Bayesian Statistics VIII

Matthieu Vignes

April 2024

Monte Carlo sampling from a Beta Posterior distribution

For review, if q, the frequency of the `1` allele in a population has a Beta prior with parameters a and b, we say:

$$q \sim \text{Beta}(a, b)$$

And, if x counts the number of `1` alleles at a marker found in our sample of size N gene copies from the population (for example, of elephants). Then the posterior probability distribution for q is also a Beta distribution:

$$P(q|x) \propto P(x|q)P(q) = q^{x+a-1}(1-q)^{N-x+b-1}.$$

This distribution can be recognised as a Beta distribution with parameters $x + a - 1$ and $N - x + b - 1$:

$$q|x \; sim \text{Beta}(x + a, N - x + b).$$

We will consider Monte Carlo sampling in the context of having a posterior distribution for the frequency q of the `1` allele in a single population.

Imagine we are estimating the frequency of the `1` allele when we have sampled 50 diploid individuals from a population. This means we have a sample of 100 gene copies. For the purposes of estimating q, in a population that is in Hardy-Weinberg equilibrium, this is similar to sampling $N = 100$ haploid individuals.

Suppose 30 of those gene copies were of allelic type `1` and the remaining 70 were of allelic type `0`.

Further, assume a uniform prior, (i.e., a Beta$(1, 1)$ prior distribution) on q, the frequency of the `1` allele. As we saw earlier, such a setup yields a posterior distribution for the allele frequency that is a Beta$(31, 71)$ distribution.

Various summaries of this posterior (posterior mean, posterior median, credible intervals, etc.) are analytically available; however, we will practice obtaining them (and, further below, more elaborate summaries) using Monte Carlo sampling.

A sample from a Beta distribution can be obtained from R's `rbeta` function. Thus the following gives a Monte Carlo sample of size $n = 1000$ from a Beta$(31, 71)$ distribution:

```
S <- rbeta(1000, 31, 71)
```

Exercises

In the following, one solution to every even-numbered exercise is provided by way of an example. For many of the other exercises, some skeleton code is given, with ... left where you should insert additional code.

1. First get a visual approximation of the posterior distribution (as a histogram) using a Monte Carlo sample of size $1,000,000$.

```
hist(..., breaks = 100)
```

2. Find the posterior mean of the allele frequency from 5 different Monte Carlo samples of size 1000.

```
result <- rep(NA, 5)
for(i in 1:5) {
  result[i] <- mean(rbeta(1000, 31, 71))
}
result
```

```
## [1] 0.3026523 0.3069116 0.3043622 0.3033542 0.3056441
```

Note, here that for any vector x of length n, `mean(x)` is a convenient way of computing $1/n \sum_{i=1}^{n} x_i$.

3. Do the same for 5 different Monte Carlo samples of size 100

```
result <- rep(NA, 5)
for(i in 1:5) {
  result[i] <- ...
}
```

4. Get a Monte Carlo estimate ($n = 10,000$) of the posterior probability that the allele frequency is greater than 0.35.

```
result <- mean(rbeta(1e4, 31, 71) > 0.35)
result
```

```
## [1] 0.1504
```

Take a moment to look at this and see that the $g(x)$ we talked about is an indicator function.

5. Get a Monte Carlo estimate, using $n = 100,000$, that q is between 0.2 and 0.4.

```
S <- rbeta(1e5, 31, 71)
result <- mean( ... )
```

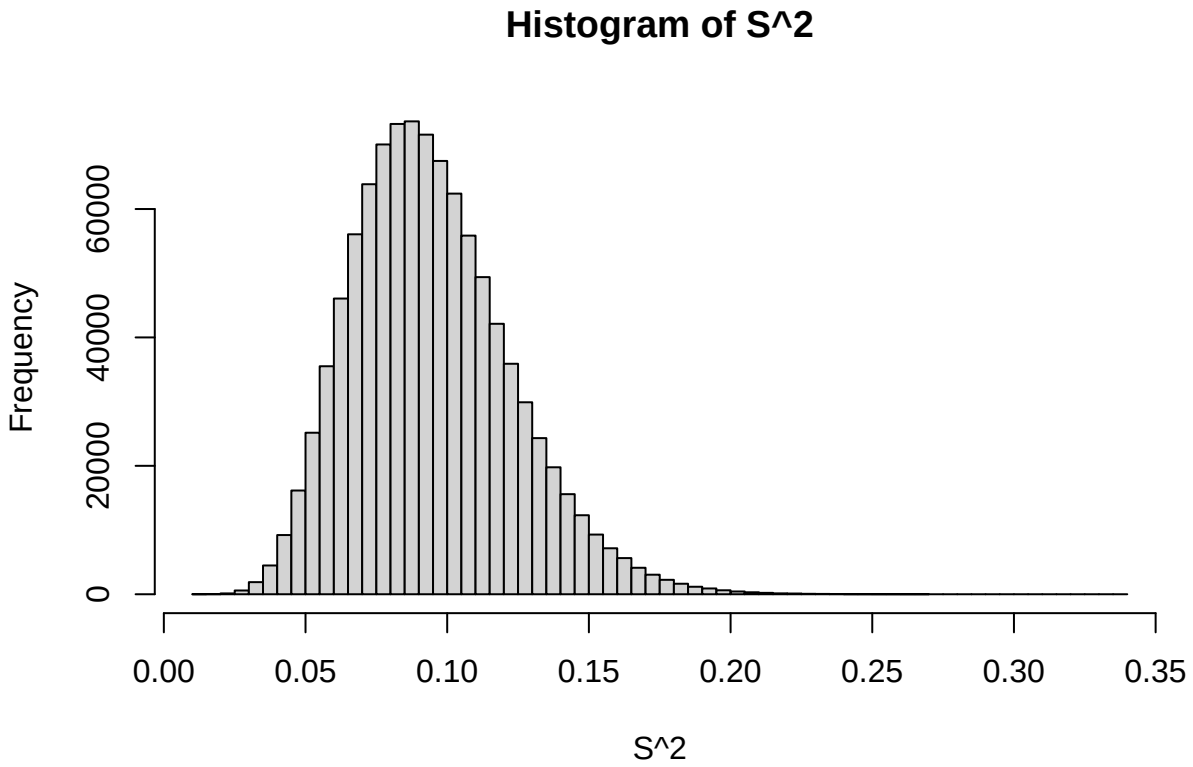
The following three exercises illustrate one of the convenient aspects of Monte Carlo sampling from a posterior. If you are interested in the posterior distribution of some function of the variable you have the posterior for, you can evaluate that directly by sampling the posterior you have in hand. This is often much easier than trying to analytically derive the posterior for your function via methods for finding the distribution of transformed random variables.

For a concrete example, here, we have the posterior distribution of the frequency of the `1` allele, but we might be more interested in the frequency of individuals with the genotype `11`; i.e., those diploid individuals that have two copies of the `1` allele.

Assume that the population is in Hardy-Weinberg equilibrium. If a population is in Hardy Weinberg equilibrium and allele `1` is at a frequency of q, then the expected frequency of `11` homozygotes is q^2.

6. Use a Monte Carlo sample of size $n = 1,000,000$ of the allele frequency to make an approximation (as a histogram) of the posterior distribution of the frequency of `11` homozygotes.

```
S <- rbeta(1e6, 31, 71)
hist(S^2, breaks = 100)
```



7. Find the posterior probability that the frequency of `11` homozygotes (assuming the HWE) is greater than 0.15.

```
S <- rbeta(1e6, 31, 71)
result <- mean( ... )
```

8. Given the posterior for the frequency of the `1` allele, and assuming the HWE, evaluate the posterior probability that a new and different sample of 10 individuals from this population will not include a `11` homozygote.

```
# if H is the frequency of a 11 homozygote, then the probability that
# a sample of 10 individuals does not have a 11 homozygote in it is
# (1 - H) ^ 10, and H is q^2, so....

# here are our simulated random variables
S <- rbeta(1e6, 31, 71)

# here is our function of simulated random variables
g <- function(q) (1 - q^2) ^ 10

# here is our Monte Carlo estimate
result <- mean(g(S))

result
```

```
## [1] 0.3862238
```

Once again, `mean(g(S))` is a quick way of writing $1/n \sum_{i=1}^{n} g(x(i)$ in R.