Computer lab 1

Kevin Neville, Gustav Sternelöv

April, 7th, 2016

1. Bernoulli ... again.

a)

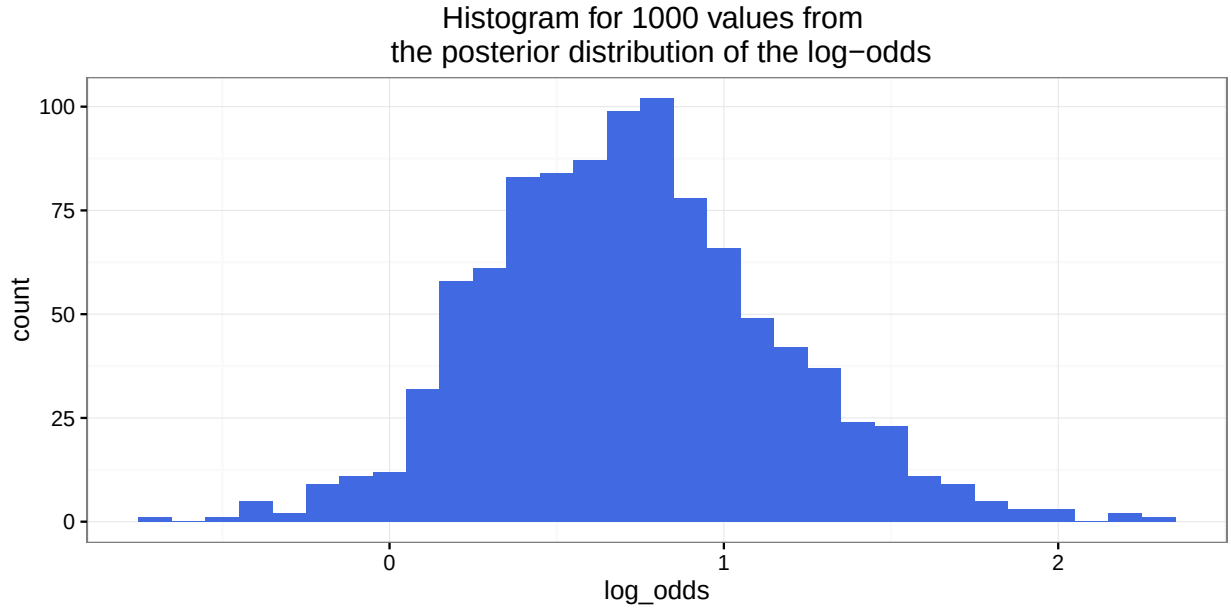
The true value for the mean of the expected beta distribution is $\alpha/(\alpha+\beta)$. When using $(\alpha_0 + \text{s})$ as α_n and $(\beta_0 + \text{f})$ as β_n, the true value of the mean becomes 0.6667.
When taking samples of different sizes from this distrubution we obtain the following values 0.6851, 0.6575, 0.666. Here we have used n= 10, 100 and 1000.
For the standard deviation we obtain the true value with the following formula: $\frac{\alpha\beta}{(\alpha\beta)^2(\alpha+\beta+1)}$. The true value then is 0.0943. This can be compared to the obtained standard deviation of the generated values which is 0.0815, 0.1089, 0.0911.
By looking at the results preseted above it is clear that the generated values converges to the true values for larger n.

b)

The true value is given by the function *pbeta* and the obtained value is 0.00397. This can be compared to the probabilities given by the generated values for different n. The obtained probabilties are: 0, 0.01, 0.003. We see that when using n=1000 the obtained probability is close to the true value.

c)

The posterior distribution of the log-odds for *n=1000* simulated θ values is visualised by the histogram below.



More information about the posterior distribution, such as the mean and median value, for the log-odds is given by the *density()* function.

```
##
## Call:
##  density.default(x = log_odds)
##
## Data: log_odds (1000 obs.);  Bandwidth 'bw' = 0.09514
##
##        x                 y
##  Min.   :-0.95267   Min.   :0.0000474
##  1st Qu.:-0.06588   1st Qu.:0.0131513
##  Median : 0.82091   Median :0.1056891
##  Mean   : 0.82091   Mean   :0.2816388
##  3rd Qu.: 1.70771   3rd Qu.:0.5371676
##  Max.   : 2.59450   Max.   :0.9356906
```

2. Log normal distribution and the Gini coefficient.

a)

Likelihood

$$\left(\frac{1}{y * \sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(log(y)-\mu)^2}$$
$$\propto \left(\frac{1}{\sqrt{\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(log(y)-\mu)^2}$$

Likelihood * prior

$$\frac{1}{\sigma^2}\left(\frac{1}{\sqrt{\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(log(y)-\mu)^2}$$

$$\propto \frac{1}{\sigma^2}\frac{1^n}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(log(y)-\mu)^2}$$

$$\propto (\sigma^2)^{-1}(\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum(log(y)-\mu)^2}$$

$$\propto (\sigma^2)^{-(n-2)/2} e^{-\frac{1}{2\sigma^2}\sum(log(y)-\mu)^2}$$

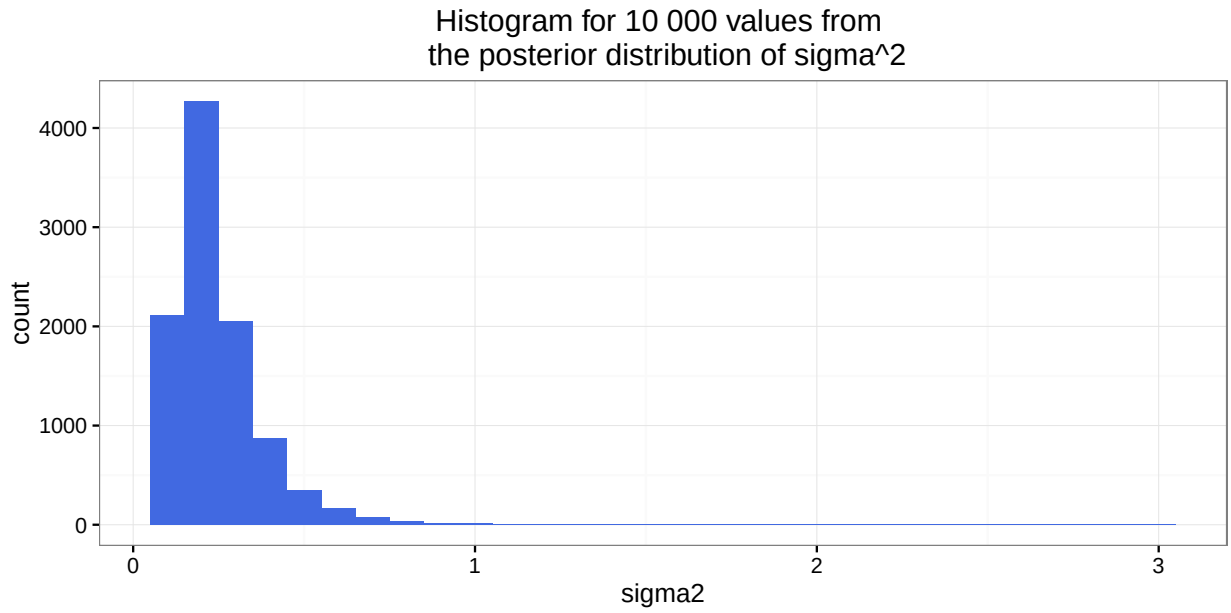
Theoretical distribution

$$p(\sigma^2) \propto (\sigma^2)^{\frac{v_n}{2}+1} e^{-\frac{v_n t_n^2}{2\sigma^2}}$$

By rearranging the expression from the multiplication of the prior and the likelihood it can be seen that the posterior for σ^2 is the $inv\text{-}\chi^2(n, \tau^2)$ distribution.

$$\propto \sigma^{-2(\frac{n}{2}+1)} e^{-\frac{1}{2\sigma^2}\sum(log(y_i)-\mu)^2}$$

b)




```
## [1] 2.979245
```

When simulating 10000 draws from our posterior we get the following mean and variance:

```
##      Mean
## 0.2467178
```

```
##   Variance
## 0.02072043
```

The theoretical mean and variance is given by the following formulas:

$$Mean: \frac{v}{v-2}s^2$$

$$Variance: \frac{2v^2}{(v-2)^2(v-4)}$$

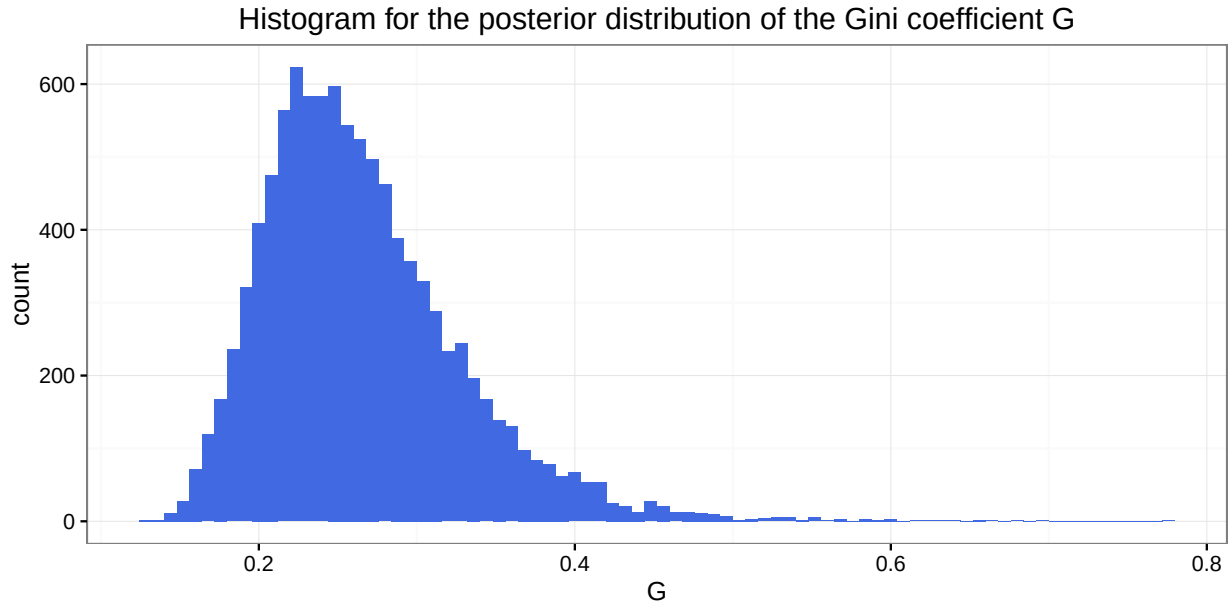
```
## theoretical_mean
##        0.2473497
```

```
## theoretical_variance
##           0.02039396
```

The theoretical mean and variance are very close to the ones generated from our posterior.

c)

The posterior distribution of the Gini coefficient G for the current set is visualised with a histogram.



The mean of the histogram above is 0.2655, which tells us that the income is more equal than inequal.

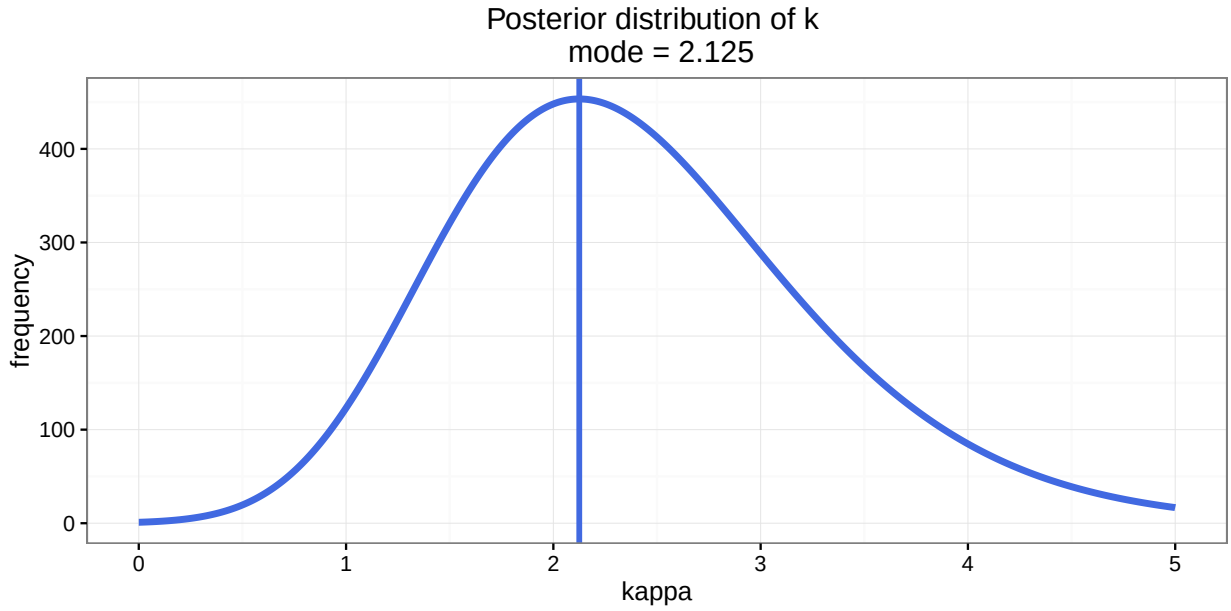
Assignment 3

a-b)

The posterior distribution for κ is proportional to

$$(\frac{1}{I_0(\kappa)})^n exp[\kappa * \sum cos(y_i - \mu) - \kappa]$$

This posterior distribution of κ is plotted over a grid of κ values going from 0 to 5 by steps of 0.001.



The mode value for the posterior distribution is 2.125. This is the value with the maximum posterior probability for the posterior distribution of κ for the wind direction data.

Assignment 4

a)

The prior distribution of λ is proportional to

$$\lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda}$$

Y follows a Poisson($\frac{n_i \lambda}{100000}$) distribution and its likelihood is proportional to

$$\lambda^{\sum Y_i} e^{-\lambda n}$$

Where λ is equal to $\frac{n_i \lambda}{100000}$.

The posterior is equal to the prior times the likelihood

$$\lambda^{\alpha - 1} e^{-\beta \lambda} * \lambda^{\sum Y_i} e^{-\lambda n}$$

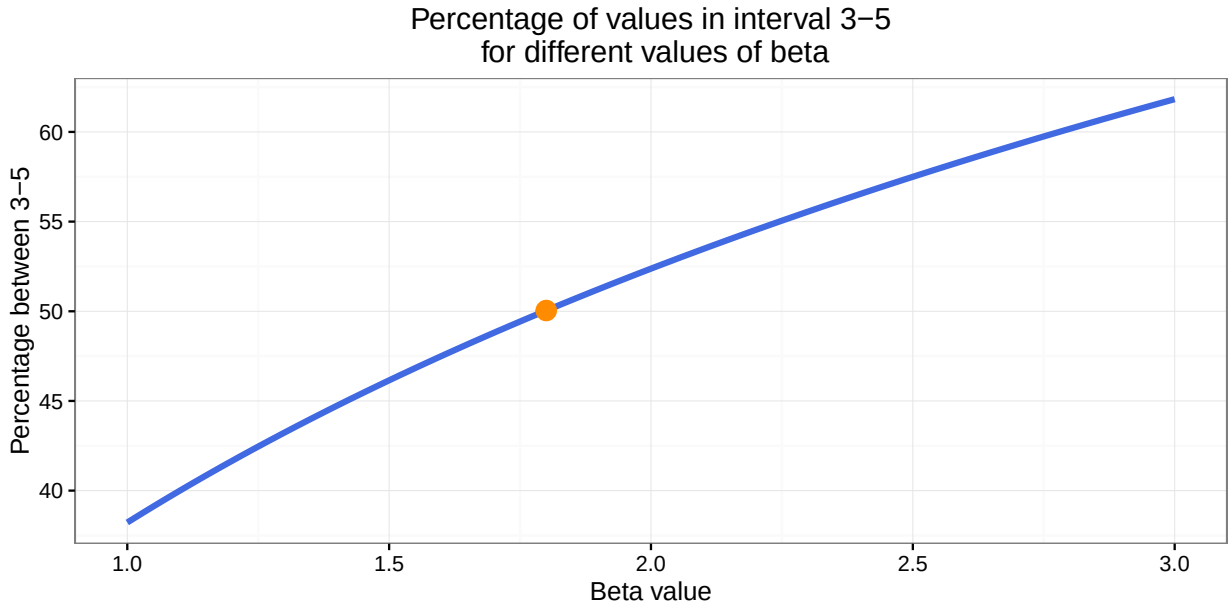
Which can be written as (remember that λ is equal to $\frac{n_i \lambda}{100000}$)

$$\lambda^{\alpha_0 + \sum Y_i - 1} e^{-\lambda (b_0 + \frac{\sum n_i}{100000})}$$

It is then easy to see that the posterior distribution also is a gamma distribution but with the parameters $\alpha_0 + \sum Y_i$ and $b_0 + \frac{\sum n_i}{100000}$

b)

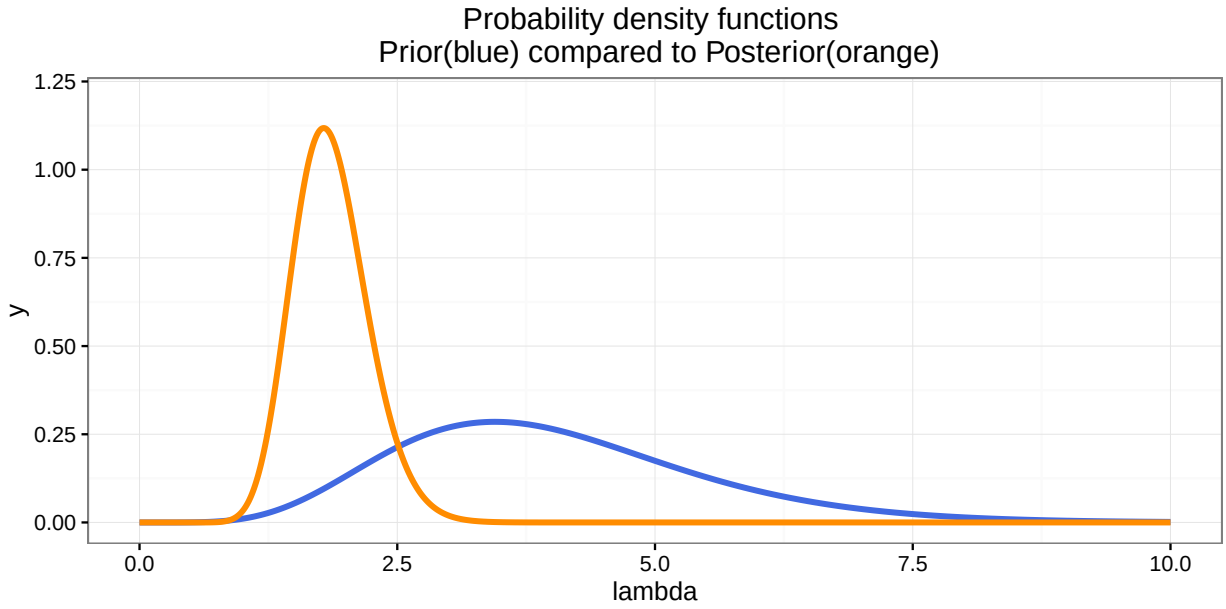
A trial and error approach is used for finding the parameter values for the prior. λ follows a Gamma(4β, β) distribution and the aim is to find a value of β such that $\Pr(3 \leq \lambda \leq 5) \approx 0.5$. Values in the range 1 to 3 by steps of 0.01 is tested and the results are presented in the plot below.



The value closest to 0.5 is given when β is equal to 1.80. Hence, the chosen prior distribution for λ is a Gamma(4*1.8, 1.8).

c)

The prior for λ is updated in line with the results obtained in *b)*. With the updated information about the prior is it a gamma distribution with the following parameter settings: $\lambda \sim Gamma(7.2, 1.8)$. The posterior distribution is a also a gamma distribution but with updated parameters, $\lambda \sim Gamma(7.2+19, 1.8+12.31663)$. The probability density function of λ for both the prior and the posterior are compared with the following plot.



The observed data has had a rather heavy impact on the distribution specified by the prior. As shown by the plot has the curve for the probability density function of the posterior shifted to the left by quite some margin. For example is the posterior probability that λ is between 3 and 5 just 0.313 %.

Appendix

R-code

```
library(ggplot2)
# 1.a)
set.seed(12345)
beta10 <- rbeta(10, 16, 8)
beta100 <- rbeta(100, 16, 8)
beta1000 <- rbeta(1000, 16, 8)
#sd(beta10); sd(beta100) ; sd(beta1000)
true_sd <- sqrt((16*8)/((16+8)^2*(16+8+1)))
#mean(beta10); mean(beta100) ; mean(beta1000)
true_mean <- 16/(16+8) # True value, (alpha + s) / ((alpha + s) + (beta + f))
# 1.b)
beta_10_val <-length(subset(beta10, beta10 < 0.4)) / 10
beta_100_val <- length(subset(beta100, beta100 < 0.4)) / 100
beta_1000_val <- length(subset(beta1000, beta1000 < 0.4)) / 1000
true_prob <- pbeta(0.4, 16, 8)
log_odds <- log(beta1000 / (1-beta1000))
log_odds_frame <- data.frame(log_odds=log_odds)
ggplot(log_odds_frame, aes(log_odds)) + geom_histogram(binwidth=0.1, fill="royalblue") + theme_bw() + gg
density(log_odds)
y <- c(14,25,45,25,30,33,19,50,34,67)
sigma2 <- 0
set.seed(12345)
for(i in 1:10000){
```

```
  chisq <- rchisq(1, 10)
  sigma2[i] <- (sum((log(y)-3.5)^2)) / chisq
}
sigma2_frame <- data.frame(sigma2=sigma2)
ggplot(sigma2_frame, aes(sigma2)) + geom_histogram(binwidth=0.1, fill="royalblue") + theme_bw() + ggtitl
max(sigma2)
#mean(sigma2)
#var(sigma2)
# theoretical mean
#c(theoretical_mean=(sum((log(y)-3.5)^2))/(10-2))
# theoretical variance
#c(theoretical_variance=(2*10^2 * ((sum((log(y)-3.5)^2))/10)^2 ) / (8^2 * 6))
c(Mean=mean(sigma2))
c(Variance=var(sigma2))
c(theoretical_mean=(sum((log(y)-3.5)^2))/(10-2))
c(theoretical_variance=(2*10^2 * ((sum((log(y)-3.5)^2))/10)^2 ) / (8^2 * 6))
G <- 2*pnorm((sqrt(sigma2)/sqrt(2)))-1
G_frame <- data.frame(G=G)
ggplot(G_frame, aes(G)) + geom_histogram(binwidth=0.008, fill="royalblue") + theme_bw() + ggtitle("Hist
# a)
radians <- c(-2.44,2.14,2.54,1.83,2.02,2.33,-2.79,2.23,2.07,2.02)
kappa <- seq(from=0, to=5, by=0.001)
distKappa <- (1/besselI(kappa, 0))^10 * exp(kappa * sum(cos(radians-2.39))-kappa)
kappaFrame <- data.frame(x=kappa, y=distKappa)
maxkappa <- which.max(distKappa)
modeKappa <- kappa[2126]
ggplot(kappaFrame, aes(x=x,y=y)) + geom_line(col="royalblue", size=1.5) + geom_vline(xintercept=modeKapp
  ggtitle("Posterior distribution of k \n mode = 2.125") + theme_bw() +ylab("frequency") + xlab("kappa")
bseq <- seq(1,3, 0.01)
prob <- 0
j <- 0
for(i in bseq){
  j <- j+1
  prob[j] <-  (pgamma(q = 5, 4*i, i) - pgamma(q = 3, 4*i, i) ) *100
}
prob <- data.frame(x=bseq, y=prob)
ggplot(prob, aes(x=x, y=y)) + geom_line(col="royalblue", size=1.35) + ggtitle("Percentage of values in
# c
ysum <- sum(120342 + 235967 + 243745 + 197452 + 276935 + 157222) / 100000
 # prior
prior <- data.frame(x=seq(0,10, 0.01), y=dgamma(seq(0, 10, by=0.01), 4*1.81, 1.81))
posterior <- data.frame(x=seq(0,10, 0.01),y=dgamma(seq(0, 10, by=0.01), 4*1.81 + 19, 1.81 + 12.31663))
ggplot(prior, aes(x=x, y=y)) + geom_line(col="royalblue", size=1.3) + ylim(0,1.2) + geom_line(data=poste
## NA
```