Solutions to take home exam

Shashank Sule

10/11/2019

Problem 1

a) Here I use the Beta-Binomial conjugacy model and the fact that if the priors p_i are independent and $y_i \mid p_i$ are independent, then the posteriors $p_i \mid y_i$ are independent. Then the density for the joint posterior distribution is a product of the densities of the posteriors $p_i \mid y_i$. Assuming that the likelihood function is Binomial and the prior is Beta distributed, the posteriors are given as follows:

$$p_{2005} \mid y_{2005} \sim Beta(a + y_{2005}, b + n_{2005} - y_{2005})$$
$$p_{2015} \mid y_{2015} \sim Beta(a + y_{2015}, b + n_{2015} - y_{2015})$$

Here $y_{2005} = 264$, $y_{2015} = 437$, $n_2005 = 1760$ and $n_{2015} = 1932$

```
264 + 1496
```

```
## [1] 1760
```

```
437 + 1495
```

```
## [1] 1932
```

Consequently,

$$p_{2005} \mid y_{2005} \sim Beta(265, 1497)$$

and

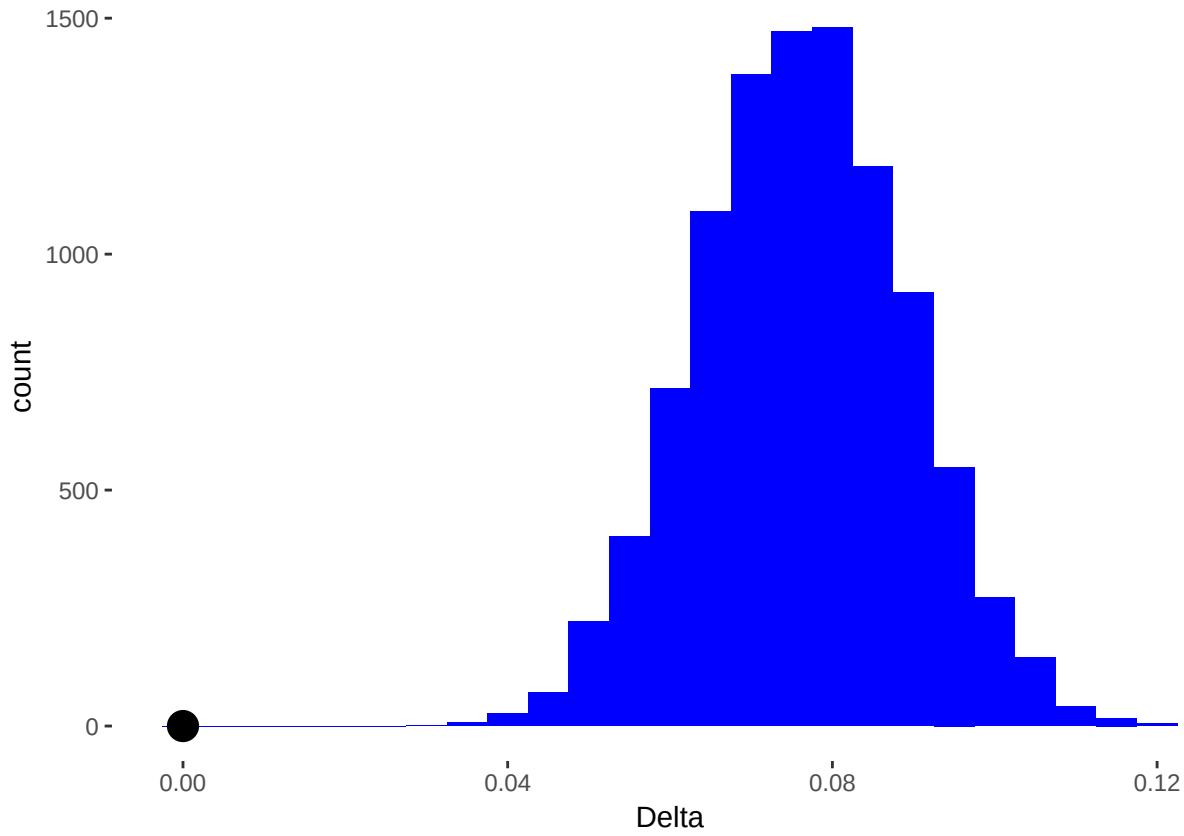
$$p_{2015} \mid y_{2015} \sim Beta(438, 1496)$$

The density of the joint posterior will be a product of the densities for Beta$(265, 1497)$ and Beta$(438, 1496)$.

b)

```
S <- 10000
delta <- -(rbeta(S, 265,1497) - rbeta(S,438,1496))
ggplot(as.data.frame(delta), aes(x=delta))+
  geom_histogram(fill="blue", binwidth = 0.005) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
```

```
panel.background = element_blank())+
  annotate("point", x=0, y=0, color="black", size=5)+
  xlab("Delta")
```




```
mean(delta <= 0)
```

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

Thus, the proportions of science majors indeed seem to have changed over time.

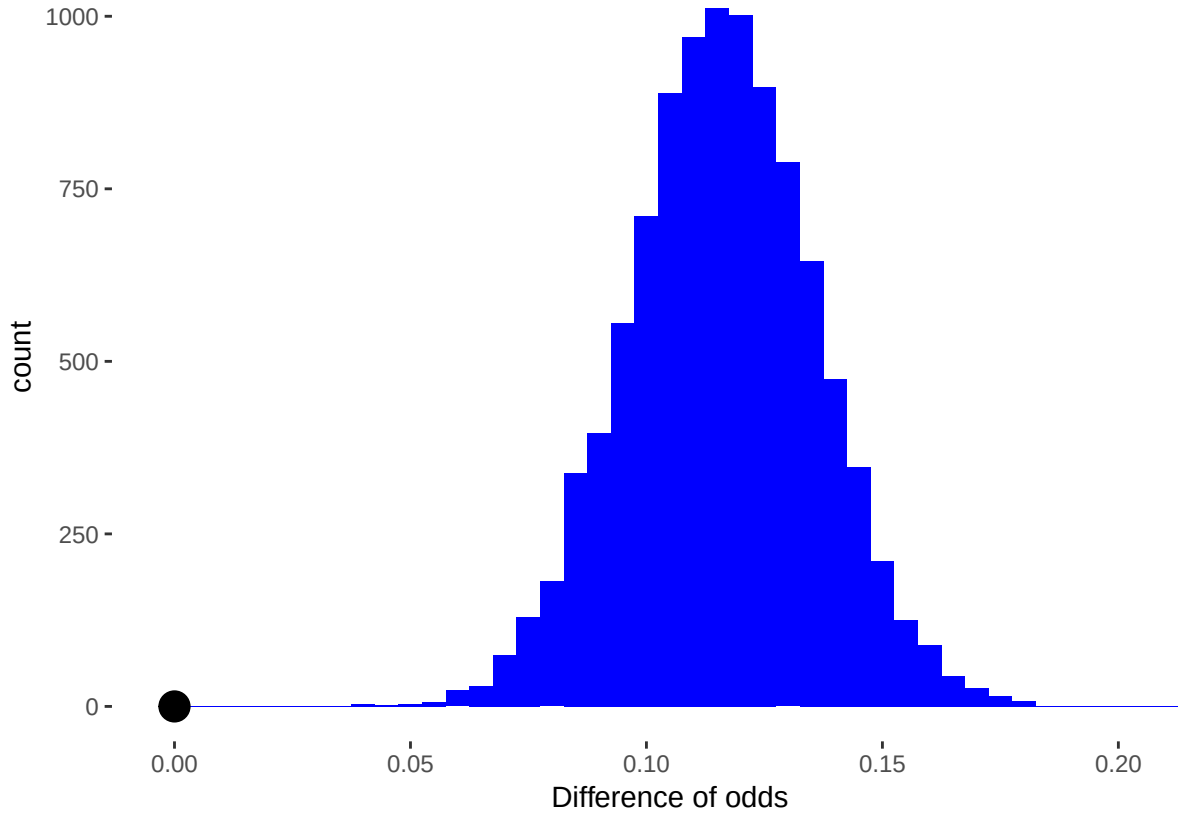
c) The odds formula is the following:

$$o = \frac{p}{1-p}$$

I use the formula to convert the samples of proportion from the posterior to samples of odds. Then I define the difference vector between 2015 odds and 2005 odds and find the proportion above 0.

```
p_2015 <- rbeta(S,438,1496)
p_2005 <- rbeta(S, 265,1497)
o_2015 <- p_2015/(1 - p_2015)
o_2005 <- p_2005/(1 - p_2005)
delta_o <- o_2015 - o_2005
ggplot(as.data.frame(delta_o), aes(x=delta_o))+
```

```
  geom_histogram(fill="blue", binwidth = 0.005) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank())+
  annotate("point", x=0, y=0, color="black", size=5)+
  xlab("Difference of odds")
```




```
mean(delta_o >= 0)
```

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

Thus, the odds of being a science student have increased.

d) We assume that the proportions in 2005 do not affect those in 2015. I do not think this assumption is justified. We could have used a sequential update where we would have used the posterior in 2005 as a prior for 2015, which would have been a different prior than Beta$(1, 1)$. Furthermore, the people taking the survey in 2005 were not the same as the people taking the survey in 2015 so we cannot conclusively put down the change in proportions to the colleges' push for science majors.

Problem 2

a) Here I use the Normal-Normal conjugacy model for the mean where standard deviation is given. Note that $\mu_0 = 8$, $\phi_0 = 1$, and $\phi = (1/1.5)^2$. Additionally, $n = 14$ and

$\bar{y} = 7.285714$

```
Y <- c(9.0,7.5,7.0,8.0,5.0,6.5,8.5,7.0,9.0,7.0,5.5,6.0,8.5,7.5)
ybar <- mean(Y)
ybar
```

```
## [1] 7.285714
```

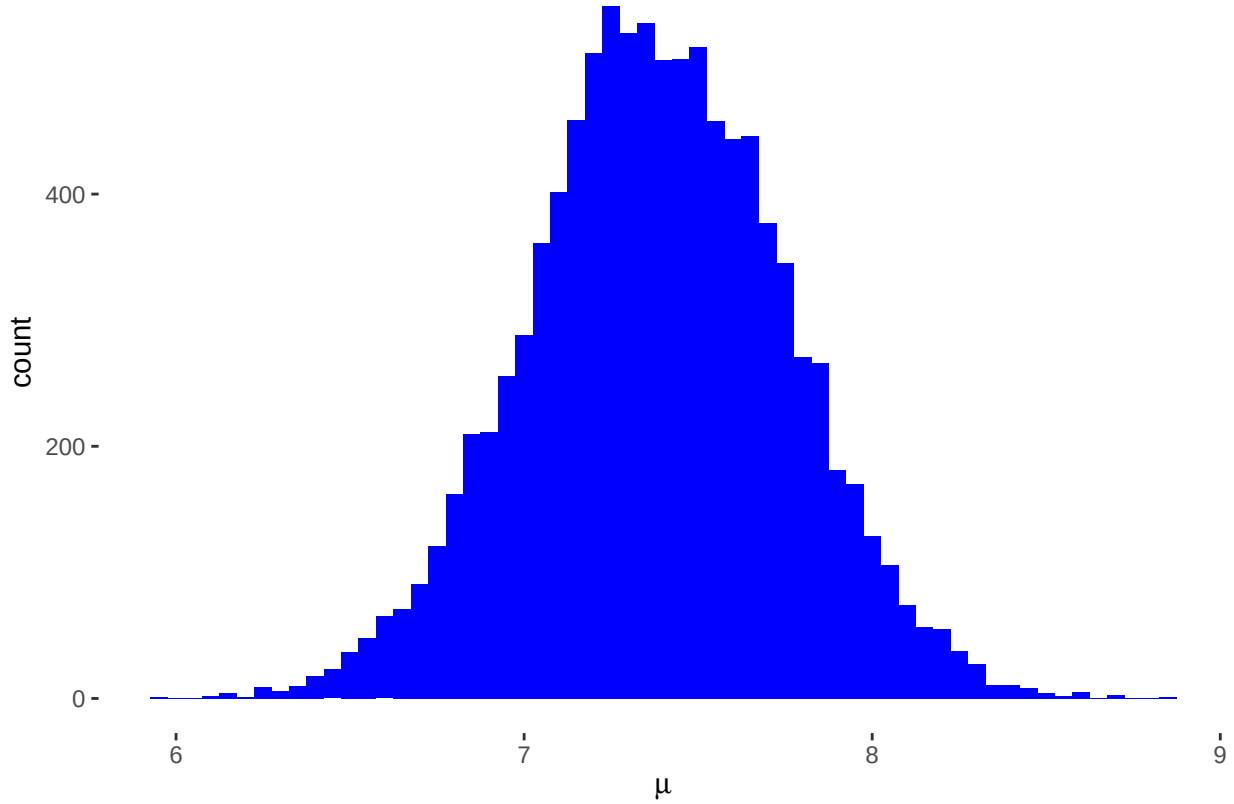
Then the posterior is as follows:

$$\mu \mid y_1, \ldots, y_n, \sigma \sim N\left(\frac{\phi_0\mu_0 + n\phi\bar{y}}{\phi_0 + n\phi}, \sqrt{\frac{1}{\phi_0 + n\phi}}\right)$$

$$= N\left(\frac{(1)(8) + 14((1/1.5)^2)7.285714}{1 + (14)(7.285714)}, \sqrt{\frac{1}{1 + (14)(7.285714)}}\right)$$

b) To create the 90% intervals, I use Monte carlo sampling from the posterior and find the credible interval for my sample:

```
S <- 10000
mu_0 <- 8
phi_0 <- 1
phi <- (1/1.5)^2
n <- 14
mu_posterior <- rnorm(S, mean = (phi_0*mu_0 + n*phi*ybar)/(phi_0 + n*phi),
                      sd = sqrt(1/(phi_0 + n*phi)))
ggplot(as.data.frame(mu_posterior), aes(x=mu_posterior))+
  geom_histogram(fill="blue", binwidth = 0.05) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank())+
  xlab(expression(mu))
```



The 90% credible interval for the mean is the following

```
quantile(mu_posterior, c(0.05, 0.95))
```

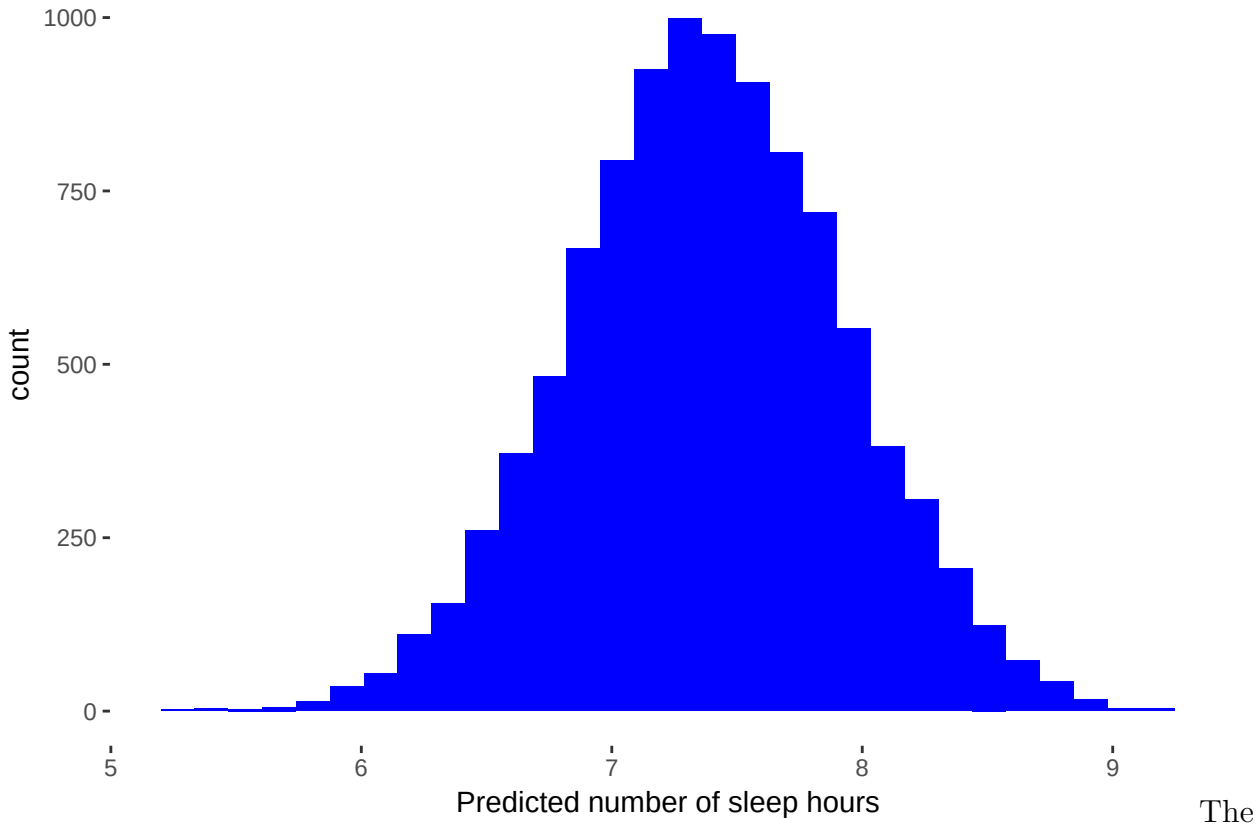
```
##       5%      95%
## 6.773580 7.985746
```

c) To obtain the 90% credible interval for the predictive distribution I sample the mean from the posterior, and then using that mean sample from the likelihood. The 90% credible interval for this sampling distribution will be the one desired.

```
samples <- numeric(length = S)
for(i in 1:S){
  mu <- rnorm(1, mean = (phi_0*mu_0 + n*phi*ybar)/(phi_0 + n*phi),
              sd = sqrt(1/(phi_0 + n*phi)))
  samples[i] <- mean(rnorm(n,mean=mu, sd=1.5))
}

ggplot(as.data.frame(samples), aes(x=samples))+
  geom_histogram(fill="blue") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank())+
  xlab("Predicted number of sleep hours")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The 90% credible interval for the predicted number of sleep hours is

```
quantile(samples, c(0.05,0.95))
```

```
##       5%      95%
## 6.481846 8.286122
```

Problem 3

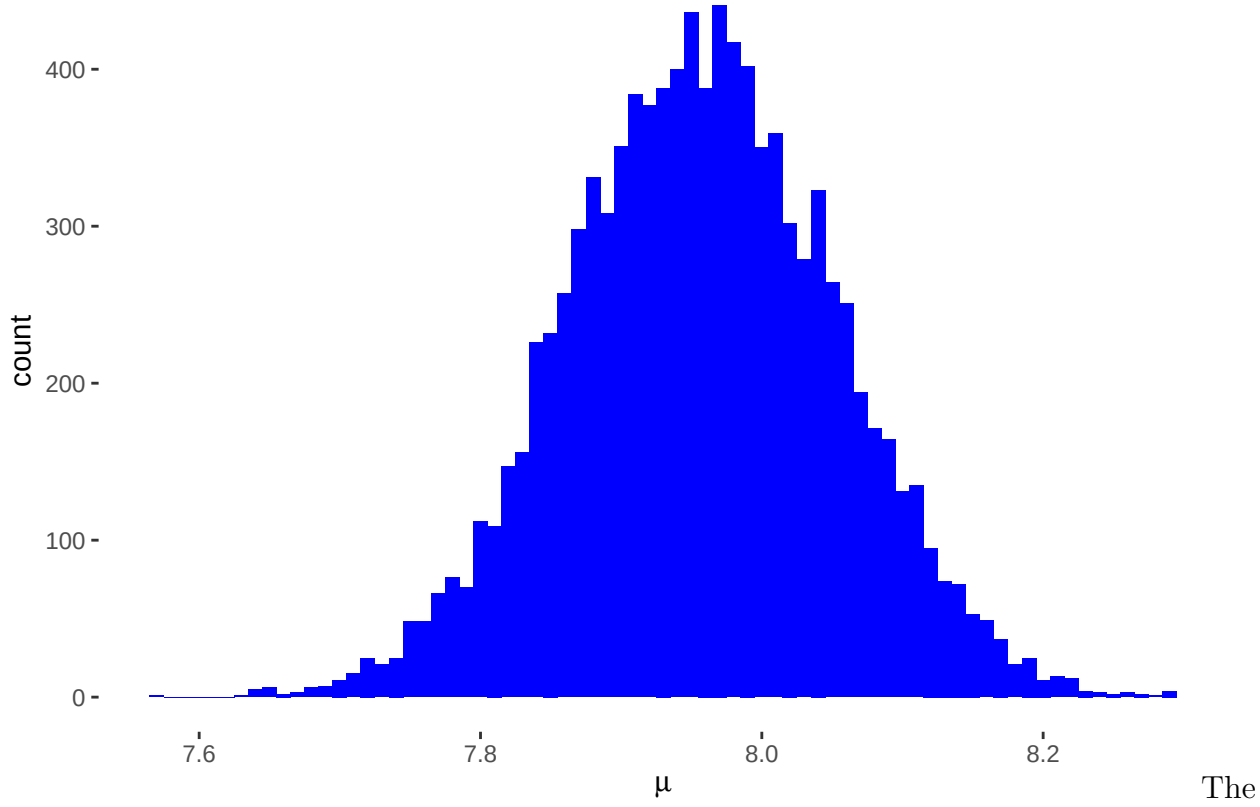
a) The posterior distribution is normally distributed just as John's was. The difference in this case is that $\phi_0 = (1/0.1)^2 = 100$. Thus,

$$\mu \mid y_1, \ldots, y_n, \sigma \sim N\left(\frac{(100)(8) + 14((1/1.5)^2)7.285714}{100 + (14)(7.285714)}, \sqrt{\frac{1}{1 + (14)(7.285714)}}\right)$$

b) Constructing the middle posterior credible posterior follows the same process as the previous problem:

```
phi_0 <- 100
mu_posterior <- rnorm(S, mean = (phi_0*mu_0 + n*phi*ybar)/(phi_0 + n*phi),
                      sd = sqrt(1/(phi_0 + n*phi)))
ggplot(as.data.frame(mu_posterior), aes(x=mu_posterior))+
  geom_histogram(fill="blue", binwidth = 0.01) +
```

```
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank())+
  xlab(expression(mu))
```



The 90% credible interval is thus

```
quantile(mu_posterior, c(0.05,0.95))
```

```
##       5%      95%
## 7.800974 8.113239
```

c) The 90% credible intervals calculated are quite different: The length of John's credible interval is ≈ 1.7 while Mary's is ≈ 0.3. Hence, it is sensitive to the choice of prior. The mean and standard deviation of John's and Mary's posteriors are the following:

```
means <- numeric(length=2)
sds <- numeric(length=2)
phi_0 <- 1
means[1] <- (phi_0*mu_0 + n*phi*ybar)/(phi_0 + n*phi)
sds[1] <- sqrt(1/(phi_0 + n*phi))
phi_0 <- 100
means[2] <- (phi_0*mu_0 + n*phi*ybar)/(phi_0 + n*phi)
sds[2] <- sqrt(1/(phi_0 + n*phi))
comps <- matrix(data = c(means, sds), nrow = 2, ncol=2, byrow = TRUE, dimnames=list(c("M
comps
```

```
##                            John       Mary
## Mean                  7.3846154 7.95815900
## Standard deviation 0.3721042 0.09702693
```

Actually the mean is more insensitive to perturbations in ϕ_0 because the function for the mean

$$\frac{\phi_0\mu_0 + n\phi\bar{y}}{\phi_0 + n\phi}$$

is $O(1)$ when considered as a function of ϕ_0. However,

$$\sqrt{\frac{1}{\phi_0 + n\phi}}$$

is $O(\phi_0^{-1/2})$ when considered as a function of ϕ_0. In fact, as precision is increased by 100 times, the standard deviation in the posterior should drop by about an order of magnitude (i.e $\sqrt{100} = 10$), which is what we see in the above computation. In other words, because Mary was already more assured of her beliefs about the mean number of hours, it reflected in the posterior distribution which represented the sharpened belief about the mean after collecting data. Note that the posterior distribution always has smaller standard deviation (i.e higher precision) than the prior in the Normal-Normal model for the mean:

$$\sigma_{posterior} = \sqrt{\frac{1}{\phi_0 + n\phi}} \leq \sqrt{\frac{1}{\phi_0}} = \sigma_{prior}$$

(This follows because $n\phi > 0$). Thus, by collecting data one's beliefs can only become sharper, and if Mary already has sharper beliefs than John and they look at the same data, Mary will remain sharper about her beliefs (but sharper by how much depends on how much data they look at!).