Chapter 5. Bayesian Statistics (II)

Bayesian for multi-parameter models

The principle remains the same. The (joint) posterior distribution given data y is once again

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

where $\theta = (\theta_1, \ldots, \theta_d)$ are the parameters of interest.

For illustration, consider the special case of $\theta = (\theta_1, \theta_2)$.

1. The joint posterior distribution
$$p(\theta_1, \theta_2|y) \propto \pi(\theta_1, \theta_2) \cdot p(y|\theta_1, \theta_2)$$
2. The marginal posterior distribution of θ_2
$$p(\theta_2|y) = \int p(\theta_1, \theta_2|y)\, d\theta_1 \propto \int \pi(\theta_1, \theta_2) \cdot p(y|\theta_1, \theta_2)\, d\theta_1$$
3. The conditional posterior distribution of θ_1 given θ_2 is
$$p(\theta_1|\theta_2, y) = \frac{p(\theta_1, \theta_2|y)}{p(\theta_2|y)} \propto \pi(\theta_1, \theta_2) \cdot p(y|\theta_1, \theta_2)$$
Note the difference with joint posterior distribution is that here θ_2 is regarded as *fixed and known*.

Remark: The following relation is useful for the simulation of posterior distribution
$$p(\theta_1, \theta_2|y) = p(\theta_1|\theta_2, y) \cdot p(\theta_2|y)$$

EXAMPLES

Normal model. Suppose that $y = \{y_1, \ldots, y_n\}$ are iid samples from $N(\theta, \sigma^2)$ such that $(\theta, \log(\sigma^2))$ has a flat prior, or

$$\pi(\theta, \sigma^2) \propto 1/\sigma^2.$$

The joint posterior distribution $p(\theta, \sigma^2|y)$.

$$p(\theta, \sigma^2|y) \propto \left(\sigma^2\right)^{-1-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left[n(\theta-\bar{y})^2+(n-1)s^2\right]}$$

where s^2 is the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The marginal posterior distribution $p(\sigma^2|y)$.

$$\begin{aligned}
p(\sigma^2|y) &= \int p(\theta, \sigma^2|y) d\theta \\
&\propto \int \left(\sigma^2\right)^{-1-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left[n(\theta-\bar{y})^2+(n-1)s^2\right]} d\theta \\
&= \left(\sigma^2\right)^{-1-\frac{n}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \sqrt{2\pi\sigma^2/n} \\
&\propto \left(\sigma^2\right)^{-\frac{n+1}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}}
\end{aligned}$$

It follows that the posterior distribution of

$$\left(\left. \frac{(n-1)s^2}{\sigma^2} \right| y \right) = \chi^2(n-1)$$

The marginal posterior distribution $p(\theta|y)$.

$$
\begin{aligned}
p(\theta|y) &= \int p(\theta, \sigma^2|y) d\sigma^2 \\
&\propto \int \left(\sigma^2\right)^{-1-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left[n(\theta-\bar{y})^2+(n-1)s^2\right]} d\sigma^2 \\
&\propto \left[n(\theta - \bar{y})^2 + (n-1)s^2\right]^{-\frac{n}{2}} \\
&\propto \left[1 + \left(\frac{\theta - \bar{y}}{s/\sqrt{n}}\right)^2 \frac{1}{n-1}\right]^{-\frac{n}{2}}
\end{aligned}
$$

It follows that the posterior distribution of

$$
\left(\left.\frac{\theta - \bar{y}}{s/\sqrt{n}}\right| y\right) = t(n-1)
$$

The conditional posterior distribution $p(\theta|\sigma^2, y)$.

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

The conditional posterior distribution $p(\sigma^2|\theta, y)$.

$$\left(\left.\frac{(n-1)s^2 + n(\bar{y}-\theta)^2}{\sigma^2}\right| \theta, y\right) = \chi^2(n)$$

Remark: To simulate from the posterior distribution $p(\theta, \sigma^2|y)$, one can first simulate σ^2 from marginal posterior distribution $p(\sigma^2|y)$, then simulate θ from the conditional posterior distribution $p(\theta|\sigma^2, y)$.

Example. Suppose a stock's daily return Y was recorded for $n = 22$ consecutive business days, with $\bar{y} = 5\%$ and $s = 4\%$. Assume that the daily return Y follows $N(\theta, \sigma^2)$ with prior $\pi(\theta, \sigma^2) \propto 1/\sigma^2$. Find the 95% posterior interval for θ. Also use simulation to approximate $E[\theta/\sigma|y]$.

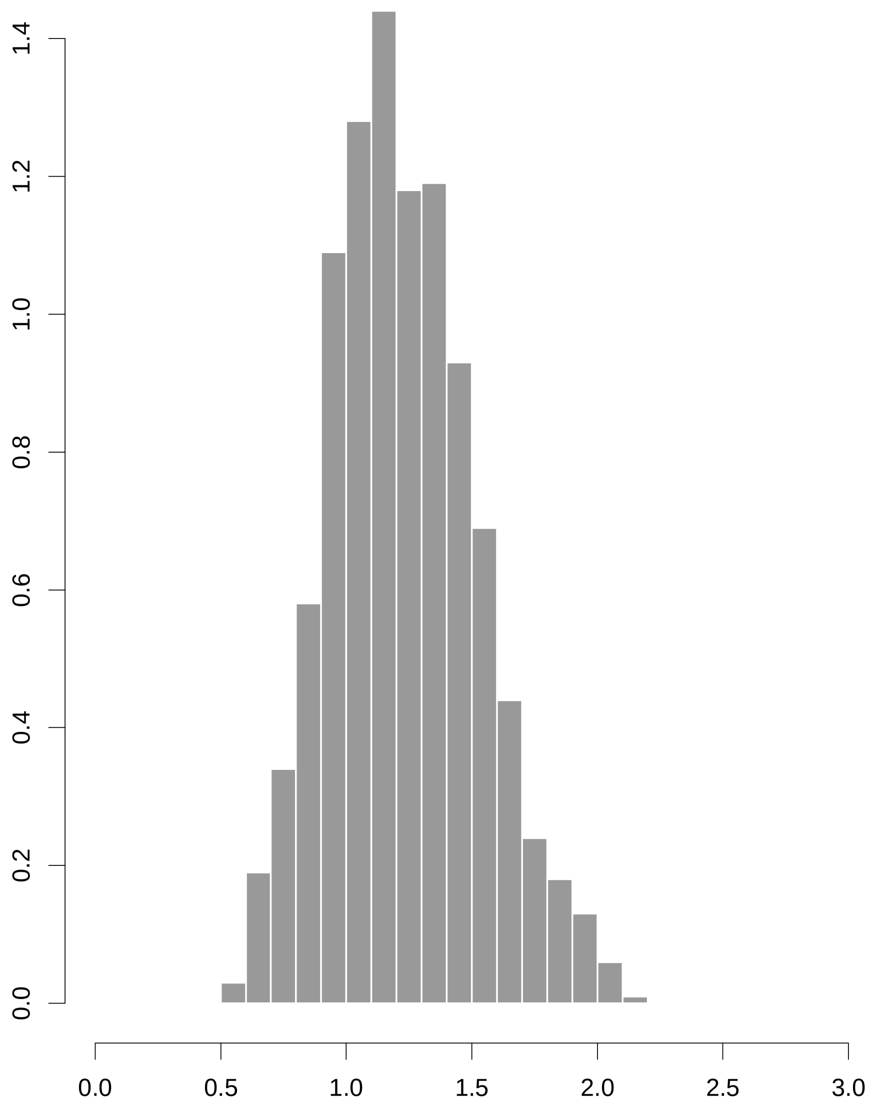
Solution: Since

$$\left(\frac{\theta - \bar{y}}{s/\sqrt{n}} \middle| y \right) = t(n-1)$$

The 95% posterior interval is (in %)

$$\bar{y} \pm t_{0.025}(n-1)\frac{s}{\sqrt{n}} = 5 \pm 2.080 * \frac{4}{\sqrt{21}} = [3.2, 6.8]$$

Below is the histogram of 1000 draws of θ/σ. For each draw, we (1) draw a sample of σ: draw a sample say u from $\chi^2(n-1)$, then let $\sigma = \sqrt{(n-1)s^2/u}$; (2) given σ, draw a sample θ from $N(\bar{y}, \sigma^2/n)$; (3) θ/σ is a data point. The sample average of θ/σ is 1.23.


0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Multinomial model. Let $Y = (Y_1, \ldots, Y_d)$ be multinomial with parameter $(n; \theta_1, \ldots, \theta_d)$ where

$$\theta_1 + \cdots + \theta_d = 1.$$

Consider prior distribution (Dirichlet distribution)

$$\pi(\theta) \propto \prod_{i=1}^{d} \theta_i^{\alpha_i - 1}$$

restricted to non-negative θ_i's with $\theta_1 + \cdots + \theta_d = 1$.

The joint posterior distribution $p(\theta|y)$.

$$p(\theta|y) \propto \pi(\theta) \cdot p(y|\theta) \propto \prod_{i=1}^{d} \theta_i^{\alpha_i - 1} \cdot \prod_{i=1}^{d} \theta_i^{y_i} = \prod_{i=1}^{d} \theta_i^{\alpha_i + y_i - 1}$$

That is, $p(\theta|y)$ is a Dirichlet distribution with parameter $(\alpha_1 + y_1, \ldots, \alpha_d + y_d)$.

The marginal posterior distribution $p(\theta_1|y)$.

$$p(\theta_1|y) \propto \int_{\{\sum_{i=2}^{d} \theta_i = 1 - \theta_1\}} \theta_1^{\alpha_1 + y_1 - 1} \prod_{i=2}^{d} \theta_i^{\alpha_i + y_i - 1} d\theta_2 \cdots d\theta_{d-1}$$

It follows that $p(\theta_1|y)$ is Beta$(\alpha_1 + y_1, \sum_{i=2}^{d} [\alpha_i + y_i])$.

The conditional posterior distribution $p(\theta_1|y)$.

$$p(\theta_2, \ldots, \theta_d|\theta_1, y) \propto \theta_1^{\alpha_1+y_1-1} \prod_{i=2}^{d} \theta_i^{\alpha_i+y_i-1}$$

restricted to $\{\theta_2 + \cdots + \theta_d = 1 - \theta_1\}$. It follows that

$$\left(\frac{\theta_2}{1-\theta_1}, \ldots, \frac{\theta_d}{1-\theta_1} \middle| \, \theta_1, y \right) = \text{Dirichlet}(\alpha_2 + y_2, \ldots, \alpha_d + y_d).$$

Remark on simulation: One way to simulate $(\theta_1, \ldots, \theta_d)$ from posterior distribution is to simulate sequentially θ_1 from $p(\theta_1|y)$, and then θ_2 from $p(\theta_2|\theta_1, y)$, ..., and θ_{d-1} from $p(\theta_{d-1}|\theta_1, \ldots, \theta_{d-2}, y)$, and finally set $\theta_d = 1-(\theta_1+\cdots+\theta_{d-1})$. Note that all these conditional distributions are Beta distributions [up to a multiplicative constant]. Another way to simulate $(\theta_1, \ldots, \theta_d)$ from posterior Dirichlet distribution is to simulate x_i from Gamma($\alpha_i + y_i$,1/2) for each $i = 1, \ldots, d$ and let $\theta_i = x_i/(x_1 + \cdots + x_d)$.

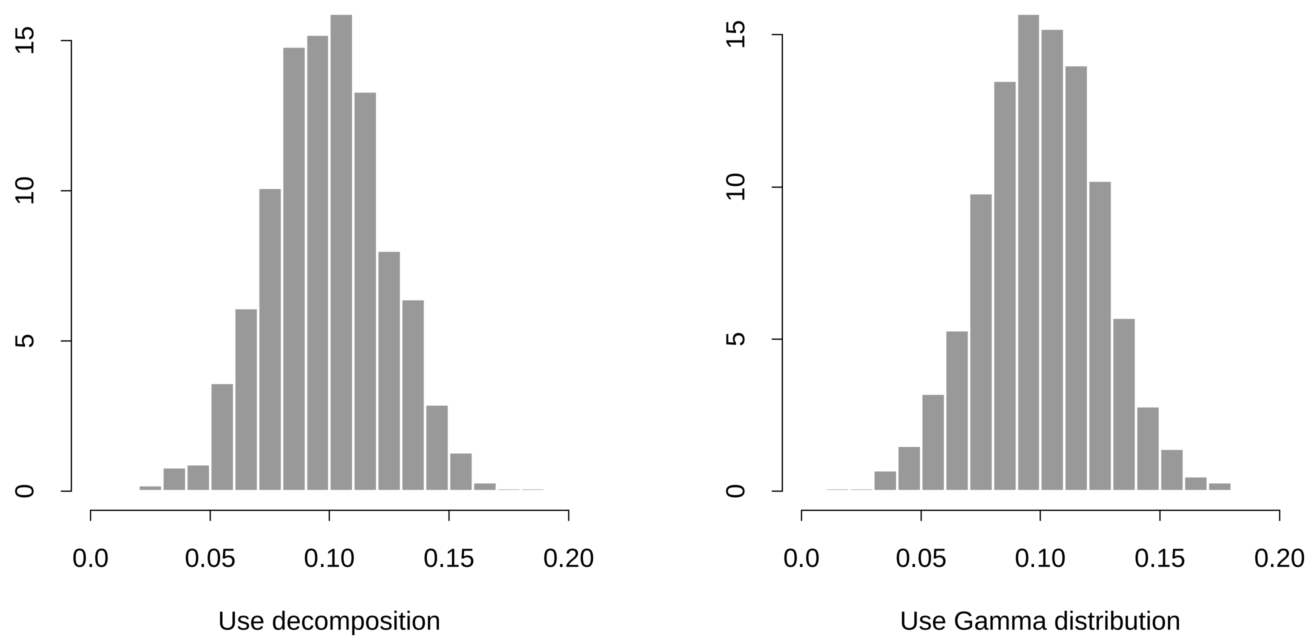
Example. In late October 1988, a pre-election poll was conducted by CBS news of 1447 adults in US to find out their preferences in the upcoming Presidential election. Out of 1447 persons, $y_1 = 727$ supported George Bush, $y_2 = 583$ supported Michael Dukakis, and $y_3 = 137$ supported other candidates or expressed no opinion. Assume that the samples are randomly selected from the population, then the data follows multinomial distribution with parameters $(\theta_1, \theta_2, \theta_3)$. The quantity of interest is $\theta_1 - \theta_2$.

Solution: Assume a non-informative prior with $\alpha_1 = \alpha_2 = \alpha_3 = 1$. The posterior distribution for $(\theta_1, \theta_2, \theta_3)$ is Dirichlet(728, 584, 138). We will draw 1000 samples of $(\theta_1, \theta_2, \theta_3)$ from the posterior Dirichlet distribution, and compute $\theta_1 - \theta_2$ for each sample. We will simulate using two equivalent approaches.

- Using conditional distribution decomposition. Simulate θ_1 from Beta(728, 584+138). Given θ_1, simulate u from Beta(584, 138) and let $\theta_2 = (1 - \theta_1)u$. Let $\theta_3 = 1 - \theta_2 - \theta_3$. Record $\theta_1 - \theta_2$.

- Using Gamma distribution. Simulate independent x_1, x_2, x_3 from, respectively, Gamma(728, 1/2)=χ^2(728*2), Gamma(584, 1/2)=χ^2(584*2), and Gamma(138, 1/2)=χ^2(138*2). Let $\theta_i = x_i/(x_1 + x_2 + x_3)$. Record $\theta_1 - \theta_2$.

The histograms are attached below, the sample means are 0.099 and 0.100 respectively. None of the sample points of $\theta_1 - \theta_2$ are below zero.



0
5
10
15
0.0
0.05
0.10
0.15
0.20
Use decomposition
Use Gamma distribution

Comparison of two populations

Comparison of two proportions. Suppose Y_1 has distribution $B(n_1; \theta_1)$, Y_2 has distribution $B(n_2; \theta_2)$, and Y_1 and Y_2 are independent. We are interested in $\theta_1 - \theta_2$, given the data $Y_1 = y_1$ and $Y_2 = y_2$.

Assuming a non-informative prior $\pi(\theta_1, \theta_2) \propto 1$ on $[0, 1]^2$. The joint posterior distribution $p(\theta_1, \theta_2|y)$ is

$$p(\theta_1, \theta_2|y) \propto \theta_1^{y_1}(1-\theta_1)^{n_1-y_1}\theta_2^{y_2}(1-\theta_2)^{n_2-y_2}$$

Thus the posterior distributions of θ_1 and θ_2 are independent and

$$p(\theta_1|y) = \text{Beta}(y_1+1, n_1-y_1+1)$$

$$p(\theta_2|y) = \text{Beta}(y_2+1, n_2-y_2+1)$$

One can use simulation to draws samples of $\theta_1 - \theta_2$ or use normal approximations (when n_1 and n_2 large) of $\theta_1 - \theta_2$.

Comparison of two normal means. Suppose $x = (x_1, \ldots, x_{n_1})$ are iid samples from $N(\theta_1, \sigma^2)$, $y = (y_1, \ldots, y_{n_2})$ are iid samples from $N(\theta_2, \sigma^2)$, and that the two samples are independent. We are interested in $\theta_1 - \theta_2$. All the parameters $(\theta_1, \theta_2, \sigma)$ are unknown.

Assume a non-informative prior $\pi(\theta_1, \theta_2, \sigma^2) \propto 1/\sigma^2$. The posterior is

$$p(\theta_1, \theta_2, \sigma | x, y) \propto (\sigma^2)^{-1-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}[n_1(\bar{x}-\theta_1)^2+n_2(\bar{y}-\theta_2)^2+(n-2)s_p^2]}$$

where

$$n = n_1 + n_2, \quad s_p^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{(n_1 - 1) + (n_2 - 2)}$$

Analogously, one have the marginal posterior distribution

$$p(\sigma^2 | x, y) \propto \left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{(n-2)s_p^2}{2\sigma^2}}$$

or

$$\left(\frac{(n-2)s_p^2}{\sigma^2} \middle| x, y \right) = \chi^2(n-2).$$

The conditional posterior distributions of θ_1, θ_2 given σ are independent, and

$$p(\theta_1|\sigma, x, y) = N(\bar{x}, \sigma^2/n_1), \quad p(\theta_2|\sigma, x, y) = N(\bar{y}, \sigma^2/n_2).$$

Remark on simulation. To draw samples of $(\theta_1, \theta_2, \sigma)$. One can draw u from $\chi^2(n-2)$ and let $\sigma^2 = (n-2)s_p^2/u$, then draw θ_1, θ_2 independently from $N(\bar{x}, \sigma^2/n_1)$ and $N(\bar{y}, \sigma^2/n_2)$ respectively. If one is interested in $\theta_1 - \theta_2$, for each sample point of $(\theta_1, \theta_2, \sigma)$ compute $\theta_1 - \theta_2$. If one is interested $\theta_1\theta_2$, for each sample point compute $\theta_1\theta_2$. And so on so forth.

The theoretical posterior distribution of $\theta_1 - \theta_2$ can be obtained as follows. Note that the conditional posterior distribution of $\theta_1 - \theta_2$

given σ is

$$p(\theta_1 - \theta_2 | \sigma, x, y) = N(\bar{x} - \bar{y}, \sigma^2[1/n_1 + 1/n_2]).$$

Therefore

$$p(\theta_1 - \theta_2, \sigma^2 | x, y) = p(\theta_1 - \theta_2 | \sigma^2, x, y) \cdot p(\sigma^2 | x, y)$$

$$\propto \left(\sigma^2\right)^{-\frac{n+1}{2}} e^{-\frac{1}{2\sigma^2}[(1/n_1+1/n_2)^{-1}((\theta_1-\theta_2)-(\bar{x}-\bar{y}))^2+(n-2)s_p^2]}$$

Integrating out σ^2, we have similarly

$$\left(\left. \frac{(\theta_1 - \theta_2) - (\bar{x} - \bar{y})}{s_p \cdot \sqrt{1/n_1 + 1/n_2}} \right| x, y \right) = t(n-2)$$

Example. Who is a better hitter, Ted Williams (Boston Red Sox) or Joe DiMaggio (NY Yankees)? Their major league career statistics are given below.

Player	At-bats	Hits	Batting Average	Home Run	Home Run Average
T.W.	7706	2654	.3444	521	.0676
J.D.	6821	2214	.3246	361	.0529

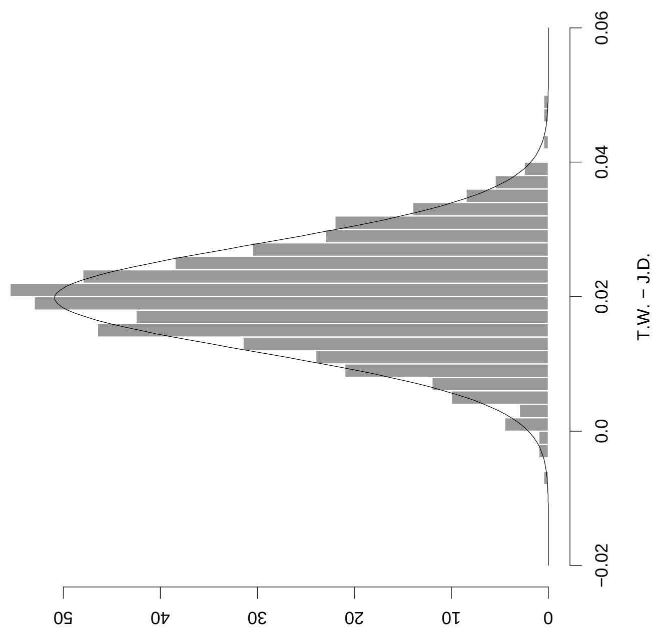
Find the posterior probability that Ted Williams is a better hitter than Joe Dimaggio.

Solution: We consider the hits, and leave the home runs as exercise. Let θ_1 be the hit proportion for T.W. and θ_2 for that of J.D. Assume a non-informative prior $\pi(\theta_1, \theta_2) \propto 1$. Then the posterior is

$$p(\theta_1, \theta_2|y) \propto \theta_1^{2654}(1-\theta_1)^{5052} \cdot \theta_2^{2214}(1-\theta_2)^{4607}$$

We are interested in $P(\theta_1 - \theta_2 > 0|y)$. We simulate 1000 draws of $\theta_1 - \theta_2$ [we simulate θ_1 and θ_2 independently from Beta(2655,5053) and Beta(2215, 4608), respectively, and compute $\theta_1 - \theta_2$ for each (θ_1, θ_2).]

Below is the histogram of $\theta_1 - \theta_2$. Among 1000 draws, 995 are positive. Therefore the posterior probability $P(\theta_1 - \theta_2 > 0|y) \approx 0.995$.



If we use normal approximation, $\theta_1 - \theta_2$ are approximately distributed as

$$N\left(\frac{2654}{2654+5052} - \frac{2214}{2214+4607}, \frac{2654*5052}{(2654+5052)^2(2654+5052+1)} + \frac{2214*4607}{(2214+4607)^2(2214+4607+1)}\right) = N(0.0198, 0.0078^2).$$

Its density is super-imposed on the histogram.

Example. Does birth weight increase when a mother quits smoking? Below is a data set.

Smokes					Quit	
4.5	6.1	6.9	7.5	9.9	5.4	7.2
5.4	6.4	6.9	7.6		6.6	7.3
5.6	6.6	7.1	7.6		6.8	7.4
5.9	6.6	7.1	7.8		6.8	
6.0	6.6	7.2	8.0		6.9	

Assume the birth weight of a baby whose mother who smokes is $N(\theta_1, \sigma^2)$ and the birth weight of a baby whose mother once smoked but quit is $N(\theta_2, \sigma^2)$. Find the posterior probability of $\theta_1 - \theta_2 > 0$, and give a 95% posterior interval for $\theta_1 - \theta_2$.

Solution: The data $n_1 = 21$, $n_2 = 8$, and (for smoke) $\bar{x} = 6.824$, $s_x = 1.093$, (for quit) $\bar{y} = 6.800$, $s_y = 0.589$. The pooled estimate

$$s_p^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2} = 0.9749, \quad s_p = 0.987$$

To simulate $\theta_1 - \theta_2$, we first draw u from $\chi^2(n-2)$ and let $\sigma^2 = (n-2)s_p^2/u$, and then simulate θ_1 and θ_2 independently from $N(\bar{x}, \sigma^2/n_1)$ and $N(\bar{y}, \sigma^2/n_2)$. The histogram of 1000 draws are below. The 95% posterior interval from simulation is $[-0.807, 0.863]$. Out of these 1000 draws of $\theta_1 - \theta_2$, 499 are positive. So the posterior probability of $\theta_1 - \theta_2 > 0$ is 0.499.

Note that theoretically

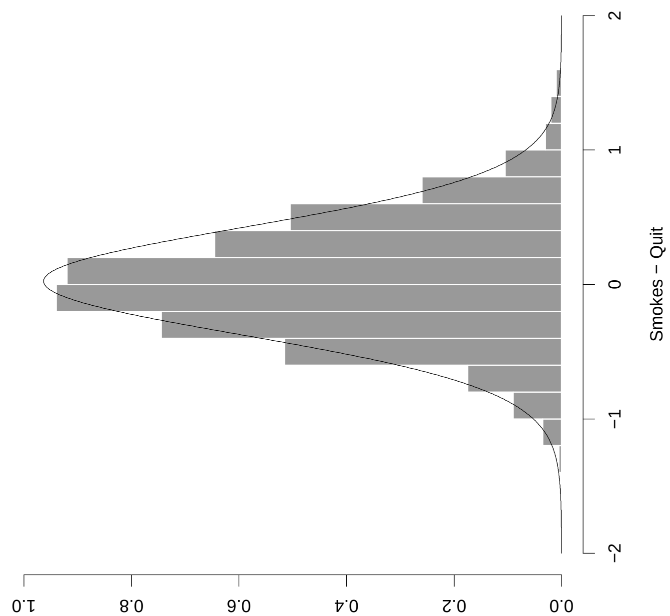
$$\left(\left. \frac{(\theta_1 - \theta_2) - (\bar{x} - \bar{y})}{s_p\sqrt{1/n_1 + 1/n_2}} \right| x, y \right) = t(n-2).$$

Therefore the theoretical 95% posterior interval is

$$(\bar{x} - \bar{y}) \pm t_{0.025}(n-2) * s_p\sqrt{1/n_1 + 1/n_2} = [-0.818, 0.866]$$

and

$$P(\theta_1 - \theta_2 > 0|x, y) = P\left[t(n-2) \geq -\frac{(\bar{x} - \bar{y})}{s_p\sqrt{1/n_1 + 1/n_2}}\right] = 0.523.$$



An example of generalized linear model

It is rare that multiparameter models allow simple calculation of posterior distribution. Simulation is often the only available tool for data analysis. In this section we discuss in detail a two-parameter generalized linear model for a bioassay experiment.

The problem and the data. In the development of drugs, acute toxicity test or bioassay are commonly performed on animals. The animal responses are typically dichotomous: alive or dead, tumor or no tumor, and so on. The experiments are often administered by injecting various dose levels of the compound to batches of animals, which generate data of form (x_i, n_i, y_i), where x_i is the dose level (often measured in logarithmic scale), n_i is the size of the batch of animals receiving dose x_i, and y_i is the number of animals with positive response. The specific real data set is shown below.

Dose x_i (log g/ml)	Size of batch n_i	Number of deaths y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Statistical model. Assume that y_i is Binomial (n_i, θ_i), with θ_i the population death rate for animals receiving dose x_i. We would like θ_i to be dependent on x_i, and by definition $\theta_i \in [0, 1]$. The following *logistic regression* model is adopted.

$$\text{logit}(\theta_i) = \alpha + \beta x_i$$

where $\text{logit}(\theta) \doteq \log(\theta/(1-\theta))$. The inverse function of $\text{logit}(\cdot)$ is

$$\text{logit}^{-1}(u) = e^u/(1+e^u).$$

Note that in this model x_i's are explanatory variables and regarded as fixed.

Prior and likelihood. We use a flat prior $\pi(\alpha, \beta) \propto 1$ and the likelihood

$$p(y_i|\alpha, \beta) \propto \left[\text{logit}^{-1}(\alpha + \beta x_i)\right]^{y_i} \cdot \left[1 - \text{logit}^{-1}(\alpha + \beta x_i)\right]^{n_i - y_i}.$$

The posterior $p(\alpha, \beta|y)$. We have

$$p(\alpha, \beta|y) \propto \pi(\alpha, \beta) \cdot \prod_{i=1}^{4} p(y_i|\alpha, \beta) \propto \prod_{i=1}^{4} p(y_i|\alpha, \beta)$$

Discretization of the posterior distribution. There is no analytical expression to the posterior distribution, and we will use simulation to obtain numerical summaries. Since the problem is only two dimensional, it is reasonable to expect that simulating from a discretized approximation of the continuous posterior distribution will do a good job. We will restrict the region to $(\alpha, \beta) \in [-2, 6] \times [-5, 30]$. The contour plot is shown below.

The discretization is done on a uniform 400×700 grid. For each grid point, we compute the unnormalized posterior density. Afterwards we normalize these quantities such that their sum over all the grid points become one. In other words, we now have a discrete approximation of the posterior distribution.

Remark. A very popular methodology to simulate the posterior distribution is the so-called Markov Chain Monte Carlo (MCMC) method. It is very different from the discretization method we used in this example. When the dimension gets higher, discretization becomes obviously much more difficult.

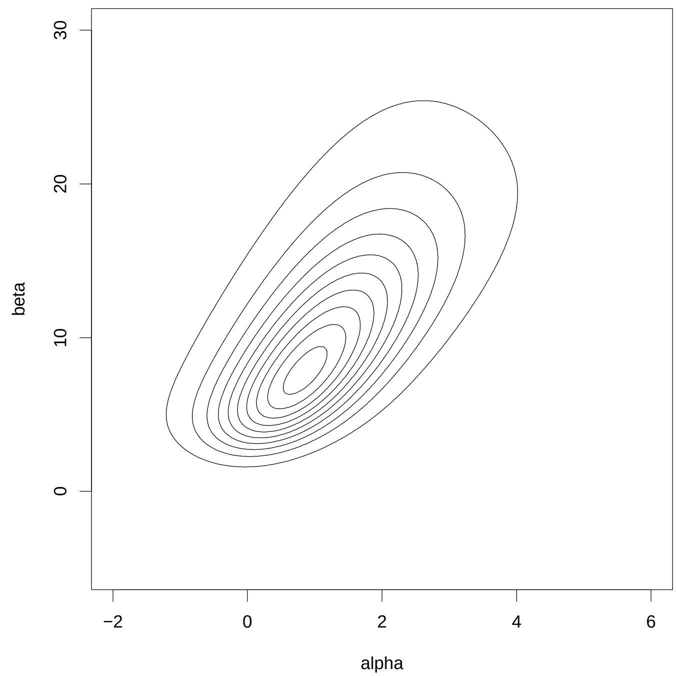


Figure 1: contour plot for the posterior distribution

Simulating from the discrete approximation of the posterior distribution.

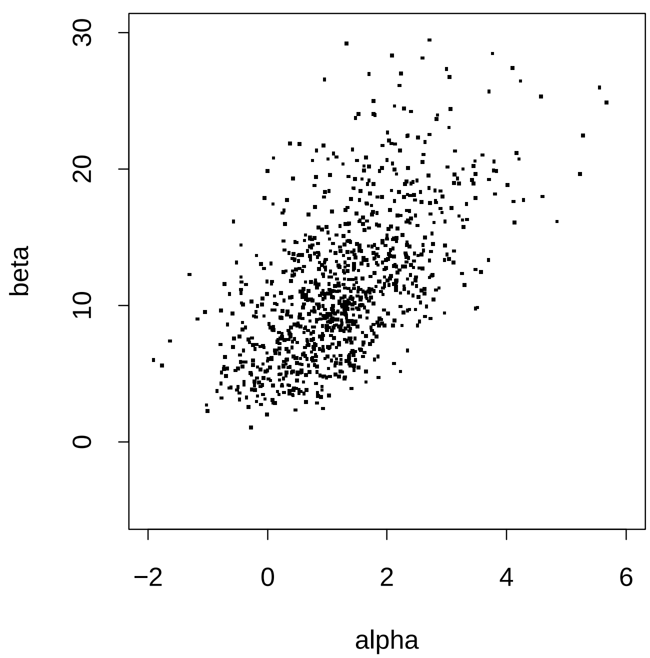
1. Draw α from its discrete marginal distribution $p(\alpha|y)$.
2. Given α, draw β from the discrete conditional distribution $p(\beta|\alpha, y)$.
3. Jitter the sample α and β by adding a uniform random perturbation centered at zero with a width equal to the spacing of the sampling grid.
4. Repeated these three steps 1000 times to obtain 1000 samples of (α, β).

The histogram is attached below

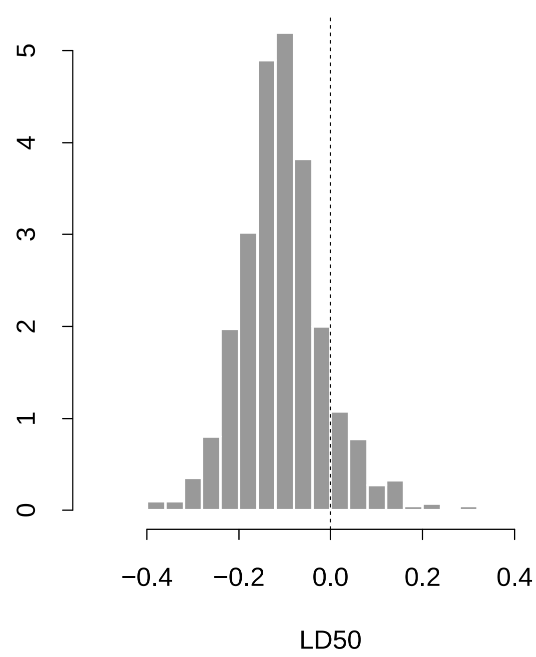
The quantities of interest. The sign of β is important. For all the 1000 samples we have $\beta > 0$, which indicates the compound is harmful. Another quantity of interest in LD50 – the dose level at which the probability of death is 50%, or

$$\alpha + \beta \cdot \text{LD50} = \text{logit}^{-1}(0.5) = 0 \quad \Rightarrow \quad \text{LD50} = -\alpha/\beta.$$

The histogram of LD50 is attached.



30
20
10
0
beta
−2
0
2
4
6
alpha



5
4
3
2
1
0
−0.4
−0.2
0.0
0.2
0.4
LD50