Bayesian Statistics: A Primer for Data Scientists - PART II

antonietta.mira@usi.ch

Thanks to: Alan Agresti and Maria Kateri

PLS DO NOT CIRCULATE THESE SLIDES

Beta Family of priors

Bayesian priors for π are defined over the interval [0, 1].
A common choice is the ***beta distribution***.
With hyperparameters $\alpha_1 > 0$ and $\alpha_2 > 0$, it has *pdf*

$$p(\pi \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1 - 1}(1 - \pi)^{\alpha_2 - 1}, \quad 0 \le \pi \le 1, \qquad (1)$$

which we denote by $\text{Beta}(\alpha_1, \alpha_2)$.
The initial term with the gamma functions provides the appropriate constant so that the *pdf* integrates to 1.

Whenever we see a function of x of the form $x^a(1 - x)^b$ for x in the interval [0, 1] and with exponents that exceed -1, it is the kernel of a beta *pdf*, in the sense that we obtain a beta *pdf* by multiplying it by a certain constant.

Beta distribution

The family of beta *pdf*s has a wide variety of shapes, and this makes it a versatile family of distributions that is able to provide a functional form for a wide range of prior informations.

When $\alpha_1 = \alpha_2$, the beta *pdf* is symmetric around $\mu = 0.50$.
The *uniform distribution* over [0, 1] results when $\alpha_1 = \alpha_2 = 1$.

The *pdf* has a bimodal U-shape when $\alpha_1 = \alpha_2 < 1$
and a bell shape when $\alpha_1 = \alpha_2 > 1$.

See Figure 1.

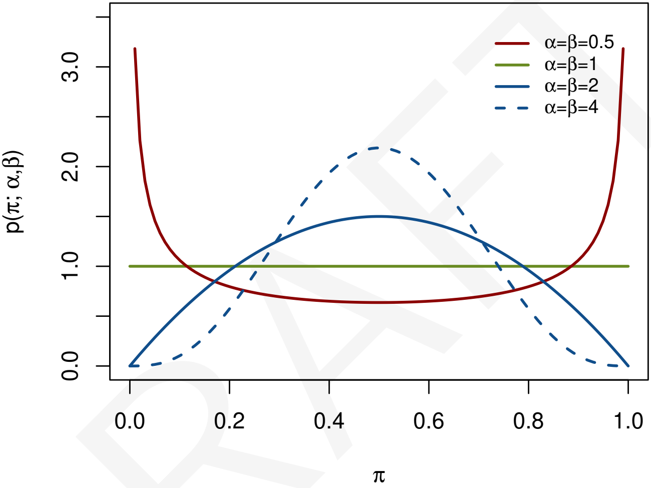


Figure: The *pdf* of a beta distribution with hyperparameters $\alpha_1 = \alpha_2$ is symmetric, with variance decreasing as α_1 increases.

The mean of the Beta(α_1, α_2) distribution is

$$\mu = E(\pi) = \frac{\alpha_1}{\alpha_1 + \alpha_2}. \tag{2}$$

The *pdf* is unimodal skewed to the left when $\alpha_1 > \alpha_2 > 1$, in which case $\mu > 0.50$, and it is skewed to the right when $\alpha_2 > \alpha_1 > 1$, in which case $\mu < 0.50$.

The mode in these cases is $(\alpha_1 - 1)/(\alpha_1 + \alpha_2 - 2)$.

Beta distribution

The variance of the beta distribution is

$$\sigma^2 = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)} = \frac{\mu(1 - \mu)}{\alpha_1 + \alpha_2 + 1}. \tag{3}$$

From the second expression, for a fixed value of $\alpha_1 + \alpha_2$, the variance decreases as the mean μ approaches 0 or 1;
for a fixed value of μ, the variance decreases as $\alpha_1 + \alpha_2$ increases.

When $\alpha_1 = \alpha_2 = \alpha$, $\mu = 1/2$ and $\sigma^2 = 1/[4(2\alpha + 1)]$, decreasing as α increases.

Jeffreys prior

The Beta(0.5, 0.5) prior is called the ***Jeffreys*** prior for π.

The name refers to University of Cambridge professor Harold Jeffreys,[1] who proposed this prior in 1946.

For a single parameter θ, the Jeffreys prior is the one that is proportional to $\sqrt{I(\theta)}$ for the ***information*** $I(\theta)$ for a single observation,

$$I(\theta) = E\left(\frac{\partial \log f(Y \,|\, \theta)}{\partial \theta}\right)^2,$$

where the expectation is taken with respect to Y for fixed θ.

[1] In 1939 Jeffreys published a foundational book on the Bayesian statistical approach, *Theory of Probability*, and he also proposed the Bayes factor

A Bernoulli random variable with $\pi = P(Y = 1)$ and $1 - \pi = P(Y = 0)$ has probability mass function that is the binomial formula for $n = 1$, $f(y \mid \pi) = \pi^y (1 - \pi)^{1-y}$ for $y = 0, 1$.

For it,

$$E(Y) = \pi \quad \text{and} \quad \text{var}(Y) = E(Y - \pi)^2 = \pi(1 - \pi),$$

$$\log f(y \mid \pi) = y \log(\pi) + (1 - y) \log(1 - \pi)$$

and

$$\frac{\partial}{\partial \pi}[\log f(y \mid \pi)] = \frac{y}{\pi} - \frac{1 - y}{1 - \pi} = \frac{y - \pi}{\pi(1 - \pi)},$$

so that

$$I(\pi) = E\left[\frac{\partial \log f(Y \mid \pi)}{\partial \pi}\right]^2 = E\left[\frac{Y - \pi}{\pi(1 - \pi)}\right]^2 = \frac{E(Y - \pi)^2}{[\pi(1 - \pi)]^2}$$

$$I(\pi) = \frac{\pi(1 - \pi)}{[\pi(1 - \pi)]^2} = \frac{1}{\pi(1 - \pi)}.$$

The Jeffreys prior in this case is proportional to $\sqrt{I(\pi)} = \sqrt{1/[\pi(1 - \pi)]} = \pi^{-1/2}(1 - \pi)^{-1/2}$, which is the kernel of a beta *pdf* with $\alpha_1 = \alpha_2 = 0.5$.

When used with a one-dimensional parameter θ, the Jeffreys prior is an example of a noninformative prior called the ***reference prior***, which is the prior that has the least possible influence on the posterior in a certain sense.[2]

A reference prior yields posterior inferential results that depend almost entirely on the likelihood function and are similar to those of good frequentist methods.

[2] It maximizes the expected value of a distance measure, called *Kullback–Leibler divergence*, between the prior and the posterior distribution, $KL = \int \log[g(\theta \mid \boldsymbol{y})/p(\theta)]g(\theta \mid \boldsymbol{y})d\theta$.

Another rationale of Jeffreys in proposing the formula for this prior is that it is **invariant** to the parameterization: It provides equivalent results when applied for different scales of measurement for the parameter.

For example, the probability for a particular interval (a, b) of values of π is the same when we find it using the Jeffreys Beta(0.5, 0.5) prior for π by integrating over (a, b) or when we find it using the Jeffreys prior for $\phi = \log[\pi/(1-\pi)]$ (which is the *logit* parameter used in modeling binary response variables) and integrate over the interval $(\log[a/(1-a)], \log[b/(1-b)])$.

In the context of Bayesian inference about a binomial parameter π, Bayes' theorem states that the posterior $g(\pi \mid y)$ for π uses the binomial probability mass function $f(y \mid \pi; n)$ for the number of successes in n trials and the beta prior *pdf* $p(\pi \mid \alpha_1, \alpha_2)$ through

$$g(\pi \mid y) = \frac{f(y \mid \pi; n) p(\pi \mid \alpha_1, \alpha_2)}{f(y)}.$$

As explained earlier the numerator product determines the posterior distribution, because the denominator is the marginal probability function of the data and does not involve the parameter π.

That is,

$$\begin{aligned} g(\pi \mid y) \propto & \quad [\pi^y (1-\pi)^{n-y}][\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}] \\ = & \quad \pi^{y+\alpha_1-1}(1-\pi)^{n-y+\alpha_2-1}, \quad 0 \le \pi \le 1. \end{aligned}$$

This posterior also is a beta distribution, once we multiply by the appropriate gamma functions so it integrates to 1.

This beta distribution is indexed by hyperparameter values $\alpha_1^* = y + \alpha_1$ and $\alpha_2^* = n - y + \alpha_2$.

That is, the posterior falls in the same family of probability distributions as the prior, but its hyperparameters are updated, based on the data.

The chosen prior is a ***conjugate prior***.

Conjugate prior

A prior such that the posterior comes from the same family when combined with a certain likelihood function is called a ***conjugate prior*** for that likelihood function.

For binomial sampling and its corresponding likelihood function, the beta distribution is the conjugate prior.

When a conjugate prior exists, an advantage is the explicit form generated for the posterior distribution.

However, conjugate priors are usually not available for more complex models, in which case we can use simulation methods to approximate the posterior distribution.

Bayesian Point Estimation Using Posterior Mean

The most common Bayesian point estimate of a parameter is the **mean of its posterior** distribution.

We have seen that this estimate results from minimizing a squared-error *loss function*.

For the beta posterior distribution, this estimate of π is

$$\begin{aligned}\tilde{\pi} = E(\pi \mid y) &= \frac{\alpha_1^*}{\alpha_1^* + \alpha_2^*} = \frac{y+\alpha_1}{(y+\alpha_1)+(n-y+\alpha_2)} = \frac{y+\alpha_1}{n+\alpha_1+\alpha_2} \\ &= \left(\frac{n}{n+\alpha_1+\alpha_2}\right)\frac{y}{n} + \left(\frac{\alpha_1+\alpha_2}{n+\alpha_1+\alpha_2}\right)\frac{\alpha_1}{\alpha_1+\alpha_2}. \qquad (4)\end{aligned}$$

For a binomial outcome y for a number of successes, y/n is the sample proportion, $\hat{\pi}$, which is also the MLE of π.

The posterior mean $\tilde{\pi} = E(\pi \mid y)$ is a weighted average of the sample proportion y/n and the mean of the prior, $\alpha_1/(\alpha_1+\alpha_2)$.

When $\alpha_1 = \alpha_2$, the Bayesian estimate shrinks the sample proportion toward 0.50.

The amount of information in the posterior mean estimate $\tilde{\pi}$ is summarized by n for the data and $\alpha_1 + \alpha_2$ for the beta prior.

The effect of the prior is to add $\alpha_1 + \alpha_2$ **imaginary observations**, of which α_1 are successes.

In this weighted average, the weight $n/(n + \alpha_1 + \alpha_2)$ given to the sample proportion increases toward 1 as n increases.

By contrast, as we take **larger values for $\alpha_1 + \alpha_2$** for fixed n, the **influence of the prior on the posterior mean estimate increases**.

Example: Bayes Estimates with Three priors and the Same Data

In this example (based on a 1962 letter exchange between Leonard Jimmy Savage and Jerome Cornfield) we find Bayesian posterior mean estimates of the probability π of success for each of three binomial experiments that have the same data, $y = 10$ successes in $n = 10$ trials:

(1) **A British woman** claims that in tasting a cup of tea with milk, she can tell whether the milk was poured before or after the tea;

(2) **A professor** of 18th century musicology claims to be able to tell for any pair of pages of music, one composed by Mozart and one by Haydn, who composed each.

(3) A **person in a drunken state** claims to be able to predict whether the flip of a balanced coin will result in a head or in a tail.

In each case, the frequentist MLE of the probability π of a correct prediction is $\hat{\pi} = y/n = 10/10 = 1.0$.

We shall next find Bayesian posterior mean estimates of π for each situation.

For the **tea taster**, if we have no *a priori* reason to expect any particular value for π, we could select a **uniform prior**, which is the Beta(1, 1) distribution.

From equations (2) and (3), this prior has mean 0.50 and standard deviation $1/\sqrt{12} = 0.29$.

By contrast, for **coin flips** we may be highly skeptical of any claim about predictions and decide to use a beta prior with **mean 0.50 but with small standard deviation, say 0.03**, so that the beta distribution is highly concentrated near 0.50.

We might have greater faith in the **musicologist's** claim, and use a beta prior with a **mean of 0.90 and standard deviation 0.10**.

For a chosen mean μ, from the formula $\mu = \alpha_1/(\alpha_1 + \alpha_2)$ for a beta distribution, any particular value α_1 has a corresponding $\alpha_2 = \alpha_1(1 - \mu)/\mu$.

By trial and error for a particular value of μ we can find a pair of (α_1, α_2) values that have a particular standard deviation σ, or we could solve simultaneously for (α_1, α_2) based on (μ, σ).

The following `R` code shows that for a beta prior with $(\mu, \sigma) =$ (0.50, 0.03) for the coin-tossing predictions, we can take hyperparameter values $(\alpha_1, \alpha_2) \approx$ (138, 138)

For $(\mu, \sigma) =$ (0.90, 0.10) for the musicologist, we can take $(\alpha_1, \alpha_2) \approx$ (7, 7/9).

```
alpha1 = 138; alpha2 = 138
        # beta hyperparameters for predictor of coin flips

alpha1/(alpha1+alpha2);
sqrt((alpha1*alpha2)/((alpha1+alpha2)^2 * (alpha1 + alpha2 + 1)))

[1] 0.5          # mean of beta prior for coin flips

[1] 0.03004209          # std deviation of beta prior for coin flips
==========================================

alpha1 = 7; alpha2 = 7/9
        # beta hyperparameters for musicologist

alpha1/(alpha1+alpha2);
sqrt((alpha1*alpha2)/((alpha1+alpha2)^2 * (alpha1 + alpha2 + 1)))

[1] 0.9          # mean of beta prior musicologist

[1] 0.1012579          # std deviation of beta prior musicologist
```

With $y = 10$ successes in $n = 10$ trials, the posterior beta distribution has hyperparameter values $\alpha_1^* = y + \alpha_1 = 10 + \alpha_1$ and $\alpha_2^* = n - y + \alpha_2 = 10 - 10 + \alpha_2 = \alpha_2$.

The posterior mean is
$\tilde{\pi} = \alpha_1^*/(\alpha_1^* + \alpha_2^*) = (10 + \alpha_1)/(10 + \alpha_1 + \alpha_2)$.

We now find $\tilde{\pi}$ for the three chosen priors:

```
> alpha1post <- 10 + alpha1; alpha2post <- alpha2
       # hyperpara's for posterior when y=n=10
> alpha1post/(alpha1post + alpha2post)
       # beta posterior mean
[1] 0.9166667
       # for British tea taster
> qbeta(0.5, alpha1post, alpha2post)
[1] 0.9389309
       # posterior median for British tea taster
```

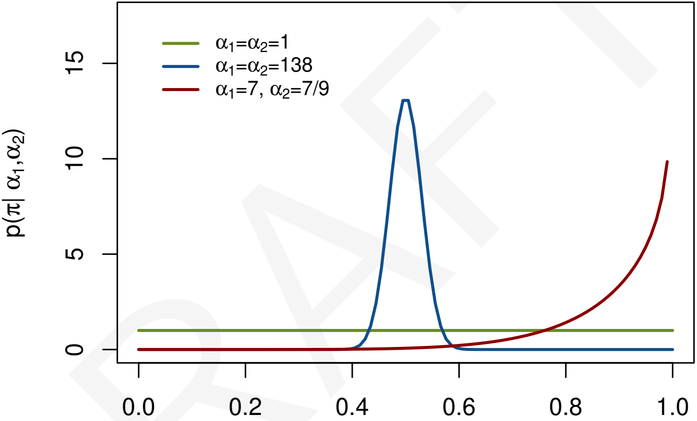
```
> alpha1 = 138; alpha2 = 138
# beta hyperparameters with mean 0.50, std dev. 0.03
> alpha1post <- 10 + alpha1; alpha2post <- alpha2
# hyperpara's for posterior when y=n=10
> alpha1post/(alpha1post + alpha2post)
       # posterior mean for coin flips
[1] 0.5174825
```

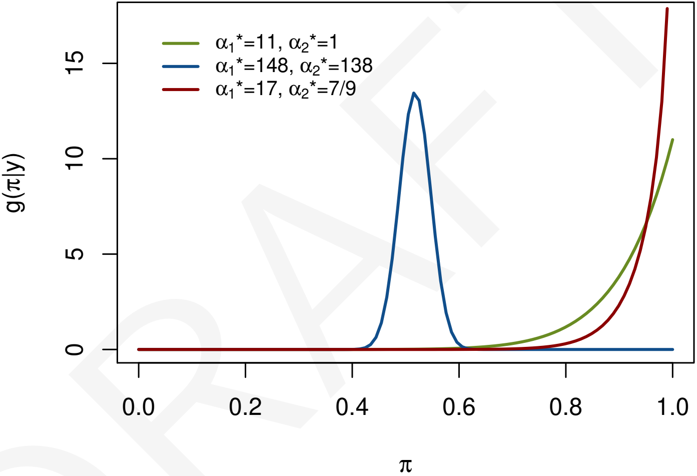
```
> alpha1 = 7; alpha2 = 7/9
# hyperparameters with mean 0.90, standard dev. 0.10
> alpha1post <- 10 + alpha1; alpha2post <- alpha2
       # hyperpara's for posterior when y=n=10
> alpha1post/(alpha1post + alpha2post)
       # posterior mean for musicologist
[1] 0.95625
> qbeta(0.5, alpha1post, alpha2post)
[1] 0.971962

       # posterior median for musicologist
```

With these priors, the posterior mean estimates of π are:
0.917 for the tea taster and
0.956 for the musicologist but only
0.517 for the predictor of coin flips.

The posterior mean for coin flipping is very close to the prior mean of 0.50, the data having relatively little influence when the prior is so sharp.
Figure 2 shows the beta prior and posterior distributions for these three cases.

The posterior distributions are highly skewed for the tea taster and the musicologist, and the code shows that the posterior medians are even closer to 1.0.


α_1=α_2=1
α_1=α_2=138
α_1=7, α_2=7/9
$p(\pi| \alpha_1,\alpha_2)$
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0


α_1*=11, α_2*=1
α_1*=148, α_2*=138
α_1*=17, α_2*=7/9
$g(\pi|y)$
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
π

Bayesian Updating:
Posterior Becomes Prior for Future Data

Bayesian analyses can easily be updated as a data scientist obtains additional data.

Suppose an initial study having a prior $p(\pi \mid \alpha_1, \alpha_2)$ for a binomial parameter π has data $\boldsymbol{y}_1$ and yields the posterior $g(\pi \mid \boldsymbol{y}_1)$.

That posterior can then serve as a prior to combine with additional data in the future about the same parameter.

Let $\ell(\pi \mid \boldsymbol{y}_1)$ denote the likelihood function for the initial study.

The posterior is proportional to the product of the likelihood function and the prior, $g(\pi \mid \boldsymbol{y}_1) \propto p(\pi \mid \alpha_1, \alpha_2)\ell(\pi \mid \boldsymbol{y}_1)$.

Let $\ell(\pi \mid \boldsymbol{y}_2)$ denote the likelihood function for a later study with an independent sample $\boldsymbol{y}_2$.

Using the posterior from the initial study as the prior for the new study,

$$g(\pi \mid \boldsymbol{y}_1, \boldsymbol{y}_2) \propto g(\pi \mid \boldsymbol{y}_1)\ell(\pi \mid \boldsymbol{y}_2) \propto p(\pi \mid \alpha_1, \alpha_2)\ell(\pi \mid \boldsymbol{y}_1)\ell(\pi \mid \boldsymbol{y}_2).$$

Now $\ell(\pi \mid \boldsymbol{y}_1)\ell(\pi \mid \boldsymbol{y}_2)$ is the likelihood function $\ell(\pi \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$ for the two studies combined.

Therefore $g(\pi \mid \boldsymbol{y}_1, \boldsymbol{y}_2) \propto p(\pi \mid \alpha_1, \alpha_2)\ell(\pi \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$ is the same as the posterior obtained using the initial prior with the data together from both studies.

That posterior could in turn serve as a prior for future studies.

The example in the previous section used a uniform prior, Beta(1, 1), for the British tea taster who claimed to be able to detect whether tea or milk was poured first in the cup.

With $n_1 = 10$ taste trials and binary observations $\{y_{i1} = 1,\ i = 1, \ldots, 10\}$ having $y_1 = \sum_{i=1}^{10} y_{i1} = 10$ successful guesses, we found in the discussion of that example that the posterior is beta with $\alpha_1^* = y_1 + \alpha_1 = 10 + 1 = 11$ and $\alpha_2^* = n_1 - y_1 + \alpha_2 = 10 - 10 + 1 = 1$, for which the posterior mean estimate for π is $\tilde{\pi} = \alpha_1^*/(\alpha_1^* + \alpha_2^*) = 11/(11 + 1) = 0.917$.

This posterior now serves as the prior for the next set of trials.

If in this set, she tastes $n_2 = 5$ more cups and has outcomes $\{y_{i2}\}$ with $y_2 = \sum_{i=1}^{5} y_{i2} = 4$ successful guesses, the new posterior has hyperparameters $\alpha_1^{**} = y_2 + \alpha_1^* = 4 + 11 = 15$ and $\alpha_2^{**} = n_2 - y_2 + \alpha_2^* = 5 - 4 + 1 = 2$.

This results in a second-stage posterior mean estimate of $\alpha_1^{**}/(\alpha_1^{**} + \alpha_2^{**}) = 15/17 = 0.882$, compared with the original prior mean of $\alpha_1/(\alpha_1 + \alpha_2) = 0.50$ and the first posterior mean of 0.917.

The second-stage posterior mean is a bit lower than the first-stage one because the second tea-testing experiment had a lower success rate than the first experiment.

The second-stage posterior mean is the same as we would have obtained if the two experiments had been conducted in a single stage, with 15 cups tested and a cumulative success total of 14 combined with a Beta(1,1) prior.

Bayesian Inference for a Proportion

Now that we have learned more about beta priors for a binomial parameter π and found the corresponding beta distribution for the posterior distribution, we can conduct statistical inference about π.

For estimation, we can form point estimators and posterior intervals.

For significance testing, we can find posterior probabilities that summarize evidence about the hypotheses of interest.

Bayesian Estimators and the Bias/Variance Tradeoff

For point estimation of the binomial parameter π, we can compare Bayesian and frequentist estimators.

The MLE $\hat{\pi} = Y/n$ of π is unbiased, that is, $E(\hat{\pi}) = \pi$ for all possible values of π.

In fact, it has the minimum variance among all the possible unbiased estimators of π.

We found the posterior mean estimate of π for a beta prior.

From expression (4), this Bayesian estimator $\tilde{\pi} = E(\pi \mid y)$ has expectation

$$E\left[\left(\frac{n}{n+\alpha_1+\alpha_2}\right)\frac{Y}{n} + \left(\frac{\alpha_1+\alpha_2}{n+\alpha_1+\alpha_2}\right)\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right)\right] =$$

$$= \left(\frac{n}{n+\alpha_1+\alpha_2}\right)\pi + \left(\frac{\alpha_1+\alpha_2}{n+\alpha_1+\alpha_2}\right).\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right)$$

This expectation is a weighted average of π and the prior mean $\alpha_1/(\alpha_1 + \alpha_2)$.

The estimator is biased, as Bayes estimators typically are, because $E(\tilde{\pi}) \neq \pi$.

But good estimators need not be exactly unbiased, but merely asymptotically unbiased, with the bias decreasing toward 0 as n increases.

For fixed α_1 and α_2, the Bayes estimator has expected value converging toward π as n increases, so it satisfies this property.

For example, when $\alpha_1 = \alpha_2 = \alpha$, the bias $[E(\tilde{\pi}) - \pi] = \alpha(1 - 2\pi)/(n + 2\alpha)$.

This is greatest in absolute value as π approaches 0 or 1, but regardless of the value of π, it is small when n is large.

Recall that the *mean squared error* summarizes how close, on average, a frequentist or Bayesian estimator $\hat{\theta}$ tends to be to the parameter θ that it estimates.

A formula for the MSE has as a consequence the important result that some bias in an estimator can be a good thing: The MSE decomposes into the variance of the estimator and its squared bias,

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2 =$$

$$= \mathrm{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 = \text{variance} + (\text{bias})^2.$$

By using an estimator that has some bias, if its variance decreases sufficiently, we may have the benefit of smaller MSE.

This result is called the ***bias/variance tradeoff***.

For example, although the Bayes estimator of π is biased, its variance

$$\mathrm{var}\left[\left(\frac{n}{n+\alpha_1+\alpha_2}\right)\frac{Y}{n}\right] =$$

$$= \left(\frac{n}{n+\alpha_1+\alpha_2}\right)^2 \mathrm{var}\left(\frac{Y}{n}\right) = \left(\frac{n}{n+\alpha_1+\alpha_2}\right)^2 \frac{\pi(1-\pi)}{n},$$

is smaller than $\pi(1-\pi)/n$, the variance of the MLE (which is unbiased), for all possible values of π and n.

If π is somewhat near the prior mean $\alpha_1/(\alpha_1+\alpha_2)$, so the bias is relatively small, MSE will be smaller for the Bayes estimator than for the MLE.

But, if π is far from the prior mean, such as very close to 0 or 1 when the prior mean is 0.50, MSE will be smaller for the MLE.

Posterior Intervals for a Proportion: Percentile and Highest Posterior Density

We can now build a *posterior interval*, also called a *credible interval* for π based on the posterior *pdf* $g(\pi \mid y)$.

As discussed we can do this in two ways, using *percentiles* of the posterior or the region of *highest posterior density* (HPD) values.

For example, the 95% equal-tail percentile posterior interval for π has as endpoints the 2.5 and 97.5 percentiles of the posterior of π.

The HPD region of π values has posterior probability 0.95 such that the posterior *pdf* is higher over all values in the region than over all the values not in it.

The HPD is the shortest region that contains the relevant probability.

For example, the 2018 General Social Survey in the U.S. asked "Do you believe in hell?" The percentage who responded *yes* was 74.0% for the 727 subjects with a high school education or less, 70.1% for the 304 subjects with a junior college or bachelor, and 56.8% for the 111 subjects with a graduate degree.

We shall find posterior intervals for the population proportion π of those with a graduate degree who would respond *yes*.

For the counts of 63 for the *yes* response and 48 for the *no* response,
the MLE is $\hat{\pi} = 63/111 =$ **0.568**, and
the 95% Wald interval of $\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}$ is **(0.475, 0.660)**.

For the uniform Beta(1, 1) prior with $y = 63$ and $n - y = 48$, the posterior is Beta(α_1^*, α_2^*) with $\alpha_1^* = y + \alpha_1 = 64$ and $\alpha_2^* = n - y + \alpha_2 = 49$.

The posterior mean estimate of π is $\alpha_1^*/(\alpha_1^* + \alpha_2^*) = 64/(64+49) =$ **0.566**, and the 0.025 and 0.975 quantiles of the beta posterior are **(0.474, 0.656)**.

```
> qbeta(c(0.025, 0.975), 63 + 1, 48 + 1)


[1] 0.4744648 0.6560523
      # quantiles of beta posterior for
      # 95% equal-tail percentile int. uniform prior

> qbeta(c(0.025, 0.975), 63.5, 48.5)


[1] 0.4746530 0.6570094
      # quantiles of beta posterior for
      # 95% equal-tail percentile int. with Jeffreys prior
```

```
> library(binom)

> binom.bayes(63, 111, conf.level=0.95, type="central",
alpha1=1, alpha2=1)

method x n shape1 shape2 mean lower upper
1 bayes 63 111 64 49 0.5663717 0.4744648 0.6560523
      # 95% equal-tail percentile int.
      # for uniform prior dist.

> binom.bayes(63, 111, conf.level=0.95, alpha1=1, alpha2=1
method x n shape1 shape2 mean lower upper


1 bayes 63 111 64 49 0.5663717 0.4752644 0.6568256
      # 95% HPD interval
      # for uniform prior dist.
```

```
> binom.bayes(63, 111, conf.level=0.95, alpha1=0.5, alpha2=


method x n shape1 shape2 mean lower upper
1 bayes 63 111 63.5 48.5 0.5669643 0.4754671 0.6577965

      # 95% HPD interval for Jeffreys prior
> binom.confint(63, 111, conf.level=0.95, method="asymptoti

      # 95% Wald confidence int.
method x n mean lower upper
1 asymptotic 63 111 0.5675676 0.475405 0.6597301
> binom.confint(63, 111, conf.level=0.95, method="wilson")
method x n mean lower upper
1 wilson 63 111 0.5675676 0.4746712 0.6559436
      # 95% score confidence int.
      # better than Wald if n small or pi near 0 or 1
```

The Bayesian and the frequentist intervals are very similar, although **interpretations differ**.

Frequentist intervals are justified by the property that in repeatedly taking random samples of size 111 from the population in the U.S. having a graduate degree, in the long run 95% of the confidence intervals would contain the actual value of π.

For **Bayesian posterior intervals**, we can infer directly that the probability is 0.95 that π falls in the interval obtained.
We recommend using the HPD interval rather than the equal-tail percentile interval when the posterior *pdf* is monotone increasing or decreasing from the boundary of the parameter space.

For example, in estimating a binomial parameter π, suppose that all n trials are successes.

When $y = n$ and we use a uniform or a Jeffreys prior for π, the posterior *pdf* $g(\pi \mid y)$ is monotone increasing from 0 to 1.

It is not then sensible to exclude 1.0 and nearby values from the posterior interval.

Influence of Sample Size and prior on Posterior Intervals

As the sample size n increases, the influence of the prior weakens and the posterior has appearance closer to that of the likelihood function.

In the limit, Bayes estimates of π become more similar to the MLE and Bayesian posterior intervals become more similar to frequentist confidence intervals.

Specifically, the beta posterior has hyperparameter values $\alpha_1^* = y + \alpha_1$ and $\alpha_2^* = n - y + \alpha_2$.

With fixed values of the prior hyperparameters α_1 and α_2, as n increases, from equation (4) the posterior mean $E(\pi \mid y) = \alpha_1^*/(\alpha_1^* + \alpha_2^*) = (y + \alpha_1)/(n + \alpha_1 + \alpha_2)$ is approximately $\hat{\pi} = y/n$, which is the sample proportion and MLE of π.

Also, $\alpha_1^* + \alpha_2^* = n + \alpha_1 + \alpha_2$ is close to n, so the posterior variance (3) is

$$\frac{\alpha_1^*\alpha_2^*}{(\alpha_1^*+\alpha_2^*)^2(\alpha_1^*+\alpha_2^*+1)} = \left[\frac{\alpha_1^*}{\alpha_1^*+\alpha_2^*}\right]\left[\frac{\alpha_2^*}{\alpha_1^*+\alpha_2^*}\right]\left[\frac{1}{\alpha_1^*+\alpha_2^*+1}\right]$$

$$\approx \left[\frac{y}{n}\right]\left[\frac{n-y}{n}\right]\left[\frac{1}{n}\right] = \frac{\hat{\pi}(1-\hat{\pi})}{n},$$

which is the MLE of the variance of $\hat{\pi}$.

As n increases for fixed values of α_1 and α_2, like the sampling distribution of $\hat{\pi}$, the beta distribution is more nearly normal.

Thus, for large n, posterior intervals for the posterior beta distribution are quite close to the frequentist confidence interval, $\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$.

We illustrate this with a sequence of cases in which $\hat{\pi} = 0.70$, computing for increasing values of *n* the frequentist Wald confidence interval and the Bayesian HPD interval obtained with uniform prior ($\alpha_1 = \alpha_2 = 1$):

```
> library(binom)
> binom.confint(7,10,conf.level=0.95, method="asymptotic")

      # n = 10
method lower upper 1 asymptotic 0.41597 0.98403
      # frequentist Wald CI
> binom.bayes(7, 10, conf.level=0.95, alpha1=1, alpha2=1)
method lower upper
      # Bayesian HPD interval for 1 bayes 0.41205 0.90663

      # uniform prior
> binom.confint(70, 100, conf.level=0.95,
                method="asymptotic")
      # n = 100
method lower upper 1 asymptotic 0.61018 0.78982
      # frequentist Wald CI
```

```
method lower upper

1 bayes 0.60657 0.78345
      # Bayesian HPD interval


> binom.confint(7000, 10000, conf.level=0.95,
                method="asymptotic")
      # n=10000
method lower upper
1 asymptotic 0.69102 0.70898
      # frequentist Wald CI
> binom.bayes(7000, 10000, conf.level=0.95,
                alpha1=1, alpha2=1)
method lower upper
1 bayes 0.69097 0.70893
      # Bayesian HPD interval
```

For a particular sample size n (set to 100 in the `R` code below), a Bayesian interval is more similar to the frequentist confidence interval as the prior is more diffuse, that is, as $\alpha_1 = \alpha_2 = \alpha$ decreases toward 0:

```
> library(binom)
> binom.bayes(70, 100, conf.level=0.95,
      alpha1=100, alpha2=100)
      # beta hyperpara's=100
method lower upper
1 bayes 0.5106053 0.6224795
      # narrow interval when prior has small variance
> binom.bayes(70, 100, conf.level=0.95,
                alpha1=10, alpha2=10)
      # beta hyperpara's=10
method lower upper
1 bayes 0.5821577 0.7496266

> binom.bayes(70, 100, conf.level=0.95,
      alpha1=1, alpha2=1)
      # beta hyperpara's=1
method lower upper
1 bayes 0.6065663 0.7834458
```

```
> binom.bayes(70, 100, conf.level=0.95,
      alpha1=0.1, alpha2=0.1)
      # beta hyperpara's=0.1
method lower upper
1 bayes 0.6095769 0.787402

> binom.bayes(70, 100, conf.level=0.95,
                alpha1=0.01, alpha2=0.01)
      # beta hyperpara's=0.01
method lower upper
1 bayes 0.6098849 0.7878043

> binom.confint(70, 100, conf.level=0.95,
                method="asymptotic")
      # frequentist Wald CI
method lower upper
1 asymptotic 0.6101832 0.7898168
```

When $\alpha_1 = \alpha_2$ is close to 0 (e.g., 0.1 or 0.01 each), the Bayesian HPD interval is similar to the frequentist confidence interval.

Posterior Probability Analogs of *P*-values

As explained earlier with a null hypothesis such as H_0: $\pi = 0.50$, it is often relevant to summarize the evidence that $\pi < 0.50$ versus $\pi > 0.50$.

We can do this by reporting the posterior tail probabilities, $P(\pi < 0.50 \mid y)$ and $P(\pi > 0.50 \mid y)$.

For instance, for the summary of opinions about the existence of hell in of the subjects having a graduate degree, 63 said *yes* and 48 said *no*.

With the uniform prior, the posterior is beta with hyperparameters $\alpha_1^* = y + 1 = 64$ and $\alpha_2^* = n - y + 1 = 49$.

We next use this posterior to find $P(\pi < 0.50 \mid y)$ and $P(\pi > 0.50 \mid y)$:

```
> pbeta(0.50, 64, 49)
      # cumulative probability at 0.50 for beta posterior
[1] 0.07803
> 1 - pbeta(0.50, 64, 49)
      # posterior P(pi > 0.50 | y)
[1] 0.92197
```

The posterior $P(\pi > 0.50 \mid y) = 0.922$ and $P(\pi < 0.50 \mid y) = 0.078$ provide substantial but not overly strong evidence that the majority believe in hell.

A corresponding frequentist statistical inference reports the P-value for testing H_0: $\pi = 0.50$ (implicitly $\pi \leq 0.50$) against H_1: $\pi > 0.50$.

Since $y = 63$, the P-value is the binomial probability of observing $Y \geq 63$ in $n = 111$ trials when actually H_0 is true, that is, $1 - P(Y \leq 62 \mid \pi = 0.50) = 0.092$.

It is plausible that this particular H_0 is true (i.e., $\pi = 0.50$), because the P-value is small but not overly so.

```
> 1 - pbinom(62, 111, 0.50)        # one-sided (right-tail)
P-value for Binom(111, 0.50) dist.
[1] 0.09182859       # when y = 63 and n = 111
```

Bayes Factors

We now describe another Bayesian method for summarizing evidence about hypotheses.

For a parameter θ and for subsets Θ_0 and Θ_1 of the parameter space, consider the hypotheses

$$H_0 : \ \theta \in \Theta_0, \quad H_1 : \ \theta \in \Theta_1.$$

In practice, the hypotheses often correspond to two possible models for the data that do not need to be nested, that is, one a special case of the other.

Suppose H_0 specifies that the observations $\boldsymbol{Y}$ have joint *pdf* $f_0(\boldsymbol{y} \mid \theta)$ and H_1 specifies that $\boldsymbol{Y}$ have joint *pdf* $f_1(\boldsymbol{y} \mid \theta)$.

We summarize the data's evidence about the two hypotheses by comparing the values of the marginal distributions of $\boldsymbol{Y}$ at the observed value $\boldsymbol{y}$, under each hypothesis.

With prior *pdf* $p(\theta)$, the joint distribution of $(\boldsymbol{Y}, \theta)$ has probability function $h(\boldsymbol{y}, \theta) = f(\boldsymbol{y} \mid \theta) p(\theta)$, and we obtain the marginal distribution $m(\boldsymbol{y})$ by integrating out θ.

The marginal distributions under the two hypotheses are

$$m_0(\boldsymbol{y}) = \int_{\Theta_0} f_0(\boldsymbol{y} \mid \theta) p_0(\theta) d\theta \quad \text{and} \quad m_1(\boldsymbol{y}) = \int_{\Theta_1} f_1(\boldsymbol{y} \mid \theta) p_1(\theta) d\theta.$$

After observing $\boldsymbol{y}$, the ***Bayes factor*** in support of H_0 relative to H_1 is

$$BF = \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})}. \tag{5}$$

The value $BF = 1$ indicates that the data provide equal support for each hypothesis, whereas $BF > 1$ indicates greater support for H_0 than H_1.

The greater the value of BF above 1, the stronger the evidence provided by the data in support of H_0.

The overall marginal distribution of the data, averaged over prior probab. $P(\theta \in \Theta_0)$ for H_0 and $P(\theta \in \Theta_1)$ for H_1 that sum to 1 is

$$m(\boldsymbol{y}) = m_0(\boldsymbol{y}) P(\theta \in \Theta_0) + m_1(\boldsymbol{y}) P(\theta \in \Theta_1).$$

By Bayes' theorem, the marginal probability function for the data, given each hypothesis, can be expressed in terms of the probability of each hypothesis, given the data, as

$$m_0(\boldsymbol{y}) = P(\theta \in \Theta_0 \mid \boldsymbol{y}) m(\boldsymbol{y}) / P(\theta \in \Theta_0), \quad m_1(\boldsymbol{y}) =$$

$$= P(\theta \in \Theta_1 \mid \boldsymbol{y}) m(\boldsymbol{y}) / P(\theta \in \Theta_1).$$

Therefore, we can express the Bayes factor as

$$BF = \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} = \frac{P(\theta \in \Theta_0 \mid \boldsymbol{y})m(\boldsymbol{y})/P(\theta \in \Theta_0)}{P(\theta \in \Theta_1 \mid \boldsymbol{y})m(\boldsymbol{y})/P(\theta \in \Theta_1)} = \tag{6}$$

$$= \frac{P(\theta \in \Theta_0 \mid \boldsymbol{y})/P(\theta \in \Theta_1 \mid \boldsymbol{y})}{P(\theta \in \Theta_0)/P(\theta \in \Theta_1)}. \tag{7}$$

Now, the prior *odds* of H_0 relative to H_1 are

$$\frac{P(\theta \in \Theta_0)}{P(\theta \in \Theta_1)} = \frac{\int_{\Theta_0} p(\theta)\, d\theta}{\int_{\Theta_1} p(\theta)\, d\theta}.$$

After observing the data $\boldsymbol{y}$, the corresponding posterior odds equal

$$\frac{P(\theta \in \Theta_0 \mid \boldsymbol{y})}{P(\theta \in \Theta_1 \mid \boldsymbol{y})} = \frac{\int_{\Theta_0} g(\theta \mid \boldsymbol{y})\, d\theta}{\int_{\Theta_1} g(\theta \mid \boldsymbol{y})\, d\theta}.$$

Thus, from (7), $BF =$ (posterior odds)/(prior odds), or equivalently,

$$\text{Posterior odds } = BF(\text{Prior odds}).$$

When the prior probabilities are identical, in which case the prior odds equal 1, BF simplifies to the ratio of the posterior probabilities under H_0 and under H_1.

With frequentist significance testing, P-values ≤ 0.05 make one skeptical of H_0, P-values ≤ 0.02 provide strong evidence against H_0, and P-values ≤ 0.01 provide very strong evidence.

When $P(\theta \in \Theta_0) = P(\theta \in \Theta_1) = 0.50$, P-values of 0.05, 0.02, and 0.01 correspond to BF values of $0.95/0.05 = 19$, $0.98/0.02 = 49$, and $0.99/0.01 = 99$.

Thus, roughly speaking, BF values exceeding about 20 provide *fairly strong* evidence against H_0, BF values exceeding 50 provide *strong* evidence, and BF values exceeding 100 provide *very strong* evidence.

Table 1 summarizes the BF strength of evidence against H_0.

Table: Strength of evidence against H_0 provided by various Bayes factor values

BF	Evidence against H_0
$1-3$	negligible
$3-20$	positive
$20-50$	fairly strong
$50-100$	strong
>100	very strong

We illustrate the Bayes factor by finding BF for H_0: $\pi > 0.50$ and H_1: $\pi < 0.50$ about a binomial parameter π when we use a beta prior for π with $\alpha_1 = \alpha_2$.

This prior is symmetric around 0.50, so the prior $P(\pi < 0.50) = P(\pi > 0.50)$.

For the example in the previous subsection with the uniform prior for the probability π of belief in hell for those with a graduate degree, given the binomial outcome y, $P(\pi > 0.50 \mid y) = 0.922$ and $P(\pi < 0.50 \mid y) = 0.078$.

Thus, the Bayes factor is 0.922/0.078 = 11.8.

Conditional on the binomial observation, we judge that the probability that $\pi > 0.50$ is 11.8 times the probability that $\pi < 0.50$.

For belief in hell by those with a graduate degree, there is positive evidence that $\pi > 0.50$ but the evidence is not strong.

A criticism of the Bayes factor is that its value can be highly sensitive to some assumption or aspect of the model on which the hypotheses are based that is not easily checked.

For example, its value is quite highly dependent on the choice for the prior.

Also, it is not available with improper priors.

Another criticism is that the Bayes factor is not useful when an application has a large or a continuous set of potential models rather than a small, discrete set.

Bayesian Prediction of Future Observations

To predict future observations, we can use a Bayesian posterior *predictive distribution*, that is, the probability distribution of a future observation Y_f.

As explained given the data, we obtain the posterior predictive *pdf* by integrating the probability function for y, given θ, with respect to the information we have about θ in its posterior $g(\theta \mid \boldsymbol{y})$,

$$h(y_f \mid \boldsymbol{y}) = \int_{\Theta} f(y_f \mid \theta) g(\theta \mid \boldsymbol{y}) d\theta.$$

A natural prediction for a future observation is the **mean of the posterior predictive** distribution.

For predicting a future binary observation that can take value 0 or 1, $P(Y_f = 1 \mid \pi) = \pi$.

So, for $y_f = 1$, $f(y_f \mid \pi) = \pi$, so the predictive *pdf* takes value at 1,

$$h(1 \mid \boldsymbol{y}) = \int_0^1 f(1 \mid \pi) g(\pi \mid \boldsymbol{y}) d\pi = \int_0^1 \pi g(\pi \mid \boldsymbol{y}) d\pi.$$

This is the mean of the beta posterior *pdf* $g(\pi \mid \boldsymbol{y})$.

With the uniform prior, we have

$$P(Y_f = 1 \mid y) = h(1 \mid \boldsymbol{y}) = E(\pi \mid y) = (y + 1)/(n + 2).$$

Since y_f can only take values 0 and 1, this is also the mean of the posterior predictive distribution.

For the example, the predictive probability that another randomly selected person with a graduate degree believes in hell is the posterior mean for π of 0.567.
Some applications naturally focus on the posterior predictive distribution for the *number of successes* Y_f in some future number n_f of observations.

Averaging the binomial conditional distribution of Y_f for a given value of π with respect to the beta posterior of π yields a ***beta-binomial distribution***. That distribution is specified by its sample size index and by parameters that are the hyperparameters of the beta distribution.

With posterior beta hyperparameters α_1^* and α_2^* and mean $\mu^* = \alpha_1^*/(\alpha_1^* + \alpha_2^*)$, the mean and variance of Y_f for n_f future observations are

$$E(Y_f) = n_f\mu^*, \quad \mathrm{var}(Y_f) = n_f\mu^*(1-\mu^*)\big[1 + (n_f - 1)/(\alpha_1^* + \alpha_2^* + 1)\big]$$

The beta-binomial distribution has the same mean but larger variance than the binomial distribution with index n_f and success probability μ^*.

Having greater variance is natural, because of the additional sampling variability from taking a new sample.

However, as $(\alpha_1^* + \alpha_2^*)$ increases, its variance decreases toward the binomial variance $n_f\mu^*(1 - \mu^*)$.

In fact, as $(\alpha_1^* + \alpha_2^*)$ increases, the beta-binomial distribution converges to the binomial distribution.

We found that the posterior beta hyperparameters relate to the hyperparameters of the beta prior and to the number of successes and sample size for the original sample by $\alpha_1^* = y + \alpha_1$ and $\alpha_2^* = n - y + \alpha_2$, so that $\alpha_1^* + \alpha_2^* = n + \alpha_1 + \alpha_2$.

Therefore, the beta-binomial distribution converges toward the binomial distribution as the sample size n increases.

We can obtain probabilities for a beta-binomial distribution using software.

We illustrate by returning to the example of the British tea taster.

After observing 10 successful guesses in 10 cups, combined with a uniform prior we obtained a posterior beta distribution for the probability of a successful prediction, with $\alpha_1^* = 11$ and $\alpha_2^* = 1$, for which the posterior mean is $\mu^* = 11/12 = 0.9167$.

For the next $n_f = 5$ cups, here is the beta-binomial distribution for the number of correct guesses Y_f:

```
> library(extraDistr)

> dbbinom(y, 5, alpha = 11, beta = 1)

       # displays P(Y_f = 0), P(Y_f = 1), ..., P(Y_f = 5)

[1] 0.0002289377 0.0025183150 0.0151098901 0.0654761905
0.2291666667 0.6875000000
```

The expected number of successful guesses is $n_f\mu^* = 5(11/12) = 4.58$, and the probability is 0.6875 of getting all five correct.