e. a "uniform prior".

But this π is not a probability distribution since $\theta \in (-\infty, \infty)$ and we can't have $\int_{-\infty}^{\infty} c \, d\theta = 1$.

Definition A prior $\pi(\theta)$ is called **proper** if $\int \pi(\theta)\,d\theta = 1$, and is called **improper** if the integral can't be normalised to equal 1.

An improper prior can lead to a proper posterior (e.g. uniform prior $\pi(\theta) \propto c$ for $\theta \in \mathbb{R}$ above) and we can use the posterior for inference.

But we can't use an improper posterior for meaningful inference.

Prior ignorance

If no reliable prior information is available we might want a prior which has minimal effect on our inference.

E.g. if $\Theta = \{\theta_1, \dots, \theta_m\}$ then $\pi(\theta_i) = \frac{1}{m}$, $i = 1 \dots m$ does not favour any value of θ, is "non-informative".

But things are not so simple when θ is continuous.

Example If $\Theta = (0,1)$ we might think $\theta \sim U(0,1)$ represents ignorance

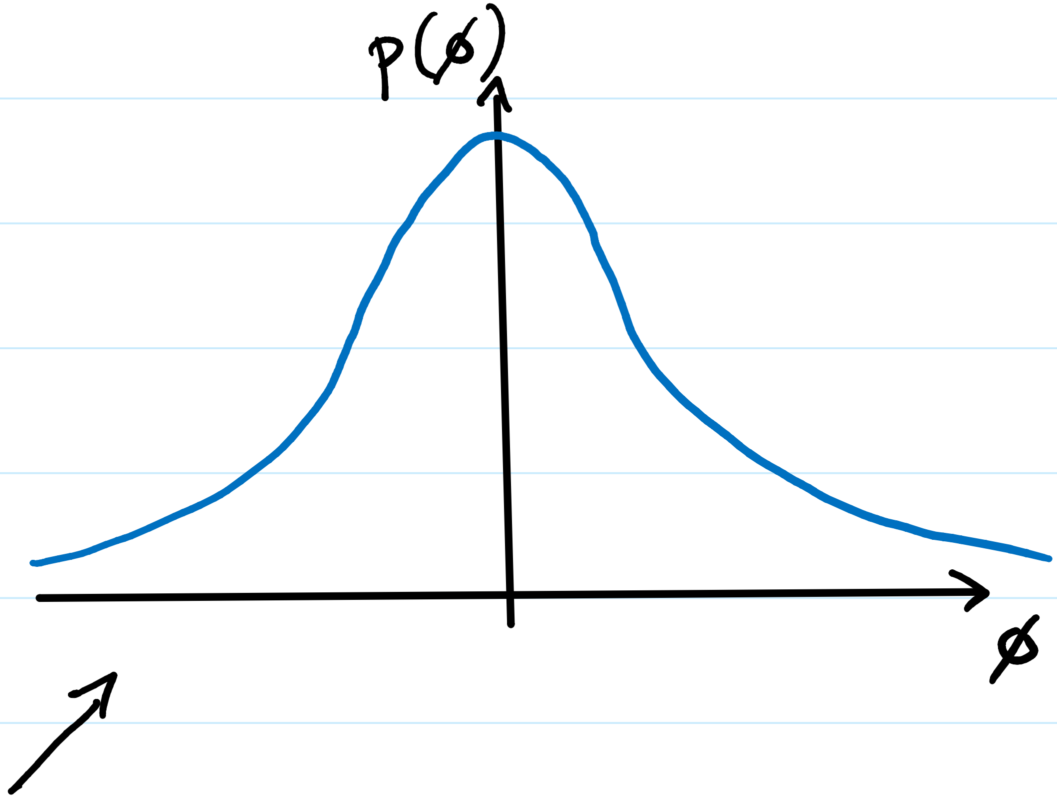
However, if we are ignorant about θ

then we are also ignorant about $\phi = \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{\text{log odds}}$

θ has pdf $\pi(\theta) = 1, \quad 0 < \theta < 1.$

So ϕ has pdf $p(\phi) = \pi(\theta(\phi)) \frac{d\theta}{d\phi}$ $\qquad \theta = \frac{e^{\phi}}{1+e^{\phi}}$

$$= 1 \times \frac{e^{\phi}}{(1+e^{\phi})^2}, \quad \phi \in \mathbb{R}.$$



this does not seem consistent with ignorance about ϕ.

Jeffreys priors

The problem with the ϕ-example above is that the representation of "ignorance" changes if we change parametrisation from θ to ϕ.

Suppose θ is a scalar.
A solution to the issue is the Jeffreys prior defined by $\pi(\theta) \propto I(\theta)^{1/2}$

square root of expected information

If $X_1 \dots X_n$ are from $f(x|\theta)$, this is $\pi(\theta) \propto i(\theta)^{1/2}$.

In what sense is Jeffreys prior a "solution"?

Suppose $\phi = h(\theta)$.

Consider:

(i) Find $\pi(\theta)$ using Jeffreys rule, then transform this pdf to a pdf $p(\phi)$ for ϕ.

(ii) Determine prior for ϕ using $p(\phi) \propto I(\phi)^{1/2}$.

Then (i) and (ii) give the <u>same</u> prior for ϕ.

<u>Example</u> Suppose $X_1 \ldots X_n \sim \text{Bernoulli}(\theta)$.

Then $i(\theta) = \dfrac{1}{\theta(1-\theta)}$.

So Jeffrey prior is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, $0<\theta<1$.

This is a $\text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$.

Jeffreys priors:

- can be improper

- can be defined for vector θ by

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

(determinant of I)$^{1/2}$

BUT a simpler approach is more common: find the Jeffreys prior for each 1-dim. component of θ and take the product to get the whole prior (i.e. assume prior independence).

4.4 Hypothesis testing and Bayes factors

Suppose we want to compare two hypotheses H_0 and H_1, exactly one of which is true.
The Bayesian approach attaches prior probabilities $P(H_0)$, $P(H_1)$ to H_0, H_1 (where $P(H_0) + P(H_1) = 1$).

The prior odds of H_0 relative to H_1 is

$$\text{prior odds} = \frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{1 - P(H_0)}.$$

[Odds of event $A = P(A) / (1 - P(A))$.]

We can compute posterior probabilities $P(H_i | \underset{\sim}{x})$, $i=0,1$ and compare them.

By Bayes theorem,

$$P(H_i | \underset{\sim}{x}) = \frac{P(\underset{\sim}{x} | H_i) P(H_i)}{P(\underset{\sim}{x} | H_0) P(H_0) + P(\underset{\sim}{x} | H_1) P(H_1)} \quad i=0,1 \qquad (1)$$

Note: $P(H_i | \underset{\sim}{x})$ is the probability of H_i conditioned on data $\underset{\sim}{x}$, whereas p-values <u>can't</u> be interpreted this way.

The <u>posterior odds</u> of H_0 relative to H_1 is

$$\text{posterior odds} = \frac{P(H_0 | \underset{\sim}{x})}{P(H_1 | \underset{\sim}{x})}.$$

Using ①,

$$\underbrace{\frac{P(H_0 \mid \underline{x})}{P(H_1 \mid \underline{x})}}_{\text{posterior odds}} = \underbrace{\frac{P(\underline{x} \mid H_0)}{P(\underline{x} \mid H_1)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{prior odds}}$$

posterior odds = Bayes factor × prior odds

where the <u>Bayes factor</u> of H_0 relative to H_1 is

$$B_{01} = \frac{P(\underline{x} \mid H_0)}{P(\underline{x} \mid H_1)} \qquad ②$$

The change from prior odds to posterior odds depends on $\underline{x}$ only via the Bayes factor B_{01}.

B_{01} tells us how $\underline{x}$ shifts our strength of belief in H_0 relative to H_1.

General setup

We are assuming we have

(i) prior probabilities $P(H_i)$, $i=0,1$, $\quad P(H_0)+P(H_1)=1$

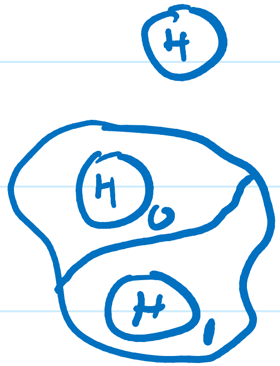
(ii) a prior distribution for θ_i under H_i,

i.e. $\pi(\theta_i \mid H_i)$ for $\theta_i \in \Theta_i$, $\quad i=0,1$.

(iii) a model under H_i for data $\underline{x}$ given by $f(\underline{x} \mid \theta_i, H_i)$

The two $\left\{ \begin{array}{l} \text{priors in (ii) could be of different forms} \\ \text{models in (iii) could be of different forms.} \end{array} \right.$

Sometimes (see example later) (i) and (ii) might be combined. The prior might be $\pi(\theta)$ for $\theta \in \Theta$ where



- $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \phi$

- prior probabilities are $P(H_i) = \int_{\theta \in \Theta_i} \pi(\theta)\, d\theta$

- and $\pi(\theta_i \mid H_i)$ is the conditional density of θ given H_i,

i.e. $$\pi(\theta_i \mid H_i) = \frac{\pi(\theta_i)}{\int_{\theta \in \Theta_i} \pi(\theta)\, d\theta}.$$

Consider ②: conditioning on θ_i (law of total prob) we have

$$P(\underset{\sim}{x} \mid H_i) = \int_{\Theta_i} f(\underset{\sim}{x} \mid \theta_i, H_i)\, \pi(\theta_i \mid H_i)\, d\theta_i \qquad ③.$$

$P(\underset{\sim}{x} \mid H_i)$ is called the marginal likelihood for H_i: it is the likelihood $f(\underset{\sim}{x} \mid \theta_i, H_i)$ averaged over Θ_i, weighted according to the prior $\pi(\theta_i \mid H_i)$.

So here we average over θ. The Bayes factor is the ratio of two such averages: $B_{01} = \dfrac{P(\underset{\sim}{x} \mid H_0)}{P(\underset{\sim}{x} \mid H_1)}$.

This is somewhat similar to the likelihood ratio of Sec. 3, except for LR we maximised over H_0, H_1 to find LR statistic Λ.

Note: 1. We are treating H_0, H_1 in the same way, whereas in Sec 3 we treated H_0, H_1 asymmetrically.

2. Bayes factor of H_1 relative to H_0 is just $B_{10} = B_{01}^{-1}$.

3. Bayes factors can only be used with proper priors: from ②, ③ B_{01} depends on two constants of proportionality (one for each $\pi(\theta_i | H_i)$) so these constants must be known.

Assume our model is $f(\underline{x} \mid \theta)$.

If $H_i: \theta = \theta_i$, $i = 0, 1$, are both simple, then

$$B_{01} = \frac{f(\underline{x} \mid \theta_0)}{f(\underline{x} \mid \theta_1)}$$

← Lik ratio

If $H_i: \theta \in \Theta_i$, $i = 0, 1$, are both composite, then

$$B_{01} = \frac{\int_{\Theta_0} f(\underline{x} \mid \theta) \pi(\theta \mid H_0) \, d\theta}{\int_{\Theta_1} f(\underline{x} \mid \theta) \pi(\theta \mid H_1) \, d\theta}.$$

Interpretation of Bayes factor:

B_{01}	Evidence for H_0
<1	negative (i.e. evidence supports H_1)
$1-3$	hardly worth a mention
$3-20$	positive
$20-150$	strong
>150	very strong

Example ("IQ") Suppose $X \sim N(\theta, \sigma^2)$ where $\sigma^2 = 100$.

So $f(x|\theta) = \dfrac{1}{\sqrt{200\pi}} e^{-\frac{1}{200}(x-\theta)^2}$.

Let $H_0: \theta = 100$, $H_1: \theta = 130$.

Suppose we observe $x = 120$.

Then $B_{01} = \dfrac{f(120|100)}{f(120|130)} = 0.223$.

$B_{10} = 1/0.223 = 4.48$, so positive evidence for H_1

Let prior probabilities be $P(H_0) = 0.95$, $P(H_1) = 0.05$.

Using post. odds = Bayes factor × prior odds,

$$\frac{p_0}{1-p_0} = B_{01} \times \frac{0.95}{0.05} \qquad \text{where } p_0 = P(H_0 | \underline{x})$$

Solving, $p_0 = \dfrac{19B_{01}}{1+19B_{01}} = 0.81$, so still a high

posterior probability of H_0.

<u>Example</u> ("Weight") $X_1, \dots, X_n \mid \theta \sim N(\theta, \sigma^2)$, $\sigma^2 = 3^2$

Let $H_0 : \theta \leq 175$, $H_1 : \theta > 175$

Prior: $\theta \sim N(\mu_0, \sigma_0^2)$, $\mu_0 = 170$, $\sigma_0^2 = 5^2$.

Prior prob: $P(H_0) = P\left(N(\mu_0, \sigma_0^2) \leq 175\right) = \Phi\left(\frac{175 - \mu_0}{\sigma_0}\right) = 0.84$

Prior odds: $\frac{P(H_0)}{P(H_1)} = \frac{0.84}{0.16} = 5.3$.

Observe $x_1, \dots, x_n$, $n = 10$, $\bar{x} = 176$.

Posterior $N(\mu_1, \sigma_1^2)$, $\mu_1 = \dots = 175.8$, $\sigma_1^2 = \dots = 0.869$.

Posterior prob: $P(H_0 \mid \underline{x}) = \Phi\left(\dfrac{175 - 175.8}{\sqrt{0.869}}\right) = 0.198.$

Post odds $= \dfrac{0.198}{0.802} = 0.24$

So Bayes factor $B_{01} = \dfrac{\text{post. odds}}{\text{prior odds}} = 0.0465.$

and $B_{10} = B_{01}^{-1} = 21.5$

Data provide strong evidence in favor of H_1

Example

[Example from Carlin and Louis (2008).]

Product P_0 – old, standard.

Product P_1 – newer, more expensive.

Assumptions:

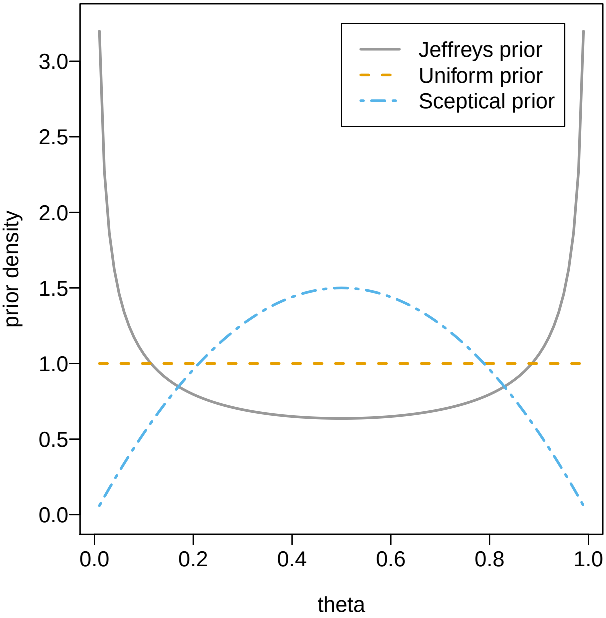
- the probability θ that a customer prefers P_1 has prior $\pi(\theta)$ which is Beta(a, b)
- the number of customers X (out of n) that prefer P_1 is $X \sim \text{Binomial}(n, \theta)$.

Let's say $\theta \geqslant 0.6$ means that P_1 is a substantial improvement over P_0. So take

$$H_0 : \theta \geqslant 0.6 \quad \text{and} \quad H_1 : \theta < 0.6.$$

We consider 3 possibile priors:

- Jeffreys' prior: $\theta \sim \text{Beta}(0.5, 0.5)$.
- Uniform prior: $\theta \sim \text{Beta}(1, 1)$.
- Sceptical prior: $\theta \sim \text{Beta}(2, 2)$, i.e. favours values of θ near $\frac{1}{2}$.


Jeffreys prior
Uniform prior
Sceptical prior
prior density
theta
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0

Prior odds $= P(H_0)/P(H_1)$ where

$$P(H_0) = \int_{0.6}^{1} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta$$

$$P(H_1) = \int_{0}^{0.6} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta.$$

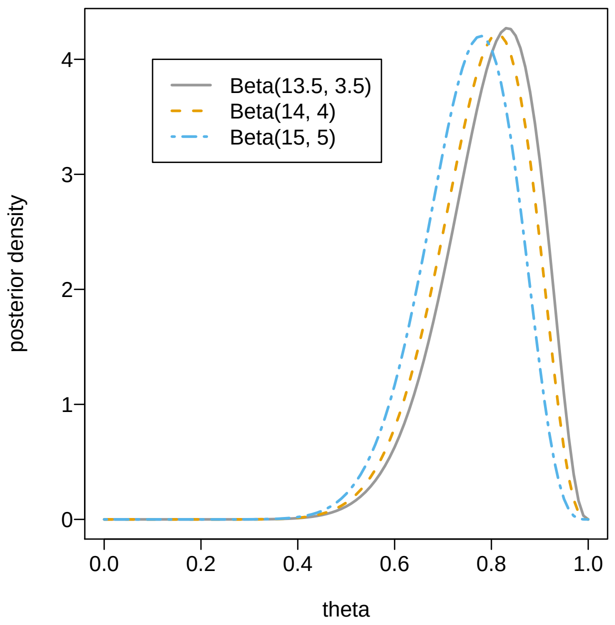
Suppose we have $x = 13$ "successes" from $n = 16$ customers.

Then (Section 4.1) the posterior $\pi(\theta \mid x)$ is $\text{Beta}(x + a, n - x + b)$ with $x = 13$ and $n = 16$.

Posterior odds $= P(H_0 \mid x)/P(H_1 \mid x)$ where

$$P(H_0 \mid x) = \int_{0.6}^{1} \frac{1}{B(x+a, n-x+b)} \theta^{x+a-1}(1-\theta)^{n-x+b-1} \, d\theta$$

$$P(H_1 \mid x) = \int_{0}^{0.6} \frac{1}{B(x+a, n-x+b)} \theta^{x+a-1}(1-\theta)^{n-x+b-1} \, d\theta.$$


Beta(13.5, 3.5)
Beta(14, 4)
Beta(15, 5)
posterior density
theta
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

Prior	Prior odds	Posterior odds	Bayes factor
$\text{Beta}(0.5, 0.5)$	0.773	26.6	34.4
$\text{Beta}(1, 1)$	0.667	20.5	30.8
$\text{Beta}(2, 2)$	0.543	13.4	24.6

Conclusion: strong evidence for H_0.

4.5 Asymptotic normality of posterior distribution

We have $\pi(\theta|\underline{x}) \propto L(\theta)\,\pi(\theta)$

Let $\tilde{l}(\theta) = \log \pi(\theta|\underline{x})$

$$= \text{constant} + \underbrace{l(\theta)}_{\sum_{i=1}^{n} \log f(x_i|\theta),\ n \text{ terms}} + \underbrace{\log \pi(\theta)}_{\text{one term}}$$

expect likelihood contribution to dominate for large n

Let $\tilde{\theta}$ be the posterior mode, assume $\tilde{l}'(\tilde{\theta}) = 0$.

Then

$$\tilde{l}(\theta) \approx \tilde{l}(\tilde{\theta}) + (\tilde{\theta} - \theta)\underbrace{\tilde{l}'(\tilde{\theta})}_{=0} + \frac{1}{2}(\theta - \tilde{\theta})^2 \tilde{l}''(\tilde{\theta})$$

$$= \tilde{l}(\tilde{\theta}) - \frac{1}{2}(\theta - \tilde{\theta})^2 \tilde{J}(\tilde{\theta})$$

where $\tilde{J}(\theta) = -\tilde{l}''(\theta)$.

So $\pi(\theta | \underset{\sim}{x}) = \exp(\tilde{l}(\theta)) \propto \exp\left(-\frac{1}{2}(\theta - \tilde{\theta})^2 \tilde{J}(\tilde{\theta})\right)$

i.e. $\theta \,\big|\, \underset{\sim}{x} \approx N\left(\tilde{\theta}, \tilde{J}(\tilde{\theta})^{-1}\right)$

$$\theta \mid \underset{\sim}{x} \approx N(\tilde{\theta}, \tilde{J}(\tilde{\theta})^{-1}) \qquad ①$$

In large samples the likelihood contribution will dominate, resulting in $\tilde{\theta}$ and $\tilde{J}(\tilde{\theta})$ being close to the MLE $\hat{\theta}$ and observed information $J(\hat{\theta})$. Hence

$$\theta \mid \underset{\sim}{x} \approx N(\hat{\theta}, J(\hat{\theta})). \qquad ②$$

①, ② look similar to the corresponding frequentist results, but note:

in ①, ②, θ is a random variable and $\tilde{\theta}(\underset{\sim}{x})$, $\hat{\theta}(\underset{\sim}{x})$ constants whereas in frequentist $\hat{\theta}(\underset{\sim}{X})$ is a random variable and θ constant.

Using the asymptotic results:

(i) frequentist $\hat{\theta} \approx N(\theta, J(\theta)^{-1})$ leads to 95% confidence interval of $\left(\hat{\theta} \pm 1.96\, J(\hat{\theta})^{-1/2}\right)$

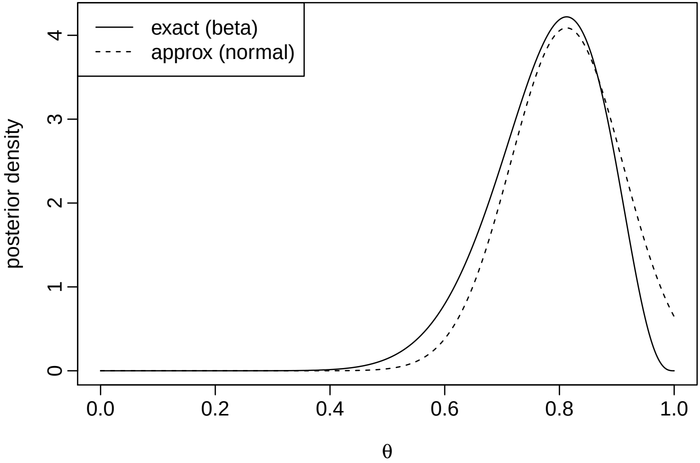
(ii) Bayesian (2) $\theta \mid \underset{\sim}{x} \approx N(\hat{\theta}, J(\hat{\theta})^{-1})$ leads to 95% credible interval of $\left(\hat{\theta} \pm 1.96\, J(\hat{\theta})^{-1/2}\right)$.

That is, the <u>same</u> interval of θ-values in both cases, but with <u>different</u> interpretations.

Normal approx to posterior (1)

Prior $\theta \sim U(0,1)$.

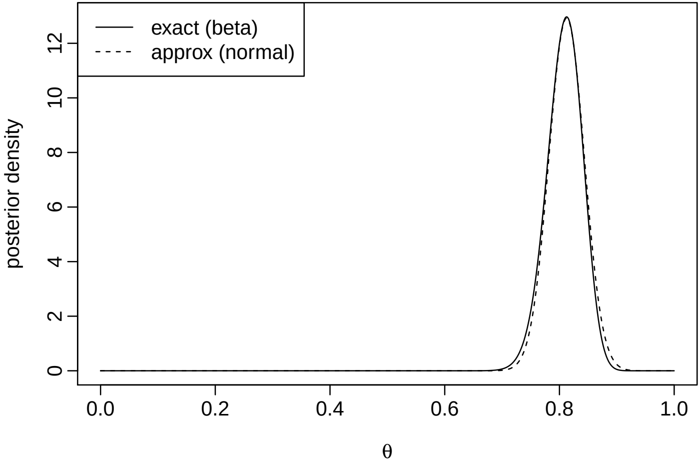
Bernoulli likelihood: $x = 13$ successes out of $n = 16$ trials.



Normal approx to posterior (2)

Prior $\theta \sim U(0, 1)$.

Bernoulli likelihood: $x = 130$ successes out of $n = 160$ trials.



Part B courses

double unit, practicals, R

SB1 : applied, computational, regression models

SB2.1 : statistical inference, frequentist and Bayesian

SB2.2 : machine learning

SB3.1 : applied probability

SB3.2 : lifetime models