



Bayes Theorem Recap

- We derive the posterior distribution via **Bayes theorem**:

$$p(\theta|y) = \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)}. \tag{1}$$

- The denominator:

$$\Pr(y) = \int \Pr(y|\theta) \times p(\theta) d\theta$$

is a **normalizing constant** to ensure the RHS of (1) integrates to 1 (we assume a continuous parameter θ).

- More colloquially:

$$\begin{aligned} \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \\ &= \Pr(y|\theta) \times p(\theta) \end{aligned}$$

since in considering the posterior we only need to worry about terms that depend on the parameter θ .



Overview of Bayesian Inference

To carry out inference, **integration** is required, and a large fraction of the Bayesian research literature focusses on this aspect. Bayesian approaches to:

1. **Estimation:** **marginal posterior distributions** on parameters of interest.
2. **Hypothesis Testing:** **Bayes factors** give the evidence in the data with respect to two or more hypotheses, and provide one approach.
3. **Prediction:** via the **predictive distribution**.

These three endeavors will now be described in the context of a **binomial model**.

Elements of Bayes Theorem for a Binomial Model

We assume independent responses with a common “success” probability θ .

In this case, the contribution of the data is through the binomial probability distribution:

$$\Pr(Y = y|\theta) = \binom{N}{y} \theta^y (1 - \theta)^{N-y} \quad (2)$$

and tells us the probability of seeing $Y = y$, $y = 0, 1, \dots, N$ given the probability θ .

For fixed y , we may view (2) as a function of θ – this is the **likelihood function**.

The **maximum likelihood estimate** (MLE) is that value

$$\hat{\theta} = y/N$$

that gives the highest probability to the observed data, i.e. maximizes the likelihood function.

The Beta Distribution as a Prior Choice for Binomial θ

- Bayes theorem requires the likelihood, which we have already specified as binomial, and the prior.
- For a probability $0 < \theta < 1$ an obvious candidate prior is the uniform distribution on $(0,1)$: but this is too restrictive in general.
- The **beta distribution**, $\text{beta}(a, b)$, is more flexible and so may be used for θ , with a and b specified **in advance**, i.e., *a priori*. The uniform distribution is a special case with $a = b = 1$.
- The form of the beta distribution is

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}$$

for $0 < \theta < 1$, where $\Gamma(\cdot)$ is the gamma function¹.

- The distribution is valid² for $a > 0, b > 0$.

¹ $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$

²A distribution is valid if it is non-negative and integrates to 1



The Beta Distribution as a Prior Choice for Binomial θ

How can we think about specifying a and b ?

For the normal distribution the parameters μ and σ^2 are just the mean and variance, but for the beta distribution a and b have no such simple interpretation.

The mean and variance are:

$$E[\theta] = \frac{a}{a+b}$$

$$\text{var}(\theta) = \frac{E[\theta](1 - E[\theta])}{a + b + 1}.$$

Hence, increasing a and/or b **concentrates** the distribution about the mean.

The quantiles, e.g. the median or the 10% and 90% points, are not available as a simple formula, but are easily obtained within software such as R using the function `qbeta(p, a, b)`.

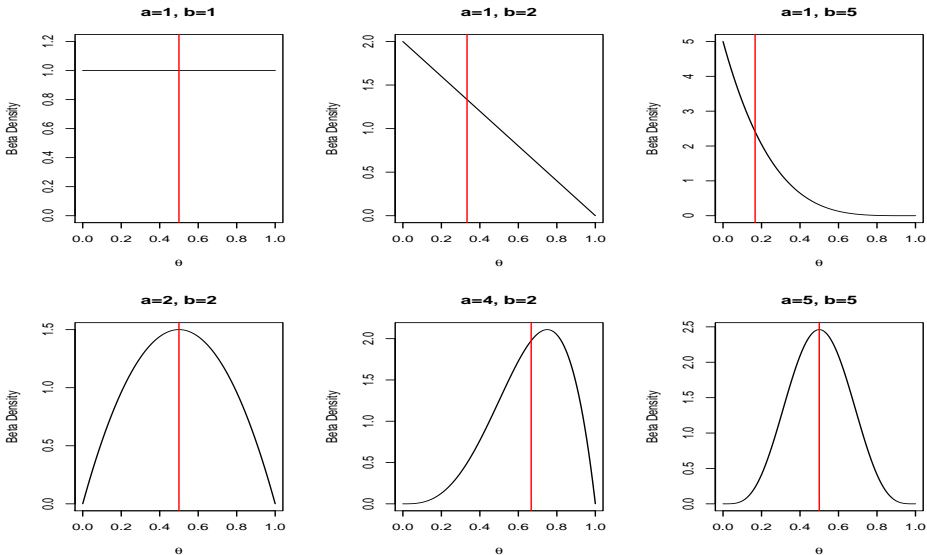


Figure 3: Beta distributions, $\text{beta}(a, b)$, the red lines indicate the means.

Samples to Summarize Beta Distributions

Probability distributions can be investigated by generating samples and then examining histograms, moments and quantiles.

In Figure 4 we show histograms of beta distributions for different choices of a and b .

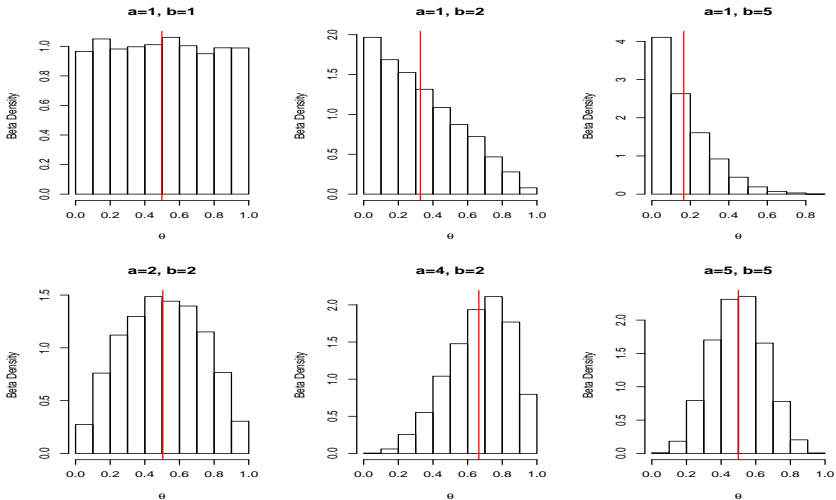


Figure 4: Random samples from beta distributions; sample means as red lines.



Samples for Describing Weird Parameters

- So far the samples we have generated have produced summaries we can easily obtain anyway.
- But what about **functions** of the probability θ , such as the odds $\theta/(1 - \theta)$?
- Once we have samples for θ we can simply **transform** the samples to the functions of interest.
- We may have clearer prior opinions about the odds, than the probability.
- The histogram representation of the prior on the odds $\theta/(1 - \theta)$ when θ is **beta(10,10)**.

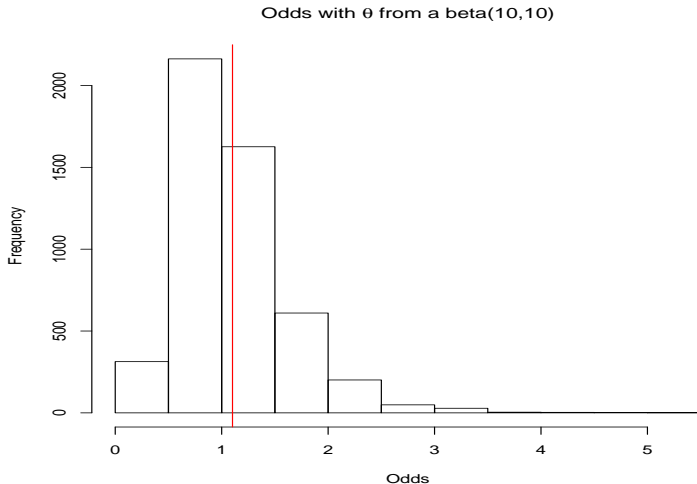


Figure 5: Samples from the prior on the odds $\theta/(1 - \theta)$ with $\theta \sim \text{beta}(10, 10)$, the red line indicates the sample mean.

Issues with Uniformity

We might think that if we have little prior opinion about a parameter then we can simply assign a **uniform prior**, i.e. a prior

$$p(\theta) \propto \text{const.}$$

There are two problems with this strategy:

- We can't be uniform on all scales since, if $\phi = g(\theta)$:

$$\underbrace{p_\phi(\phi)}_{\text{Prior for } \phi} = \underbrace{p_\theta(g^{-1}(\phi))}_{\text{Prior for } \theta} \times \underbrace{\left| \frac{d\theta}{d\phi} \right|}_{\text{Jacobian}}$$

and so if $g(\cdot)$ is a nonlinear function, the Jacobian will be a function of ϕ and hence not uniform.

- If the parameter is not on a finite range, an **improper** distribution will result (that is, the form will not integrate to 1). This can lead to an improper posterior distribution, and without a proper posterior we can't do inference.



Are Priors Really Uniform?

- We illustrate the first (non-uniform on all scales) point.
- In the binomial example a uniform prior for θ seems a natural choice.
- But suppose we are going to model on the logistic scale so that

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

is a quantity of interest.

- A uniform prior on θ produces the very non-uniform distribution on ϕ in Figure 6.
- Not being uniform on all scales is not necessarily a problem, and is correct probabilistically, but one should be aware of this characteristic.

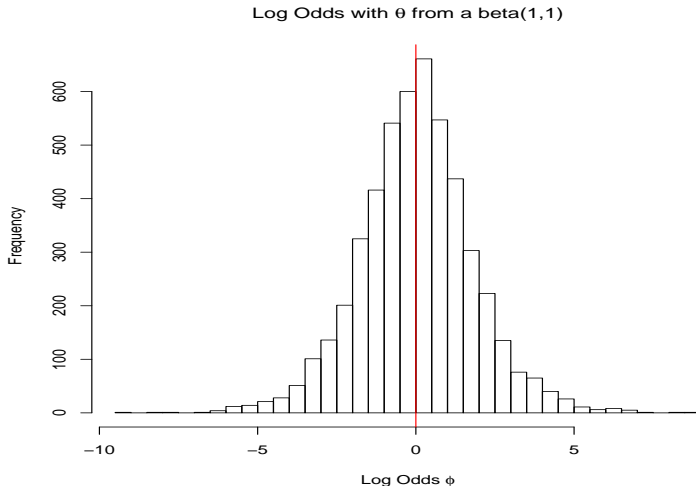


Figure 6: Samples from the prior on the odds $\phi = \log[\theta/(1 - \theta)]$ with $\theta \sim \text{beta}(1, 1)$, the red line indicates the sample mean.

Posterior Derivation: The Quick Way

- When we want to identify a particular probability distribution we **only** need to concentrate on terms that involve the random variable.
- For example, if the random variable is X and we see a density of the form

$$p(x) \propto \exp(c_1 x^2 + c_2 x),$$

for constants c_1 and c_2 , then we **know** that the random variable X **must** have a normal distribution.

Posterior Derivation: The Quick Way

- For the binomial-beta model we concentrate on terms that only involve θ .
- The **posterior** is

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

- We recognize this as the important part of a **beta($y + a$, $N - y + b$)** distribution.
- We know what the **normalizing constant** must be, because we have a distribution which must integrate to 1.

Posterior Derivation: The Long (Unnecessary) Way

- The posterior can also be calculated by keeping in all the normalizing constants:

$$\begin{aligned}
 p(\theta|y) &= \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)} \\
 &= \frac{1}{\Pr(y)} \binom{N}{y} \theta^y (1-\theta)^{N-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}. \quad (3)
 \end{aligned}$$

- The normalizing constant is

$$\begin{aligned}
 \Pr(y) &= \int_0^1 \Pr(y|\theta) \times p(\theta) d\theta \\
 &= \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{y+a-1} (1-\theta)^{N-y+b-1} d\theta \\
 &= \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}
 \end{aligned}$$

- The integrand on line 2 is a beta($y+a$, $N-y+b$) distribution, up to a normalizing constant, and so we know what this constant has to be.

Posterior Derivation: The Long (and Unnecessary) Way

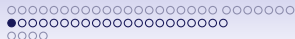
- The normalizing constant is therefore:

$$\Pr(y) = \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}$$

- This is a probability distribution, i.e. $\sum_{y=0}^N \Pr(y) = 1$ with $\Pr(y) > 0$.
- For a particular y value, this expression tells us the probability of that value **given** the model, i.e. the likelihood and prior we have selected: this will reappear later in the context of **hypothesis testing**.
- Substitution of $\Pr(y)$ into (3) and canceling the terms that appear in the numerator and denominator gives the posterior:

$$p(\theta|y) = \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \theta^{y+a-1} (1-\theta)^{N-y+b-1}$$

which is a **beta**($y+a, N-y+b$).



The Posterior Mean: A Summary of the Posterior

- Recall the mean of a beta(a, b) is $a/(a + b)$.
- The posterior mean of a beta($y + a, N - y + b$) is therefore

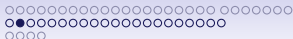
$$\begin{aligned}
 E[\theta|y] &= \frac{y + a}{N + a + b} \\
 &= \frac{y}{N + a + b} + \frac{a}{N + a + b} \\
 &= \frac{y}{N} \times \frac{N}{N + a + b} + \frac{a}{a + b} \times \frac{a + b}{N + a + b} \\
 &= \text{MLE} \times W + \text{Prior Mean} \times (1-W).
 \end{aligned}$$

- The **weight** W is

$$W = \frac{N}{N + a + b}.$$

- As N increases, the weight tends to 1, so that the posterior mean gets closer and closer to the MLE.
- Notice that the **uniform** prior $a = b = 1$ gives a posterior mean of

$$E[\theta|y] = \frac{y + 1}{N + 2}.$$



The Posterior Mode

- First, note that the mode of a beta(a, b) is

$$\text{mode}(\theta) = \frac{a - 1}{a + b - 2}.$$

- As with the posterior mean, the posterior mode takes a weighted form:

$$\begin{aligned} \text{mode}(\theta|y) &= \frac{y + a - 1}{N + a + b - 2} \\ &= \frac{y}{N} \times \frac{N}{N + a + b - 2} + \frac{a - 1}{a + b - 2} \times \frac{a + b - 2}{N + a + b - 2} \\ &= \text{MLE} \times W^* + \text{Prior Mode} \times (1 - W^*). \end{aligned}$$

- The **weight** W^* is

$$W^* = \frac{N}{N + a + b - 2}.$$

- Notice that the **uniform** prior $a = b = 1$ gives a posterior mode of

$$\text{mode}(\theta|y) = \frac{y}{N},$$

the MLE. Which makes sense, right?

Other Posterior Summaries

- We will rarely want to report a point estimate alone, whether it be a posterior mean or posterior median.
- Interval estimates are obtained in the obvious way.
- A simple way of performing testing of particular parameter values of interest is via examination of interval estimates.
- For example, does a 95% interval contain the value $\theta_0 = 0.5$?

Other Posterior Summaries

- In our beta-binomial running example, a 90% posterior **credible interval** (θ_L, θ_U) results from the points

$$0.05 = \int_0^{\theta_L} p(\theta|y) d\theta$$

$$0.95 = \int_0^{\theta_U} p(\theta|y) d\theta$$

- The quantiles of a beta are not available in closed form, but easy to evaluate in R:

```
y <- 7; N <- 10; a <- b <- 1
qbeta(c(0.05, 0.5, 0.95), y+a, N-y+b)
[1] 0.4356258 0.6761955 0.8649245
```

- The 90% credible interval is (0.44,0.86) and the posterior median is 0.68.

Choosing a Prior, Approach One

- To select a beta, we need to specify two quantities, a and b .
- The posterior mean is

$$E[\theta|y] = \frac{y + a}{N + a + b}.$$

- Viewing the denominator as a **sample size** suggests a method for choosing a and b within the prior.
- We need to specify two numbers, but rather than a and b , which are difficult to interpret, we may specify the mean $m_{\text{prior}} = a/(a + b)$ and the prior sample size $N_{\text{prior}} = a + b$
- We then solve for a and b via

$$a = N_{\text{prior}} \times m_{\text{prior}}$$

$$b = N_{\text{prior}} \times (1 - m_{\text{prior}}).$$

- Intuition:** a is like a prior number of successes and b like the prior number of failures.

An Example

- Suppose we set $N_{\text{prior}} = 5$ and $m_{\text{prior}} = \frac{2}{5}$.
- It is **as if** we saw 2 successes out of 5.
- Suppose we obtain data with $N = 10$ and $\frac{y}{N} = \frac{7}{10}$.
- Hence $W = 10/(10 + 5)$ and

$$\begin{aligned} E[\theta|y] &= \frac{7}{10} \times \frac{10}{10+5} + \frac{2}{5} \times \frac{5}{10+5} \\ &= \frac{9}{15} = \frac{3}{5}. \end{aligned}$$

- Solving:

$$a = N_{\text{prior}} \times m_{\text{prior}} = 5 \times \frac{2}{5} = 2$$

$$b = N_{\text{prior}} \times (1 - m_{\text{prior}}) = 5 \times \frac{3}{5} = 3$$

- This gives a $\text{beta}(y + a, N - y + b) = \text{beta}(7 + 2, 3 + 3)$ posterior.

Choosing a Prior, Approach Two

- An alternative convenient way of choosing a and b is by specifying **two quantiles** for θ with associated (prior) probabilities.
- For example, we may wish $\Pr(\theta < 0.1) = 0.05$ and $\Pr(\theta > 0.6) = 0.05$.
- The values of a and b may be found numerically.
- For example, we may solve

$$[p_1 - \Pr(\theta < q_1 | a, b)]^2 + [p_2 - \Pr(\theta < q_2 | a, b)]^2 = 0 \quad (4)$$

for a, b .

Bayesian Sequential Updating

- We show how probabilistic beliefs are updated as we receive more data.
- Suppose the data arrives sequentially via two experiments:
 1. Experiment 1: (y_1, N_1) .
 2. Experiment 2: (y_2, N_2) .
- **Prior 1:** $\theta \sim \text{beta}(a, b)$.
- **Likelihood 1:** $y_1 | \theta \sim \text{binomial}(N_1, \theta)$.
- **Posterior 1:** $\theta | y_1 \sim \text{beta}(a + y_1, b + N_1 - y_1)$.
- This posterior forms the prior for experiment 2.
- **Prior 2:** $\theta \sim \text{beta}(a^*, b^*)$ where $a^* = a + y_1$, $b^* = b + N_1 - y_1$.
- **Likelihood 2:** $y_2 | \theta \sim \text{binomial}(N_2, \theta)$.
- **Posterior 2:** $\theta | y_1, y_2 \sim \text{beta}(a^* + y_2, b^* + N_2 - y_2)$.
- Substituting for a^*, b^* :

$$\theta | y_1, y_2 \sim \text{beta}(a + y_1 + y_2, b + N_1 - y_1 + N_2 - y_2).$$

Bayesian Sequential Updating

- Schematically:

$$(a, b) \rightarrow (a + y_1, b + N_1 - y_1) \rightarrow (a + y_1 + y_2, b + N_1 - y_1 + N_2 - y_2)$$

- Suppose we obtain the data in one go as $y^* = y_1 + y_2$ successes from $N^* = N_1 + N_2$ trials.
- The posterior is

$$\theta|y^* \sim \text{beta}(a + y^*, b + N^* - y^*),$$

which is the same as when we receive in two separate instances.

Predictive Distribution

- Suppose we see y successes out of N trials, and now wish to obtain a **predictive distribution** for a future experiment with M trials.
- Let $Z = 0, 1, \dots, M$ be the number of successes.
- Predictive distribution:

$$\begin{aligned} \Pr(z|y) &= \int_0^1 p(z, \theta|y) d\theta \\ &= \int_0^1 \Pr(z|\theta, y) p(\theta|y) d\theta \\ &= \int_0^1 \underbrace{\Pr(z|\theta)}_{\text{binomial}} \times \underbrace{p(\theta|y)}_{\text{posterior}} d\theta \end{aligned}$$

where we move between lines 2 and 3 because z is **conditionally independent** of y given θ .

Difference in Binomial Proportions

- It is straightforward to extend the methods presented for a single binomial sample to a pair of samples.
- Suppose we carry out two binomial experiments:

$$Y_1 | \theta_1 \sim \text{binomial}(N_1, \theta_1) \quad \text{for sample 1}$$

$$Y_2 | \theta_2 \sim \text{binomial}(N_2, \theta_2) \quad \text{for sample 2}$$

- Interest focuses on $\theta_1 - \theta_2$, and often in examining the possibility that $\theta_1 = \theta_2$.
- With a sampling-based methodology, and independent beta priors on θ_1 and θ_2 , it is straightforward to examine the posterior $p(\theta_1 - \theta_2 | y_1, y_2)$.

Calibration of Bayes Factors

- Kass and Raftery (1995) suggest **intervals** of Bayes factors for reporting:

1/Bayes Factor	Evidence Against H_0
1 to 3.2	Not worth more than a bare mention
3.2 to 20	Positive
20 to 150	Strong
>150	Very strong

- These provide a guideline, but should not be followed without question.

Three Approaches to Inference for the ASE Data

1. Posterior Probabilities:

- A simple approach to testing is to calculate the posterior probability that $\theta < 0.5$.
- We can then pick a threshold for indicating worthy of further study, e.g. if $\Pr(\theta < 0.5|y) < 0.01$ or $\Pr(\theta < 0.5|y) > 0.99$

2. Bayes Factors:

- Calculating the Bayes factor.
- Pick a threshold for indicating worthy of further study, e.g. if reciprocal of the Bayes factor is greater than 150.

3. Decision theory:

- Place priors on the null and alternative hypotheses.
- Calculate the posterior odds:

$$\frac{\Pr(H_0|y)}{\Pr(H_1|y)} = \frac{\Pr(y|H_0)}{\Pr(y|H_1)} \times \frac{\Pr(H_0)}{\Pr(H_1)}$$

$$\text{Posterior Odds} = \text{Bayes Factor} \times \text{Prior Odds}$$

- Pick a threshold R , so that if the Posterior Odds $< R$ we choose H_1 .

Bayesian Analysis of the ASE Data

- In Figure 13 we give a histogram of the posterior probabilities $\Pr(\theta < 0.5|y)$ and we see large numbers of genes have probabilities close to 0 and 1, indicating allele specific expression (ASE).
- In Figure 14 we plot $\Pr(\theta < 0.5|y)$ versus the p-values and the general pattern is what we would expect — small p-values have posterior probabilities close to 0 and 1.
- The strange lines in this plot are due to the discreteness of the outcome y .
- In Figure 15 we plot the -Log Bayes Factor against $\Pr(\theta < 0.5|y)$. Large values of the former correspond to strong evidence of ASE; again we see an agreement in inference, with large values of the negative log Bayes factor corresponding with $\Pr(\theta < 0.5|y)$ close to 0 and 1.

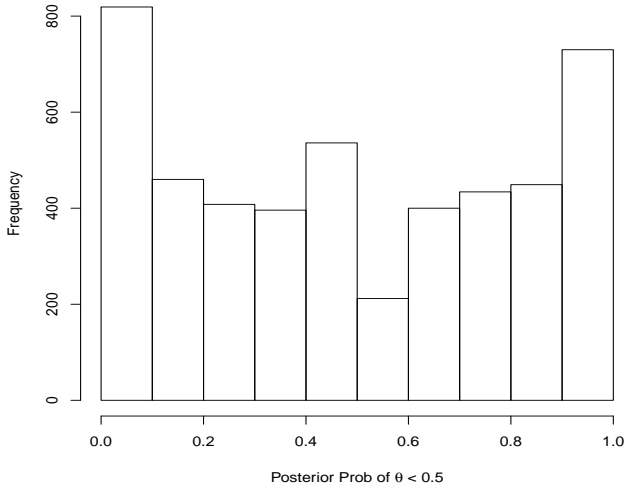


Figure 13: Histogram of 4,844 posterior probabilities of $\theta < 0.5$.



ASE Output Data

- Below are some summaries from the ASE analysis – we order with respect to the variable $\log\text{BFr}$, which is the reciprocal Bayes factor (so that high numbers correspond to strong evidence against the null).
- The `postprob` variable is the posterior probability of $\theta < 0.5$.

```
allvals <- data.frame(Nsum, ysum, pvals, postprob, logBFr)
oBF <- order(-logBFr)
orderallvals <- allvals[oBF,]
head(orderallvals)
  Nsum ysum pvals postprob logBFr
4751  437   6 5.340324e-119 1.000000e+00 267.9572
4041  625  97 1.112231e-72 1.000000e+00 161.1355
2370  546 468 8.994944e-69 2.621622e-69 152.2517
2770  256 245 1.127211e-58 2.943484e-59 129.6198
tail(orderallvals)
  Nsum ysum pvals postprob logBFr
824   761  382 0.9422103 0.4567334 -2.086604
2163  776  390 0.9142477 0.4429539 -2.091955
3153  769  384 1.0000000 0.5143722 -2.097079
2860 1076  546 0.6474878 0.3129473 -2.146555
```