Parameter estimation

Dr. Jarad Niemi

STAT 544 - Iowa State University

January 24, 2024

Outline

- Parameter estimation
  - Beta-binomial example
  - Point estimation
  - Interval estimation
  - Simulation from the posterior
- Priors
  - Subjective
  - Conjugate
  - Default
  - Improper

Parameter estimation

For point or interval estimation of a parameter θ in a model M based on data y, Bayesian inference is based off

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta)$$

where

- $p(\theta)$ is the prior distribution for the parameter,
- $p(\theta|y)$ is the posterior distribution for the parameter,
- $p(y|\theta)$ is the statistical model (or likelihood), and
- $p(y)$ is the prior predictive distribution (or marginal likelihood).

Obtaining the posterior

The hard way:

1. Derive $p(y)$.
2. Derive $p(\theta|y) = p(y|\theta)p(\theta)/p(y)$.

The easy way:

1. Derive $f(\theta) = p(y|\theta)p(\theta)$.
2. Recognize $f(\theta)$ as the kernel of some distribution.

Definition

The kernel of a probability density (mass) function is the form of the pdf (pmf) with any terms not involving the random variable omitted.

For example, $\theta^{a-1}(1-\theta)^{b-1}$ is the kernel of a beta distribution.

Derive the posterior - the hard way

Suppose $Y \sim Bin(n,\theta)$ and $\theta \sim Be(a,b)$, then

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta \\ &= \int \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\theta^{a-1}(1-\theta)^{b-1}}{\text{Beta}(a,b)}d\theta \\ &= \binom{n}{y}\frac{1}{\text{Beta}(a,b)}\int \theta^{a+y-1}(1-\theta)^{b+n-y-1}d\theta \\ &= \binom{n}{y}\frac{\text{Beta}(a+y,b+n-y)}{\text{Beta}(a,b)} \end{aligned}$$

which is known as the Beta-binomial distribution.

$$\begin{aligned} p(\theta|y) &= p(y|\theta)p(\theta)/p(y) \\ &= \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\theta^{a-1}(1-\theta)^{b-1}}{\text{Beta}(a,b)} \Big/ \binom{n}{y}\frac{\text{Beta}(a+y,b+n-y)}{\text{Beta}(a,b)} \\ &= \frac{\theta^{a+y-1}(1-\theta)^{b+n-y-1}}{\text{Beta}(a+y,b+n-y)} \end{aligned}$$

Thus $\theta|y \sim Be(a+y, b+n-y)$.

Derive the posterior - the easy way

Suppose $Y \sim Bin(n, \theta)$ and $\theta \sim Be(a, b)$, then

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{a+y-1}(1-\theta)^{b+n-y-1} \end{aligned}$$

Thus $\theta|y \sim Be(a + y, b + n - y)$.

Interpretation of prior parameters

When constructing the $Be(a,b)$ prior with the binomial likelihood which results in the posterior

$$\theta|y \sim Be(a+y, b+n-y),$$

we can interpret the prior parameters in the following way:

- a: prior successes
- b: prior failures
- $a+b$: prior sample size
- $a/(a+b)$: prior mean

These interpretations may aid in construction of this prior for a given application.

Posterior mean is a weighted average of prior mean and the MLE

The posterior is $\theta|y \sim Be(a+y, b+n-y)$. The posterior mean is

$$\begin{array}{rl} E[\theta|y] & = \frac{a+y}{a+b+n} \\ & = \frac{a}{a+b+n} + \frac{y}{a+b+n} \\ & = \frac{a+b}{a+b+n}\left(\frac{a}{a+b}\right) + \frac{n}{a+b+n}\left(\frac{y}{n}\right) \end{array}$$

Thus, the posterior mean is a weighted average of the prior mean $a/(a+b)$ and the MLE y/n with weights equal to the prior sample size $(a+b)$ and the data sample size (n).

Example data

Assume $Y \sim Bin(n, \theta)$ and $\theta \sim Be(1, 1)$ (which is equivalent to Unif(0,1)). If we observe three successes ($y = 3$) out of ten attempts ($n = 10$). Then our posterior is $\theta|y \sim Be(1 + 3, 1 + 10 - 3) \stackrel{d}{=} Be(4, 8)$. The posterior mean is

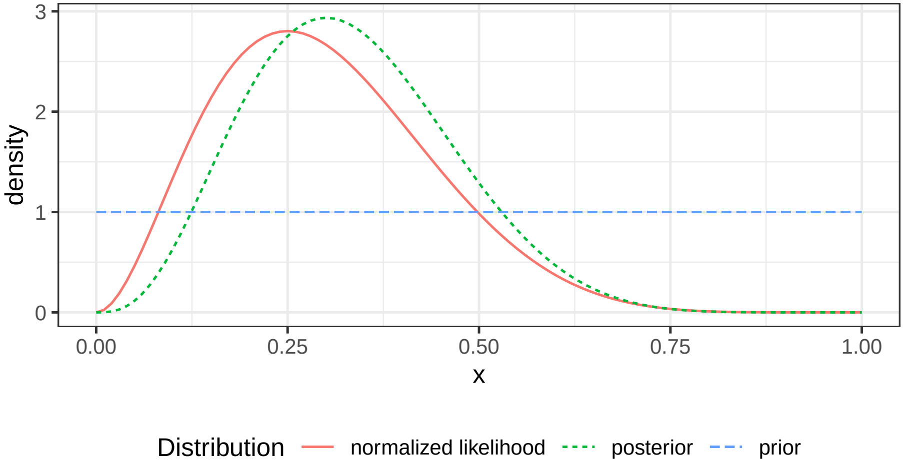
$$E[\theta|y] = \frac{2}{12} \times \frac{1}{2} + \frac{10}{12} \times \frac{3}{10} = \frac{4}{12}.$$

Remark Note that a $Be(1, 1)$ is equivalent to $p(\theta) = \mathrm{I}(0 < \theta < 1)$, i.e.

$$p(\theta|y) \propto p(y|\theta)p(\theta) = p(y|\theta)\mathrm{I}(0 < \theta < 1)$$

so it may seem that a reasonable approach to a default prior is to replace $p(\theta)$ by a 1 (times the parameter constraint). We will see later that this depends on the parameterization.

Posterior distribution



Try it yourself at `https://jaradniemi.shinyapps.io/one_parameter_conjugate/`.

Point and interval estimation

Nothing inherently Bayesian about obtaining point and interval estimates.

Point estimation requires specifying a loss (or utility) function.

A $100(1-a)\%$ credible interval is any interval in the posterior that contains the parameter with probability $(1-a)$.

Point estimation

Define a loss (or utility) function $L\left(\theta, \hat{\theta}\right) = -U\left(\theta, \hat{\theta}\right)$ where

- θ is the parameter of interest
- $\hat{\theta} = \hat{\theta}(y)$ is the estimator of θ.

Find the estimator that minimizes the expected loss:

$$\hat{\theta}_{Bayes} = \text{argmin}_{\hat{\theta}}\, E\left[L\left(\theta, \hat{\theta}\right)\middle| y\right]$$

or maximizes expected utility.

Common estimators:

- Mean: $\hat{\theta}_{Bayes} = E[\theta|y]$ minimizes $L\left(\theta, \hat{\theta}\right) = \left(\theta - \hat{\theta}\right)^2$
- Median: $\int_{\hat{\theta}_{Bayes}}^{\infty} p(\theta|y)d\theta = \frac{1}{2}$ minimizes $L\left(\theta, \hat{\theta}\right) = \left|\theta - \hat{\theta}\right|$
- Mode: $\hat{\theta}_{Bayes} = \text{argmax}_{\theta}\, p(\theta|y)$ is obtained by minimizing $L\left(\theta, \hat{\theta}\right) = -\mathrm{I}\left(|\theta - \hat{\theta}| < \epsilon\right)$ as $\epsilon \to 0$, also called maximum a posterior (MAP) estimator.

Mean minimizes squared-error loss

Theorem

The mean minimizes expected squared-error loss.

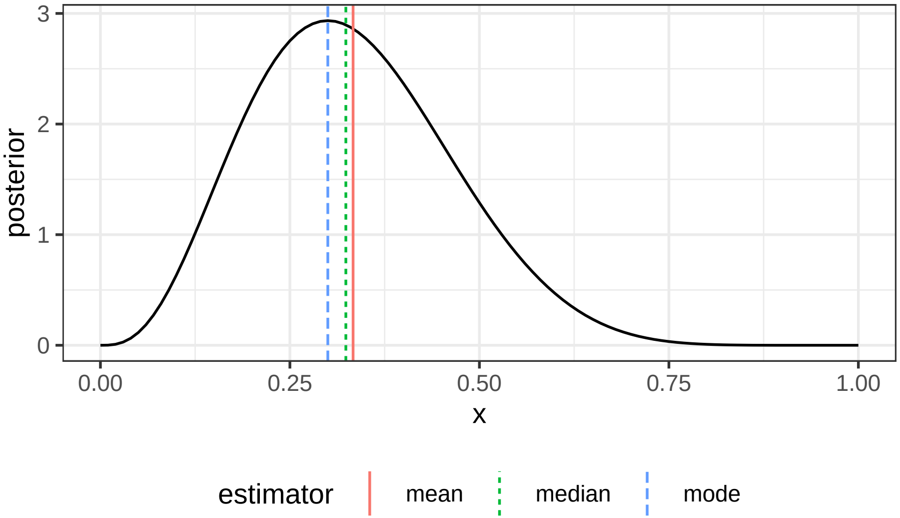
Proof.

Suppose $L\left(\theta,\hat\theta\right) = \left(\theta-\hat\theta\right)^2 = \theta^2 - 2\theta\hat\theta + \hat\theta^2$, then

$$
\begin{aligned}
E\left[L\left(\theta,\hat\theta\right)\middle| y\right] &= E\left[\theta^2|y\right] - 2\hat\theta E[\theta|y] + \hat\theta^2 \\
\frac{d}{d\hat\theta}E\left[L\left(\theta,\hat\theta\right)\middle| y\right] &= -2E[\theta|y] + 2\hat\theta \overset{set}{=} 0 \implies \hat\theta = E[\theta|y] \\
\frac{d^2}{d\hat\theta^2}E\left[L\left(\theta,\hat\theta\right)\middle| y\right] &= 2
\end{aligned}
$$

So $\hat\theta = E[\theta|y]$ minimizes expected squared-error loss. □

Point estimation



Interval estimation

Definition

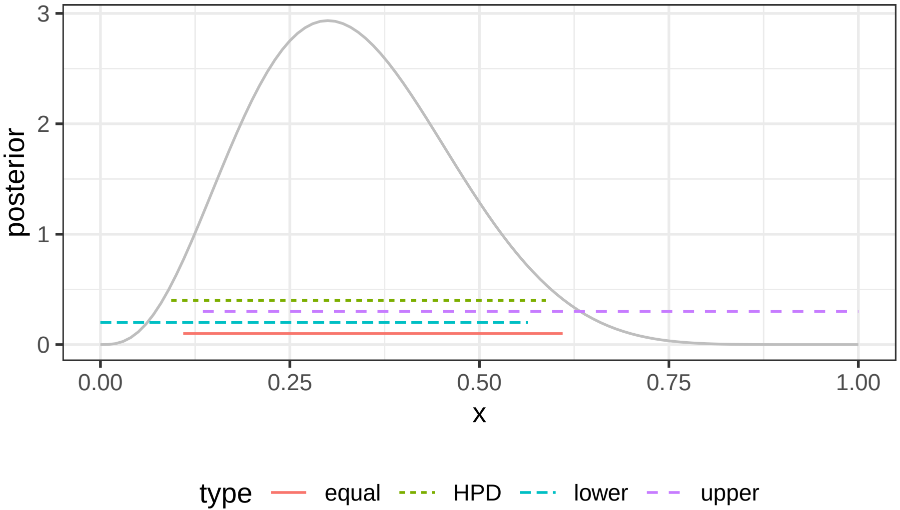
A $100(1-a)\%$ credible interval is any interval (L,U) such that

$$1-a = \int_L^U p(\theta|y)d\theta.$$

Some typical intervals are

- Equal-tailed: $a/2 = \int_{-\infty}^L p(\theta|y)d\theta = \int_U^\infty p(\theta|y)d\theta$
- One-sided: either $L=-\infty$ or $U=\infty$
- Highest posterior density (HPD): $p(L|y) = p(U|y)$ for a uni-modal posterior which is also the shortest interval

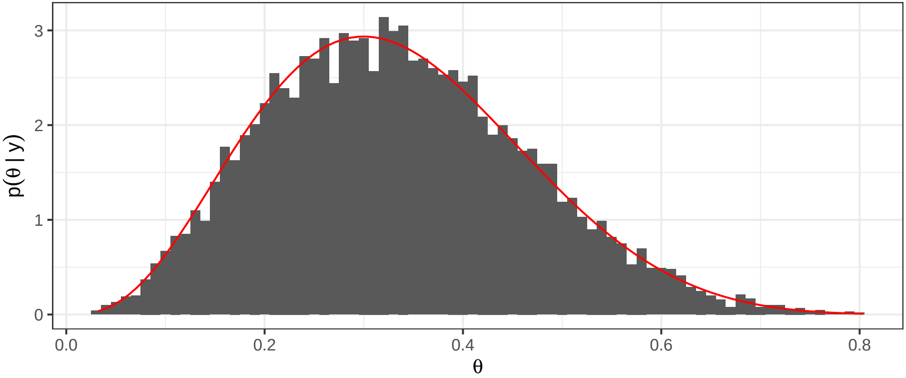
Interval estimation



Simulation from the posterior

An estimate of the full posterior can be obtained via simulation, i.e.

```
sim = data.frame(x = rbeta(10000, shape1 = a + y, shape2 = b + n - y))
```



Estimates via simulation

We can also obtain point and interval estimates using these simulations

```
round(c(mean = mean(sim$x), median = median(sim$x)),2)
```

```
  mean median 
  0.34   0.33 
```

```
round(quantile(sim$x, c(.025,.975)),2) # Equal-tail
```

```
 2.5% 97.5% 
 0.11  0.61 
```

```
round(c(quantile(sim$x, .05),1),2) # Upper
```

```
  5%      
0.13 1.00 
```

```
round(c(0,quantile(sim$x, .95)),2) # Lower
```

```
      95% 
0.00 0.57 
```

Guess the probability

What do you think the probability is?

- A 6-sided die lands on 1.
- The first base pair in my chromosome 1 is A.
- Kansas City Chiefs win 2023 Super Bowl.

What are priors?

Definition

A prior probability distribution, often called simply the prior, of an uncertain quantity θ is the probability distribution that would express one's uncertainty about θ before the "data" is taken into account.

http://en.wikipedia.org/wiki/Prior_distribution

Priors

Definition

A prior $p(\theta)$ is conjugate if for $p(\theta) \in \mathcal{P}$ and $p(y|\theta) \in \mathcal{F}$, $p(\theta|y) \in \mathcal{P}$ where $\mathcal{F}$ and $\mathcal{P}$ are families of distributions.

For example, the beta distribution ($\mathcal{P}$) is conjugate to the binomial distribution with unknown probability of success ($\mathcal{F}$) since

$$\theta \sim \text{Be}(a, b) \qquad \text{and} \qquad \theta|y \sim \text{Be}(a + y, b + n - y).$$

Definition

A natural conjugate prior is a conjugate prior that has the same functional form as the likelihood.

For example, the beta distribution is a natural conjugate prior since

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \qquad \text{and} \qquad L(\theta) \propto \theta^{y}(1-\theta)^{n-y}.$$

Discrete priors are conjugate

Theorem

Discrete priors are conjugate.

Proof.

Suppose $p(\theta)$ is discrete, i.e.

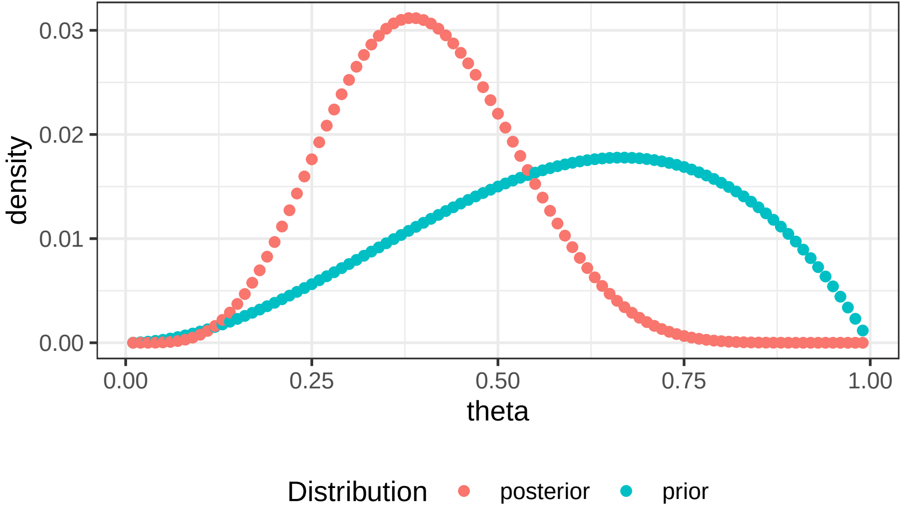
$$P(\theta = \theta_i) = p_i \qquad \sum_{i=1}^{\mathrm{I}} p_i = 1$$

and $p(y|\theta)$ is the model. Then, $P(\theta = \theta_i|y) = p_i'$ is the posterior with

$$p_i' = \frac{p_i p(y|\theta_i)}{\sum_{j=1}^{\mathrm{I}} p_j p(y|\theta_j)} \propto p_i p(y|\theta_i).$$

□

Discrete prior



Discrete mixtures of conjugate priors are conjugate

Theorem

Discrete mixtures of conjugate priors are conjugate.

Proof.

Let $p_i = P(H_i)$ and $p_i(\theta) = p(\theta|H_i)$,

$$\theta \sim \sum_{i=1}^{\mathrm{I}} p_i p_i(\theta) \qquad \sum_{i=1}^{\mathrm{I}} p_i = 1,$$

and $p_i(y) = \int p(y|\theta)p_i(\theta)d\theta$, then

$$\begin{aligned} p(\theta|y) &= \frac{1}{p(y)}p(y|\theta)p(\theta) = \frac{1}{p(y)}p(y|\theta)\sum_{i=1}^{\mathrm{I}} p_i p_i(\theta) = \frac{1}{p(y)}\sum_{i=1}^{\mathrm{I}} p_i p(y|\theta)p_i(\theta) \\ &= \frac{1}{p(y)}\sum_{i=1}^{\mathrm{I}} p_i p_i(y)p_i(\theta|y) = \sum_{i=1}^{\mathrm{I}} \frac{p_i p_i(y)}{p(y)}p_i(\theta|y) = \sum_{i=1}^{\mathrm{I}} \frac{p_i p_i(y)}{\sum_{j=1}^{\mathrm{I}} p_j p_j(y)}p_i(\theta|y) \end{aligned}$$

Mixtures of conjugate priors are conjugate

Bottom line: if

$$\theta \sim \sum_{i=1}^{I} p_i p_i(\theta) \qquad \sum_{i=1}^{I} p_i = 1$$

and $p_i(y) = \int p(y|\theta)p_i(\theta)d\theta$, then

$$\theta|y \sim \sum_{i=1}^{I} p_i' p_i(\theta|y) \qquad p_i' \propto p_i p_i(y)$$

where $p_i(\theta|y) = p(y|\theta)p_i(\theta)/p_i(y)$.

Mixture of beta distributions

Recall, if $Y \sim Bin(n,\theta)$ and $\theta \sim \text{Be}(a,b)$, then the marginal likelihood is

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta = \int \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\theta^{a-1}(1-\theta)^{b-1}}{\text{Beta}(a,b)} \\ &= \binom{n}{y}\frac{1}{\text{Beta}(a,b)}\int \theta^{a+y-1}(1-\theta)^{b+n-y-1}d\theta \\ &= \binom{n}{y}\frac{\text{Beta}(a+y,b+n-y)}{\text{Beta}(a,b)} \quad y=0,\ldots,n \end{aligned}$$

which is called the beta-binomial distribution with parameters $a+y$ and $b+n-y$.
If $Y \sim Bin(n,\theta)$ and

$$\theta \sim p\,\text{Be}(a_1,b_1) + (1-p)\text{Be}(a_2,b_2),$$

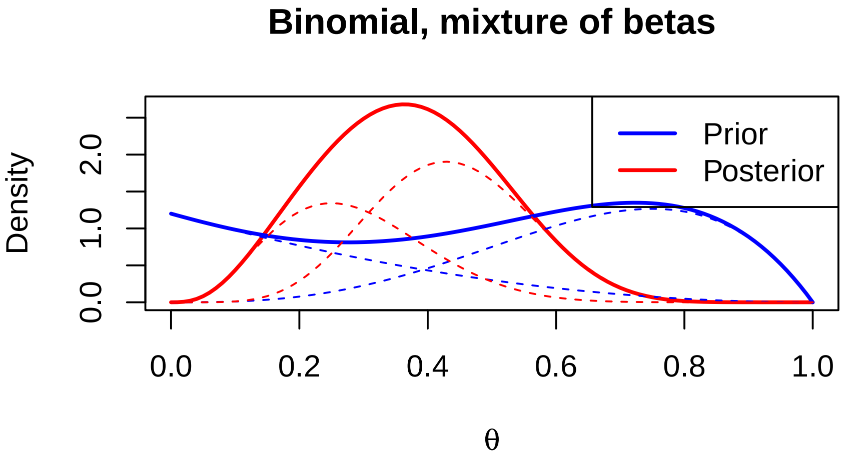
then

$$\theta|y \sim p'\,\text{Be}(a_1+y,b_1+n-y) + (1-p')\text{Be}(a_2+y,b_2+n-y)$$

with

$$p' = \frac{p\,p_1(y)}{p\,p_1(y)+(1-p)p_2(y)} \qquad p_i(y) = \binom{n}{y}\frac{\text{Beta}(a_i+y,b_i+n-y)}{\text{Beta}(a_i,b_i)}$$

Mixture priors



Default priors

Definition

A default prior is used when a data analyst is unable or unwilling to specify an informative prior distribution.

Default priors

Can we always use $p(\theta) \propto 1$?

Suppose we use $\phi = \log(\theta/[1-\theta])$, the log odds as our parameter, and set $p(\phi) \propto 1$, then the implied prior on θ is

$$
\begin{aligned}
p_\theta(\theta) \propto \; & 1 \left| \tfrac{d}{d\theta} \log(\theta/[1-\theta]) \right| \\
& = \tfrac{1-\theta}{\theta} \left[\tfrac{1}{1-\theta} + \tfrac{\theta}{[1-\theta]^2} \right] \\
& = \tfrac{1-\theta}{\theta} \left[\tfrac{[1-\theta]+\theta}{[1-\theta]^2} \right] \\
& = \theta^{-1}[1-\theta]^{-1}
\end{aligned}
$$

a Be(0,0), if that were a proper distribution, and is different from setting $p(\theta) \propto 1$ which results in the Be(1,1) prior. Thus, the constant prior is not invariant to the parameterization used.

Fisher information background

Definition

Fisher information, $\mathcal{I}(\theta)$, for a scalar parameter θ is the expectation of the second derivative of the log-likelihood, i.e.

$$\mathcal{I}(\theta) = E\left[\frac{\partial^2}{\partial \theta^2} \log p(y|\theta) \middle| \theta \right].$$

Theorem (Casella & Berger (2nd ed) Lemma 7.3.11)

For exponential families,

$$\mathcal{I}(\theta) = -E\left[\left(\frac{\partial}{\partial \theta} \log p(y|\theta) \right)^2 \middle| \theta \right].$$

If $\theta = (\theta_1, \ldots, \theta_n)$, then the Fisher information is the expectation of the Hessian matrix, which has the ith row and jth column that is the partial derivative with respect to θ_i followed by the partial derivative with respect to θ_j, of the log-likelihood.

Jeffreys prior

Definition

Jeffreys prior is a prior that is invariant to parameterization and is obtained via

$$p(\theta) \propto \sqrt{\det \mathcal{I}(\theta)}$$

where $\mathcal{I}(\theta)$ is the Fisher information.

For example, for a binomial distribution $\mathcal{I}(\theta) = \frac{n}{\theta[1-\theta]}$, so

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

a Be(1/2,1/2) distribution.

Fisher information

Theorem

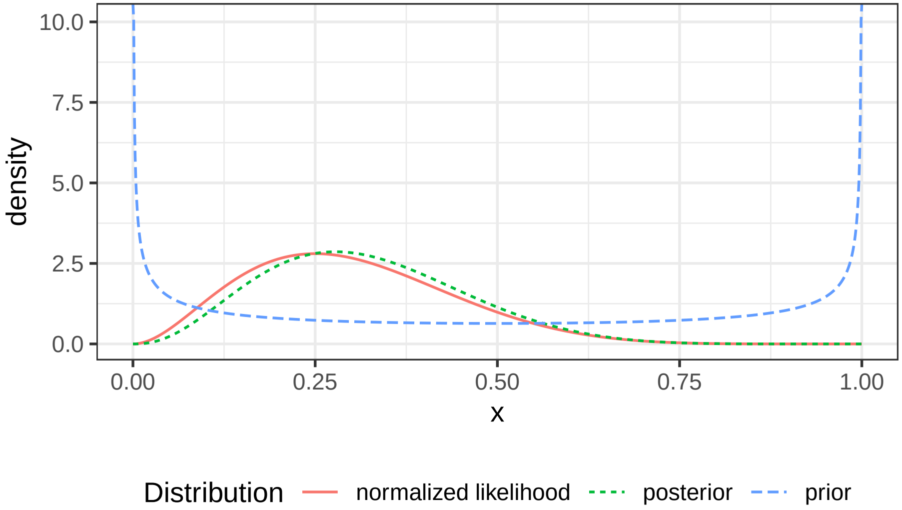
The Fisher information for $Y \sim Bin(n,\theta)$ *is* $\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}$.

Proof.

Since the binomial is an exponential family,

$$\begin{aligned}
\mathcal{I}(\theta) &= -E_{y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\right] = -E_{y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log\binom{n}{y} + y\log\theta + (n-y)\log(1-\theta)\right] \\
&= -E_{y|\theta}\left[\frac{\partial}{\partial\theta}\frac{y}{\theta} - \frac{n-y}{1-\theta}\right] = -E_{y|\theta}\left[-\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}\right] = -\left[-\frac{n\theta}{\theta^2} - \frac{n-n\theta}{(1-\theta)^2}\right] = \frac{n}{\theta} + \frac{n}{(1-\theta)} \\
&= \frac{n}{\theta(1-\theta)}
\end{aligned}$$

□



10.0
7.5
5.0
2.5
0.0
density
0.00
0.25
0.50
0.75
1.00
x
Distribution
normalized likelihood
posterior
prior

Non-conjugate priors

If $Y \sim Bin(n, \theta)$ and $p(\theta) = e^\theta/(e-1)$, then

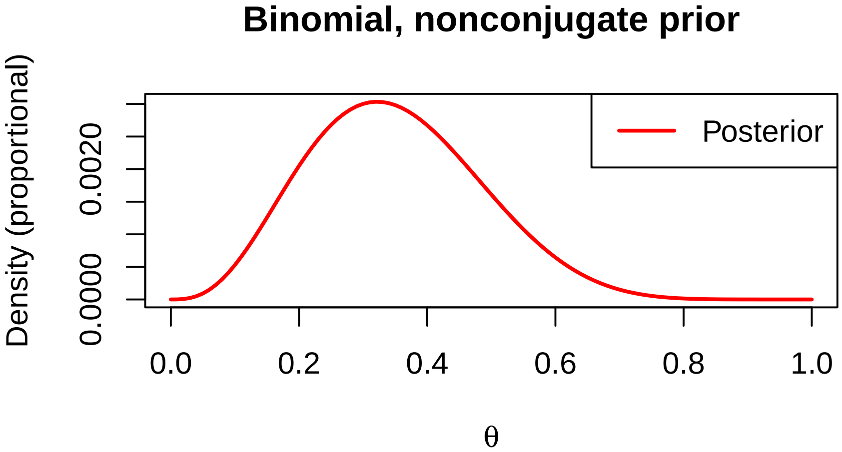
$$p(\theta|y) \propto f(\theta) = \theta^y(1-\theta)^{n-y}e^\theta$$

which is not a known distribution.

Options

- Plot $f(\theta)$ (possibly multiplying by a constant).
- Find $i = \int f(\theta)d\theta$, so that $p(\theta|y) = f(\theta)/i$.
- Evaluate $f(\theta)$ on a grid and normalize by the grid spacing.

Plot of $f(\theta)$



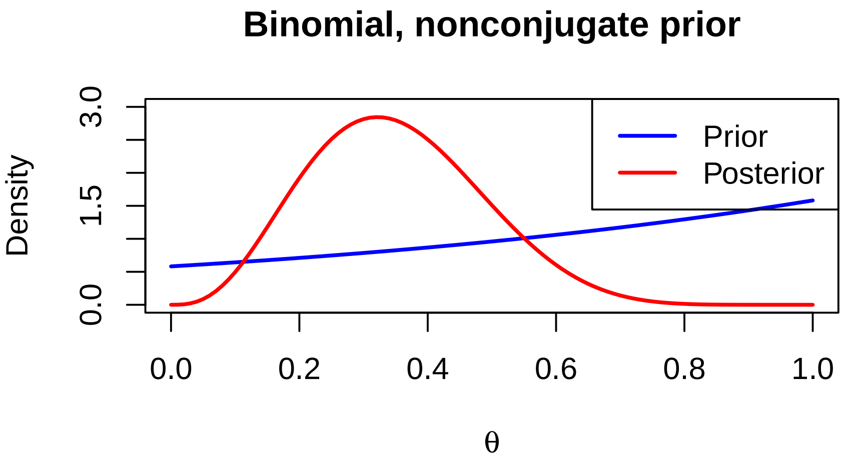
Numerical integration

Find $i = \int f(\theta)d\theta$, so that $p(\theta|y) = f(\theta)/i$.

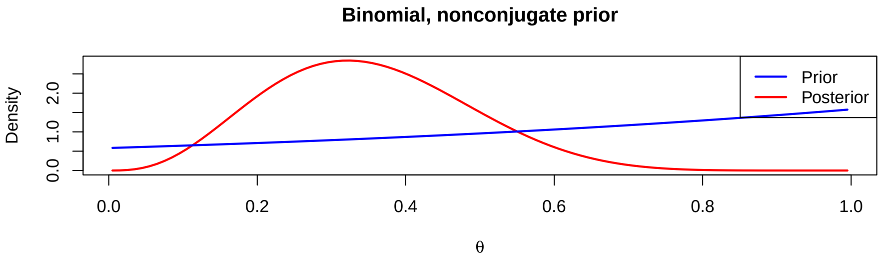
```
(i = integrate(f, 0, 1))

0.001066499 with absolute error < 1.2e-17
```

Nonconjugate prior, numerical integration



Nonconjugate prior, evaluated on a grid




```
theta[c(which(cumsum(d)*w>0.025)[1]-1, which(cumsum(d)*w>0.975)[1])] # 95\% CI

[1] 0.105 0.625
```

Improper priors

Definition

An unnormalized density, $f(\theta)$, is proper if $\int f(\theta)d\theta = c < \infty$, and otherwise it is improper.

To create a normalized density from a proper unnormalized density, use

$$p(\theta|y) = \frac{f(\theta)}{c}$$

to see that $p(\theta|y)$ is a proper normalized density note that $c = \int f(\theta)d\theta$ is not a function of θ, then

$$\int p(\theta|y)d\theta = \int \frac{f(\theta)}{\int f(\theta)d\theta} d\theta = \int \frac{f(\theta)}{c} d\theta = \frac{1}{c}\int f(\theta)d\theta = \frac{c}{c} = 1$$

Be(0,0) prior

Recall that Be(a, b) is a proper probability distribution if $a > 0, b > 0$.

Suppose $Y \sim Bin(n, \theta)$ and $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, i.e. the kernel of a $Be(0, 0)$ distribution. This is an improper distribution.

The posterior, $\theta|y \sim Be(y, n-y)$, is proper if $0 < y < n$.