Example female birth rate

Background

Here we revisit the analyses from Chapter 2 of BDA 3 related to estimating the proportion of female births in European-race populations. This is an example of estimating a population parameter. Hence, the first thing we need to do is to define more carefully what do we mean by *population*. For example, are we interested in the sampled population (that is those 241945 + 251527 babies Laplace had in his data) or are we interested in some larger *superpopulation*. If we are interested in the former case, we don't need to do any statistical analysis since the ratio of girls to boys in the sample is exactly

```
241945/(241945 + 251527)
```

```
## [1] 0.4902912
```

In the latter case we need to define the superpopulation. Here, we are interested in the ratio of girls to boys in all new born babies in Paris to be interpreted broadly to include past, current and future babies of Laplace's time (This still leaves some vagueness in how "Laplace's time" is defined but let's not go to that). Now, we can denote by θ the proportion of all female births in this superpopulation of *all babies*. Clearly we have not observed all the babies and, thus, cannot just calculate simple ratio.

However, the available data is a random sample from this superpopulation with sample size $n = 241945 + 251527$ and $y = 241945$ girls. This allows for statistical inference on θ. Hence, let's first assume that our prior for the sex ratio is uninformative uniform prior

$$p(\theta) = \begin{cases} 1 & \text{if } 0 \geq \theta \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

which corresponds to Beta(α, β) distribution with parameters $\alpha = \beta = 1$ (see Chapter 2 in BDA 3). The observation model is then Binomial

$y \sim \text{Binomial}(\theta, n)$.

We have now defined the necessary pieces for Bayesian inference with posterior

$p(\theta|y, n) \propto p(y|\theta, n)p(\theta) = \text{Beta}(y + 1, n - y + 1)$

Where the last equality is a special result related Binomial observation model and Beta prior. See BDA 3, Chapter 2.

Hence, we can easily examine the posterior distribution

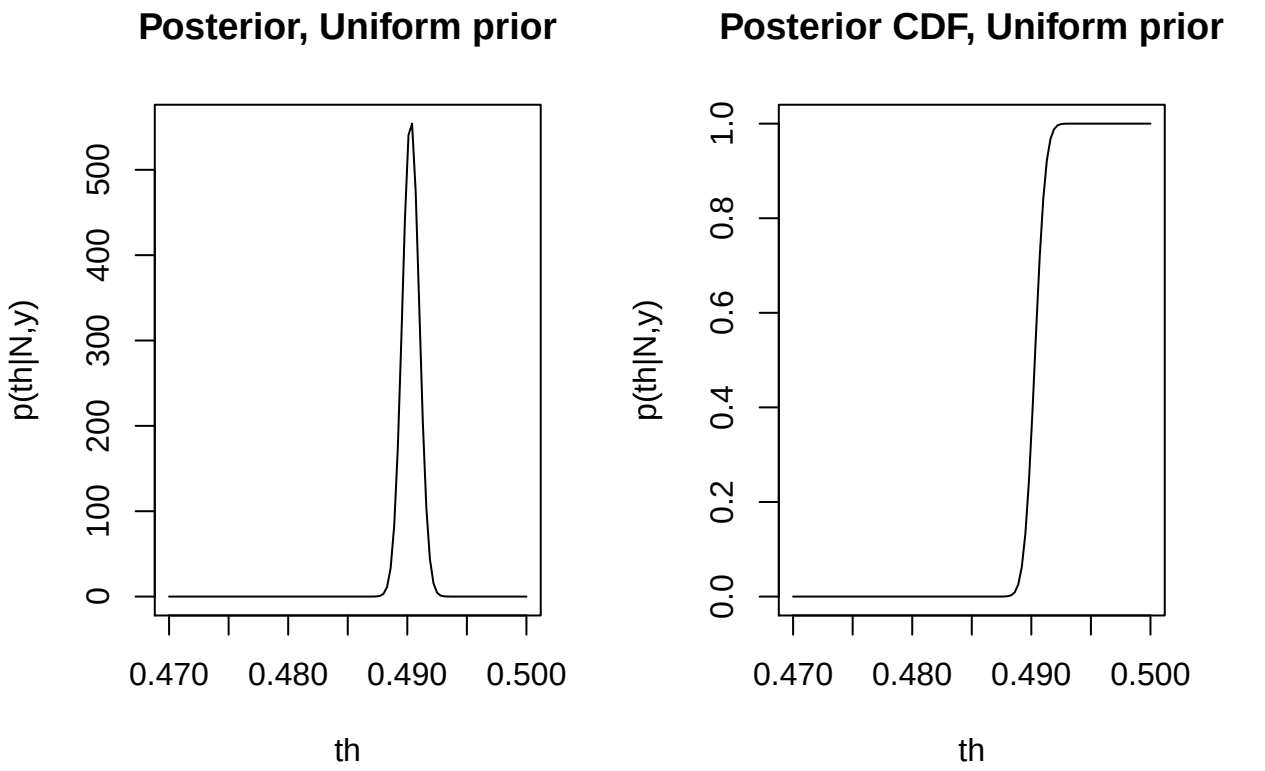
```
# Set the data into variables
n <- 241945 + 251527
y <- 241945

# construct a vector of theta values for visualization
# Note! we draw the probability density function only for
#       short interval of theta values since it is so peaked
x <- seq(0.47, 0.50, length=101)  # the evaluation points

# plot the posterior probability density p(th|n,y)
#  (type "?dbeta" to get help on how to use the dbeta function)
par(mfrow=c(1,2))           # divide plot into 2 subplots
p <- dbeta(x, y+1, n-y+1)   # calculate the probability density
```

```r
plot(x,p, main="Posterior, Uniform prior", xlab="th", ylab="p(th|N,y)", type="l")

# --- plot cumulative density function
cdf <- pbeta(x, y+1, n-y+1)  # calculate the probability density
plot(x,cdf, main="Posterior CDF, Uniform prior", xlab="th", ylab="p(th|N,y)", type="l")
```



Let's next calculate some statistics from the posterior distribution and calculate the posterior median and central 95% posterior interval for θ as well as theprobability that $\theta < 0.485$

```r
#
qbeta(c(0.025,0.5,0.975), y+1, n-y+1)
```

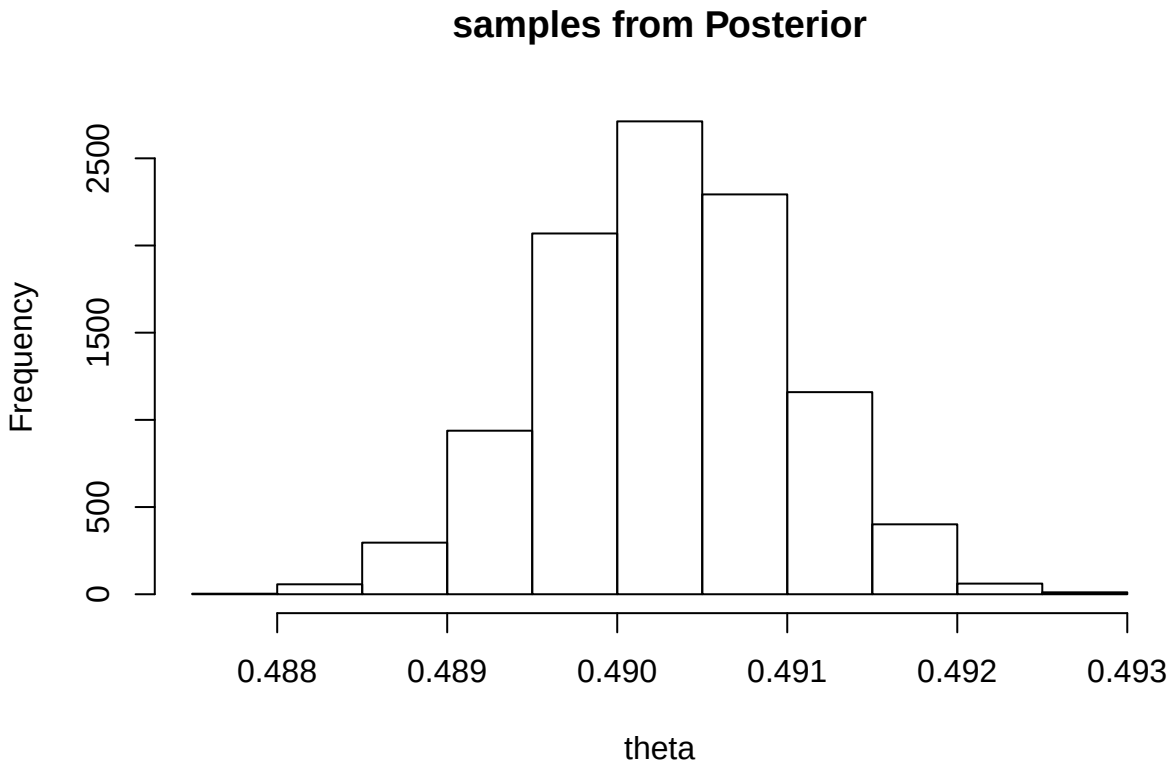
```
## [1] 0.4888965 0.4902913 0.4916861
```

```r
pbeta(0.485, y+1, n-y+1)
```

```
## [1] 5.164892e-14
```

Next, we calculate some statistics using random samples of θ values from the posterior.

```r
# Sample random draws from the posterior distribution and use them
# to visualize the distribution and to calculate the posterior mean
# and variance
th <- rbeta(10000, y+1, n-y+1)
hist(th, main="samples from Posterior", xlab="theta")
```




```
# --- calculate posterior mean and variance and 95% interval ---
mean(th)
```

```
## [1] 0.4903022
```

```
var(th)
```

```
## [1] 5.071819e-07
```

```
quantile(th, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.4888953 0.4916958
```

```
# --- Calculate probability p(th < 0.485)           ---
sum(th<0.485)/length(th)
```

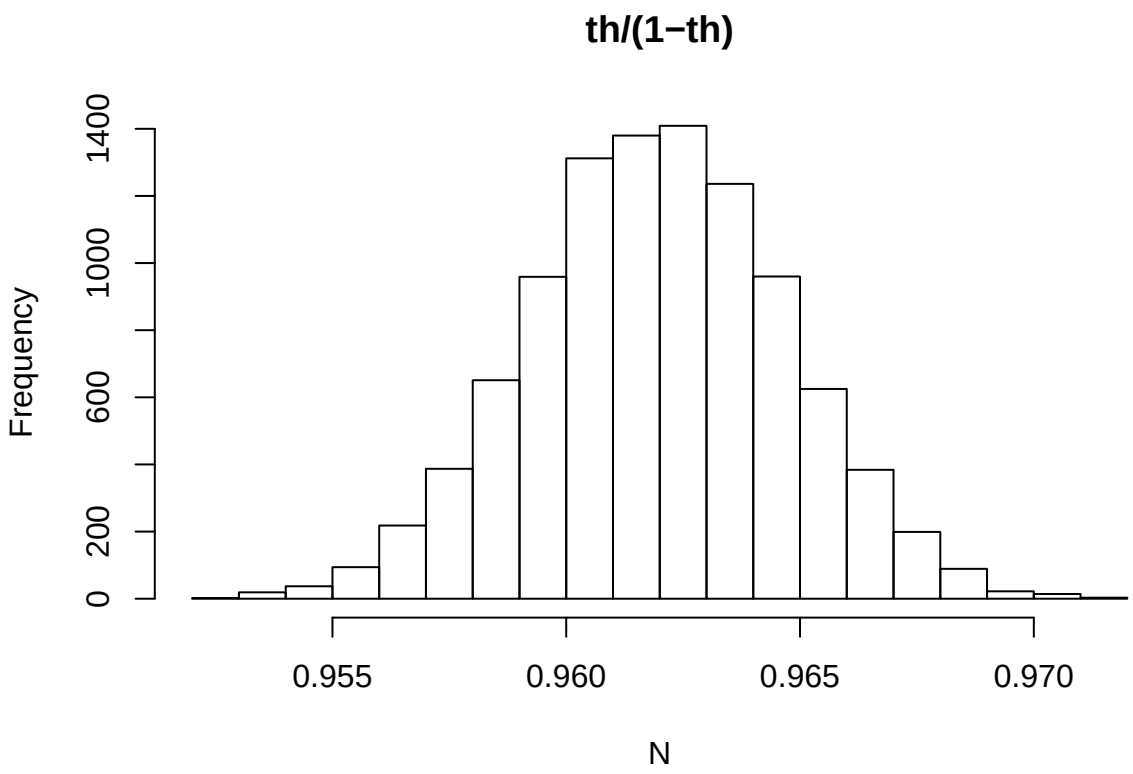
```
## [1] 0
```

Whenever we are able to produce random draws from the posterior distribution we can utilize a very powerfull approximation technique called Monte Carlo approximation. The above calculations very elementary examples of Monte Carlo approximations but the method can be extended to much more. Consider that we are not interested in θ as such but, for example, on the odds ratio

$\phi = \theta/(1-\theta)$

The posterior distribution for ϕ is not available in closed form so we cannot use any built in R functions to visualize it or calculate it summary statistics. However, once we have random samples from the posterior of θ we can easily generate random samples from ϕ by just calculating the odds ratio with each of the samples.

```r
#  Visualize the posterior distribution of odds ratio theta/(1- theta)
hist(th/(1-th), main="th/(1-th)", xlab="N")
```




```r
# --- Calculate the probability that odds ratio is less than 1.0
sum(th/(1-th)<1.0)/length(th)
```

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

```r
# calculate the central 95% posterior interval of odds ratio
quantile(th/(1-th), c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.9565461 0.9673260
```