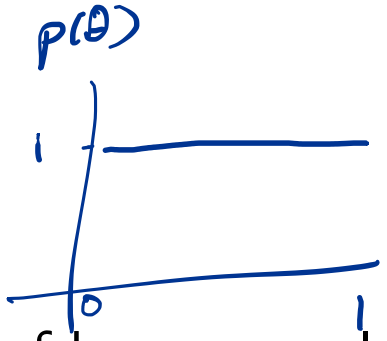
STAT3016/4116/7016:
Introduction to Bayesian Data Analysis

RSFAS, College of Business and Economics, ANU

One parameter models

The binomial model



"flat prior"

Happiness Data

A survey was conducted to assess the proportion of happy people in the female 65+ population. The population size is N. The sample size was n=129 females aged 65+. Each person was asked if they were generally happy or not. Let $Y_i = 1$ if respondent i reported being generally happy, and let $Y_i = 0$ otherwise. Let θ be the true proportion that are happy. That is $\theta = \sum_{i=1}^{N} Y_i / N$.

$\theta \sim Beta(1, 1)$

$\Rightarrow \theta \sim Unif(0, 1)$

$Y_i \sim Bern$

- Conditional on θ what would be an appropriate distribution of the binary random variables Y_i?
- Suppose the observed outcomes are $\{y_1, ..., y_{129}\}$, what is the likelihood function?

The binomial model - prior distribution

What are our prior beliefs on the value of θ?

Any restrictions on the values that the parameter θ can take?

What does the prior $p(\theta) = 1$ for all $\theta \in [0, 1]$ imply about our prior beliefs?

Observed data: - 129 individuals surveyed; 118 individuals report being generally happy; 11 individuals do not report being generally happy

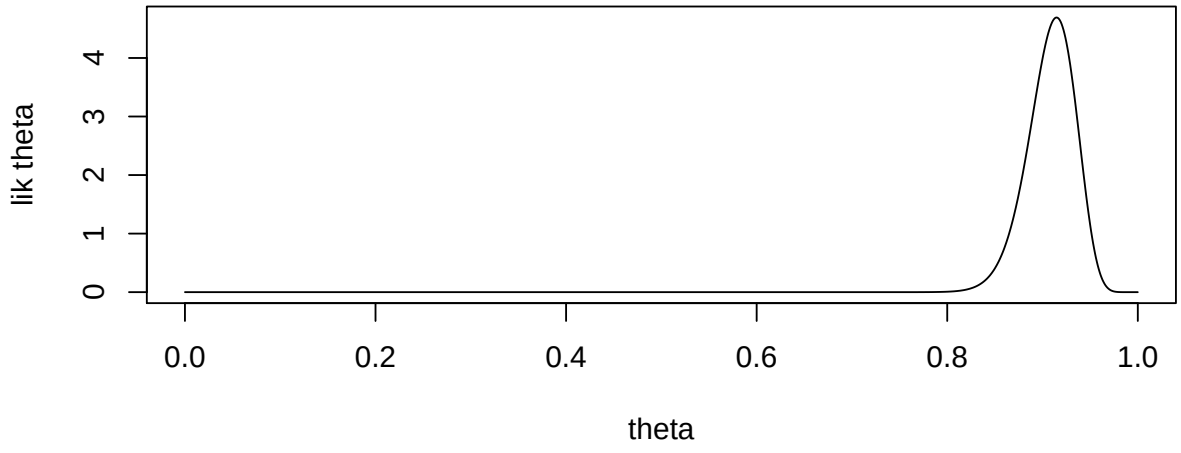
$\theta | y \sim \text{Beta}(118+1, 11+1) = \text{Beta}(119, 12)$

If $p(\theta) = 1$ what is the posterior distribution of θ?

Under the uniform prior, compute the relative posterior probabilities of two values of θ, say θ_1 and θ_2.

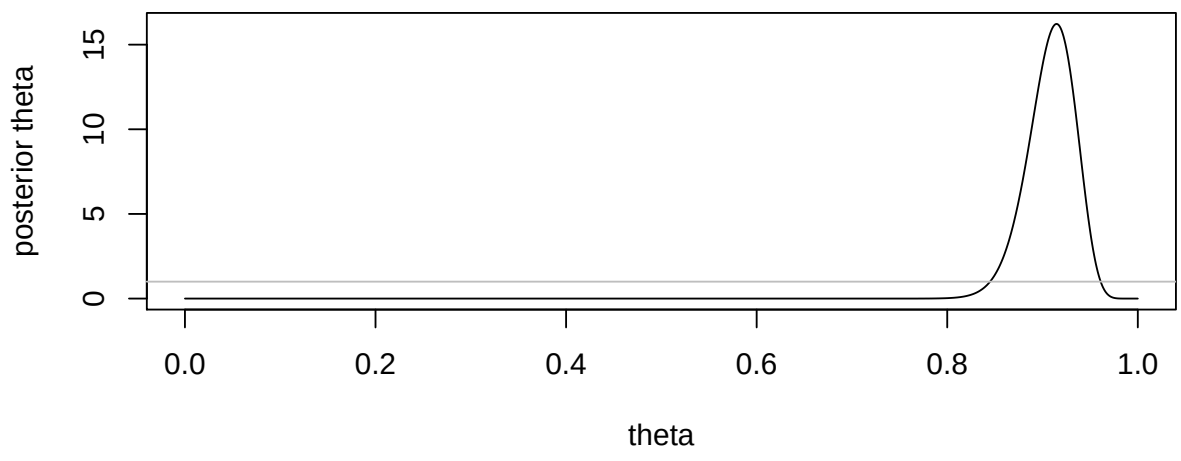
$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \left(\frac{\theta_1}{\theta_2}\right)^{119-1} \left(\frac{1-\theta_1}{1-\theta_2}\right)^{12-1}$$

The binomial model - posterior distribution



same shape but different scale.

assume a flat prior ⇒ posterior is dominated by likelihood.



The beta distribution

Suppose $\theta \sim \mathrm{Beta}(\alpha, \beta)$. The pdf of θ is

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \text{ for } 0 \leq \theta \leq 1$$

where $\Gamma(x)$ is the Gamma function which is the integral $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$

$E[\theta] = \frac{\alpha}{\alpha+\beta}$

$Var[\theta] = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

$\mathrm{Mode}[\theta] = \frac{\alpha-1}{\alpha+\beta-2}$

Binomial model - conjugate prior

For the binomial model, we know the likelihood as a function of θ is of the form

$$p(y|\theta) \propto \theta^y (1-\theta)^{n-y}$$

If the prior density is also of this form, then the posterior density will also be of this form. Let's parametize the prior density as

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

We recognise this distribution as Beta(α, β). Comparing $p(\theta)$ and $p(y|\theta)$ suggests that the prior density is equivalent to α prior successes and β prior failures. (nb: we refer to α and β as *hyperparameters*). actually we can assign probabilities to α & β.

Binomial model - conjugate prior

So the posterior density for θ is

$$\begin{aligned} p(\theta|y) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{y}(1-\theta)^{n-y} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &\propto \text{Beta}(\alpha+y, \beta+n-y) \end{aligned}$$

The prior and the posterior follow the same parametric form - this is known as CONJUGACY.

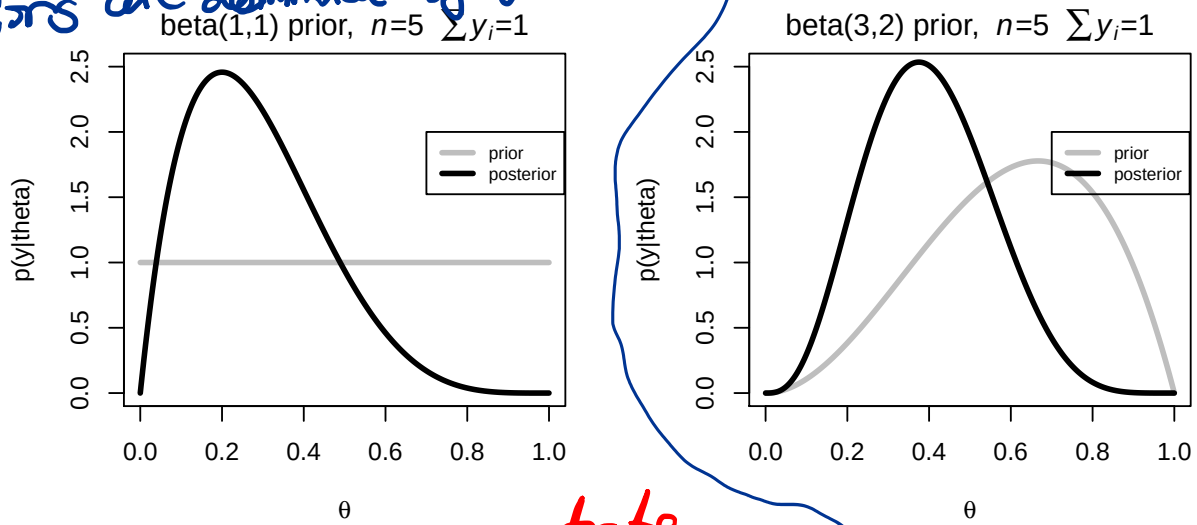
The beta prior is a conjugate family for the binomial likelihood.

Definition: If $\mathscr{F}$ is a class of sampling distributions $p(y|\theta)$, and $\mathscr{P}$ is a class of prior distributions for θ, then the class $\mathscr{P}$ is conjugate for $\mathscr{F}$ if

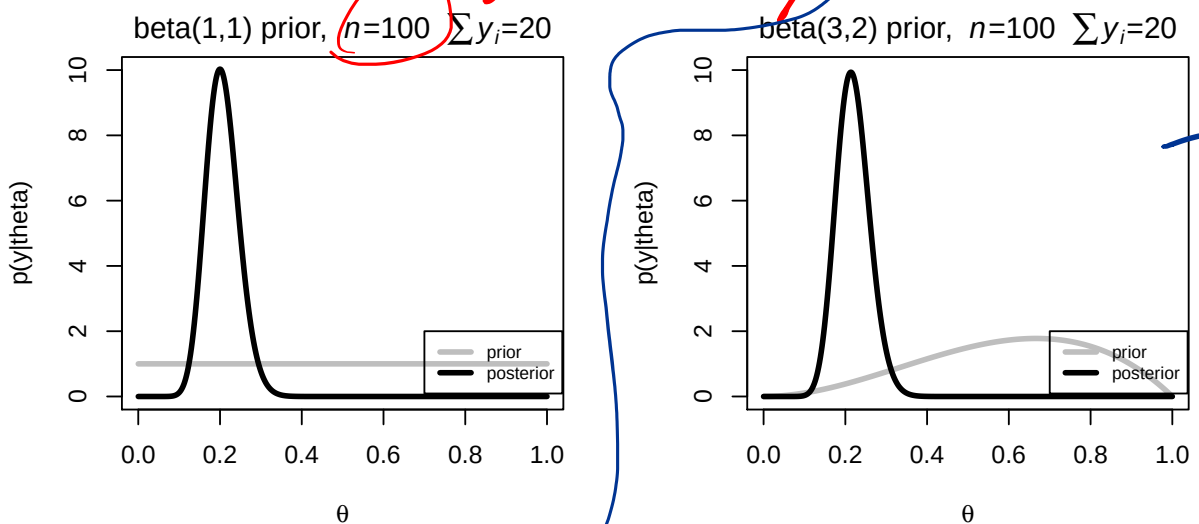
$$p(\theta|y) \in \mathscr{P} \text{ for all } p(y|\theta) \in \mathscr{F} \text{ and } p(\theta) \in \mathscr{P}$$

Binomial model - conjugate prior

LHS 2 plots (with flat priors) posteriors are dominated by likelihoods



more concentrate
less unstability



even though we have prior (but) the posterior is so dominated by likelihood b/c ~~sample size~~ population size is large.

Beta conjugate prior - binomial model

sample proportion

posterior mean

$$\frac{y}{n} < E[\theta|y] = \frac{\alpha + y}{\alpha + \beta + n} < \frac{\alpha}{\alpha+\beta}$$

which always lies between the sample proportion y/n and the prior mean $\alpha/(\alpha + \beta)$.

$$Var(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

What happens if $n >> \alpha + \beta$?

$E(\theta|y) \to \frac{y}{n}$

$Var(\theta|y) \to 0$

Beta conjugate prior - binomial model

In fact we can show using the Central Limit Theorem

$$\left(\frac{\theta - E[\theta|y]}{\sqrt{var(\theta|y)}} \Big| y \right) \rightarrow N(0,1)$$

We often use the above result to justify approximating the posterior distribution with a normal distribution.

For the binomial parameter, the normal approximation is more accurate if we transform θ to the logit scale first.

both in $(0, \infty)$

Binomial model - prediction

Suppose after collection of the Happiness data (that is n=129 females are surveyed and the data are analysed), it was discovered that data was not collected from one female. That is, the sample size should have been 130. What is the probability that this unobserved female is happy??

Binomial model - prediction

Let $\tilde{Y}_{130}$ be the unobserved outcome. We would like to evaluate the posterior predictive distribution
$P(\tilde{Y}_{130}|Y_1 = y_1, ..., Y_{129} = y_{129})$.

$$
\begin{aligned}
Pr(\tilde{Y}_{130} = 1|y_1, ..., y_n) &= \int Pr(\tilde{Y}_{130} = 1, \theta|y_1, ..., y_n)d\theta \\
&= \int Pr(\tilde{Y}_{130} = 1|\theta, y_1, ..., y_n)p(\theta|y_1, ..., y_n)d\theta \\
&= \int \theta p(\theta|y_1, ..., y_n)d\theta \\
&= E[\theta|y_1, ..., y_n]
\end{aligned}
$$

⇒ which is just the posterior mean.

Notes:

- The predictive distribution does not depend on any unknown quantities
- The predictive distribution depends on our observed data (so $\tilde{Y}$ is not independent of $Y_1, ..., Y_n$ - recall discussion on exchangeability). If they were independent, we could never infer anything about unsampled cases from sampled cases.

Binomial model - prediction

$Pr(\tilde{Y}=1 \mid y_1,\dots,y_n) = E(\theta \mid y_1,\dots,y_n) = \frac{1}{(n+1)+1} = \frac{1}{n+2}$

$\theta \mid y \sim Beta(1, 1+n)$

$(\sum y_i = 0)$

$Mode(\theta \mid y) = 0$

Exercise 1: Conditional on θ, suppose Y_i (i=1,...,n) are i.i.d binary random variables with expectation θ. Assuming an uninformative prior on θ, find $Pr(\tilde{Y} = 1 | Y = y)$ (where $Y = \sum_{i=1}^{n} Y_i$). Also find mode$(\theta | Y = y)$. Discuss the two posterior summaries. In particular, consider the case where Y=0.

Confidence regions

Bayesian coverage

$$Pr(\mathrm{l}(y) < \theta < u(y)|Y = y) = 0.95$$

Frequentist coverage

$$Pr(\mathrm{l}(Y) < \theta < u(Y)|\theta) = 0.95$$

Contrast the interpretation of the above two confidence intervals.

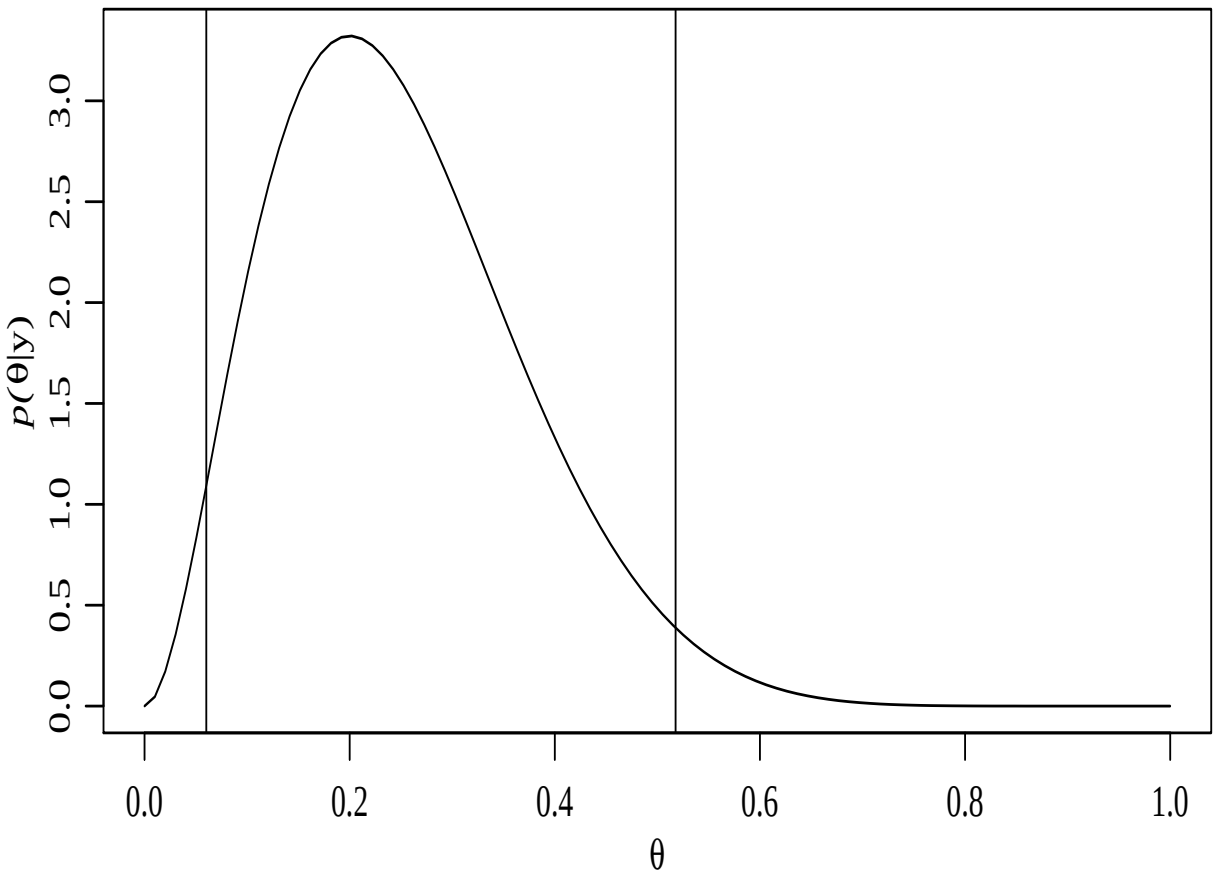
Confidence regions

Quantile-based interval

To make a $100 \times (1-\alpha)\%$ quantile-based confidence interval, find numbers $\theta_{\alpha/2}$, $\theta_{1-\alpha/2}$ such that

1. $Pr(\theta < \theta_{\alpha/2} | Y = y) = \alpha/2$
2. $Pr(\theta > \theta_{1-\alpha/2} | Y = y) = \alpha/2$

qbeta(c(0.025, 0.975), a+y, b+n-y)



Confidence regions

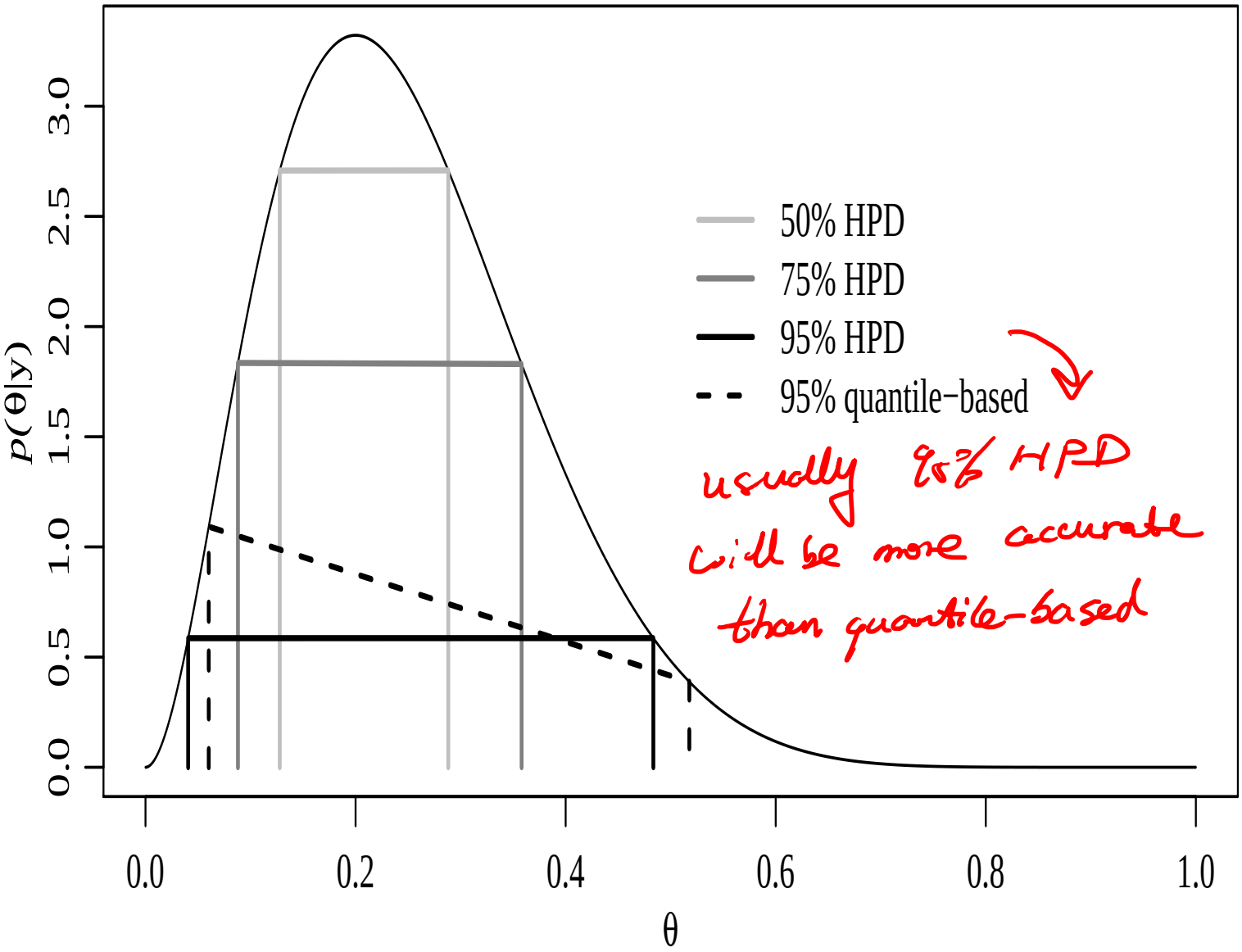
Highest posterior density (HPD) region

A $100 \times (1-\alpha)\%$ HPD region consists of the parameter space, $s(y) \subset \Theta$ such that

1. $Pr(\theta \in s(y)|Y = y) = 1 - \alpha$
2. If $\theta_\alpha \in s(y)$ and $\theta_b \notin s(y)$ then $p(\theta_a|Y = y) > p(\theta_b|Y = y)$

Confidence regions

Highest posterior density (HPD) region



HPD interval estimation - R code

```
> hpd<-function(x,dx,p){
+ md<-x[dx==max(dx)]
+ px<-dx/sum(dx)
+ pxs<--sort(-px)
+ ct<-min(pxs[cumsum(pxs)< p])
+ list(hpdr=range(x[px>=ct]),mode=md) }
```

Binomial model - Interval estimation

$\hat{\beta}=0.2$ $\quad n=10$ $\quad \hat{\beta} \pm 1.96\sqrt{\frac{\hat{\beta}(1-\hat{\beta})}{n}} \Rightarrow (-0.048, 0.45)$

(Frequentist)

Exercise 2: Suppose out of n=10 conditionally independent draws of a binary random variable we observe Y=2 ones. Using a uniform prior distribution for θ, find a 95% posterior confidence interval for θ. Also find a 95% frequentist confidence interval for θ. Discuss the similarities/differences in the values and interpretation of your two interval estimates.

$\theta|y \sim Beta(1+2, 1+8) \sim Beta(3, 9)$

qbeta(c(0.025, 0.975), a+y, b+n-y)

0.06021773, 0.51775585

$a=1$ $\quad b=1$ $\quad n=10$ $\quad y=2$

$a+y=3$ $\quad b+n-y=9$

Poisson model

The Poisson distribution arises naturally in the study of data taking the form of counts or rare events; eg the number of airplane accidents or the incidence of diseases.

If a random variable Y follows the Poisson distribution with rate θ, the probability distribution is

$$Pr(Y = y|\theta) = \theta^y e^{-\theta}/y! \text{ for } y \in \{0, 1, 2, ...\}$$

For a vector $y_1, ..., y_n$ of independent and identically distributed observations conditional on θ, the likelihood is $\prod_{i=1}^{n} p(y_i|\theta)$.

$p(\theta|y_1, \cdots, y_n) \propto p(\theta) \cdot p(y|\theta)$

$= p(\theta) \cdot \prod_{i=1}^{n} \theta^{y_i} e^{-\theta}$

$= p(\theta)\, \theta^{\sum y_i} e^{-n\theta}$

$= \theta^{a-1} e^{-b\theta} \theta^{\sum y_i} e^{-n\theta}$

$= \theta^{a+\sum y_i - 1} e^{-(b+n)\theta}$

$\sim \text{Gamma}(a + \sum y_i, b+n)$

key is to find $p(\theta) \propto \theta^{c_1} e^{-c_2\theta}$

$\sim \text{Gamma}(c_1, c_2)$

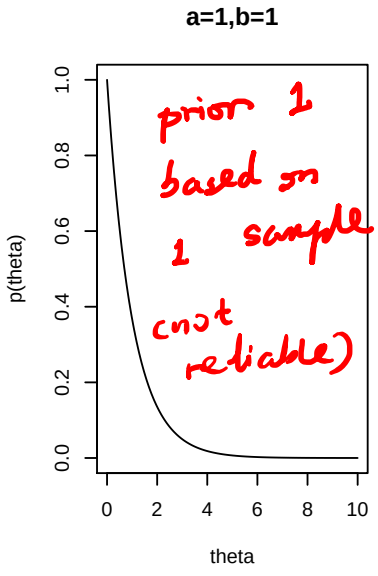
$(c_1 = a,\ c_2 = b)$

Poisson model - conjugate prior

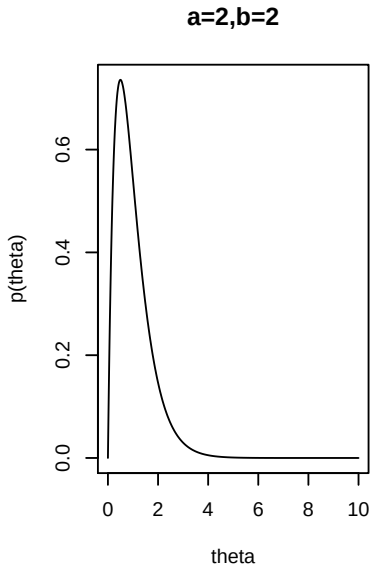
$\text{Gamma}(a+\sum y_i, b+n)$

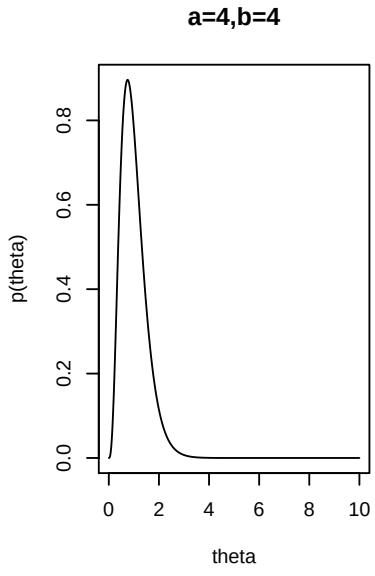
- What form will the conjugate prior take? Do you recognise the corresponding probability distribution? Using you conjugate prior, what is the posterior density of θ given $y_1, ..., y_n$?
- Express the posterior expectation of θ as a combination of the prior expectation and the sample average. $E(\theta|y) = \frac{a+\sum y_i}{b+n} = \frac{b}{b+n} \times \frac{a}{b} + \frac{n}{b+n} \times \frac{\sum y_i}{n}$
- Provide an interpretation of the parameters of the prior distribution.
- What happens to your posterior mean and variance when n is really large?? (holding the prior distribution parameters fixed)
- Modify your model if $Y_i|\theta, x_i \sim Pois(x_i\theta)$ where x_i is an exposure variable (eg x_i is a time or size variable). (see Tutorial)

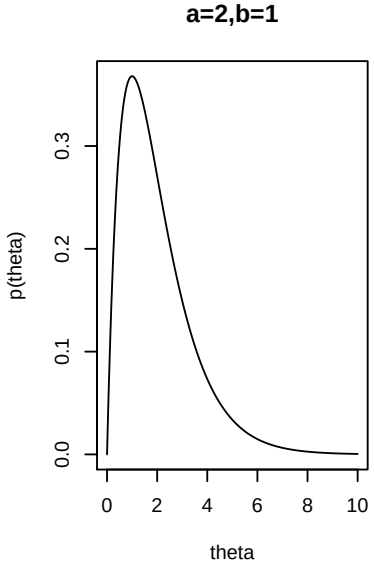
Poisson model - conjugate prior

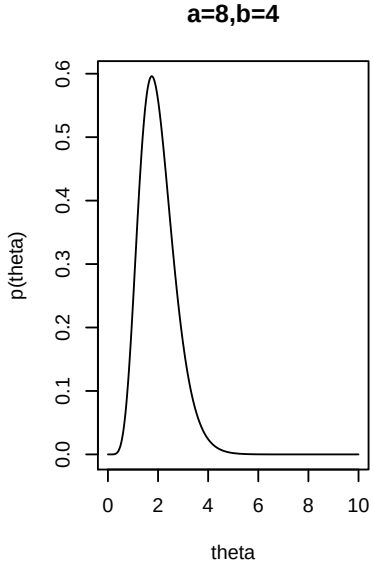


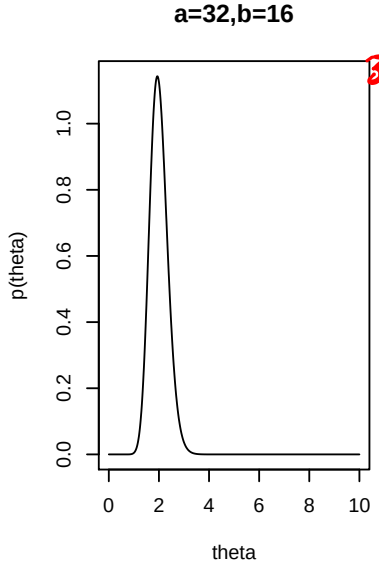
prior 1 based on 1 sample (not reliable)











16 sample size
32 observations ...

Poisson model

Exercise 3: Derive the posterior predictive distribution $p(\tilde{y}|y_1, ..., y_n)$ for the Poisson model and assuming a conjugate prior with parameters α and β (see Tutorial).

Exercise 4: Derive the prior predictive distribution $p(y)$ (assuming a single observation y) for the Poisson model and assuming a conjugate prior with parameters α and β.

negative binomial

$y \sim NB(a, \frac{b}{b+1})$

$a :=$ # of successes

$\frac{b}{b+1} :=$ prob of success

$y :=$ r.v. of number of failures

Poisson model - birth rates

Example: A study was conducted in the US in the 1970s to compare women with college degrees to those without in terms of their numbers of children. Let $Y_{1,1}, ..., Y_{n_1,1}$ denote the numbers of children for the n_1 women without college degrees and $Y_{1,2}, ..., Y_{n_2,2}$ be the data for women with degrees. We will use the following sampling models:

$$Y_{1,1}, ..., Y_{n_1,1}|\theta_1 \overset{\text{iid}}{\sim} Pois(\theta_1)$$

$$Y_{1,2}, ..., Y_{n_2,2}|\theta_2 \overset{\text{iid}}{\sim} Pois(\theta_2)$$

The data are $n_1 = 111; \sum_{i=1}^{n_1} Y_{i,1} = 217, \bar{Y}_1 = 1.95$
$n_2 = 44; \sum_{i=1}^{n_2} Y_{i,2} = 66, \bar{Y}_2 = 1.5$

Assume the following prior distribution
$(\theta_1, \theta_2) \overset{\text{iid}}{\sim} Gamma(a = 2, b = 1)$.
(Interpretation of prior parameters?)

Poisson model - birth rates

Gamma(1, 1)

use $a=1$, $b=1$ prior guess

Exercise 5a: Find posterior means, modes and 95% quantile-based confidence intervals for θ_1 and θ_2.

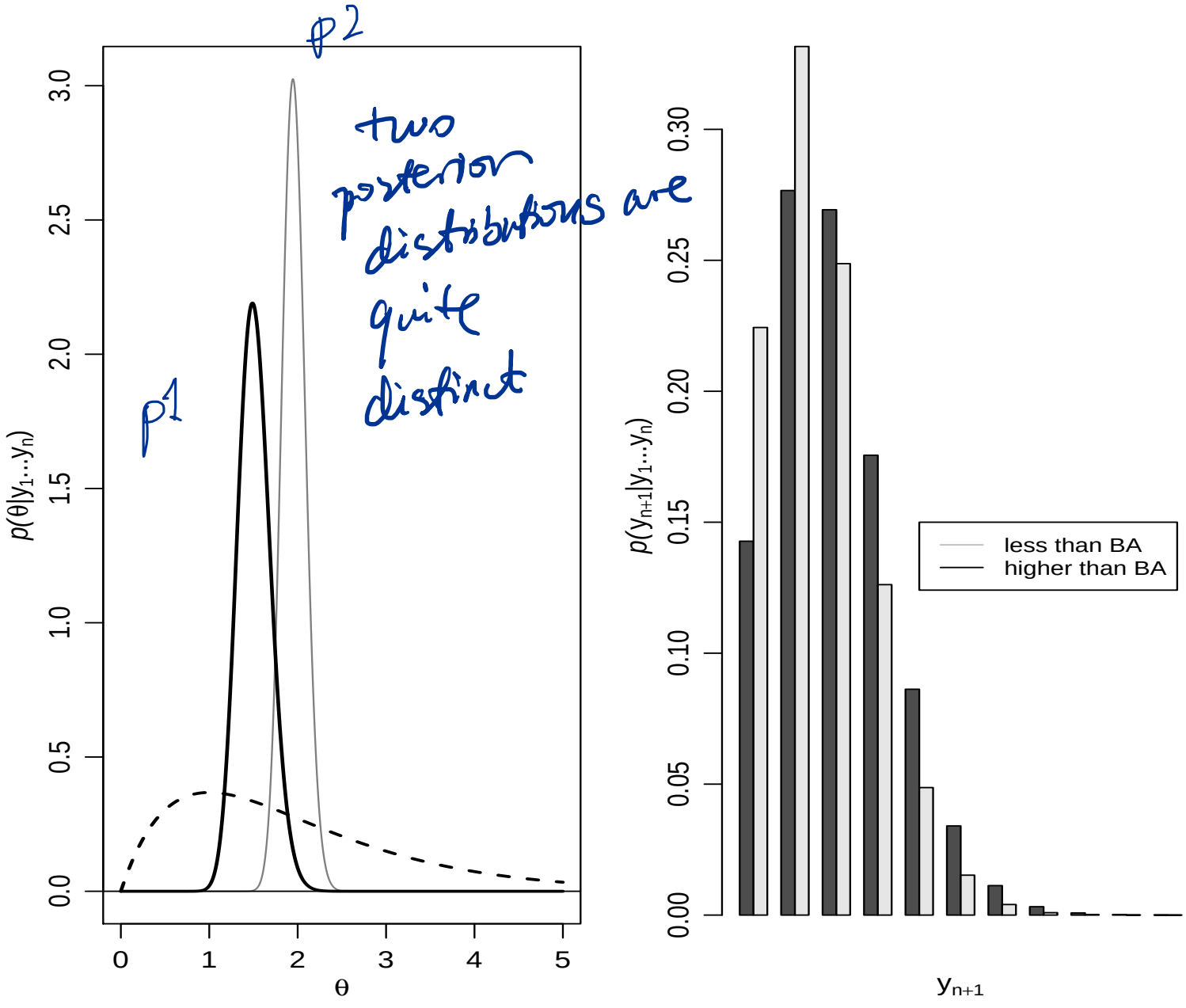
Exercise 5b: Estimate
$Pr(\theta_1 > \theta_2 | \sum_{i=1}^{n_1} Y_{i,1} = 217, \sum_{i=1}^{n_2} Y_{i,2} = 66)$ and compare to the posterior predictive probability that $\tilde{Y}_1 > \tilde{Y}_2$

$$\int_0^\infty \int_0^{\theta_1} p(\theta_1, \theta_2 | y_1, y_2) d\theta_1 d\theta_2$$

use simulation instead

a lot less than 1 b/c more unpredictability

Poisson model - birth rates



Noninformative prior distributions

A prior distribution is non-informative if the prior is "flat" relative to the likelihood function.

Example: If $0 \leq \theta \leq 1$, then $\theta \sim U(0,1)$ is a non-informative prior for θ.

Rationale: let the data speak for themselves

Exercise 6: If p(logit(θ)) $\propto$ constant; then what is the form of p(θ)? Is this distribution proper?

$\text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$

$p\left(\log\left(\frac{\theta}{1-\theta}\right)\right) \propto K$

$p(\theta) = ?$

$x = \log\left(\frac{\theta}{1-\theta}\right)$, $\frac{dx}{d\theta} = \frac{1}{\theta(1-\theta)}$

$p(\theta) = 1 \times \frac{dx}{d\theta} = \theta^{-1}(1-\theta)^{-1} \Rightarrow$ improper

$\text{Beta}(0,0)$

$\int_0^1 \theta^{-1}(\theta-1)^{-1} \neq 1$

Jeffreys' invariance principle

Jeffreys' rule provides a rule for the choice of a non-informative prior, based on considering one-to-one transformations of the parameter $\phi = h(\theta)$.

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)|h'(\theta)|^{-1}$$

Jeffreys' invariance principle says that any rule for determining $p(\theta)$ should yield an equivalent result if applied to the transformed parameter ϕ. That is, if we apply the rule to $p(\theta)$ and then calculate $p(\phi)$ where $\phi = h(\theta)$, we should get the same prior as determining $p(\phi)$ directly.

Jeffreys' invariance principle

Jeffreys' principle leads to the noninformative prior density $p(\theta) \propto [J(\theta)]^{1/2}$ where $J(\theta)$ is the Fisher information for θ.

$$J(\theta) = -E\left(\frac{d^2 \log p(y|\theta)}{d\theta^2}\Big|\theta\right)$$

see extra notes

Exericse 7: Show that Jeffreys' prior is invariant to parametization

Exericse 8: What is Jeffrey's prior for the binomial model $y \sim \text{Bin}(n, \theta)$.

$p(y|\theta) = \theta^y (1-\theta)^{1-y} = \left(\frac{\theta}{1-\theta}\right)^y \cdot (1-\theta) = e^{\phi y} (1+e^{\phi})^{-1}$

when $\frac{\theta}{1-\theta} = e^{\phi}$, $\phi = \log\left(\frac{\theta}{1-\theta}\right)$

$\theta \sim \text{Beta}(0.5, 0.5)$

$p(\theta) = [J(\theta)]^{\frac{1}{2}} =$

$J(\theta) = \frac{n}{\theta(1-\theta)}$

(see solution)

$[J(\theta)]^{\frac{1}{2}} \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}} = \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$

Weakly informative prior distributions

somewhere b/w uninformative & fully informative

say Beta(1,1)

say Beta(50,50)

- fully inf ⇒ a limited range of study.
- weakly inf ⇒ work in more data situation.

A prior distribution is weakly informative if it is proper but is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available.

Use of a weakly informative prior will ensure that the posterior distribution is proper.

For example, suppose we wish to estimate the probability of a female birth. Assuming a binomial model, what would be a weakly informative prior?

Beta dist. is conjugate prior of binomial.

⇒ Beta(1,1) ✗ not weakly informative ∵ it's flat

can use Beta(2,1) no RIGHT/WRONG ANSWERS

- There is NO FIXED weakly informative prior to a given distribution.

often used in classification problem.

Mixtures of conjugate priors

"prior weight"

when to use mixture?
- multiple subgroups exist in a population.

Exercise 9: Suppose we wish to estimate the probability that a property sells at auction on the day in Canberra. Let p represent the probability that a property is successfully sold at auction on the day. We believe that either p is in the neighbourhood of 0.75 or in the neighbourhood of 0.45 (depending on which agent you ask).

not exactly

2 groups

1. **Formulate a prior density to model this prior belief**

Suppose we have data on 168 properties that went to auction in the last month. Of these 168 properties, 91 were sold at auction on the day (the other properties were passed in).

don't know 92 belong to which groups.

2. **Given the auction data, what is the posterior density of the sale at auction probability.**

3. **Sketch the prior and posterior densities.**

4. **Test whether there is evidence to suggest that p is greater than 0.5 using (a) a frequentist approach (b) a Bayesian approach**

$0.5\,\text{Beta}(9,11) + 0.5\,\text{Beta}(15,5)$

$0.5=\gamma$, $0.5=1-\gamma$ prior mean $0.45=\frac{9}{9+11}$ prior $0.75=\frac{15}{15+5}$

$\gamma g_1(p) + (1-\gamma) g_2(p)$:= mixture conjugate prior.

$n = 168, \quad y = 91$

$$g(p|y) \propto \left(\gamma g_1(p) + (1-\gamma) g_2(p)\right) \cdot g(y|p)$$

$$= \gamma\, p^{9+91-1}(1-p)^{1+77-1} + (1-\gamma)\, p^{15+91-1}(1-p)^{5+77-1}$$

$$= \gamma' \underbrace{\text{Beta}(100, 88)}_{} + (1-\gamma')\underbrace{\text{Beta}(106, 82)}_{}$$

solve for γ' s.t. $\int g(p|y)\,dp = 1$

$$\Rightarrow \gamma' = \frac{B(100, 88)}{B(100, 88) + B(106, 82)}$$

$\gamma' = 0.24 \qquad 1-\gamma' = 0.76$

BAYESIAN'S APPROACH

$\Pr(p > 0.5 | y)$

$\int_{0.5}^{1} g(p|y)\,dp$

MC simulation

$H_0: p = 0.5, \quad H_A: p > 0.5$ FREQUENTISTS' APPROACH

$$Z = \frac{\frac{91}{168} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{168}}} = 1.08 \Rightarrow \text{p-value } (0.140)$$