Lecture 6A
MTH6102: Bayesian Statistical Methods

Eftychia Solea

Queen Mary University of London

2023

Today's agenda

Today's lecture will cover

- Learn the different types of prior information.
- Be able to make a reasonable choice of prior, based on external data.

Choosing prior

- Bayesians make inferences using the posterior and therefore always need a prior.
- **Important question:** Where does one get the prior $p(\theta)$?
- If a prior is not known with certainty the Bayesian must try to make a reasonable choice. There are many ways to do this and people might make different choices.
- It is a good practice to do a sensitivity analysis to explore how posterior is affected by differences in prior.

Uninformative or noninformative prior distributions

- Suppose we have no idea of what the prior might be.
- In this case, we can define some sort of "noninformative prior" also known as vague
- No unique way of specifying an uninformative prior distribution.
- An obvious candidate for a noninformative prior is to use a flat prior, i.e., uniform over some range

$$p(\theta) \propto c$$

  where $c > 0$.

- It is flat relative to the likelihood.

Uninformative or noninformative prior distributions

$p(\theta|y) \propto p(\theta) \times p(y|\theta), \quad p(\theta) = c$

$= c \times p(y|\theta)$

$\propto p(y|\theta)$

- With a flat prior, the posterior $p(\theta|y)$ is proportional to the likelihood as functions of θ, so they have the same shape (but not necessarily the same scale)

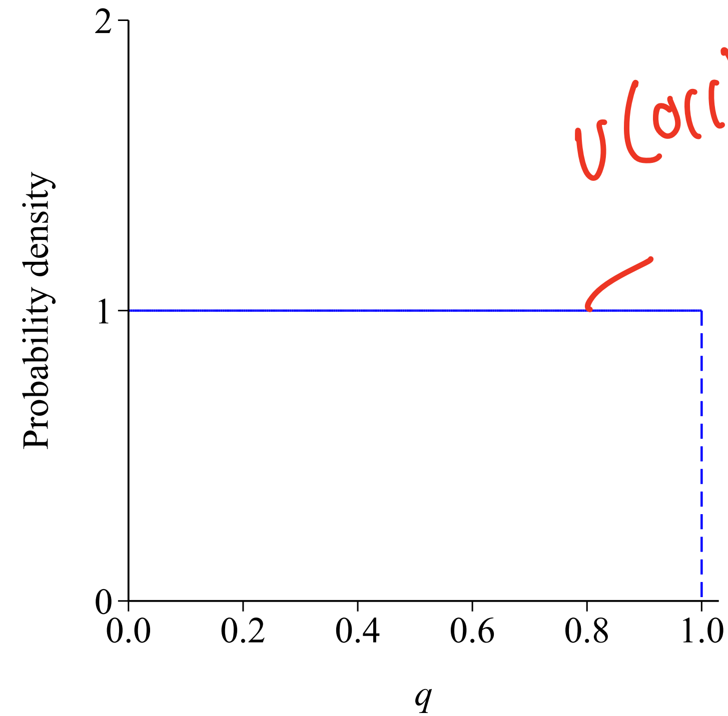
- For some simple problems e.g beta/binomial or normal/normal, a flat prior gives similar answers to likelihood-based inference (classical statistics).

Binomial data uniform prior

Uniform prior density on 0 to 1 for a probability q as an example of a flat prior. Data is k “successes” out of n trials. Recall the uniform is the beta$(1,1)$ density

$p(q|x) \sim \text{beta}(x+1, n-x+1)$

- With uniform prior, posterior mean for q is
$$\frac{k+1}{n+2}$$
- This pulls estimates away from 0 or 1 if k is close to 0 or n.



$U(0,1)$

Example: Uniform prior/binomial likelihood

Bent coin with unknown probability θ.
Flat prior: $p(\theta) = 1$ on $[0, 1]$
Data: toss 27 times and get 15 heads.

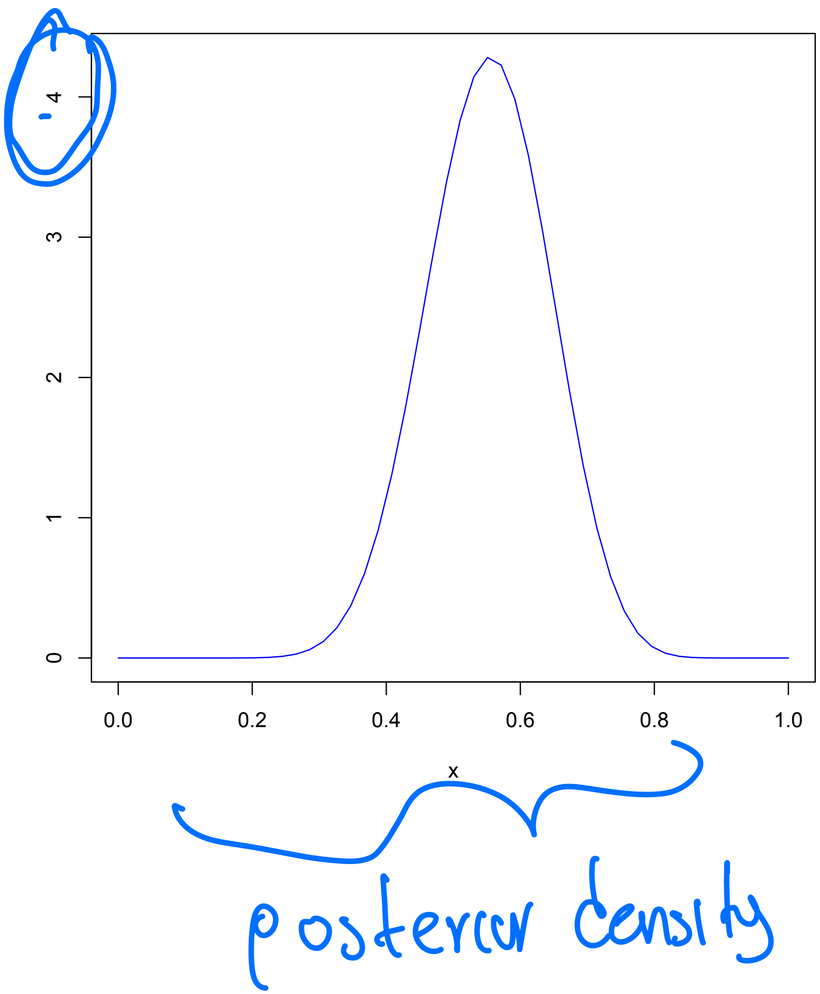
- The posterior density beta$(16, 13)$ is proportional to the binomial likelihood

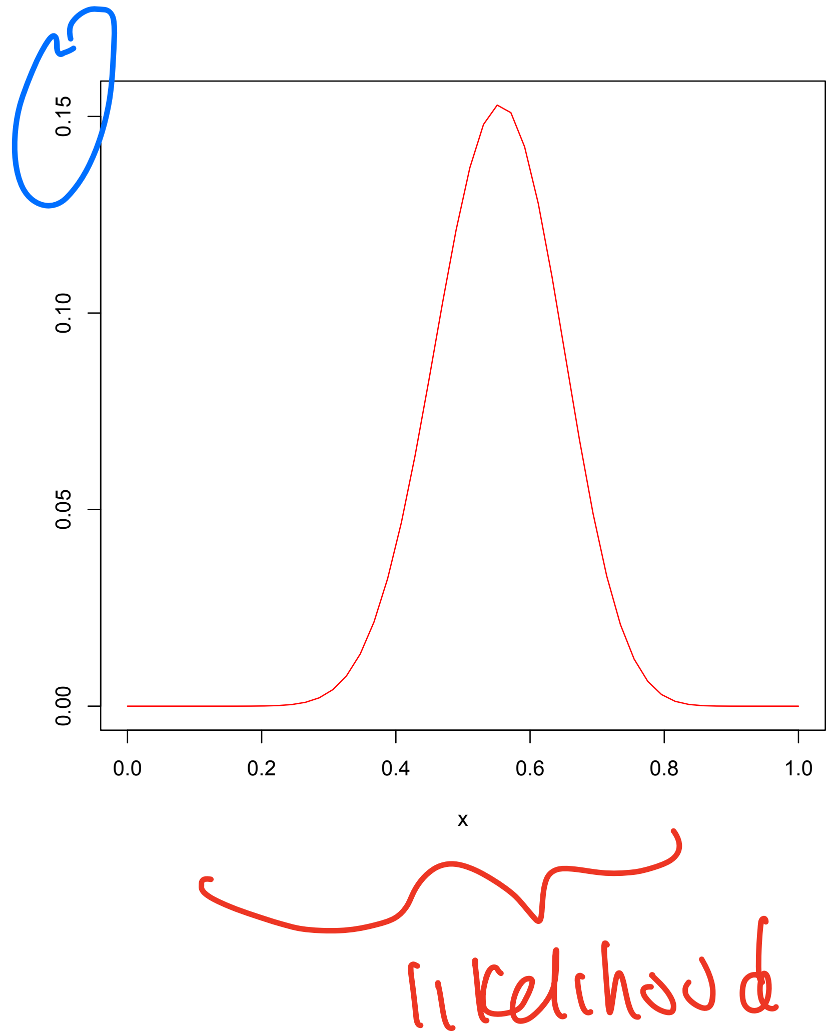
$$\boxed{p(\theta|k = 15) \propto \theta^{15}(1-\theta)^{12}}$$

- As functions of θ, $p(\theta|k = 15)$ and the binomial likelihood have the same shape.
- With n large the binomial likelihood becomes symmetric and peaked around the MLE $\hat{\theta} = \frac{x}{n} = \frac{15}{27}$
- With n large the posterior mean approaches to the MLE.

Example: Uniform prior/binomial likelihood

Left: posterior density, Right: likelihood plotted as functions of θ





Improper prior distributions

- An improper prior is one that doesn't have a finite integral which makes it improper density.
- Examples are flat priors $p(\theta) \propto c$ on 0 to ∞ since

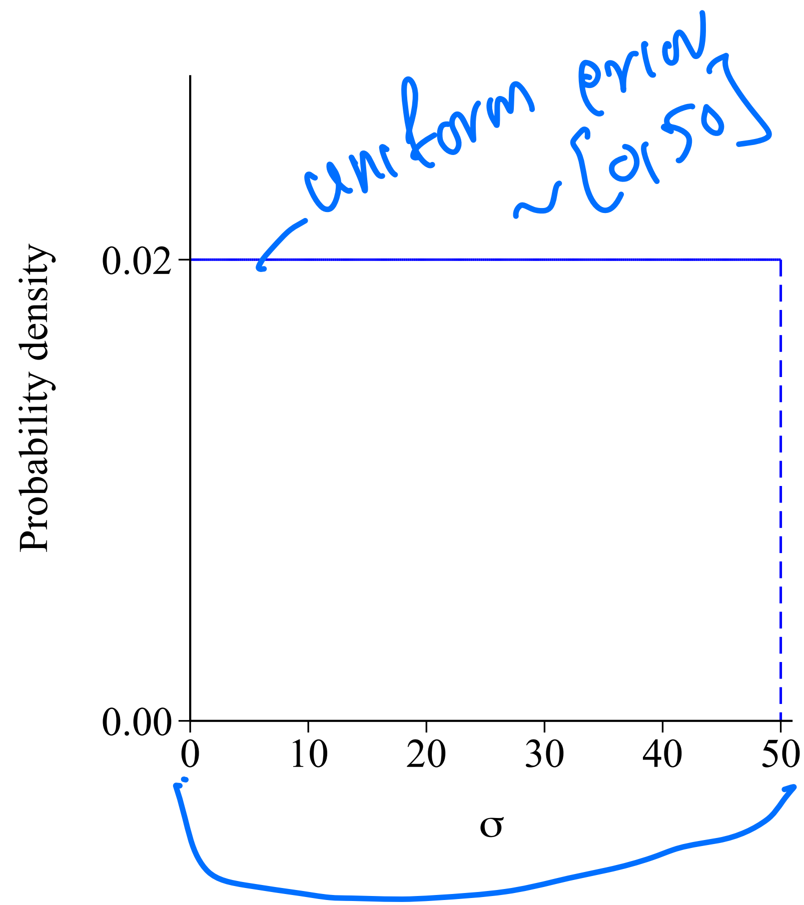
$$\int_0^\infty p(\theta)d\theta = c\int_0^\infty 1 d\theta = \infty.$$

- In many cases you can still use Bayes theorem and the resulting "posterior distribution" does have a finite integral.
- In general, improper priors are not a problem as long as the resulting posterior is a well-defined density.
- We only use proper priors in this module.

Flat priors

$\sigma \in [0, \infty)$

- Suppose a parameter must be positive, e.g. a standard deviation σ.
- We could choose a uniform prior on $[0, c]$ for some large c (otherwise this would lead to an improper prior)
- c would be chosen as larger than any plausible value for σ.



What about transformations of θ?

- If we specify a uniform prior on $[0, c]$ for σ, what is the prior for e.g. $\sigma^2 = g(\sigma)$?

- Recall, the shape of a probability density changes under non-linear monotonic transformations of the random variable.

- Suppose we have continuous random variables X and Y with pdf $f_y(x)$ and $f_y(y)$, respectively. Let $Y = g(X)$, where g is a monotonic function, then

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X(g^{-1}(y))$$

What about transformations of θ?

$$f_Y(y) = \underbrace{\left|\frac{dg^{-1}(y)}{dy}\right| \cdot f_X(g^{-1}(y))}_{} \propto \left|\frac{dg^{-1}(y)}{dy}\right| \text{ not constant}$$

$$\underset{f_X(x)=c \;\; \forall x}{1}$$

- So if f_X is constant and g is non-linear, then f_Y is not constant.
- Flat priors are not invariant under nonlinear transformations.
- A flat prior on θ does not imply a flat prior on $\psi = g(\theta)$.

Example

$\theta = \sigma$

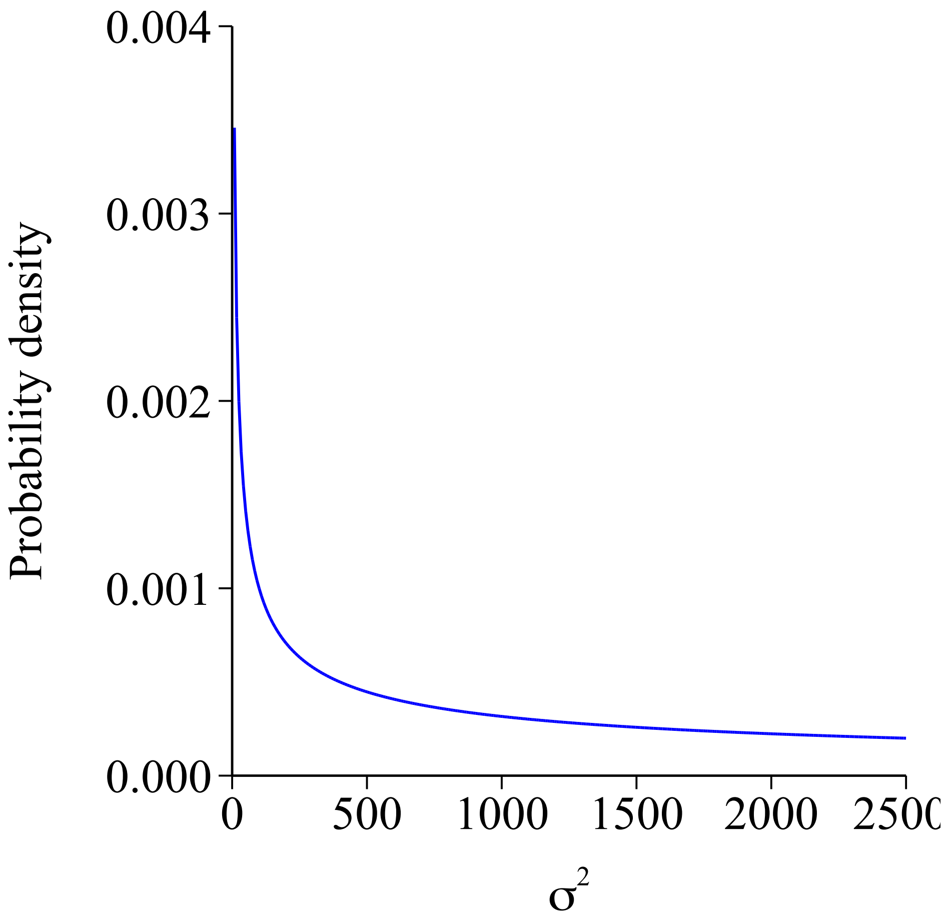
- Uniform prior for σ on $[0, c]$.
- The prior for σ^2 is not uniform.

  $\sigma^2 = g(\sigma)$
- It's proportional to

$$\frac{1}{2\sqrt{\sigma^2}}$$

on $[0, c^2]$.

$\Rightarrow$ A flat prior on σ does not imply a flat prior on σ^2.



$\theta = \sigma \qquad \sigma \in [0, c]$

$\psi = \sigma^2 = g(\theta)$

$g(x) = x^2$ is increasing with range $[0, c^2]$

Thus $\psi = g(\theta) = \theta^2$ so the inverse is $g^{-1}(\psi) = \psi^{1/2}$

$$\frac{d g^{-1}(\psi)}{d\psi} = \frac{1}{2} \psi^{-1/2}$$

Thus, the prior of $\psi = \sigma^2$ is

$$p_\psi(\psi) = \frac{1}{2} \psi^{-1/2} \cdot p_\theta\left(g^{-1}(\psi)\right)$$

But $p_\theta(\theta) = c$ $\forall\, \theta \in [0, c]$. Thus,

$$p_\psi(\psi) \propto \frac{1}{2} \psi^{-1/2} = \underbrace{\frac{1}{2\sqrt{\psi}} = \frac{1}{2\sqrt{\sigma^2}}}_{\text{The prior of } \psi.}, \quad \psi \in [0, c^2]$$

Example

- Uniform prior for σ on $[0, c]$.

$= g(\sigma)$

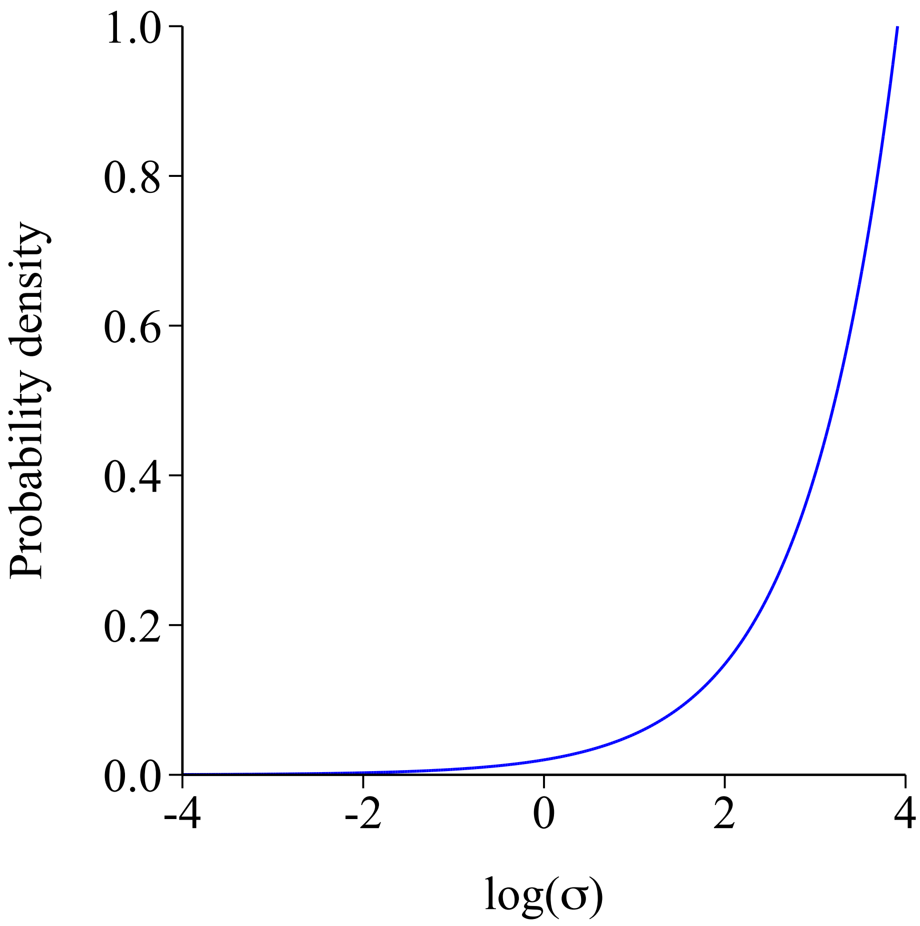
- The prior for $\log(\sigma)$ is not uniform

$\psi = \log(\sigma)$

- It's proportional to

$$e^{\log(\sigma)}$$

on $[-\infty, \log(c)]$.



Board question

- Let $x \sim \text{Bernoulli}(p)$
- Flat prior: $f(p) = 1, \quad p \in (0, 1)$
- This flat prior represents our lack of information about p before the experiment.
- Now, let $\psi = \log\left(\frac{p}{1-p}\right)$ the log of odds.
- What is the prior of ψ?

- But if we use a flat prior about θ, we would like to use a flat prior for ψ. So could we use a flat prior for ψ?

Solution

We have $f(p) = 1$ for all $p \in [0,1]$

We want to find the prior of $\underbrace{\psi = \log\left(\frac{p}{1-p}\right)}_{}= g(p)$ solve wrt p

The function is monotone with inverse

$$g^{-1}(\psi) = \frac{\exp(\psi)}{1+\exp(\psi)}$$ → the logistic function

The derivative of $g^{-1}(\psi)$ is

$$\frac{d g^{-1}(\psi)}{d\psi} = \frac{\exp(\psi)}{(1+\exp(\psi))^2}$$ (using the product rule of derivatives)

Thus,

$$p_\psi(\psi) = \frac{d g^{-1}(\psi)}{d\psi} \cdot 1$$

$$= \frac{\exp(\psi)}{(1+\exp(\psi))^2}.$$

⇒ A uniform prior on p does not imply a uniform prior on ψ

Jeffreys prior

- Jeffrey Harrison came up with a rule for creating noninformative priors that are invariant under nonlinear, smooth and monotonic transformations g.
- Let x data generated from the likelihood $p(x|\theta)$
- The Jeffreys prior $p_J(\theta)$ of θ is a noninformative prior of θ defined by

$$p_J(\theta) = c_1\sqrt{I(\theta)},$$

where $c_1 > 0$ and $I(\theta)$ is the Fisher information function given by (under some regularity conditions)

$$I(\theta) = -E\left[\frac{d^2}{d\theta^2}\log p(X|\theta)\right]$$

and $p(X|\theta)$ is the likelihood.

$X \sim p(x|\theta)$ is random variable

Jeffreys prior

- If $\int_\theta \sqrt{I(\theta)}\, d\theta < \infty$, then c_1 is taken to be $\left(\int_\theta \sqrt{I(\theta)}\, d\theta\right)^{-1}$ so that $p(\theta)$ is a proper density.

- Otherwise, if the integral is infinite, the constant c_1 is left unspecified and the prior $p(\theta)$ is an improper prior pdf of θ.

Jeffreys prior

- Jeffreys' prior is invariant to smooth monotone transformations of the parameter, $\psi = g(\theta)$, since

$$I(\psi) = I(\theta) \left(\frac{d\theta}{d\psi} \right)^2 .$$

- Hence, a Jeffreys prior for θ leads to a Jeffreys prior for $\psi = g(\theta)$ for g smooth monotone transformations

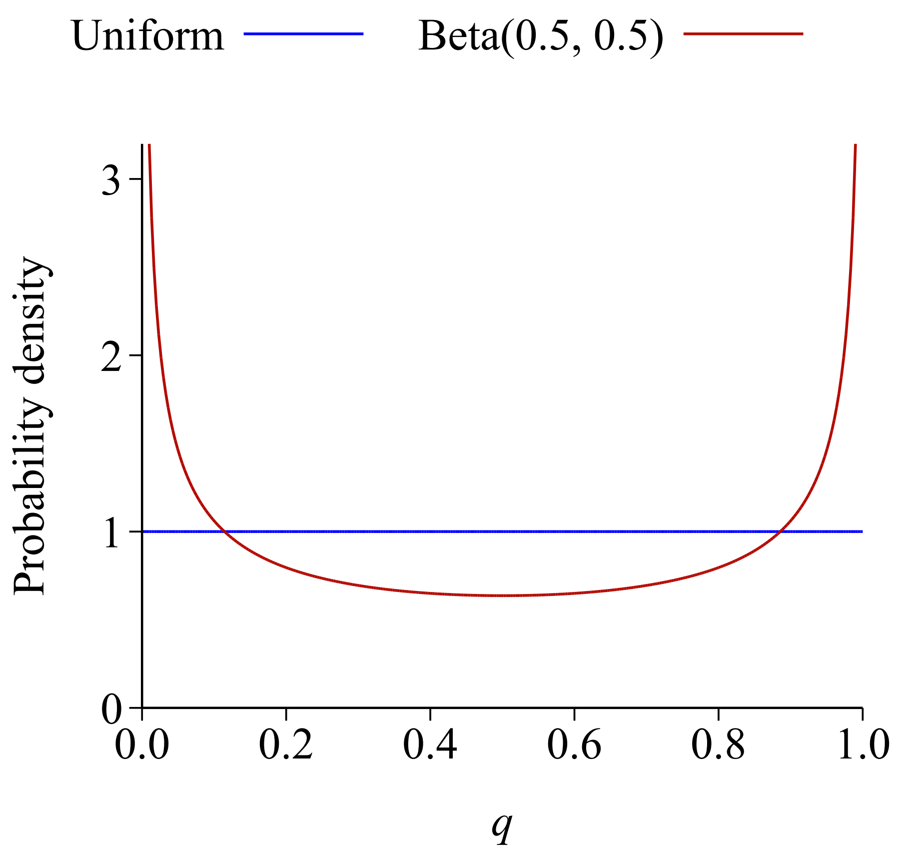
Jeffrey prior for beta/binomial model

- Let $x \sim \text{Binomial}(n, q)$, where q is the probability of success.
- Show that the Jeffreys' prior is $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ (similar to uniform or beta(1,1)),

$$p(q) \propto q^{-1/2}(1-q)^{-1/2}$$

- What is the posterior mean of q under the Binomial likelihood and Jeffreys prior?

$$\frac{k + 1/2}{n+1}$$



Board question: Jeffrey prior for beta/binomial model

Bent coin with unknown probability θ.
Jeffreys prior for θ on $[0, 1]$
Data: toss 27 times and get 15 heads.

- What is the posterior distribution and posterior mean of q under the Binomial likelihood and Jeffreys prior?

Board question: Jeffreys prior for normal/normal example

Let $x_1, \ldots, x_n$ iid from $N(\mu, \sigma^2)$ where σ^2 is known.

- Show that the Jeffreys prior for the normal likelihood is

$$p(\mu) = c_1 \sqrt{n/\sigma^2}, \quad \mu \in \mathbb{R}$$

  for some constant $c_1 > 0$.
- Is this a proper prior or improrer prior?
- Derive the posterior density for μ under the normal likelihood $N(\mu, \sigma^2)$ and Jeffreys prior for μ.

Informative prior

- Informative priors include some judgement concerning plausible values of the parameters based on external information.
- Informative priors can be based on pure judgement, a mixture of data and judgement, or external data alone.
- An informative prior distribution os one in which the probability mass is concentrated in some subset of the possible range for the parameters.

Informative prior

- There are many ways to build an informative prior. For example, using summary statistics, published estimates, intervals or standard errors.
- We can match these quantities to the mean, median standard deviation or percentiles of the prior distribution.

Example: Building an informative prior

- Let $t_1, \ldots, t_n \sim \text{Exp}(\lambda)$ denote the lifetimes of lightbulbs.
- The gamma distribution provides a conjugate prior for λ (failure rate)
- Suppose we have external information from other similar bulbs with observed failure rates $r_1, \ldots, r_K$.
- Let m and u be the mean and variance of $r_1, \ldots, r_K$, respectively.
- We want to build a gamma(α, β) distribution that for λ using this prior information.

Example: Building an informative prior

- We can use the method of moments to match the mean and the variance of the gamma distribution with the corresponding m and u
- That is

$$m = \frac{\alpha}{\beta}, \quad u = \frac{\alpha}{\beta^2}$$

- Solve for α and β

$$\beta = \frac{m}{u}, \quad \alpha = \frac{m^2}{u}.$$

- Thus, our prior for λ is gamma$(\frac{m^2}{u}, \frac{m}{u})$.

Weakly informative prior distributions

- Instead of trying to make the prior completely uniformative, an alternative is to convey some information about the plausible range of the parameters, e.g., exclude implausible values.
- Otherwise let the data speak for themselves.
- For models with large numbers of parameters, adding a little prior information may help with numerical stability.