Statistical Simulation and Computerintensive Methods

Exercise 8

Markus Kiesel | 1228952

12.01.2021

Contents

Task 1

We recalculate the estimation of the prevalence of Covid19 in spring 2020. Samples from 1279 persons were analysed with PCR testing procedures. Out of all those not a single randomly selected person was tested positively. This obviously breaks standard testing mechanisms for estimating the proportion of infected person in Austria.

However, additional information is available from similar tests in Germany which had a comparable behavior of the spread of the disease at that time. In the same time span 4 positive cases out of 4068 had been found.

```
# save variables for task
positive_aus <- 0
positive_ger <- 4
n_samples_ger <- 4068
n_samples_aus <- 1279
factor <- 1/10
```

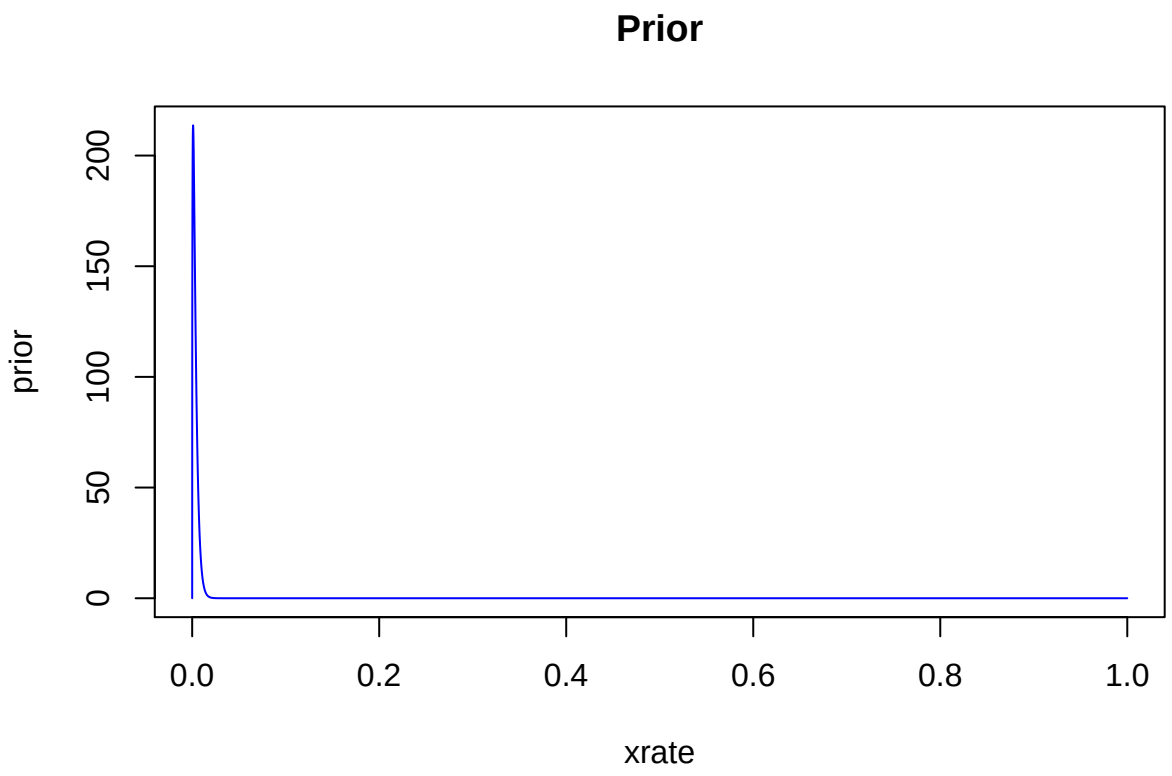
1.1

Build a Beta prior distribution for this Binomial scenario, which encodes the information of the German study. Reweight both parameters compared to the original counts with a factor of 1/10.

```
# define prior parameters
alpha <- positive_ger * factor + 1
beta <- (n_samples_ger - positive_ger) * factor + 1
xrate <- seq(0, 1, length.out = 10000)

# create prior beta distribution
prior <- dbeta(xrate, alpha, beta)

# plot prior
plot(xrate, prior, type = "l", col = "blue", main = "Prior")
```

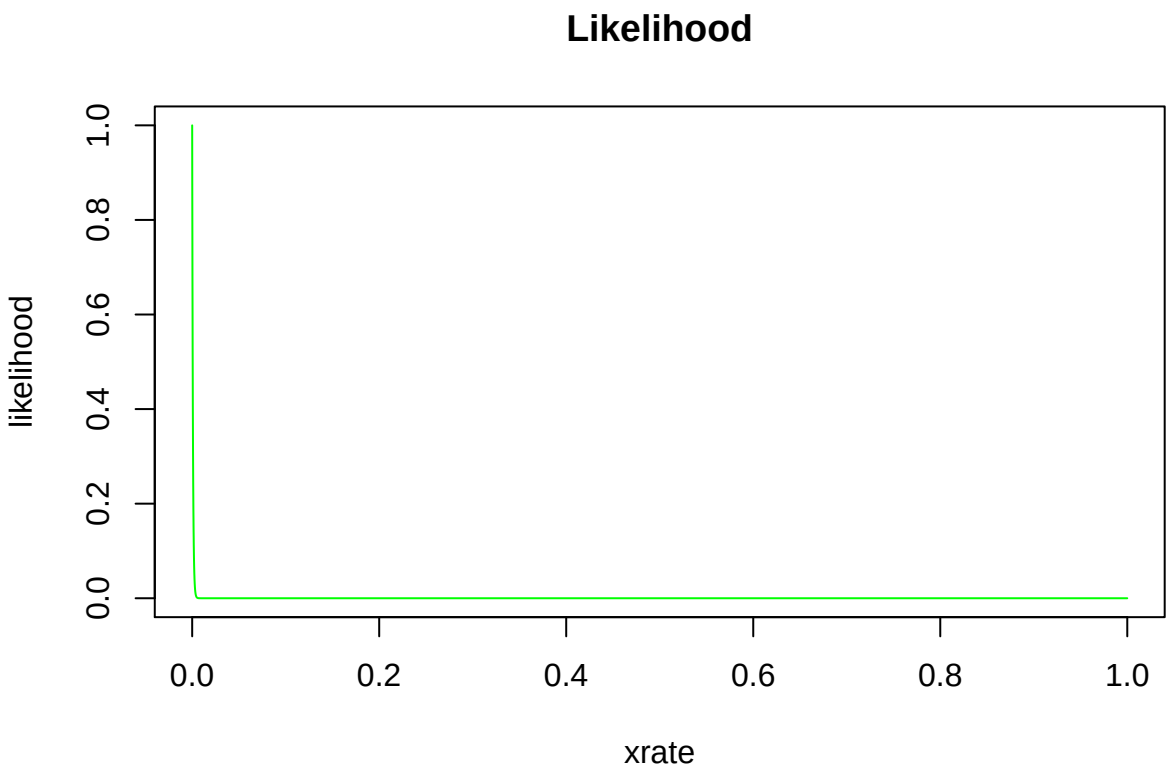


1.2

Build the corresponding Binomial model for the number of people suffering from the disease based on the 1279 test. Obtain the theoretical posterior distribution for this scenario.

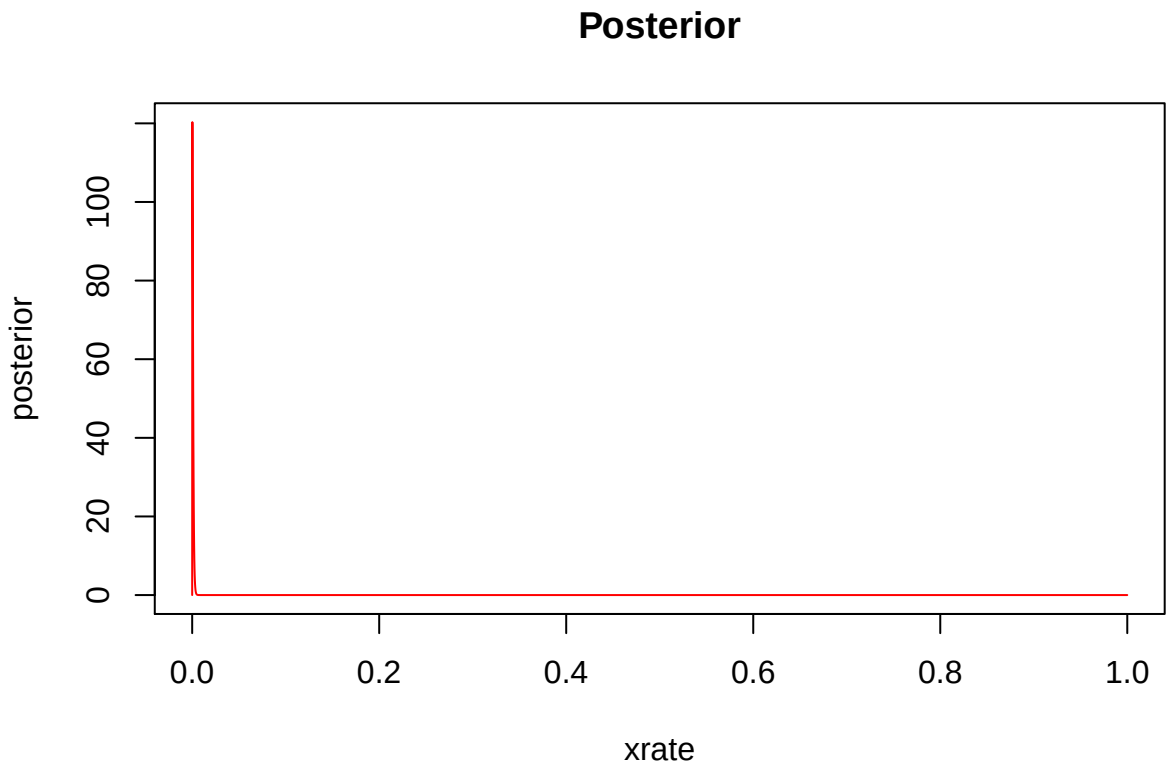
```
# create binomial likelihood function
likelihood <- dbinom(positive_aus, n_samples_aus, xrate)

# plot likelihood function
plot(xrate, likelihood, type = "l", col = "green", main = "Likelihood")
```



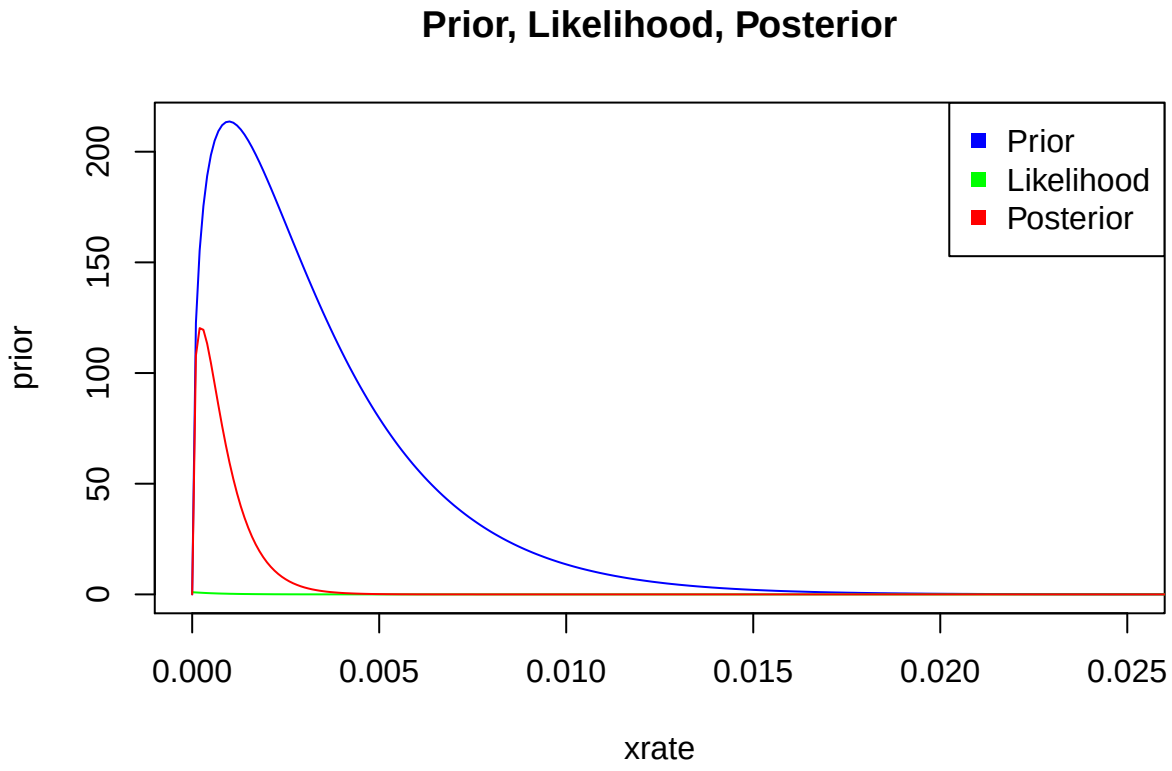

```r
# calculate posterior
posterior <- prior * likelihood

# plot
plot(xrate, posterior, type = "l", col = "red", main = "Posterior")
```



Now we plot all densities. We only show the X scale up to 0.025.

```
# plot Densities
plot(xrate, prior, type = "l", col = "blue", main = "Prior, Likelihood, Posterior",
    xlim = c(0, 0.025))
lines(xrate, likelihood, col = "green")
lines(xrate, posterior, col = "red")
legend("topright", col = c("blue", "green", "red"), legend = c("Prior", "Likelihood",
    "Posterior"), pch = c(15, 15))
```



1.3

Plot the posterior density and obtain the point estimators and 95% Highest posterior density interval of the prevalence of Covid19 (=proportion of inhabitants suffering from the disease).

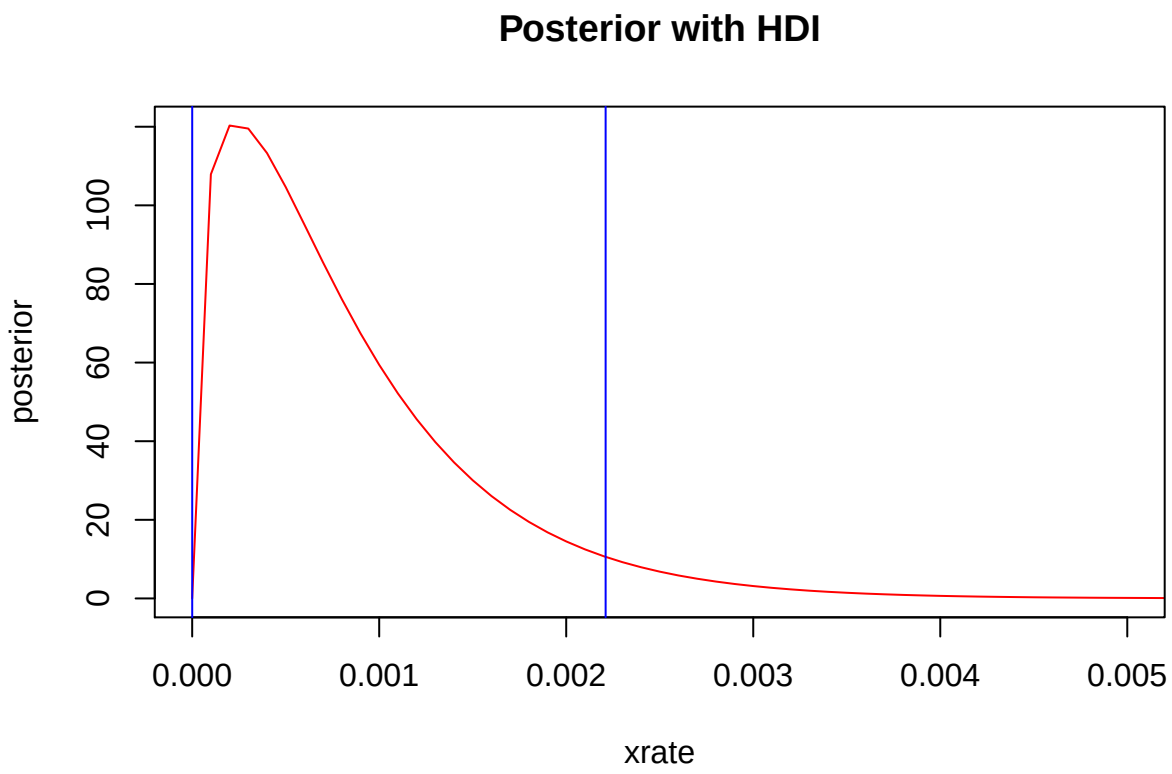
```
# shape parameters for posterior
alpha_post <- alpha + positive_aus
beta_post <- beta + n_samples_aus

# obtain the point estimators
median_post <- (alpha_post - 1/3)/(alpha_post + beta_post - 2/3)
mean_post <- alpha_post/(alpha_post + beta_post)
sd_post <- alpha_post * beta_post/((alpha_post + beta_post)^2 * (alpha_post + beta_post +
    1))

# obtain 95% HDI
hdi <- hdi(qbeta, credMass = 0.95, shape1 = alpha_post, shape2 = beta_post)

# plot posterior with hdi
plot(xrate, posterior, type = "l", col = "red", main = "Posterior with HDI", xlim = c(0,
    0.005))
```

```
abline(v = hdi[1], col = "blue")
abline(v = hdi[2], col = "blue")
```




```
# output results
vals <- c(hdi[1], hdi[2], median_post, mean_post, sd_post)
names(vals) <- c("hdi 2.5%", "hdi 97.5%", "median", "mean", "sd")
kable(vals, caption = "Point estimators and 95% Highest posterior density interval",
    col.names = "")
```

Table 1: Point estimators and 95% Highest posterior density interval

hdi 2.5%	0.0000002
hdi 97.5%	0.0022107
median	0.0006322
mean	0.0008295
sd	0.0000005

1.4

Explain why Statistik Austria chose this method instead of simulationbased or frequentist inference for obtaining intervals of the prevalence.

Because in Austria are 0 positive Covid19 cases standard simulationbased mechanisms breaks. We can't use Bootstrapping or similar techniques in this case. With Bayesian methods we are able to integrate prior knowledge into the "model". In our case we use the German study as prior knowledge.

Task 2

We revisit linear models and their residual distributions. We have already learned that the distribution of residuals is assumed to be normal. Therefore, the Bayesian linear modelling will assume a normal distribution for the data y $N(x^T\beta, \sigma^2)$ for a single explanatory variable scenario, we will therefore consider the inference of the linear model's coefficient β and the residual variance σ^2.

2.1

Define conjugate priors for the coefficient parameter and the residual means independently. Explain how the parameters can be set to be uninformative. Compare different choice of prior parameters.

Conjugate prior for mean: $\beta \sim N(m, s^2)$

Conjugate prior for variance: $\sigma^2 \sim IG(a, b)$

```
# define uninformative prior
beta_prior_uninformative <- dnorm(xrate, 0, 1e+06)

# define uninformative prior
residuals_prior_uninformative <- dinvgamma(xrate, 1e-06, 1e-06)
```

We can select a uninformative prior by setting the standard deviation to a very high number and the mean to zero for example. For selecting an informative prior we need to have domain knowledge and use a specific mean and standard deviation.

For the inverse gamma distribution we need to select very low values as shape parameters a and b to get an uninformative prior. For the informative prior domain knwoledge is needed.

2.2

Build the corresponding normal model the regression inference. Obtain the theoretical posterior distribution for both parameters separately assuming the other one to be "known".

Posterior β assuming σ known:

$$
\begin{aligned}
f(\beta|x,y) &\propto \frac{1}{\sqrt{2\sigma\pi}} \prod_{i=1}^{n} e^{-\frac{1}{2}\cdot(\frac{\beta-m}{s})^2} \cdot e^{-\frac{1}{2}\cdot(\frac{y_i - x_i\beta}{\sigma})^2} \\
&\propto \prod_{i=1}^{n} e^{-\frac{1}{2}((\frac{\beta-m}{s})^2 + (\frac{y_i - x_i\beta}{\sigma})^2)} \\
&\propto e^{\sum_{i=1}^{n} -\frac{1}{2}((\frac{\beta-m}{s})^2 + (\frac{y_i - x_i\beta}{\sigma})^2)} \\
&\propto e^{-\frac{1}{2}(\beta^2(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{1}{s^2}) - 2\beta(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i y_i + m))} \\
&\sim N\left(\frac{\sum_{i=1}^{n}\frac{x_i y_i}{\sigma^2} + \frac{m}{s^2}}{\sum_{i=1}^{n}\frac{x_i^2}{\sigma^2} + \frac{1}{s^2}}, \frac{1}{\sum_{i=1}^{n}\frac{x_i^2}{\sigma^2} + \frac{1}{s^2}}\right)
\end{aligned}
$$

Posterior σ^2 assuming μ known:

$$\begin{aligned} f(\frac{1}{\sigma^2}|x,y) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{1}{2}(\frac{y_i - x_i\beta}{\sigma})^2} \cdot \frac{b^a}{\Gamma(a)} \cdot \frac{1}{\sigma^2}^{(a-1)} \cdot e^{-b\frac{1}{\sigma^2}} \\ &\propto (\frac{1}{\sigma^2})^{a-1+n} \cdot e^{-\frac{1}{2}(\sum_{i=1}^{n}(\frac{y_i - x_i\beta}{\sigma})^2 + \frac{2b}{\sigma^2})} \\ &\propto (\frac{1}{\sigma^2})^{a-1+n} \cdot e^{-\frac{\frac{(y_i - x_i\beta)^2}{2} + b}{\sigma^2}} \\ &\sim Ga(a+n, \frac{(y_i - x_i\beta)^2}{2} + b) \end{aligned}$$

2.3

Provide the formulae for point estimators and 95% Highest posterior density interval of the regression parameters separately assuming the other one to be "known".

$$\mu = \frac{a+n}{\frac{(y_i - x_i\beta)^2}{2} + b} \tag{1}$$

$$sd = \sqrt{\frac{a+n}{(\frac{(y_i - x_i\beta)^2}{2} + b)^2}} \tag{2}$$

2.4

Test this with the data from your exercise 6: dataset Auto and model mpg ~ horsepower

```
# load data
data(Auto, package = "ISLR")
# create linear model
model_lm <- lm(mpg ~ horsepower, data = Auto)

# extract beta and residuals
beta <- model_lm$coefficients[2]
res <- model_lm$residuals
```