Bayesian Inference

Chapter 2: Conjugate models

Conchi Ausín and Mike Wiper
Department of Statistics
Universidad Carlos III de Madrid

Master in Business Administration and Quantitative Methods
Master in Mathematical Engineering

Objective

In this class we study the situations when Bayesian statistics is easy!

Conjugate models

- Yesterday we looked at a coin tossing example.
- We found that a particular, beta prior distribution lead to a beta posterior.
- This is an example of a conjugate family of prior distributions.

Coin tossing problems

In coin tossing problems, the likelihood function has the form

$$f(\mathbf{x}|\theta) = c\theta^x(1-\theta)^{n-x}$$

where x is the number of observed heads, n is the number of observed tosses and c is a constant determined by the experimental design.

Therefore, it is clear that a beta prior

$$f(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

implies that the posterior is also beta:

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto \theta^{a+x-1}(1-\theta)^{b+n-x-1} \\ &= \frac{1}{B(a+x,b+n-x)}\theta^{a+x-1}(1-\theta)^{b+n-x-1} \\ \theta &\sim \text{Beta}(a+x,b+n-x) \end{aligned}$$

Advantages of conjugate priors I: simplicity of calculation

Using a beta prior in this context has a number of advantages.

Given that we know the properties of the beta distribution, prior to posterior inference is equivalent to changing of parameter values. Prediction is also straightforward in the same way.

If $\theta \sim \text{Beta}(a, b)$ and $X|\theta \sim \text{Binomial}(n, \theta)$, then

$$P(X = x) = \frac{\binom{n}{x} B(a + x, b + n - x)}{B(a, b)} \quad \text{for } x = 0, ..., n.$$

Advantages of conjugate priors II: interpretability

- We can see that a (b) in the prior plays the same role as x ($n - x$).
- Therefore we can think of the information represented by the prior as equivalent to the information in n tosses of the coin with x heads and $n - x$ tails.
- This gives one way of thinking about how to elicit sensible values for a and b.
- To how many tosses of a coin and how many heads does my prior information equate?
- A problem is that people are often overconfident.

Prior elicitation

The previous method is a little artificial.

If we are asking a real expert to provide information it is better to ask questions about observable quantities.

For example:

- What would be the average number of heads to occur in 100 tosses of the coin?
- What about the standard deviation?

Then assuming a beta prior, we can solve

$$\mu = 100\frac{a}{a+b} \quad \sigma = 100\frac{ab}{(a+b)(a+b+1)}$$

Many people don't understand means and standard deviations so it could be even better to ask about modes or medians or quartiles.

Haldane's prior

Recalling the role of a and b also gives a reasonable way of defining a default, non-informative prior by letting $a, b \to 0$.

In this case we have a prior distribution

$$f(\theta) \propto \frac{1}{\theta(1-\theta)} \quad \text{for } 0 < \theta < 1$$

and the posterior is $\theta|\mathbf{x} \sim \text{Beta}(x, n-x)$, with mean $E[\theta|\mathbf{x}] = \frac{x}{n} = \hat{\theta}$, the MLE.

- This prior is improper!
- Should we care?
- What if we only observe a sample of heads (tails)?
- Then the posterior would be improper too!

This is a big problem in modern Bayesian statistics.

Other ways of choosing a default "objective" prior

Given the Principle of Insufficient Reason we saw yesterday, a uniform prior seems a natural selection.

However, if we know nothing about θ, shouldn't we also know nothing about $\vartheta = \log \frac{\theta}{1-\theta}$ for example?

If $\theta \sim \text{Uniform}(0, 1)$, then the laws of probability imply that the density of ϑ is

$$f(\vartheta) = \frac{e^{\vartheta}}{(1 + e^{\vartheta})^2}$$

which is clearly not uniform.

Uniform priors are sensible as default options for discrete variables but here, it is not so clear.

Jeffreys prior

Let $X|\theta \sim f(\cdot|\theta)$. Then the Jeffreys prior is

$$f(\theta) \propto \sqrt{I(\theta)}$$

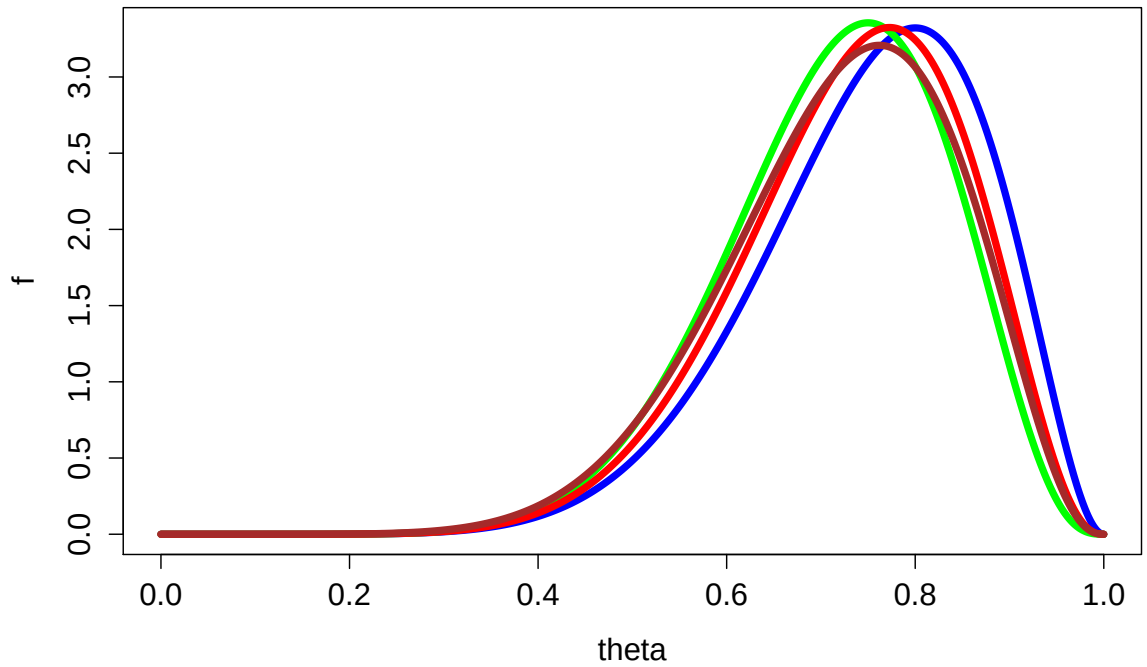
where $I(\theta) = -E_X\left[\frac{d^2}{d\theta^2} \log f(X|\theta)\right]$ is the expected Fisher information.

- Let $X|\theta \sim$ Binomial(n, θ). Then the Jeffreys prior is $\theta \sim$ Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$.
- Let $X|\theta \sim$ Negative Binomial(r, θ). The Jeffreys prior is $f(\theta) \propto \frac{1}{\theta(1-\theta)^{1/2}}$.
- The prior depends on the experimental design.
- This doesn't comply with the stopping rule principle!

There is no truly objective prior!

Example

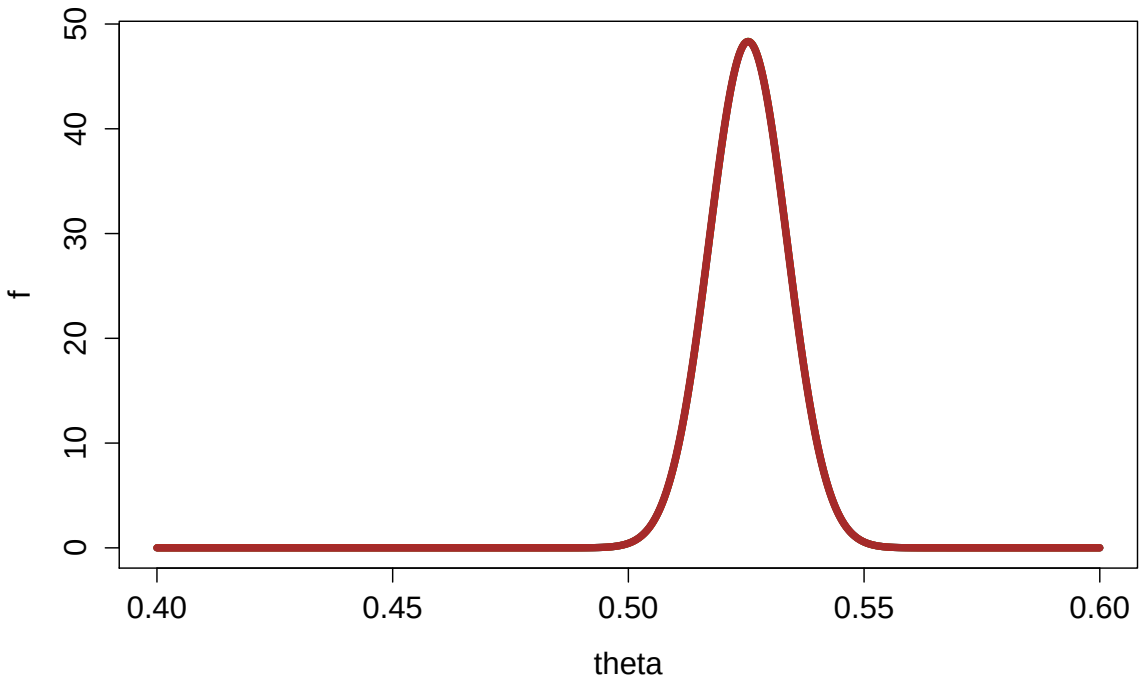
The following plot gives the posterior densities of θ for our coin tossing example given the Haldane (blue), uniform (green), Jeffreys I (red) and Jeffreys II (brown) priors.



Posterior means for θ are 0.75, 0.714, 0.731 and 0.72 respectively.

In small samples the prior can make a (small) difference ...

Example



... but in our Chinese babies example, it is impossible to differentiate between the posteriors and the posterior means are all equal to 0.5254 to 4 d.p.

Advantages of conjugate priors III: mixtures are still conjugate

- A single beta prior might not represent prior beliefs well.
- A mixture of k (sufficiently many) betas can.
- The posterior is still a mixture of k betas.

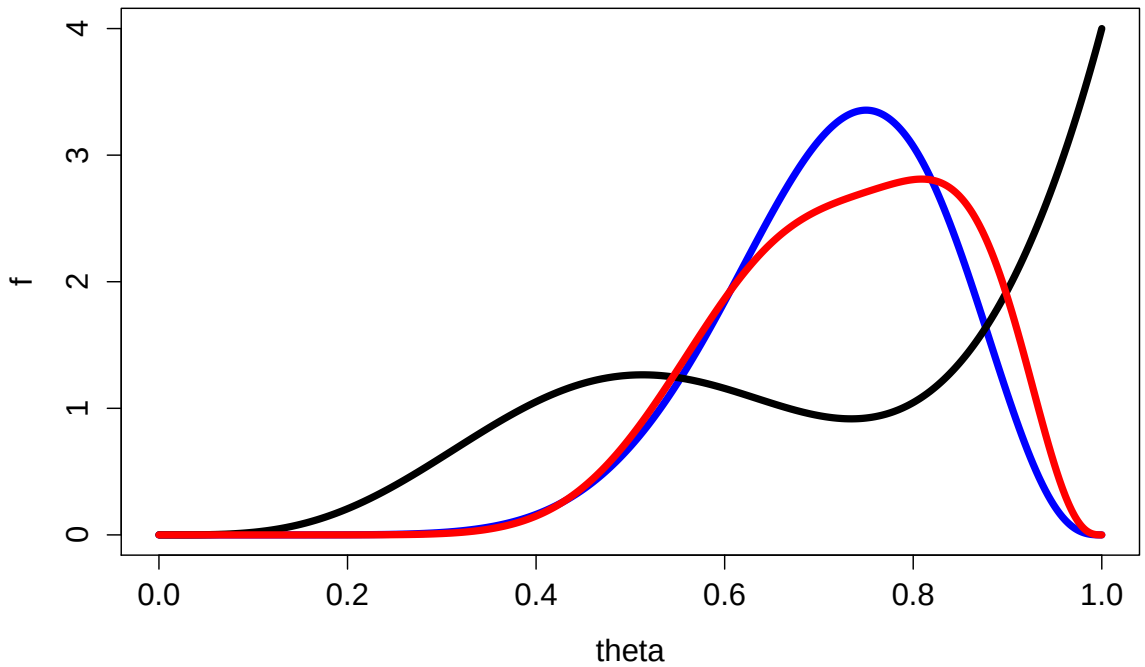
Suppose we set $f(\theta) = 0.5\text{Beta}(5,5) + 0.5\text{Beta}(8,1)$ in the coin tossing problem of yesterday. Then, given the observed data, we have

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto \theta^9(1-\theta)^3\left[0.5\frac{1}{B(5,5)}\theta^{5-1}(1-\theta)^{5-1} + 0.5\frac{1}{B(8,1)}\theta^{8-1}(1-\theta)^{1-1}\right] \\ &\propto \frac{1}{B(5,5)}\theta^{14-1}(1-\theta)^{8-1} + 0.5\frac{1}{B(8,1)}\theta^{17-1}(1-\theta)^{11-1} \\ &= w\text{Beta}(14,8) + (1-w)\text{Beta}(17,11) \end{aligned}$$

where $w = \frac{B(14,8)/B(5,5)}{B(14,8)/B(5,5)+B(17,11)/B(8,1)}$.

Example

The plot shows the prior (black), scaled likelihood (blue) and posterior (red) density.



When do conjugate priors exist?

- Conjugate priors are associated with exponential family distributions.

$$f(\mathbf{x}|\theta) = C(\mathbf{x})D(\theta)\exp\left(\mathbf{E}(\mathbf{x})^T\mathbf{F}(\theta)\right)$$

- A conjugate prior is then

$$f(\theta) \propto D(\theta)^a \exp(\mathbf{b}^T\mathbf{F}(\theta))$$

- Given a sample of size n,

$$f(\theta|\mathbf{x}) \propto D(\theta)^{a+n} \exp((\mathbf{b} + n\bar{E})^T\mathbf{F}(\theta))$$

  where $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E(\mathbf{x}_i)$ is the vector of sufficient statistics.
- Letting $a, b \to 0$ gives a natural, "objective" prior.

Rare events models

- Consider models associated with rare events (Poisson process).
- The likelihood function takes the form:

$$f(\mathbf{x}|\theta) = c\theta^n e^{-x\theta}$$

  where n represents the number of events to have occurred in a time period of length x and c depends on the experimental design.
- Therefore, a gamma distribution $\theta \sim \text{Gamma}(a, b)$, that is

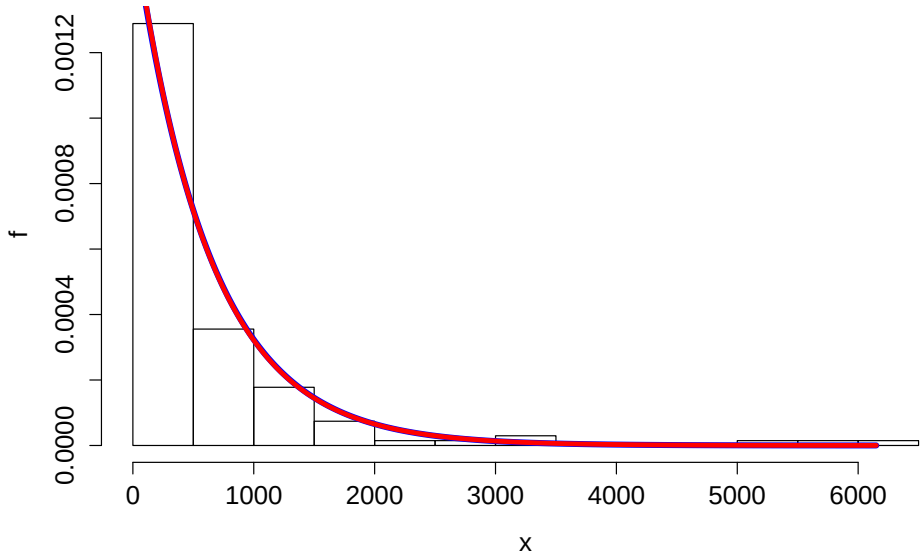
$$f(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} \quad \text{for } 0 < \theta < \infty$$

  is conjugate.
- The posterior distribution is then $\theta|\mathbf{x} \sim \text{Gamma}(a + n, b + x)$.

- The information in the prior is easily interpretable: a represents the prior equivalent of the number of rare events to occur in a time period of length b.
- Letting $a, b \to 0$ gives the natural default prior $f(\theta) \propto \frac{1}{\theta}$.
- (This is the Jeffreys prior for exponential data but not for Poisson data).
- In this case, given n observed events in time x, the posterior is $\theta|\mathbf{x} \sim \text{Gamma}(n, x)$, with mean $\frac{n}{x}$ which is equal to the MLE in experiments of this type.

Example: Software failure data

The CSIAC database provides data showing the times between 136 successive software failures. The diagram shows a histogram of the data and a classical, plug in estimator (blue) of the predictive distribution of x as well as the Bayesian posterior given a Jeffreys prior (red). The Bayesian and classical predictors are indistinguishable.



Example: Inference for a queueing system

- The M/M/1 queuing system assumes arrivals occur according to a Poisson process with rate λ.
- There is a single server.
- Service occurs on a first come first served basis.
- Service times are exponential with mean service time $1/\mu$.
- The system is stable if $\rho = \frac{\lambda}{\mu} < 1$.
- In this case, the equilibrium distribution of the number of people in the system, N, is geometric: $N \sim \text{Geometric}(1 - \rho)$.
- Time spent in the system by an arriving customer, $W \sim \text{Exponential}(\mu - \lambda)$.

Example

Hall (1991) provides collected inter-arrival and service time data for 98 users of an automatic teller machine in Berkeley, California. We shall assume that the interarrival times and service times both follow exponential distributions. The sufficient statistics were $n_a = n_s = 98$ and $x_a = 119.71$ and $x_s = 81.35$ minutes.

Given default priors for λ, μ, the posterior distributions are

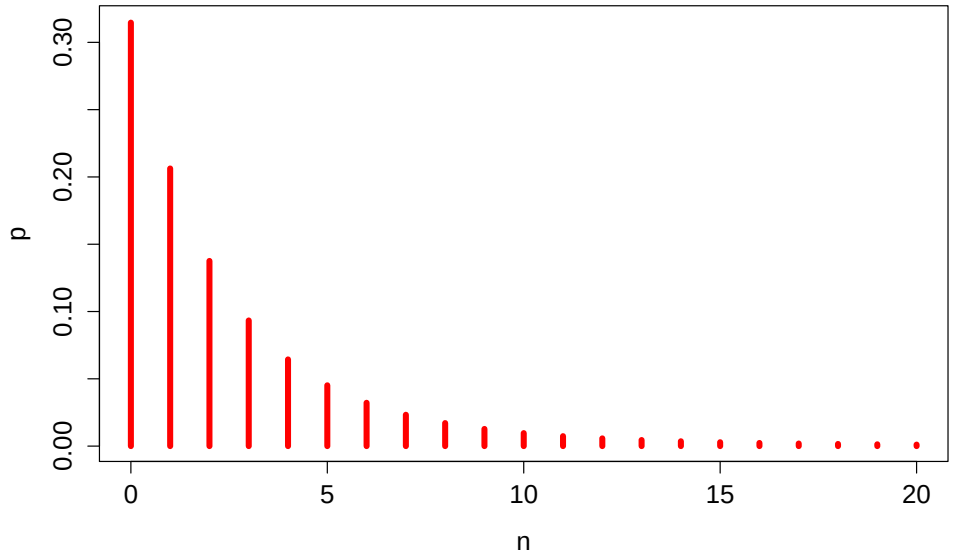
$$\lambda|\mathbf{x} \sim \text{Gamma}(98, 119.71) \quad \mu|\mathbf{x} \sim \text{Gamma}(98, 81.35).$$

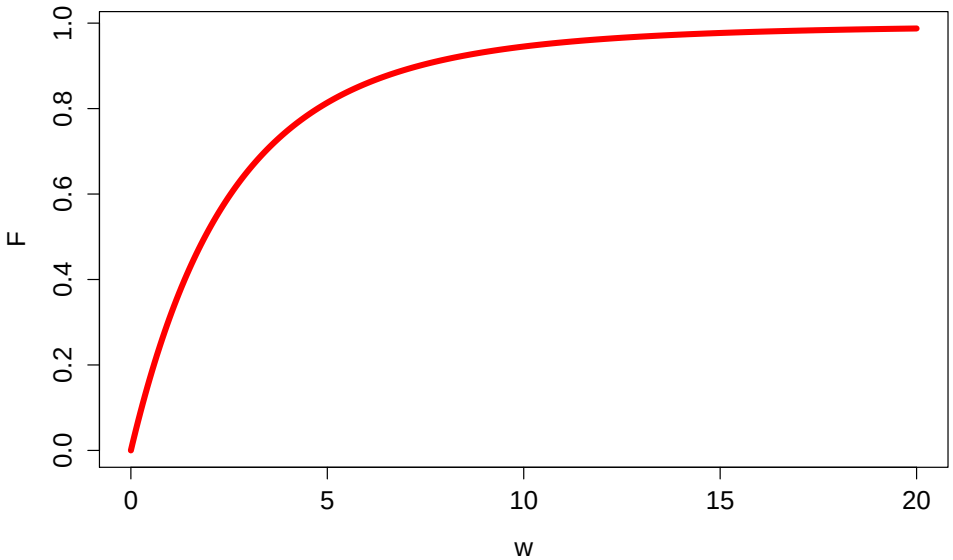
It is easy to calculate the posterior probability that the system is stable ...

... remembering that the ratio of two χ^2 distributions divided by their degrees of freedom is F distributed.

$$\begin{aligned} P(\rho < 1|\mathbf{x}) &= P\left(\frac{119.71}{81.35}\rho < \frac{119.71}{81.35} \mid \mathbf{x}\right) \\ &= P\left(F_{196}^{196} < \frac{119.71}{81.35}\right) = 0.9965. \end{aligned}$$

Given this is so high, it makes sense to consider the equilibrium distributions.


p
0.00
0.10
0.20
0.30
0
5
10
15
20
n


F
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
w

Normal models

- Consider a sample from a normal distribution $X|\mu,\sigma \sim \text{Normal}\left(\mu,\sigma^2\right)$.
- The likelihood function is

$$f(\mathbf{x}|\mu,\sigma) \propto \sigma^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{x}-\mu)^2\right]\right)$$

- Rewrite in terms of the precision, $\tau = \frac{1}{\sigma^2}$. Then

$$f(\mathbf{x}|\mu,\tau) \propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\left[(n-1)s^2 + n(\bar{x}-\mu)^2\right]\right)$$

- Define $f(\mu,\tau) = f(\tau)f(\tau|\mu)$ and assume $\tau \sim \text{Gamma}\left(\frac{a}{2},\frac{b}{2}\right)$ and $\mu|\tau \sim \text{Normal}\left(m,\frac{1}{c\tau}\right)$.
- The marginal distribution of μ is a (scaled, shifted) Student's t.

- A posteriori, we have

$$\begin{aligned} \mu|\tau, \mathbf{x} &\sim \text{Normal}\left(\frac{cm + n\bar{x}}{c+n}, \frac{1}{(c+n)\tau}\right) \\ \tau &\sim \text{Gamma}\left(\frac{a+n}{2}, \frac{b + (n-1)s^2 + \frac{cn}{c+n}(m-\bar{x})^2}{2}\right) \end{aligned}$$

- The conditional posterior precision is the sum of prior precision ($c\tau$) and precision of the MLE ($n\tau$).
- The posterior mean is a weighted average of the prior mean (m) and the MLE ($\bar{x}$).
- A default prior is obtained by letting $a, b, c \to 0$ which implies $f(\mu, \tau) \propto \frac{1}{\tau}$ and

$$\begin{aligned} \mu|\tau, \mathbf{x} &\sim \text{Normal}\left(\bar{x}, \frac{1}{n\tau}\right) \\ \tau|\mathbf{x} &\sim \text{Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \end{aligned}$$

- Then $\frac{\mu - \bar{x}}{s/\sqrt{n}} \mid \mathbf{x} \sim$ Student's t_{n-1} (boring)

One sample example

The normal core body temperature of a healthy adult is supposed to be 98.6 degrees Fahrenheit or 37 degrees Celsius on average. A normal model for temperatures, say $X|\mu, \tau \sim \text{Normal}(\mu, 1/\tau)$, has been proposed.

Mackowiak et al (1992) measured the core body temperatures of 130 individuals with mean .

The sample mean temperature is $\bar{x} = 98.2492$ Fahrenheit with standard deviation $s = 0.7332$.

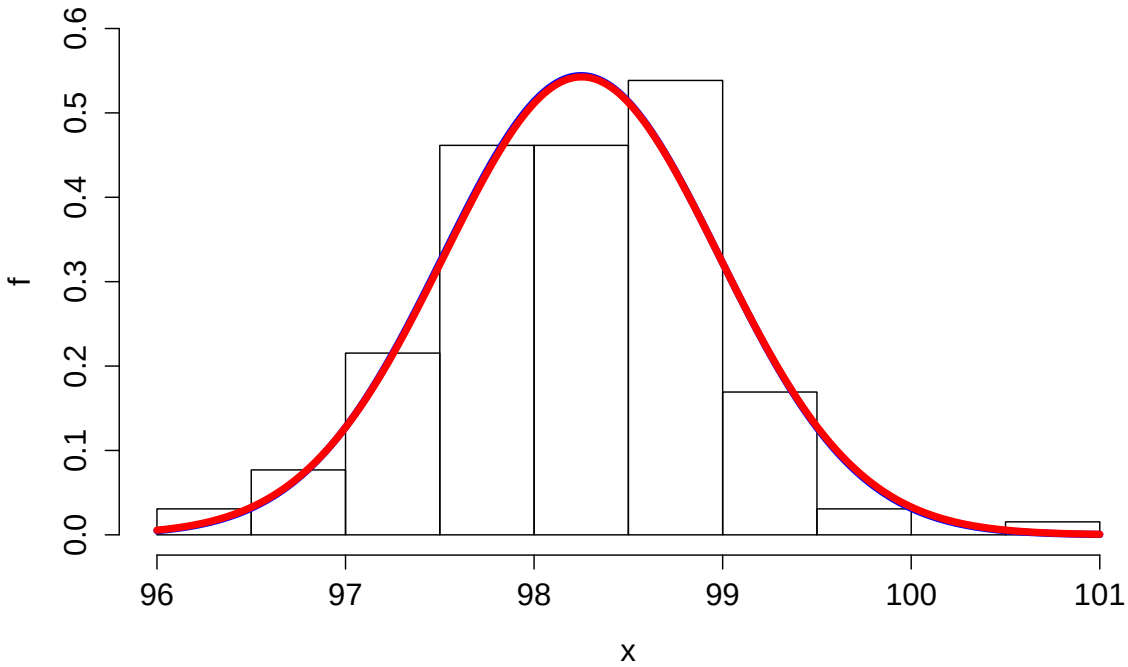
Thus, a classical 95% confidence interval for μ is

$$98.2492 \pm 1.96 \times 0.7332/\sqrt{130} = (98.1232, 98.3752)$$

and the hypothesis that the true mean is equal to 98.6 is rejected.

Consider a prior for μ centred on 98.6, for example $\mu|\tau \sim \text{Normal}(98.6, 1/\tau)$ with $f(\tau) \propto 1/\tau$. The posterior mean for μ is 98.2519 Fahrenheit and a 95% credible interval is $(98.1251, 98.3787)$ so that there still appears to be evidence against the hypothesis.

Also, the classical 'plug in' density for X (blue) and the Bayesian posterior predictive density (red) are almost identical.



An odd feature of the conjugate prior

- The model precision and the prior precision of the distribution of μ are both proportional to the model precision, τ.
- This may be restrictive and unrealistic in practical applications.
- A more natural prior for μ might be Normal $\left(m, \frac{1}{c}\right)$ independent of τ.
- Then, the joint posterior distribution looks nasty.

$$f(\mu, \tau|\mathbf{x}) \propto \tau^{\frac{a+n}{2}-1} \exp\left(-\frac{\tau}{2}\left[b + (n-1)s^2 + n(\bar{x}-\mu)^2\right] - \frac{c}{2}\left[\mu - m\right]^2\right)$$

What can we do?

- In our problem, both conditional posterior distributions are available:

$$\mu|\tau, \mathbf{x} \sim \text{Normal}\left(\frac{cm + n\tau\bar{x}}{c + n\tau}, \frac{1}{(c + n\tau)}\right)$$

$$\tau|\mu, \mathbf{x} \sim \text{Gamma}\left(\frac{a + n}{2}, \frac{b + (n-1)s^2 + n(\bar{x} - \mu)^2}{2}\right)$$

- Both these distributions are straightforward to sample from.
- Can we use this to give a Monte Carlo sample from the posterior?

Introduction to Gibbs sampling

A Gibbs sampler is a technique for sampling a multivariate distribution when it is straightforward to sample from the conditionals.

- Assume that we have a distribution $f(\theta)$ where $\theta = (\theta_1, ..., \theta_k)$.
- Let θ_{-i} represent the remaining elements of θ when θ_i is removed.
- Assume that we can sample from $\theta_i|\theta_{-i}$.

The Gibbs sampler

The Gibbs sampler proceeds by starting from (arbitrary) initial values and successively sampling the conditional distributions.

1. Set initial values $\theta^{(0)} = \left(\theta_1^{(0)}, ..., \theta_k^{(0)}\right)$. Set $t = 0$.
2. For $i = 1, ..., k$:
   1. Generate $\theta_i^{(t+1)} \sim \theta_i | \theta_{-i}^{(t)}$.
   2. Set $\theta^{(t)} = \theta_i^{(t+1)} \cup \theta_{-i}^{(t)}$.
3. $t = t + 1$
4. Go to 2.

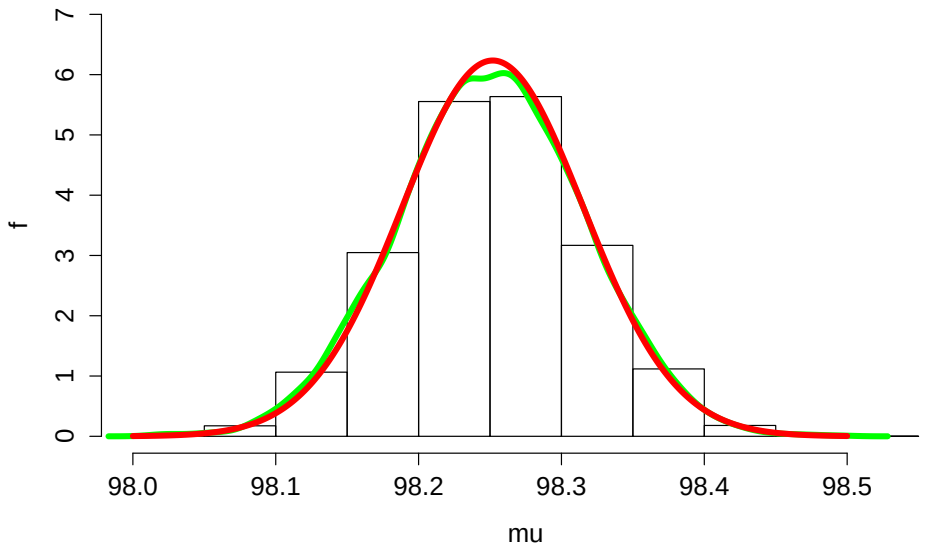
As $t \to \infty$, the sampled values approach a simple Monte Carlo sample from $f(\theta)$.

The example revisited

Consider now that we use independent priors,

$$\mu \sim \text{Normal}(98.6, 1) \quad f(\tau) \propto \frac{1}{\tau}$$

Then an estimated 95% posterior interval for μ, based on a sample of size 10000 is $(98.1222, 98.3789)$, very similar to the previous case. The diagram shows the estimated posterior density (green) and the posterior given the conjugate prior (red).



Both densities are very similar.

Two samples: the Behrens Fisher problem

For most simple one and two sample problems, when the usual default prior for μ, τ is used, posterior means and intervals for μ coincide with their frequentist counterparts. An exception is the following two sample problem:

Consider the model

$$X|\mu_1, \tau_1 \sim N\left(\mu_1, \frac{1}{\tau_1}\right), \quad Y|\mu_2, \tau_2 \sim N\left(\mu_2, \frac{1}{\tau_2}\right)$$

with priors $f(\mu_i, \tau_i) \propto \frac{1}{\tau_i}$ and independent samples of size n_i for $i = 1, 2$. Then,

$$\frac{\mu_1 - \bar{x}}{s_1/\sqrt{n_1}} \sim \text{Student's t}(n_1 - 1)$$

and similarly for μ_2.

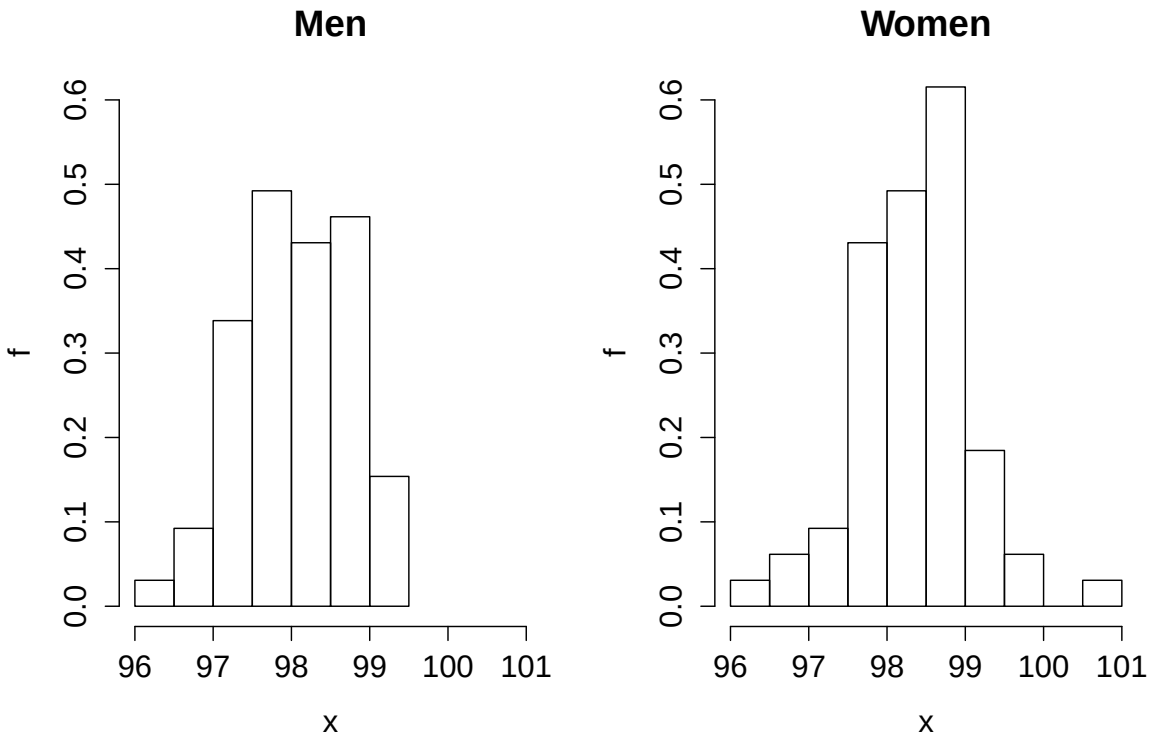
Therefore, if $\delta = \mu_1 - \mu_2$, we have

$$\delta = \bar{x} - \bar{y} + \frac{s_1}{\sqrt{n_1}} T_1 - \frac{s_2}{\sqrt{n_2}} T_2$$

- The distribution of δ is a scaled, shifted difference of two Student's t variables.
- Quantiles, ... can be calculated to a given precision by e.g. Monte Carlo.
- Writing $\delta' = \delta / \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1}}$ gives $\delta' = \sin w T_1 + \cos w T_2$ where $w = \tan^{-1} \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}$, a Behrens Fisher distribution.
- This problem is difficult to solve classically.
- Usually a t approximation to the sampling distribution of δ' is used, but ...
- the quality of the approximation depends on the true variance ratio.

Example

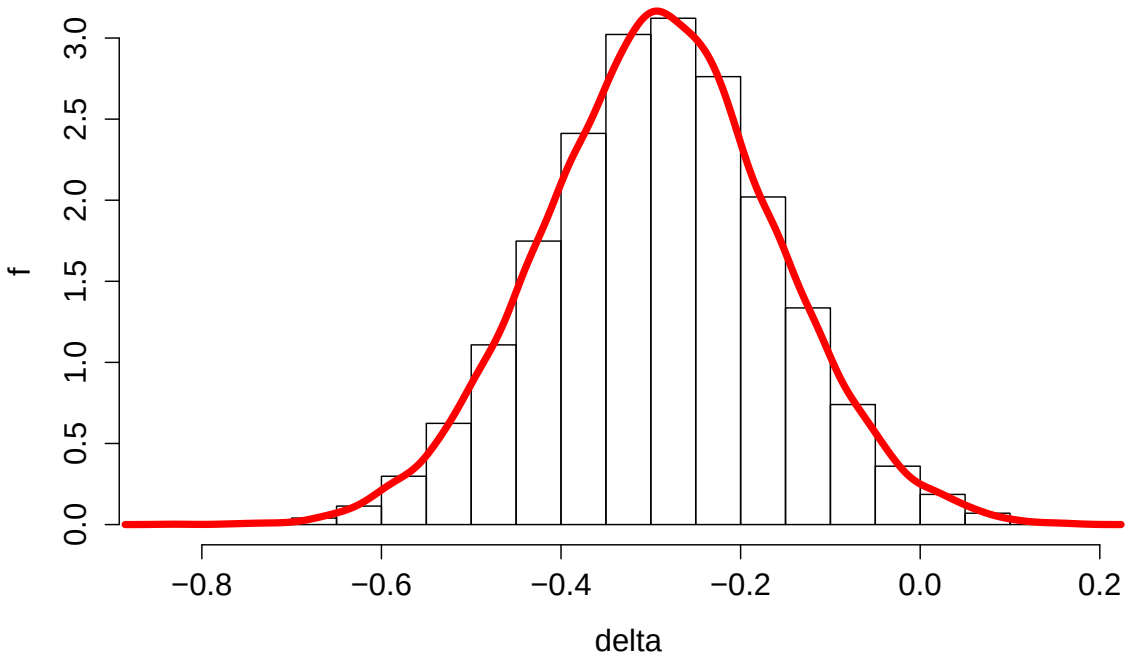
Returning to the normal body temperature example, the histograms indicate there may be a difference between the sexes.



The sample means are 98.1046 and 98.3939 respectively.

An approximate 95% confidence interval for the mean difference is $(-0.5396, -0.03881)$ suggesting that the true mean for women is higher than that for men.

Using the Bayesian approach as earlier (based on 10000 simulated values), we have an estimate of the posterior density of δ.



A Bayesian 95% credible interval is estimated as $(-0.5448, -0.0368)$.

Multinomial models

- The multinomial distribution is the extension of the binomial distribution to dice throwing problems.
- Assume a dice with k faces and probability θ_i for face i is thrown n times. Let $\mathbf{X}$ be a $k \times 1$ vector such that X_i is the number of times face i occurs. Then

$$P(\mathbf{X} = \mathbf{x}|\theta) = \frac{x!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i},$$

where $\mathbf{x} = (x_1, ..., x_k)$, $x_i \in \mathbb{Z}_+$ and $\sum_{i=1}^{k} x_i = n$ and $0 \leq \theta_i \leq 1$, $\sum_{i=1}^{k} \theta_i = 1$.

- Consider a Dirichlet prior, $\theta \sim$ Dirichlet($\mathbf{a}$), where $\mathbf{a} = (a_1, ..., a_k)$ and $a_i > 0$.

$$f(\theta) = \frac{\Gamma(\sum_{i=1}^{k} a_i)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} \theta_i^{a_i - 1}.$$

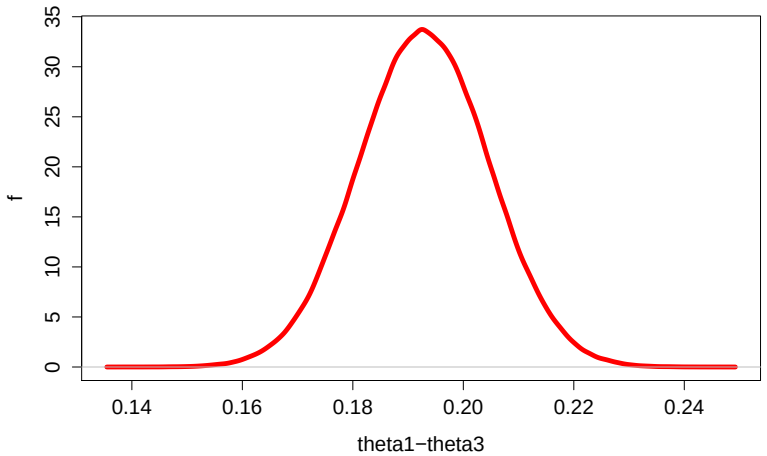
- Then $\theta|\mathbf{x} \sim$ Dirichlet $(\mathbf{a} + \mathbf{x})$.

Example

After the recent abdication of the King of Spain in favour of his son, *20minutos.es* launched a survey asking whether this was the correct decision, ($X_1 = 3698$ votes) whether the King should have waited longer ($X_2 = 347$) or whether he should have considered other options such as a referendum ($X_3 = 2446$).[1]

Let $\theta = (\theta_1, \theta_2, \theta_3)$ and assume a Dirichlet $(1/2, 1/2, 1/2)$ prior. The posterior distribution is Dirichlet$(3698.5, 347.5, 2446.5)$.

Consider the difference $\theta_1 - \theta_3$, reflecting (?) the difference between Monarchists and Republicans. We have $E[\theta_1 - \theta_3 | \mathbf{x}] = 0.193$.



[1] Votes at 12:30 on 5th May 2014.

The Dirichlet process prior

Sometimes, we do not wish to assume a parametric model for the data generating distribution. How can we do this in a Bayesian context?

- Assume $X|F \sim F$ and define a Dirichlet process prior for F.
- If the support of X is C, then for any partition, $C = C_1 \cup C_2 \cup ... \cup C_k$ and $k \in \mathbb{N}$, we suppose that

$$(F(C_1), F(C_2), ..., F(C_k)) \sim \text{Dirichlet}\,(aF_0(C_1), aF_0(C_2), ..., aF_0(C_k))$$

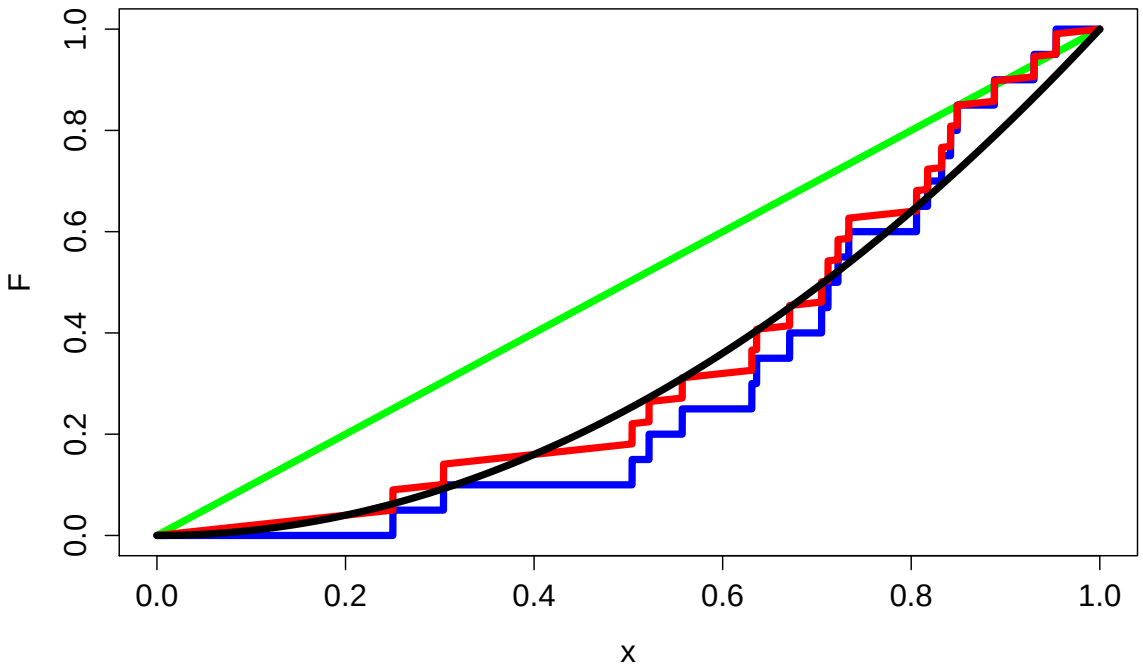
  where $a > 0$ and F_0 is a baseline, prior mean c.d.f.
- We write $F \sim$ Dirichlet process(a, F_0).
- Given a sample, $x_1, ..., x_n$, we have

$$F|\mathbf{x} \sim \text{Dirichlet process}\left(a + n, \frac{aF_0 + n\hat{F}}{a + n}\right)$$

  where $\hat{F}$ is the empirical c.d.f.
- The posterior mean is a weighted average of the empirical c.d.f. and the prior mean.

Example

The following plot shows the prior (green), posterior (red), empirical (blue) and true (black) c.d.f.s' when 20 data were generated from a Beta(2,1) distribution and a Dirichlet process prior with $a = 5$ and F_0 a uniform distribution were used.



Summary and next chapter

In this chapter we have illustrated the basic properties of conjugate models. When these exist, they allow for simple interpretation and straightforward inference.

- Unfortunately, conjugate priors do not always exist, for example if data are t or F distributed.
- Then we need numerical techniques like Gibbs sampling.
- We study these in more detail in the next chapter.