STAT 577 Midterm Exam

Tim Farkas

4/5/2020

Problem 1

On 2 April, the state Department of Health (NMDOH) reported on the results of 767 new COVID-19 tests. 40 patients tested positive for the disease, and 727 tested negative.

1a)

Epidemiologists at NM - DOH believe that the prevalence of COVID-19 among the population of New Mexicans is about 0.275%, a little less than 3 out of every thousand people. They think the prevalence of COVID-19 among the population of New Mexicans is unlikely to be any higher than 1.36%. Find a prior reflecting this information that is conjugate with the 2 April daily testing data, and explain how you obtained it.

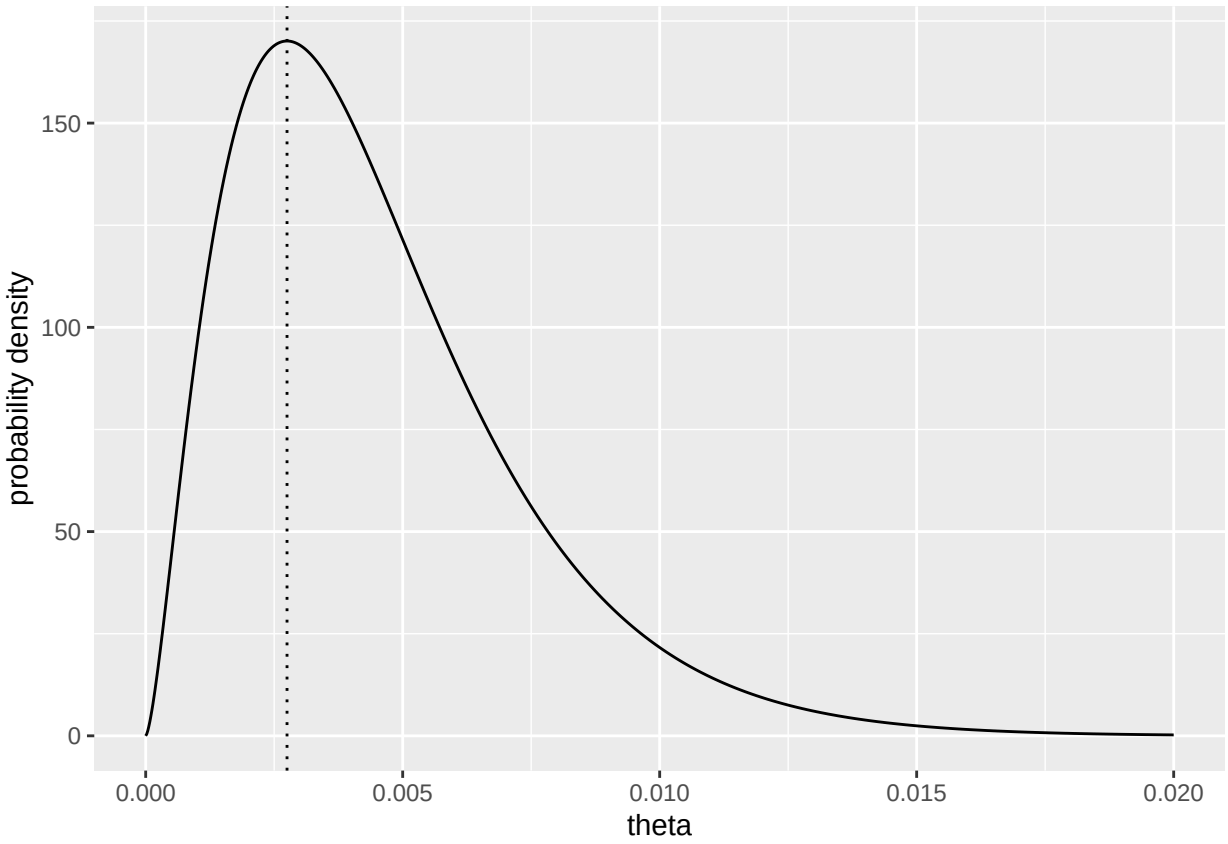
```
bb <- epi.betabuster(mode = 0.00275,
                conf = 0.99,
                greaterthan = FALSE,
                x = 0.0136)

print(c(alpha = bb$shape1,
        beta= bb$shape2))
```

```
theta_sim <- seq(0, .02, length.out = 1000)
beta_prob <- dbeta(theta_sim, bb$shape1, bb$shape2)

p <- ggplot() +
  geom_line(aes(x = theta_sim, y = beta_prob)) +
  geom_vline(xintercept = 0.00275, lty=3) +
  xlab("theta") +
  ylab("probability density")

print(p)
```



Using Wesley Johnson's betabuster R function, from the epiR package, we see that a Beta(2.5, 552.9) distribution aligns with with the prior expert knowledge about the prevalence of COVID among New Mexican residents. Checking with the qbeta function shows it worked.

1b)

NMDOH has specified criteria for who should be tested for COVID-19, and thus the people being tested can't be considered a random sample from the population of all New Mexicans — but does that mean it's not a representative sample? Consider how applicable these priors are to the data we have, and (in one paragraph) discuss which one you would prefer. (If you want to suggest your own prior as a third option, you are free to do so.)

This is a representative sample, but it doesn't represent what we need it to represent. So no, the informative prior is very much no good, because the wrong information has been elicited from the expert. For an informative prior to be useful, we must find a prior on the proportion of the population *being tested* that has COVID-19, not the population as a whole, because this is what our data are about.

Because the population being tested must satisfy some criteria in order to be tested, likely showing some minimum set of COVID-like symptoms (fever, dry cough, etc.), it would appear to me fair to argue that the point and boundary estimates given by the expert are too low for our data. The MLE for the prevalence of COVID ($\theta = \frac{40}{727} = 0.055 = 5.2\%$) agrees with this argument, as this is far, far greater than the expert's best guess of 0.275%.

In this case, I feel as though we really have neither a direct nor an inferential source of prior information for these data, and so I would not use an informative prior, but opt to use the non-informative Jeffrey's prior Beta($\frac{1}{2}$, $\frac{1}{2}$).

1c)

To establish conjugacy:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

$$= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I_{[0,1]}(\theta)}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

$$= \frac{\theta^y(1-\theta)^{n-y} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}I_{[0,1]}(\theta)}{\int_0^1 \theta^y(1-\theta)^{n-y} \times \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

$$\frac{\theta^{y+\alpha-1}(1-\theta)^{n+\beta-y-1}I_{[0,1]}(\theta)}{\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n+\beta-y-1}d\theta}$$

$$\frac{\theta^{y+\alpha-1}(1-\theta)^{n+\beta-y-1}I_{[0,1]}(\theta)}{\frac{\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha+\beta+n)}\int_0^1 \frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\theta^{y+\alpha-1}(1-\theta)^{n+\beta-y-1}d\theta}$$

$$= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\theta^{y+\alpha-1}(1-\theta)^{n+\beta-y-1}I_{[0,1]}(\theta)$$

$$\therefore p(\theta|y) \sim Beta(\alpha+y, \beta+n-y)$$

So with y = 40, n = 767, $\alpha = 2.5$, and $\beta = 552.9$

$$p(\theta|y) \sim Beta(42.5, 1239.9)$$

1d)

On 3 April, NMDOH reported results on an additional 854 tests. Analytically find the predictive distribution for the number of positive tests we would expect to see on 3 April, given the 2 April data and the informative prior you obtained in (a). This should be a Beta-Binomial distribution as discussed on page 25 of your textbook. As in part (c), you could just plug the correct values into the Beta-Binomial pdf given on that page, but I'm asking you to show me that the distribution given there is correct.

$$p(\tilde{y}|y) = \int_0^1 p(\tilde{y}|\theta)p(\theta|y)d\theta$$

$$= \int_0^1 \binom{m}{\tilde{y}}\theta^m(1-\theta)^{m-\tilde{y}} \times \frac{\Gamma(\alpha+\beta+n-y)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\theta^{y+\alpha-1}(1-\theta)^{n+\beta-y-1}I_{[0,1]}(\theta)d\theta$$

$$= \binom{m}{\tilde{y}}\frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\int_0^1 \theta^{\tilde{y}}(1-\theta)^{m-\tilde{y}} \times \theta^{y+\alpha-1}(1-\theta)^{n+\beta-y-1}(\theta)d\theta$$

$$= \binom{m}{\tilde{y}}\frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\int_0^1 \theta^{\tilde{y}+y+\alpha-1}(1-\theta)^{m-\tilde{y}+n+\beta-y-1}d\theta$$

$$= \binom{m}{\tilde{y}}\frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)}\frac{\Gamma(\tilde{y}+y+\alpha)\Gamma(m-\tilde{y}+n+\beta-y)}{\Gamma(\alpha+m+n+\beta)}\times$$

$$\int_0^1 \frac{\Gamma(\alpha+m+n+\beta)}{\Gamma(\tilde{y}+y+\alpha)\Gamma(m-\tilde{y}+n+\beta-y)}\theta^{\tilde{y}+y+\alpha-1}(1-\theta)^{m-\tilde{y}+n+\beta-y-1}d\theta$$

$$= \binom{m}{\tilde{y}} \frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n+\beta-y)} \frac{\Gamma(\tilde{y}+y+\alpha)\Gamma(m-\tilde{y}+n+\beta-y)}{\Gamma(\alpha+m+n+\beta)}$$

$$\therefore p(\tilde{y}|y)\ BetaBinomial(\alpha=2.5, \beta=552.0, y=40, n=727, m=854)$$

1e)

Find a central 90% probability interval for this predictive distribution. You can do this any way you like, though I recommend using R or OpenBUGS (or JAGS) to do it. (Analytical solutions would probably be quite difficult. Numerical integration would be fine, but I expect most of you would have an easier time just using a statistical packageto generate an appropriate sample you can use to draw inferences.) Using the same method, find a central 90% probability interval for the predictive distribution that uses the Jeffreys Beta(0.5, 0.5). (If you decided to suggest your own prior in (b), find a probability interval for its associated predictive distribution as well.)

I chose to solve this by building a probability density function for the beta-binomial predictive distriution and sampling from it. For the informative prior Beta(2.5, 522.9), a 90% credible interval ranges from 18 to 41 positive tests. For the Jeffrey's prior, the interval ranges from 32 to 65 positive test.

```
# probability density function for beta-binomial
pbb <- function(ytilde, y ,a, b, n, m) {
    choose(m, ytilde) *
    exp(
    lgamma(a + b + n) +
    lgamma(a + y + ytilde) +
    lgamma(b + n + m - y - ytilde) -
    lgamma(a + y) -
    lgamma(b + n - y) -
    lgamma(a + b + n + m)
    )
}

# get random sample from beta-binomial
rbb <-function(size, y, a, b, n, m) {
    sample(0:m, size, replace=TRUE,
       prob=pbb(ytilde=0:m, y, a, b, n, m))
}

# beta-binomial predictive distribution with informative prior
infsamp <- rbb(10000, y=40, a=2.5, b=522.9, n=727, m=854)
quantile(infsamp, probs = c(0.05, 0.95))
```

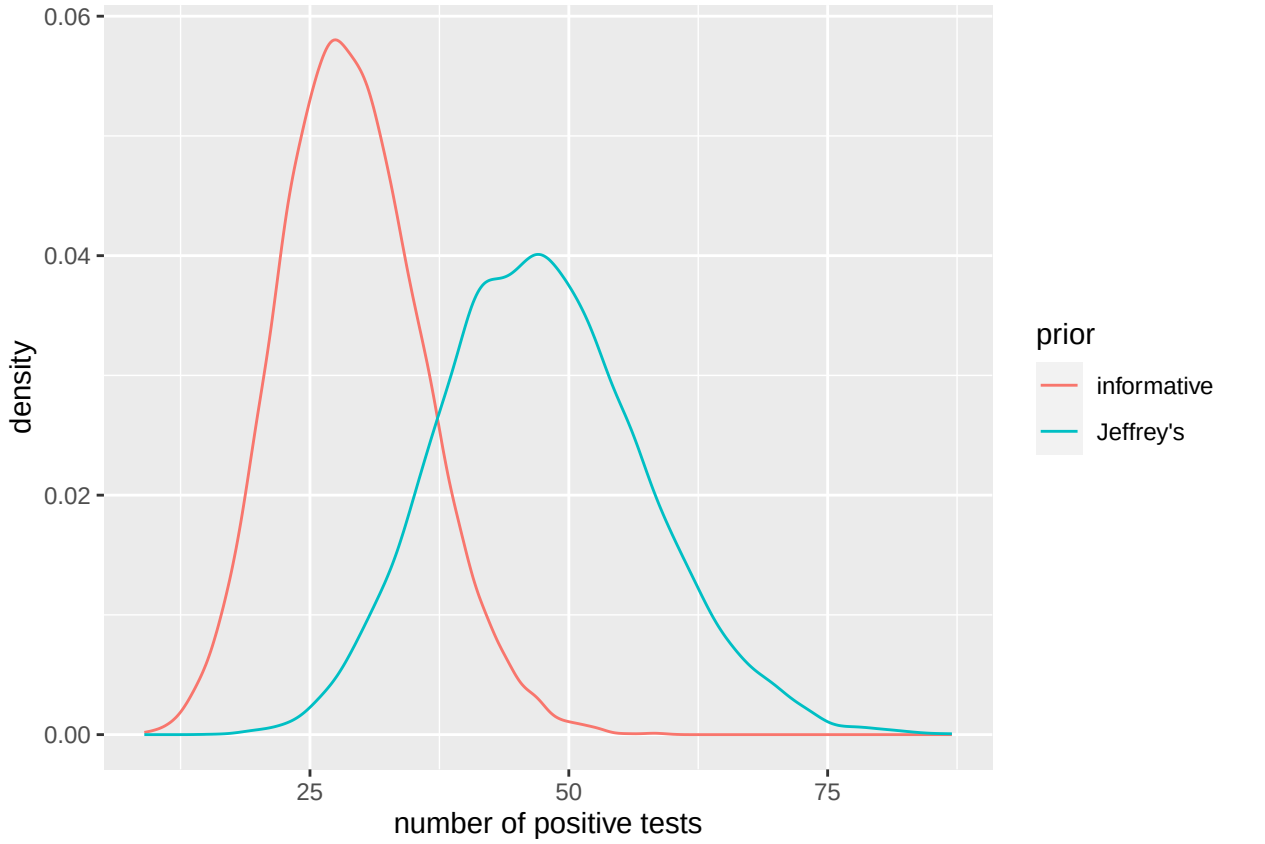
```
##  5% 95%
##  19  41
```

```
# beta-binomial predictive distribution with Jeffrey's prior
jefsamp <- rbb(10000, y=40, a=0.5, b=0.5, n=727, m=854)
quantile(jefsamp, probs = c(0.05, 0.95))
```

```
##  5% 95%
##  32  64
```

```r
all_samples <- tibble(samps=c(infsamp, jefsamp),
       prior=c(rep(c("informative", "Jeffrey's"), each=10000)))

print(ggplot(data=all_samples) +
  stat_density(aes(x=samps, color=prior),
               geom="line", position="identity") +
  xlab("number of positive tests"))
```



1f)

A very basic inferential question NMDOH epidemiologists might ask is, "Does the per- centage of New Mexicans affected by COVID-19 seem to be increasing over time?" Another very similar question is, "Is the proportion of COVID-19 tests that return a positive result changing over time?" One way we can address questions like these is to compare our predictions about what we would see on 3 April to what we actually see: if the number of positive test results is not consistent with our predictive distribution, this suggests that something is different between our 2 April data and our 3 April data (e.g. perhaps the proportion of New Mexicans affected by COVID-19 is changing over time). Of the 854 tests reported on 3 April, 92 came back positive. Are these data consistent with any of the predictive distributions you found in (e)? Why or why not?

92 positive tests is consistent with the predictive distributions based on neither the informative prior nor the Jeffrey's prior, because 92 is above the 90% credible interval in both cases. This suggests an increase in the prevalence of COVID-19 among tested patients. All other things being equal between the two days (such as patient behavior, testing criteria, and the prevalence of non-COVID diseases), this furthermore suggests an increase in the prevalence of COVID-19 in the population.

Problem 2

First the likelihood:

$$L(\mu,\tau|y) = \tau^{\frac{1}{2}} e^{\left[-\frac{n\tau}{2}(\mu-\overline{y})^2\right]} \tau^{\frac{n-1}{2}} e^{\left[-\tau\frac{(n-1)s^2}{2}\right]}$$

Thus,

$$p(\mu|\tau,y) = \tau^{\frac{1}{2}}\tau^{\frac{1}{2}}\tau^{\frac{n-1}{2}} e^{\left[-\frac{n\tau}{2}(\mu-\overline{y})^2-\tau\frac{(n-1)s^2}{2}-\frac{\omega_0\tau}{2}(\mu-\mu_0)^2\right]} \times p(\tau)$$

$$= \tau^{\frac{1}{2}}\tau^{\frac{1}{2}}\tau^{\frac{n-1}{2}} e^{-\frac{\tau}{2}\left[n(\mu-\overline{y})^2+\omega_0(\mu-\mu_0)^2\right]} e^{-\frac{\tau}{2}(n-1)s^2}$$

Then, using complete the square:

$$= \tau^{\frac{1}{2}}\tau^{\frac{1}{2}}\tau^{\frac{n-1}{2}} e^{-\frac{\tau}{2}\left[(n+\omega_0)\left(\mu-\left[\frac{n}{n+\omega_0}\overline{y}+\frac{\omega_0}{n+\omega_0}\mu_0\right]\right)^2+\left(\frac{n\omega_0}{n+\omega_0}\right)(\overline{y}-\mu_0)^2\right]} e^{-\frac{\tau}{2}(n-1)s^2}$$

$$= \tau^{\frac{1}{2}}\tau^{\frac{1}{2}}\tau^{\frac{n-1}{2}} e^{-\frac{\tau}{2}\left[n+\omega_0(\mu-\mu_B)^2\right]} e^{-\frac{\tau}{2}\frac{n\omega_0}{n+\omega_0}(\overline{y}-\mu_0)^2} e^{-\frac{\tau}{2}(n-1)s^2}$$

$$\propto \tau^{\frac{1}{2}} e^{-\frac{\tau}{2}(n+\omega_0)(\mu-\mu_0)^2}$$

$$\therefore \mu \sim N\left(\mu_B, \frac{1}{\tau(n+\omega_0)}\right)$$

Where $\mu_B = \frac{n}{n+\omega_0}\overline{y} + \frac{\omega_0}{n+\omega_0}\mu_0$

Next,

$$p(\mu|y) \propto \tau^{\frac{a+n}{2}-1}\tau^{\frac{1}{2}} e^{-\frac{\tau}{2}(\beta+BBSE)} e^{-\frac{\tau}{2}(n+\omega_0)(\mu-\mu_B)^2}$$

$$= \tau^{\frac{a+n+1}{2}-1} e^{-\frac{\tau}{2}\left[b+BBSE+(n+\omega_0)(\mu-\mu_B)^2\right]}$$

Where

$$BBSE = \sum_i^n (y_i-\overline{y})^2 + \frac{n\omega_0}{n+\omega_0}(\mu_0-\overline{y})^2$$

$$\therefore p(u|y) \sim Gamma(\frac{a+n+1}{2}, \frac{1}{2}\left(b+\sum_i^n (y_i-\overline{y})^2 + \frac{n\omega_0}{n+\omega_0}(\overline{y}-\mu_0)^2 + (n+\omega_0)(\mu-\mu_B)^2\right)$$

Hence,

$$p(\mu|y) \propto \frac{\Gamma(\frac{a+n+1}{2})}{\frac{1}{2}\left(b+\sum_i^n (y_i-\overline{y})^2 + \frac{n\omega_0}{n+\omega_0}(\overline{y}-\mu_0)^2 + (n+\omega_0)(\mu-\mu_B)^2\right)^{\frac{a+n+1}{2}}}$$

$$= \frac{\Gamma(\frac{a+n+1}{2})}{\frac{1}{2}\left(b + BSSE + (n+\omega_0)(\mu-\mu_B)^2\right)^{\frac{a+n+1}{2}}}$$

$$= \frac{\Gamma(\frac{a+n+1}{2})}{\frac{1}{2}\left(\frac{(b+BSSE)(n+a)}{(n+a)} + (n+\omega_0)(\mu-\mu_B)^2\right)^{\frac{a+n+1}{2}}}$$

$$= \frac{\Gamma(\frac{a+n+1}{2})}{\frac{1}{2}\left(\hat{\sigma}_B^2(n+a) + (n+\omega_0)(\mu-\mu_B)^2\right)^{\frac{a+n+1}{2}}}$$

$$= 2\Gamma\left(\frac{a+n+1}{2}\right)\left(\hat{\sigma}_B^2(n+a) + (n+\omega_0)(\mu-\mu_B)^2\right)^{-\frac{a+n+1}{2}}$$

$$= 2\Gamma\left(\frac{a+n+1}{2}\right)\left(\frac{\hat{\sigma}_B^2}{n+\omega_0}(n+a) + (\mu-\mu_B)^2\right)^{-\frac{a+n+1}{2}}$$

$$\therefore \mu \sim t\left(a+n, \mu_B, \sqrt{\frac{\hat{\sigma}_B^2}{n+\omega_0}}\right)$$

Hence, when n = 1:

$$\frac{\mu-\hat{\mu}}{\sqrt{\frac{\hat{\sigma^2}}{1+\omega_0}}}|y \sim t(a+1)$$

I laughed so hard when I saw your solution to HW1 6c. I literally just wrote "I am tired." on the page. Glad to see I'm in good company.