Chapter 5. Bayesian Statistics

Principles of Bayesian Statistics

Anything unknown is given a probability distribution, representing degrees of belief [subjective probability].

Degrees of belief [subjective probability] is updated by data. In this sense, Bayesian statistical inference is more like a learning process.

Degrees of belief [subjective probability] can be handled as if it were classical probability, and therefore, *mathematically* there is no difference.

Classical probability as a measurement of uncertainty interpreted as “long run relative frequency”. Extremely useful (kind of a constructive way of understanding uncertainty).

Difficulties in interpreting common-sense statements such as “What is the probability that it will rain tomorrow?” “What is the chance that Heat will repeat this season?” “What is Hillary’s chance of becoming the next President?”

Example: Horse Betting

	Amount bet on		odds	payoff for
Horse	horse (in thousands)	“probability”	against	$1 bet
H1	500			
H2	250			
H3	100			
H4	100			
H5	50			
Total	1000			

$$\text{odd against event } E : O(E) \doteq \frac{1 - P(E)}{P(E)}$$

The real track odds table, after 17% of *track take.*

Horse	Amount bet on horse (in thousands)	Adjusted "probability"	Track odds against	Track payoff for $1 bet
H1	500			
H2	250			
H3	100			
H4	100			
H5	50			
Total	1000			

The odds are set after all the money are collected. So the odds represent a concensus of the bettors' subjective perception on the abilities of horses.

Example: Coherence of bets

If you attach probability p to an event E

$$P(E) = p$$

if and only if you would exchange \$$p$ for a return of \$1 if E occurs and 0 if E does not occur.

Basic Procedure of Bayesian Statistics

1. Model setup. The parameters of interest θ [unknown].

2. Prior distribution. Assign a prior probability distribution to θ, representing your degree of belief with respect to θ.

3. Posterior distribution. Update your degree of belief with respect to θ, based on the data. The new degree of belief is called the posterior probability distribution of θ.

An observed result changes our degrees of belief in parameter values by changing a prior distribution into a posterior distribution.

Computing Posterior Distribution: Bayes Rule

Example: Suppose Y has distribution $B(n;\theta)$. What can we say about θ if the observation is $Y = y$?

1. Assign the prior π. Let $\pi(\theta) = 1$ for $\theta \in [0,1]$.
2. Compute the posterior $p(\theta|y)$.

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{p(y)} \\ &= (n+1)\binom{n}{y}\theta^{y}(1-\theta)^{n-y} \\ &\propto \theta^{y}(1-\theta)^{n-y}. \end{aligned}$$

The graphs of the posterior distribution $p(\theta|y)$ for different sets of (n, y).

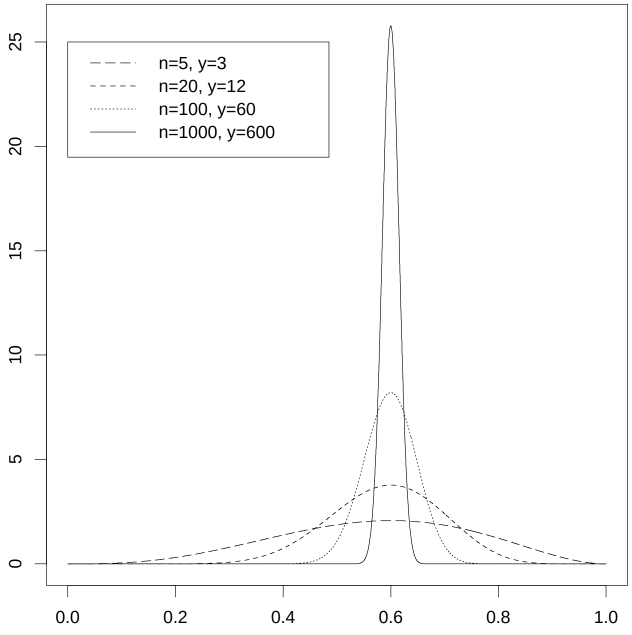


Figure 1: Posterior density $p(\theta|y)$ with uniform prior

If we assign a different prior distribution,

1. Assign the prior π. Let $\pi(\theta) = 6\theta(1-\theta)$ for $\theta \in [0,1]$.
2. Compute the posterior $p(\theta|y)$.

$$\begin{aligned}
p(\theta|y) &= \frac{p(\theta,y)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{p(y)} \\
&= (n+3)\binom{n+2}{y+1}\theta^{y+1}(1-\theta)^{n+1-y} \\
&\propto \theta^{y+1}(1-\theta)^{n+1-y}.
\end{aligned}$$

then we arrive at a different posterior distribution.

The graphs of the posterior distribution $p(\theta|y)$ for different sets of (n, y).

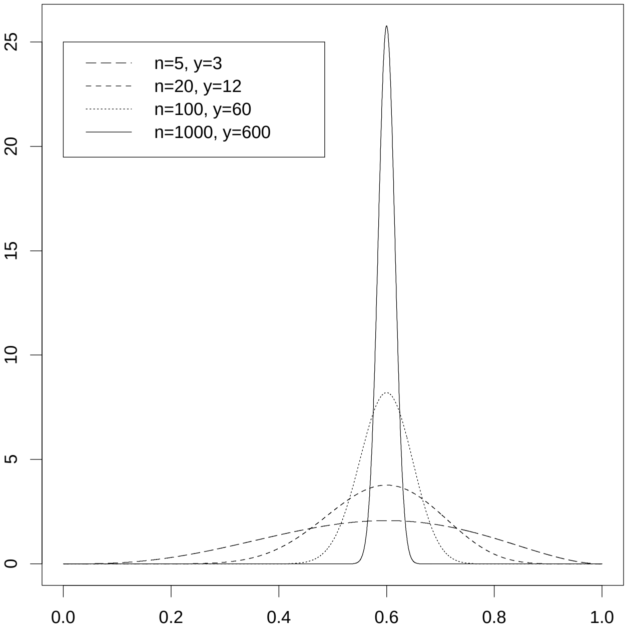


Figure 2: Posterior density $p(\theta|y)$ with Beta(2,2) prior

Comparison of the posterior distribution $p(\theta|y)$ with different prior distribution.

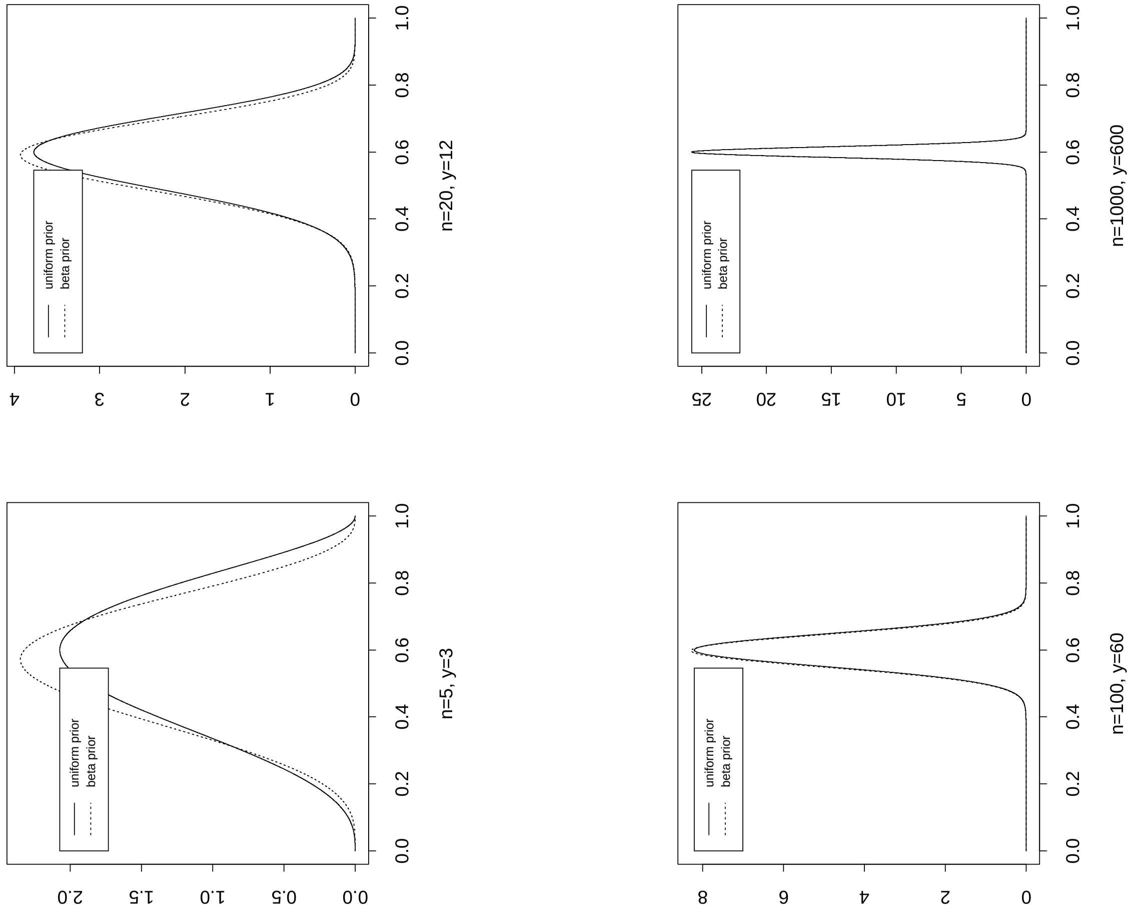


Figure 3: Posterior density $p(\theta|y)$ comparison

Digression to Beta distribution [Textbook, Section 4.7]

For $\alpha, \beta > 0$, Beta(α, β) distribution has density

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \theta \in [0,1]$$

and 0 otherwise.

Here the special function Γ is defined as

$$\Gamma(\alpha) \doteq \int_0^\infty x^{\alpha-1}e^{-x}\,dx.$$

1. For all $\alpha > 0$, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$
2. For all integers n, $\Gamma(n) = (n-1)!$.

Beta(α, β) has

$$\begin{aligned} \text{mean} &= \frac{\alpha}{\alpha+\beta} \\ \text{variance} &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

and for $\alpha \geq 1, \beta \geq 1, \alpha + \beta > 2$,

$$\text{Mode [hightest point in density curve]} = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

Back to previous example.

If the prior takes form

$$\pi(\theta) = \text{Beta}(\alpha, \beta),$$

then the posterior is

$$p(\theta|y) = \text{Beta}(\alpha + y, \beta + n - y).$$

Remark: The parameters in the prior, α and β, are often referred to as the hyperparameters.

Remark: Discussion on $n \to \infty$.

General posterior calculation.

1. $\pi(\theta)$: Prior for parameter θ.
2. $y \doteq \{y_1, y_2, \ldots, y_n\}$: observations, data.
3. $p(y|\theta)$: likelihood function [probability function or density].
4. $p(\theta|y)$: Posterior distribution.

$$p(\theta|y) \propto \pi(\theta) \cdot p(y|\theta).$$

Remark: The normalizing constant in the formula of $p(\theta|y)$ cannot be explicitly evaluated in general, which is the source of difficulties in Bayesian computation.

Some general relations between prior and posterior

1.
$$E[\theta] = E[E[\theta|y]]$$
Prior mean of θ = Average posterior mean of θ over data distribution.

2.
$$\text{Var}[\theta] = E[\text{Var}(\theta|y)] + \text{Var}(E[\theta|y])$$
Posterior variance of θ is, on average, less than prior variance of θ.

Remark: Adding more data is very convenient in Bayesian framework. [The “learning” aspect of Bayesian statistics].

[1]. Prior ⇒ Data ⇒ Posterior [Prior] ⇒ More Data ⇒ Posterior

[2]. Prior ⇒ All Data ⇒ Posterior

Same Posterior! Since

$$\begin{aligned} p(\theta|y_1, y_2) &\propto \pi(\theta) \cdot p(y_1, y_2|\theta) = \pi(\theta) \cdot p(y_1|\theta) \cdot p(y_2|\theta) \\ &\propto p(\theta|y_1) \cdot p(y_2|\theta) \end{aligned}$$

Abstract Examples

1. Normal data: unknown mean, known variance.

$y = \{y_1, \ldots, y_n\}$ are iid samples from $N(\theta, \sigma^2)$ with σ^2 known.

The prior distribution for θ is assumed to be $N(\mu, \tau^2)$.

The posterior distribution for θ is

$$\begin{aligned} p(\theta|y) &\propto \pi(\theta) \cdot p(y|\theta) \\ &= \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \cdot \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n e^{-\frac{\sum_i (y_i-\theta)^2}{2\sigma^2}} \\ &\propto e^{-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\theta^2-\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu}{\tau^2}\right)\theta} \end{aligned}$$

Thus $p(\theta|y_1,\ldots,y_n)$ is $N(\mu_n,\tau_n^2)$ with

$$\mu_n = \frac{n\bar{y}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \quad \frac{1}{\tau_n^2} = \frac{1}{\tau^2} + \frac{n}{\sigma^2}$$

Discussion on precisions and posterior mean for normal data.

Remark: Discussion on $n \to \infty$.

2. Poisson data. $y = \{y_1, \dots, y_n\}$ are iid Poisson with rate θ.

Assign prior distribution $\pi(\theta)$ as Gamma(α, β), that is,

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \theta^{\alpha-1} e^{-\beta\theta}, \qquad \theta > 0.$$

See [Textbook, Section 4.6] for Gamma distribution.

Note: The β in textbook corresponds to $1/\beta$ here.

The posterior distribution of θ is

$$\begin{aligned} p(\theta|y) &\propto \pi(\theta) \cdot p(y|\theta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \theta^{\alpha-1} e^{-\beta\theta} \cdot e^{-n\theta} \frac{\theta^{y_1+\cdots+y_n}}{y_1! \cdot y_n!} \\ &\propto \theta^{\alpha+n\bar{y}-1} e^{-(\beta+n)\theta}. \end{aligned}$$

Therefore, [Discussion on $n \to \infty$]

$$p(\theta|y) = \text{Gamma}(\alpha + n\bar{y}, \beta + n).$$

Digression to Gamma distribution: Gamma(α, β)

$$\text{mean} = \frac{\alpha}{\beta}$$

$$\text{variance} = \frac{\alpha}{\beta^2}$$

$$\text{mode} = \frac{\alpha - 1}{\beta}, \qquad \text{for } \alpha \geq 1.$$

(a)

$$\chi^2(d) = \text{Gamma}(d/2, 1/2).$$

(b)

$$\text{Exp}(\lambda) = \text{Gamma}(1, \lambda).$$

3. Exponential distribution. $y = \{y_1, \ldots, y_n\}$ are iid exponential with rate θ.

   Assign prior distribution $\pi(\theta)$ as Gamma(α, β), that is,

$$\pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}, \qquad \theta > 0.$$

   The posterior distribution of θ is

$$\begin{aligned} p(\theta|y) &\propto \pi(\theta) \cdot p(y|\theta) \\ &\propto \theta^{\alpha-1} e^{-\beta\theta} \cdot \theta^n e^{-(y_1+\cdots+y_n)\theta} \\ &= \theta^{\alpha+n-1} e^{-(\beta+n\bar{y})\theta}. \end{aligned}$$

   Therefore, [Discussion on $n \to \infty$]

$$p(\theta|y) = \text{Gamma}(\alpha + n, \beta + n\bar{y}).$$

Concrete Examples

1. A friend of yours claims to be able to toss a coin and get heads every time. You suspect he may have a two-headed coin and convince him to toss the coin 10 times (without examining it). Your probability of two-headed is 0.01 and the remaining 0.99 is associated with the coin being fair. In 10 tosses, each time he gets a heads. Now what is your probability of two-headed?

$$\frac{0.01 \times 1}{0.01 \times 1 + 0.99 \times (0.5)^{10}} = 0.91$$

2. How much do I weigh? A scale that is not perfectly accurate is used. I weighed myself 10 times in succession and the readings (y_i) were:

   $$142 \quad 132 \quad 133 \quad 136 \quad 136 \quad 140 \quad 133 \quad 134 \quad 139 \quad 135$$

   with $\bar{y} = 136$. Assume that each reading is distributed as $N(\theta, 12)$ with θ as my true weight [discussion on the variance].

   Assume that my prior of θ is $N(134, 25)$ [discussion on how this prior comes from, and its importance for small sample sizes]. Calculate the posterior.

   $$N(135.1, 1.1)$$

3. A small pilot study evaluated the effectiveness of an experimental drug. The changes in blood pressure (in mmHg, y_i) before/after drug intake are observed for 16 patients with $\bar{y} = -8$. Assume the change of blood pressure has population distribution $N(\theta, 36)$.

   Assume that your prior distribution for θ is $N(0, 100)$ [discussion on this prior distribution]. Find the posterior of θ.

$$N(-7.8, 2.2)$$

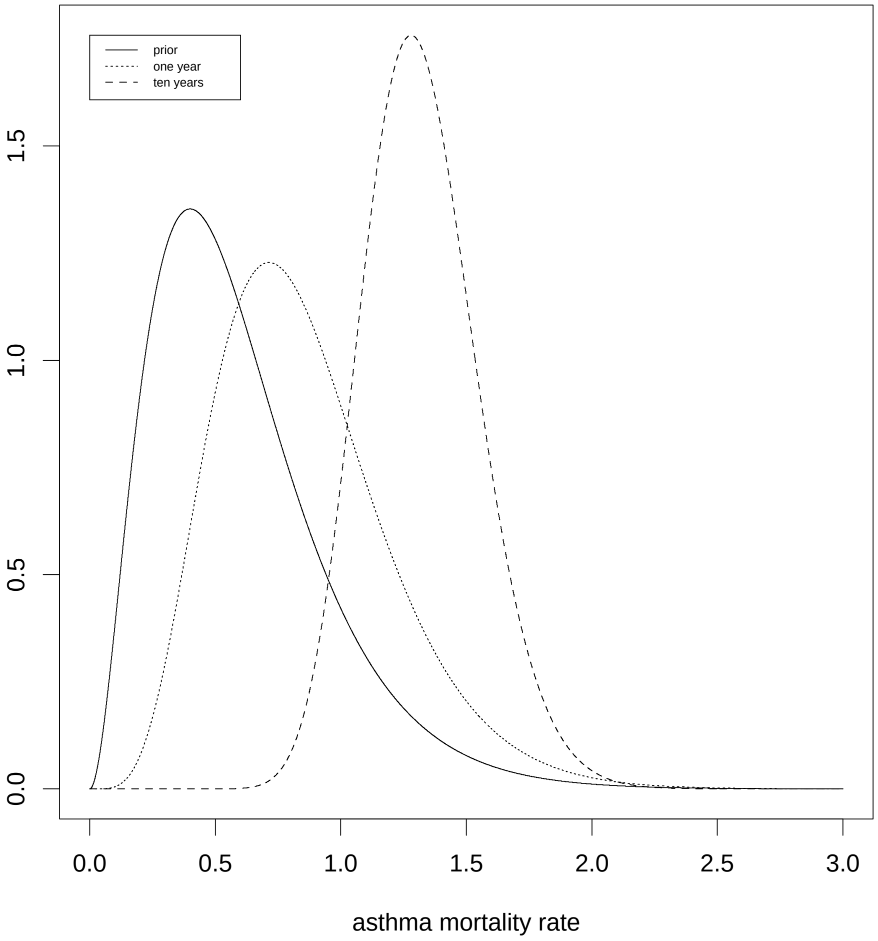
4. Suppose that for a city in USA it was found that 3 persons, out of population 200,000, died of asthma, giving a crude estimate of mortality rate in the city of 1.5 cases per 100,000, in a single year. A Poisson model is often used for epidemiological data, and we assume that the number of deaths each year from asthma is Poisson(2θ) with θ being the true underlying asthma mortality rate per 100,000 persons per year in the city.

Assume that your prior distribution for θ is Gamma$(3, 5)$ [discussion on this prior distribution]. Find the posterior of θ.

$$\text{Gamma}(6, 7).$$

Assume now we observe a mortality rate of 1.5 per 100,000 in ten years for the city, and the population size remains constant at 200,000 each year [so there are total 30 deaths in ten years]. Assume that the outcome from each year is independent with distribution Poisson(2θ). What is you posterior now?

$$\text{Gamma}(33, 25).$$


prior
one year
ten years
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
asthma mortality rate

Dependence on prior distribution

1. Different prior leads to different posterior. But this difference tends to be negligible when the data is very informative.

2. For large samples, the Bayesian inference is in a sense consistent with classical (frequency-based probability) inference, irregardless of the prior.

3. Informative priors and non-informative priors.
   (a) Informative prior: *state of knowledge.*
   (b) Non-informative prior: *"let the data speak for themselves"*. Bayes' justification of uniform prior in Binomial model and Laplace's less clear rationale "principle of insufficient reason".

Remark: Can prior be wrong?

Building Prior Distributions

Informative priors

Advantages:

1. often analytically convenient (esp for conjugate priors).
2. can take advantage of your informed understanding, beliefs, and experience.

Disadvantages:

1. not always easy to quantity the state of knowledge.
2. not always appropriate in situations such as public policy.

1. Conjugate priors: The prior and the posterior belong to the *same* parametric class. Very convenient to use.

   Many examples we have discussed indeed used conjugate prior!

2. Non-conjugate priors.

Non-informative priors

Advantages:

1. sometimes used as benchmark that will not reflect the bias of the analyst.
2. appropriate when little is known on the underlying processes.
3. useful as public policy priors.

Disadvantages:

1. may lead to improper priors (see below).
2. the definition of knowing little may depend on different parameterizations.
3. different interpretation regarding “non-informative” may lead to different prior distributions.

Improper prior distribution: "Distributions" that integrates to ∞.

For example, suppose $y = \{y_1, \ldots, y_n\}$ are iid $N(\theta, 1)$ with improper prior $\pi(\theta) \propto 1$. The posterior is

$$p(\theta|y) = \pi(\theta) \cdot p(y|\theta) \propto e^{-\frac{(\theta-\bar{y})^2}{2/n}}$$

or

$$p(\theta|y) = N(\bar{y}, 1/n)$$

Remark: For improper prior distribution, the calculation is formally the same as that for proper prior distributions. One has to be careful that the posterior distribution is proper (which is not always the case!).

Bayesian Inference

General remarks on Bayesian inference

The posterior distribution $p(\theta|y)$ including *all* the information we need for statistical inference!

Numerical summaries are still desirable for practical purpose: mean, mode, median, standard deviations, quantiles, posterior intervals (confidence intervals), to name a few.

Classical hypothesis testing is in some sense a *decision* problem – a solution which is optimal relative to certain criteria [e.g. Neyman-Pearson Lemma]. While Bayesian frameworks admits genuine inference for hypothesis, and P-value has a much more natural interpretation.

Posterior (confidence) Interval

$(1-\alpha)$ *posterior interval* is the interval $[\theta_1, \theta_2]$ such that

$$P(\theta < \theta_1) = \int_{\{\theta<\theta_1\}} p(\theta|y)\, d\theta = \frac{\alpha}{2}$$

$$P(\theta > \theta_2) = \int_{\{\theta>\theta_2\}} p(\theta|y)\, d\theta = \frac{\alpha}{2}$$

Remark: Sometimes one uses another definition of $(1-\alpha)$ posterior interval such that any value outside it is less plausible than any value in it.

Remark: The interpretation of Bayesian confidence interval is more satisfactory than that of the classical frequentist one.

How to obtain numerical summaries?

1. Analytical calculation: limited applications.

   (a) Example: Is the brain like a muscle – use it and it grows? A controlled paired experiment with 59 pairs of rats resulted in a sample mean of 5.72% increase of the cortex weight for the rats who were given "brain exercise" everyday. Assume that population percentage increase is $N(\theta, \sigma^2)$ with $\sigma = 5.13\%$, and a flat prior.

   Solution: The posterior distribution of θ is

   $$p(\theta|y) = N(\bar{y}, \sigma^2/n) = N(5.72, 0.67^2).$$

   Posterior mean, mode, standard deviation, ...

The 95% posterior interval is

$$5.75 \pm z_{0.025} 0.67 = 5.75 \pm 1.96 \times 0.67 = [4.44, 7.06].$$

Some remarks on hypothesis testing.

(b) Example: For each professional football game, experts provide a *point spread* as a measure of the difference between two teams. For $n = 672$ professional football games during 1981, 1983, and 1984, the discrepancies d_i between game outcomes and point spread are recorded. It seems reasonable to assume that the data points d_i, $i = 1, \ldots, n$ are iid samples from $N(0, \sigma^2)$. The goal is to estimate σ^2. Denote $\theta \doteq \sigma^{-2}$. The (improper) prior of θ is assumed to be such that $\ln(\theta)$ is flat on $\mathbb{R}$, or

$$\pi(\theta) \propto 1/\theta.$$

Remark: This prior is equivalent to $\pi(\sigma) \propto 1/\sigma$ or $\ln(\sigma)$ being flat on $\mathbb{R}$.

Solution: The posterior distribution is

$$p(\theta|d) \propto \pi(\theta) \cdot p(d|\theta) = \frac{1}{\theta} \cdot \left(\sqrt{\frac{\theta}{2\pi}}\right)^n e^{-\frac{\theta}{2}\sum_{i=1}^n d_i^2}$$

$$\propto \theta^{\frac{n}{2}-1} e^{-\frac{1}{2}\sum_{i=1}^n d_i^2 \cdot \theta}$$

It follows that $\theta = 1/\sigma^2$ has posterior

$$p(\theta|d) = \text{Gamma}\left(n/2, \ \sum d_i^2/2\right)$$

The data gives $n = 672$ and $\sum d_i^2 = 672 \times 13.85^2$. Then the 95% posterior interval can be obtained via software as

$$[1/214.1, \ 1/172.8].$$

Therefore the 95% posterior interval for σ^2 is $[172.8, \ 214.1]$, and the 95% posterior interval for σ is

$$[13.1, \ 14.6].$$

2. By approximation: useful, especially for large-sample inference.

A special example is about Beta distribution. For α, β large,

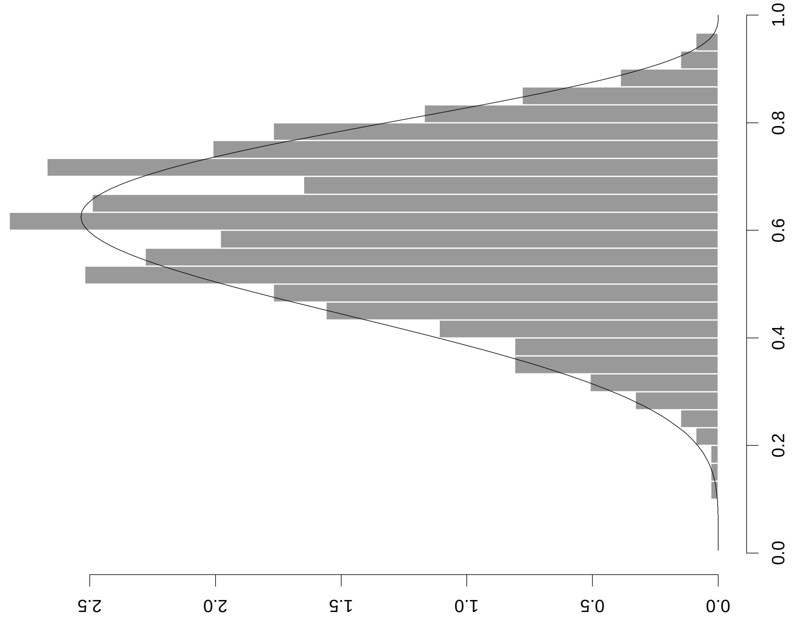
$$\text{Beta}(\alpha, \beta) \approx N\left(\frac{\alpha}{\alpha+\beta}, \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\right)$$

In general, one usually has a conditional central limit theorem for large samples.

$$\left(\frac{\theta - E[\theta|y]}{\sqrt{\text{Var}[\theta|y]}}\middle| y\right) \approx N(0,1)$$

3. By simulation: commonly used and very flexible. For many practical problems, it is very difficult to obtain analytical or analytical approximation for the posterior distribution. However it is often possible to *draw samples from the posterior distribution*, and use the sample distribution as an approximation of the posterior distribution.

Example. Suppose the posterior distribution for θ is Beta$(6, 4)$. The histogram of 1000 draws from this distribution is



The 25th sample is 0.290, and the 976th sample is 0.856. Therefore the 95% posterior interval is [0.290, 0.856]. Sample mean is 0.601, and sample standard deviation is 0.146. Note that the theoretical 95% posterior interval is [0.299, 0.863], mean 0.60, and standard deviation 0.148.

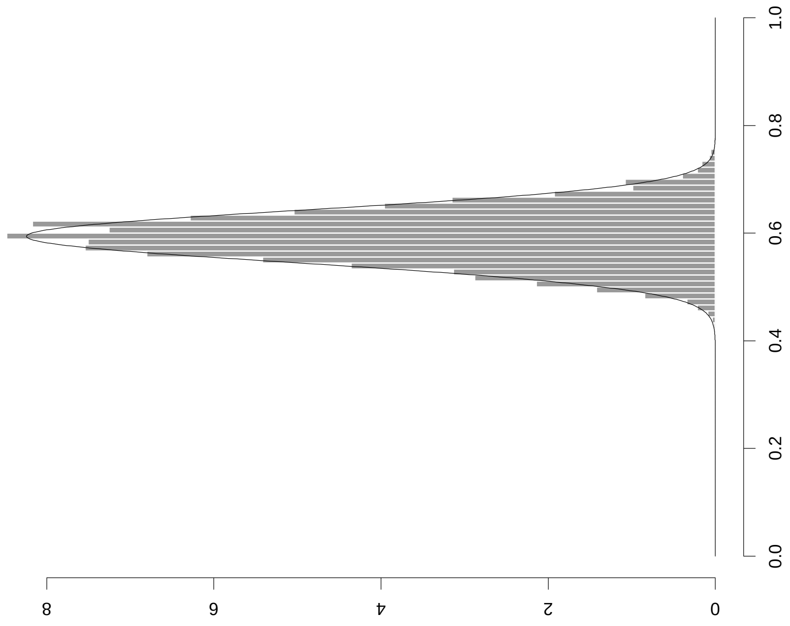
Example. Suppose Y has distribution $B(n;\theta)$ and θ has a prior density

$$\pi(\theta) = \begin{cases} \quad 2\theta \quad , & 0 \leq \theta \leq 0.5 \\ 2(1-\theta), & 0.5 < \theta \leq 1 \end{cases}$$

Suppose $n = 100$, and the observation is $y = 60$. The posterior distribution is

$$\begin{aligned} p(\theta|y) &\propto \pi(\theta)p(y|\theta) \\ &\propto \begin{cases} \theta^{y+1}(1-\theta)^{n-y}, & 0 \leq \theta \leq 0.5 \\ \theta^{y}(1-\theta)^{n+y-1}, & 0.5 < \theta \leq 1 \end{cases} \\ &= \begin{cases} \theta^{61}(1-\theta)^{40}, & 0 \leq \theta \leq 0.5 \\ \theta^{60}(1-\theta)^{41}, & 0.5 < \theta \leq 1 \end{cases} \end{aligned}$$

The histogram of 5000 draws from this distribution is as follows.



The 95% posterior interval from simulation is [0.494, 0.684]. Sample mean is 0.591, and sample standard deviation is 0.0486. Note that the theoretical 95% posterior interval is [0.299, 0.863], mean 0.593, and standard deviation 0.0479.