1. Suppose we are interested in the sleeping habits of university students. In particular, what proportion of university students get at least eight hours of sleep?

   Here we think of a population consisting of all Australian university students and let p represent the proportion of this population who sleep (on a typical night during the week) at least eight hours. We are interested in learning about the location of p.

   The value of the proportion p is unknown. In the Bayesian viewpoint a person's beliefs about the uncertainty in this proportion are represented by a probability distribution placed on this parameter. This distribution reflects the person's subjective prior opinion about plausible values of p.

   A random sample of students from a particular university will be taken to learn about this proportion. Combined with this, an expert in the field of sleep believes that university students generally get less than eight hours sleep and so p is likely to be smaller than 0.5. When pressed, she says that she is 90% confident that p is less than 0.5 and her best guess at the value of p is 0.3. She supposes that the true p is as equally likely to be above 0.3 as it is to be be below 0.3.

   A sample of 27 students is taken – in this group, 11 record that they had at least eight hours of sleep the previous night. Based on the prior information and this observed data, the person is interested in estimating the proportion p.

   Suppose that our prior density for p is denoted by $f(p)$. If we regard a "success" as sleeping at least eight hours and we take a random sample with x successes and $n-x$ failures, then the likelihood function is given by

   $$f(x|p) \propto p^x(1-p)^{n-x}, \quad 0 < p < 1.$$

   The posterior density for p, by Bayes' rule, is obtained, up to a proportionality constant, by multiplying the prior density by the likelihood:

   $$f(p|x) \propto f(p)f(x|p).$$

   Since the proportion is a continuous parameter on finite support, a natural approach is to construct a density $f(p)$ on the interval $(0,1)$ that represents the person's initial beliefs. A convenient family of densities for a proportion is the beta distribution with density

   $$f(p) \propto p^{a-1}(1-p)^{b-1}, \quad 0 < p < 1,$$

   where the hyperparameters a and b are chosen to reflect the user's prior beliefs about p. Based on our experts advice the median and the 90th percentile are given, respectively, by 0.3 and 0.5, and this can be matched (by trial and error) with a beta density with $a = 0.34$ and $b = 7.4$.

(a) Combining this beta prior with the likelihood function, show that the posterior density is also of the beta form and find the new parameters.

Solution: *The updated parameters are* $a + x$ *and* $b + (n - x)$*:*

$$g(p|x) \propto p^{a+x-1}(1-p)^{b+(n-x)-1}, \quad 0 < p < 1.$$

In our present example $x = 11$ *and* $n - x = 16$ *so our posterior density is:*

$$g(p|x) \propto p^{3.4+11-1}(1-p)^{7.4+16-1}, \quad 0 < p < 1.$$

This is an example of a conjugate prior as the prior and posterior densities have the same functional form.

(b) We want to check if it is likely that proportion of heavy sleepers is greater than 0.5. I.e. compute the posterior probability $P(p \geq 0.5|x)$ using `pbeta`.

Solution:

```
a = 3.4
b = 7.4
x = 11
n = 27
1 - pbeta(0.5, a + x, b + n-x)
[1] 0.0684257
```

This probability is small, so it is unlikely that more than half of the students are heavy sleepers.

(c) Construct a 90% interval estimate for p using `qbeta`.

Solution:

```
qbeta(c(0.05, 0.95), a + x, b + n-x)
[1] 0.2562364 0.5129274
```

(d) The above summaries are exact because they are based on R functions for the beta posterior density. We can also use simulation. Simulate 1000 values from the posterior distribution using

`sim = rbeta(1000, a+x, b+n-x)`

Find an estimate of the probability that the proportion is larger than 0.5 using the simulated data. Is this similar to what you found before?

Solution:

```
sum(sim >= 0.5)/1000
[1] 0.064
```

(e) Using the simulated data find a point estimate for p. Is this similar to what you found earlier?

Solution:

```
median(sim)
[1] 0.3763889
```

(f) Using the simulated data find a 90% confidence interval for p using your simulations from the posterior distribution. Is this similar to what you found earlier?

Solution:

```
quantile(sim, c(0.05, 0.95))
       5%       95%
0.2544360 0.5077646
```

(g) Plot the likelihood, prior and posterior on the same graph using the code below and comment on the result.

```
p = seq(0, 1, length = 500)
prior=dbeta(p,a,b)
like=dbeta(p,x+1,n-x+1)
post=dbeta(p,a+x,b+n-x)
plot(p,post,type="l",ylab="Density",lty=2,lwd=3)
lines(p,like,lty=1,lwd=3)
lines(p,prior,lty=3,lwd=3)
legend(.7,4,c("Prior","Likelihood","Posterior"),
     lty=c(3,1,2),lwd=c(3,3,3))
```

Solution:

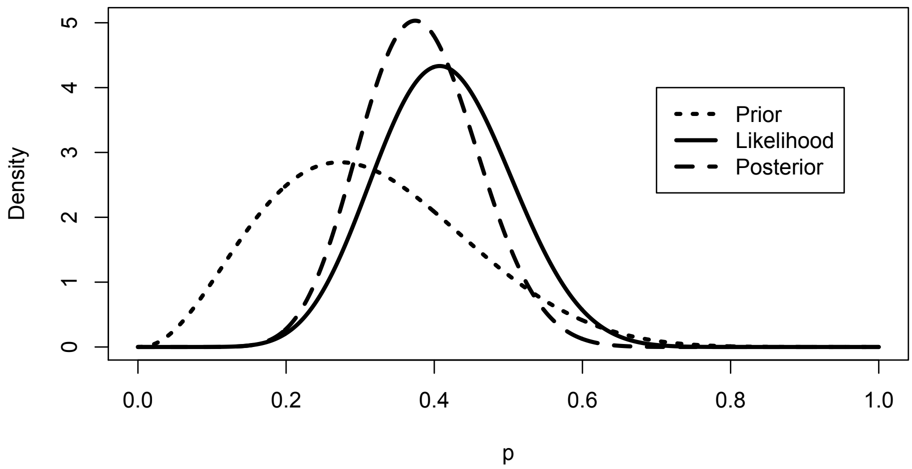

Figure 1: The posterior is located between the likelihood and the prior – reflecting how both sets of information are combined to come up with the posterior distribution (kind of like a weighted average of the two).

2. Gelman et. al. (2003) consider a problem of estimating an unknown variance using American football scores. The focus is on the difference d between a game outcome (winning score minus losing score) and a published point spread. We observe $d_1, \ldots, d_n$, the observed differenced between game outcomes and point spreads for n football games. If these differences are assumed to be a random sample from a normal distribution with mean 0 and unknown variance σ^2, the likelihood function is given by

$$f(\mathbf{d}|\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\sum_{i=1}^{n} d_i^2/(2\sigma^2)\right\}, \quad \sigma^2 > 0.$$

Suppose the non informative prior density $f(\sigma^2) \propto 1/\sigma^2$ is assigned to the variance. This is the standard vauge prior placed on variance – it is equivalent to assuming that the logarithm of the variance is uniformly distributed on the real line. Then the posterior density of σ^2 is given, up to a proportionality constant by,

$$f(\sigma^2|\mathbf{d}) \propto (\sigma^2)^{-n/2-1} \exp\left\{\frac{-v}{2\sigma^2}\right\},$$

where $v = \sum_{i=1}^{n} d_i^2$. If we define the precision parameter $P = 1/\sigma^2$, then it can be shown that P is distributed as U/v where $U \sim \chi_n^2$.

(a) Import and view the data using the following commands:

```
fball = read.table("http://www.maths.usyd.edu.au/u/gartht/fball")
attach(fball)
d = favorite - underdog - spread
hist(d)
n = length(d)
v = sum(d^2)
```

Description of the variables:

`favorite` is the score of the favourite team

`underdog` is the score of the underdog team

`spread` is the published point spread. The point spread is essentially a handicap towards the underdog – used particularly for betting, e.g. the bet might be "Will the favourite win by more than the point spread?"

Solution: *See Figure 2.*

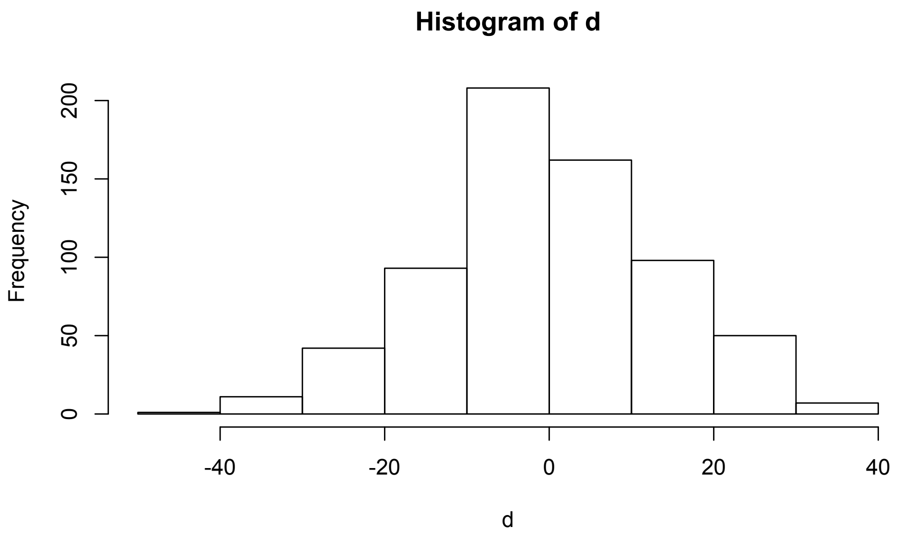

Figure 2: The distribution of the differences is roughly centered at zero with quite large variability.

(b) Simulate 1000 values from the posterior distribution of the standard deviation in two steps. First, simulate values of the precision parameter $P = 1/\sigma^2$ from the scaled χ_n^2 distribution by the command `rchisq(1000,n)/v`. Then perform the transformation $\sigma = \sqrt{1/P}$ to get values from the posterior distribution of the standard deviation σ.

Solution:

```
P = rchisq(1000, n)/v
s = sqrt(1/P)
```

(c) Construct a histogram of the draws of σ.

Solution:

```
hist(s)
```

See Figure 3.

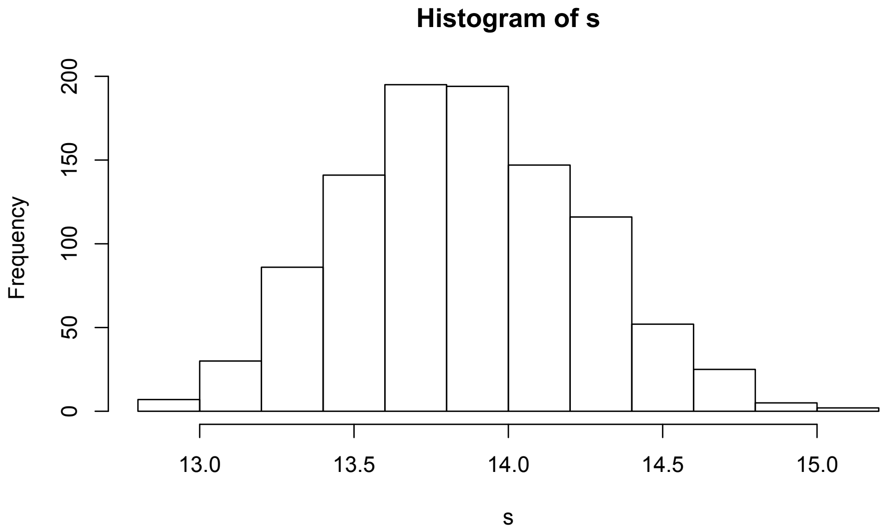

Figure 3: The distribution of the draws of σ from the posterior distribution.

(d) Find a point estimate for the standard deviation σ. I.e. find the posterior median.

Solution:

```
median(s)
[1] 13.8446
```

(e) Find a 95% confidence interval for σ.

Solution:

```
quantile(s, probs = c(0.025,0.975))
    2.5%    97.5%
13.15459 14.64796
```

References

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003), *Bayesian Data Analysis*, New York: Chapman and Hall.

Albert, J. (2009). *Bayesian computation with R.* Use R! New York: Springer.