Streptococcus heftans

A Bayesian Analysis

The Rfuncs Project

Background

Suppose that Streptococcus heftans (fictitious) is an uncommon oral bacteria that has recently been found to cause rare heart infections affecting chambers and valves, especially in the elderly and those with existing chronic heart diseases. The prevalence for S. heftans has never been rigorously estimated, but is believed to be less than 0.005 (0.5% of the population), is very likely to be less than 0.01, and is virtually certain to be less than 0.05.

Statistical Analysis

Maya Cardiya, Ph.D. plans to test $n = 500$ patients to determine if they carry S. heftans. Her protocol's statistical plan calls for the use of a beta-binomial model. The true prevalence π is unknown, so it will be modeled with a beta prior. Tailoring to this particular study, a subjective prior sets the prior median at 0.01 and the 0.99 quantile at 0.05.

a. Find the beta prior. (Use `beta.select` from `LearnBayes` to find the shape parameters)

Beta coefficients:

```
(beta_coeffs <- beta.select(quantile1 = list(p = 0.5, x = 0.01),
                            quantile2 = list(p = 0.99, x = 0.05)))
```

```
## [1]   1.38 105.62
```

As computed above, the prior takes the form $g(\pi_0) \sim Beta(\alpha = 1.38, \beta = 105.62)$.

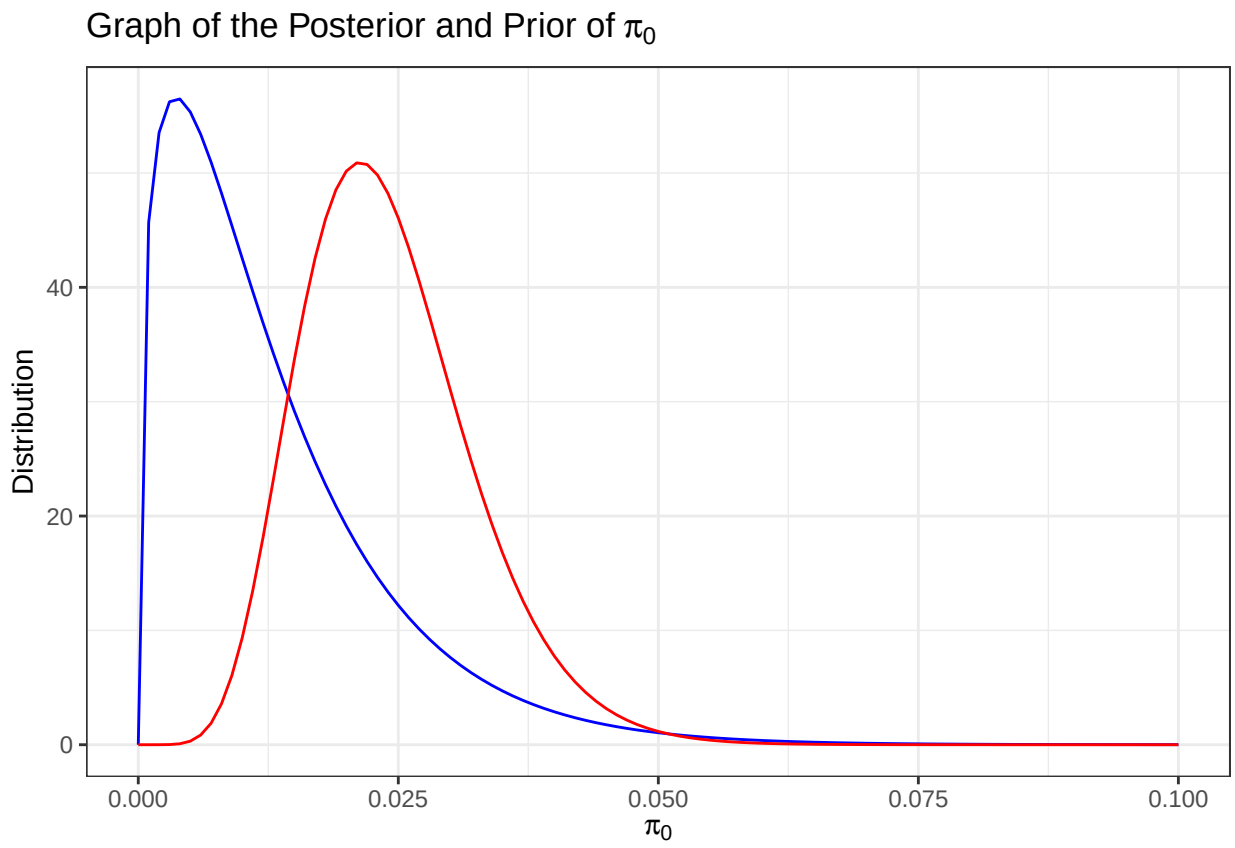
b. Dr. Cardiya's protocol calls for one interim analysis near 50% of testing. Of the first 241 subjects, 7 tested positive for S. heftans.

   i. Calculate the interim posterior distribution. Make a plot of the prior, likelihood and interim posterior distribution. Use the range $0 < \pi < 0.1$ since that's where the action is.

Based on Theorem 26.2 if the data is distributed binomial and the parameter of interest follows a beta distribution, then the posterior distribution also follows a beta distribution. In particular the posterior takes the form of $Beta(\alpha_0 + x, \beta_0 + n - x)$, where x is the number of "successes" in the n trials.

Thus, when we see 7 individuals who tested positive for S. heftans out of a total of 241 individuals, we get a posterior for π_0 of the form $h(\pi_0|x) \sim Beta(1.38 + 7, 105.62 + 234)$

```r
ggplot() +
  # Prior
  stat_function(fun = dbeta,
                args = list(shape1 = 1.38, shape2 = 105.62),
                color = "blue") +
  # Posterior
  stat_function(fun = dbeta,
                args = list(shape1 = 1.38+7, shape2 = 105.62+234),
                color = "red") +
  xlim(0, 0.1) +
  labs(title = expression("Graph of the Posterior and Prior of" ~ pi[0]),
       x = expression(pi[0]),
       y = "Distribution") +
  theme_bw()
```



ii. Contrast the prior median with the posterior median. Also contrast the middle 95% of the beta prior with the middle 95% of the interim posterior distribution.

Binomial experiment data:

```r
# (n-x) failures:
n_min_x <- 234

# x successes:
x <- 7
```

Prior median:

```
(prior_median <- qbeta(p=0.5,
                       shape1=beta_coeffs[1],
                       shape2=beta_coeffs[2]))
```

```
## [1] 0.01001174
```

Posterior median:

```
(posterior_median <- qbeta(p=0.5,
                           shape1=beta_coeffs[1] + x,
                           shape2=beta_coeffs[2] + n_min_x))
```

```
## [1] 0.02317398
```

As computed above, the prior median was 0.0100117, and the posterior median was 0.023174. As can be seen, the posterior median is a little more than 2 times as large as the prior median. After creating our interim posterior distribution our beliefs have now shifted to a larger prevalence.

Prior middle 95%:

```
(prior_middle_95 <- qbeta(p=c(0.025, 0.975),
                          shape1=beta_coeffs[1],
                          shape2=beta_coeffs[2]))
```

```
## [1] 0.0007827123 0.0411646709
```

Posterior middle 95%:

```
(post_middle_95 <- qbeta(p=c(0.025, 0.975),
                         shape1=beta_coeffs[1] + x,
                         shape2=beta_coeffs[2] + n_min_x))
```

```
## [1] 0.01072810 0.04257156
```

As computed above, the middle 95% of the prior distribution is the range was [0.0007827123, 0.0411646709], and the middle 95% of the interim posterior distribution is the range was [0.01072810, 0.04257156]. As can be seen, the middle 95% of the interim posterior distribution is slightly shifted to the right when compared to the middle 95% of the prior distribution. This reflects the same change in beliefs as described above.

iii. Contrast the prior probability that less than 1 percentage of the population have S. heftans with the posterior probability.

Prior probability that $\mathbf{P}(\pi < \mathbf{0.01})$:

```
(prior_prob <- pbeta(q=0.01,
                     shape1=beta_coeffs[1],
                     shape2=beta_coeffs[2]))
```

```
## [1] 0.499501
```

Posterior probability that $\mathbf{P}(\pi < \mathbf{0.01})$:

```
(post_prob <- pbeta(q=0.01,
                    shape1=beta_coeffs[1] + x,
                    shape2=beta_coeffs[2] + n_min_x))
```

```
## [1] 0.01714528
```

As can be seen above, the posterior probability got significantly lower than the prior probability. Yet again, this represents a shift in our beliefs, mainly due to the fact that we saw more patients have S. heftans than we would expect for such a small sample if the prevalence was truly extremely rare.

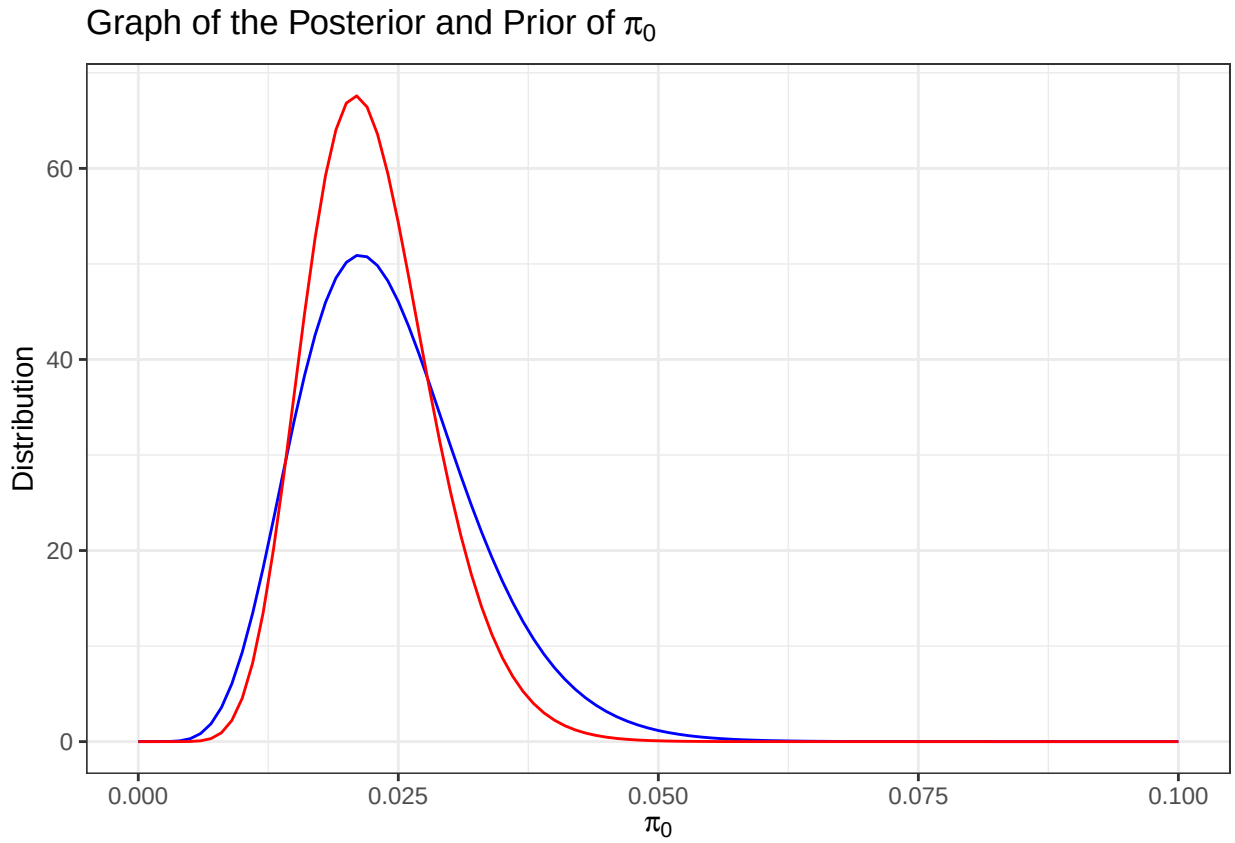
c. Suppose testing 247 more subjects yields 5 positives. Using the interim posterior distribution as your prior, repeat i,ii,iii for the final posterior distribution. Make a table showing the prior/posterior median, 95% percentile interval and spread of the interval.

We will now take the interim posterior distribution as the prior and add our new data to this prior to get our final posterior distribution.

Based on Theorem 26.2 if the data is distributed binomial and the parameter of interest follows a beta distribution, then the posterior distribution also follows a beta distribution. In particular the posterior takes the form of $Beta(\alpha_0 + x, \beta_0 + n - x)$, where x is the number of "successes" in the n trials.

Thus, when we see 5 individuals who tested positive for S. heftans out of a total of 247 individuals, we get a posterior for π_0 of the form $h(\pi_0|x) \sim Beta(1.38 + 7 + 5, 105.62 + 234 + 242)$

```
ggplot() +
  # Prior
  stat_function(fun = dbeta,
                args = list(shape1 = 1.38+7, shape2 = 105.62+234),
                color = "blue") +
  # Posterior
  stat_function(fun = dbeta,
                args = list(shape1 = 1.38+7+5, shape2 = 105.62+234+242),
                color = "red") +
  xlim(0, 0.1) +
  labs(title = expression("Graph of the Posterior and Prior of" ~ pi[0]),
       x = expression(pi[0]),
       y = "Distribution") +
  theme_bw()
```



Binomial experiment data:

```
# UPDATED DATA:

# (n-x) failures:
n_min_x_new <- n_min_x + 242

# x successes:
x_new <- x + 5
```

Prior median:

```
(prior_median <- qbeta(p=0.5,
                       shape1=beta_coeffs[1] + x,
                       shape2=beta_coeffs[2] + n_min_x))
```

```
## [1] 0.02317398
```

Posterior median:

```
(posterior_median <- qbeta(p=0.5,
                           shape1=beta_coeffs[1] + x_new,
                           shape2=beta_coeffs[2] + n_min_x_new))
```

```
## [1] 0.02195429
```

As computed above, the prior median was 0.023174, and the posterior median was 0.0219543. As can be seen, the posterior median is a little less than that of the prior median. This means that after seeing the new data our belief is that the prevalence is smaller than we had anticipated before, but still greater than our initial beliefs.

Prior middle 95%:

```
(prior_middle_95 <- qbeta(p=c(0.025, 0.975),
                          shape1=beta_coeffs[1] + x,
                          shape2=beta_coeffs[2] + n_min_x))
```

```
## [1] 0.01072810 0.04257156
```

Posterior middle 95%:

```
(post_middle_95 <- qbeta(p=c(0.025, 0.975),
                         shape1=beta_coeffs[1] + x_new,
                         shape2=beta_coeffs[2] + n_min_x_new))
```

```
## [1] 0.01217272 0.03582716
```

As computed above, the middle 95% of the prior distribution is the range was [0.01072810, 0.04257156], and the middle 95% of the posterior distribution is the range was [0.01217272, 0.03582716]. As can be seen, the middle 95% of the posterior distribution is slightly shifted to the left when compared to the middle 95% of the prior distribution. This reflects the same change in beliefs as described above.

Prior probability that $\mathbf{P}(\pi < 0.01)$:

```
(prior_prob <- pbeta(q=0.01,
                     shape1=beta_coeffs[1] + x,
                     shape2=beta_coeffs[2] + n_min_x))
```

```
## [1] 0.01714528
```

Posterior probability that $\mathbf{P}(\pi < 0.01)$:

```
(ppst_prob <- pbeta(q=0.01,
                    shape1=beta_coeffs[1] + x_new,
                    shape2=beta_coeffs[2] + n_min_x_new))
```

```
## [1] 0.005599015
```

As can be seen above, the posterior probability got significantly lower than the prior probability. Yet again, this represents a shift in our beliefs, mainly due to the fact that we saw more patients have S. heftans than we would expect for such a small sample if the prevalence was truly extremely rare. This is because even though we think that the prevalence is lower than thought from the interim prior, when compared to that of our initial assumptions we are pretty certain that the true prevalence is above 0.01.

d. Calculate the 95% Highest Posterior Density intervals for the final analysis in part c. Compare your Bayesian analysis with the results from a large sample Wald confidence interval for a binomial proportion.

95% HPDI interval:

```
hdi(qbeta,
    credMass=0.95,
    shape1=beta_coeffs[1] + x_new,
    shape2=beta_coeffs[2] + n_min_x_new)
```

```
##      lower      upper
## 0.01131527 0.03459682
## attr(,"credMass")
## [1] 0.95
```

Large Sample Wald interval:

```
prop.test(x=x_new,
          n=n_min_x_new+x_new,
          p=posterior_median,
          alternative="two.sided")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  x_new out of +x_new out of n_min_x_newx_new out of x_new
## X-squared = 0.059005, df = 1, p-value = 0.8081
## alternative hypothesis: true p is not equal to 0.02195429
## 95 percent confidence interval:
##  0.01337393 0.04376508
## sample estimates:
##          p
## 0.02459016
```

As can be seen above, the HPDI is [0.011, 0.035]. While the 95% Wald Confidence Interval was [0.13, 0.044]. The intervals have vastly different interpretations as the HPDI gives a probability that the true parameter false in that range, while the confidence interval says that over hypothetical replication about 95% of them will contain the true value of the parameter.

Acknowledgment

Thanks to Ralph O'Brien for this problem.