Labbrapport i Statistik

Laboration 1

732g43

Jakob Lindén

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet
05-11-2019

Innehåll

1. Uppgift 1

Uppgiften ämnar att undersöka data från samtliga svenska väljarbaromterar. Vidare ska endast ett parti analyseras och i denna rapport väljs Miljöpartiet.

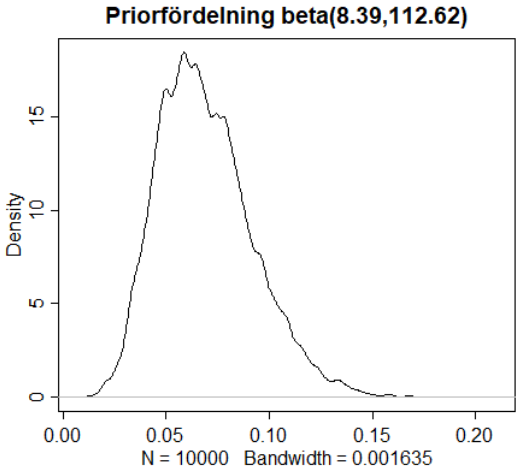
1.1 a)

```
require(rethinking)

valdata <- read.csv2("C:/Users/Jakob/Desktop/Bayesian statistics/valdata.csv")
val_andel_mp <- valdata[1:12,10]/100

alpha <- 8.389154
beta <- 112.6174

prior <- rbeta(n = 1e4, alpha,beta, ncp=0)

dens(prior,main = "Priorfördelning beta(8.39,112.62)")
```

En konjugerade prior för andelen väljare i september 2018 för Miljöpartiet ska väljas ut.

Den konjugerade priorn väljs ut genom att subjektivt undersöka tidigare års väljarbarometrar och se ungefär hur Miljöpartiet har legat till de senaste åren och ungefär hur stor avvikelse man kan förvänta sig i andelen röster. Efter undersökning av tidigare data förväntas andelen röster på Miljöpartiet ligga på ungefär 6 procent. Jag förväntar mig att andelen röster på miljöpartiet kommer ha en avvikelse på ungefär -5 procent och $+9$ procent. Således väljs den konjugerade priorn till en $Beta(\alpha = 8.39, \beta = 112.62)$. Betafördelningen som en konjugerade prior är rimlig eftersom denna fördelning endast antar värden mellan 0 och 1 vilket är rimligt när vi har med andelar att göra.

Priorfördelnigen visualiseras i figur 1.1 där vi kan se att den valda konjugerade priorn har störst sannolikhet vid ungefär 6 procent av andelen röster med den önskade avvikelsen, denna prior speglar alltså mina förkunskaper om andelen väljare som förväntas rösta på Miljöpartiet.



Figur 1.1

1.2 b)

```
n <- 1612
sumxi <- 77.376

posterior <- rbeta(n = 1e4, alpha + sumxi, beta + n - sumxi, ncp = 0)

dens(posterior,col="black", main = "Posterior- och Priorfördelning")
dens(prior,col="red",add=TRUE)
legend(0.06, 70, legend = c("Posterior", "Prior"), fill=c("black", "red"), cex = 0.8)
```

Genom användning av data för *Demoskops* väljarundersökning i september 2018 ska priorn uppdaters till posterior genom att dra 10000 värden direkt från den kända posteriorfördelningen.

För att ta reda på posteriorns exakta fördelning så multipliceras likelihood funktionen för data (Binomial) med priorfördelningen (Beta). Modellen och likelihoodfunktionen ser ut som följande:

$$\begin{aligned} X_1, \ldots, X_n|\theta &\overset{\text{iid}}{\sim} \text{Bin}(x|\theta) \\ f\left(x_1, \ldots, x_n|\theta\right) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{n-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \end{aligned} \tag{1.1}$$

Prior fördelningen ser ut som följande:

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ p(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned} \tag{1.2}$$

Genom att multiplicera likelihood och prior fås posterior fördelningen som följande:

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto p(\mathbf{x}|\theta)p(\theta) \\ &\propto \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+\sum_{i=1}^{n} x_i-1}(1-\theta)^{\beta+n-\sum_{i=1}^{n} x_i-1} \end{aligned} \tag{1.3}$$

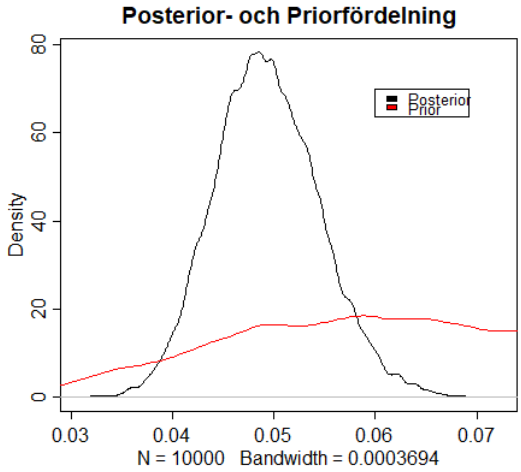
Detta leder till att posteriorfördelningens hyperparametrar är $Beta(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i)$.

n i detta fall är antalet intervjuer totalt för *Demoskops* i september 2018 vilket är 1612 stycken. Andelen som angav sig rösta på Miljöpartiet i dessa intervjuer var 4.8 procent, alltså totalt 77 stycken. Vi har

från priorfördelningen med oss att $\alpha = 8.39$ och $\beta = 112.62$, detta medför att posteriornsfördelning är $Beta(86, 1647)$.

Genom att sampla 10 000 värden från posteriorfördelningen så visualiseras den i figur 1.2 tillsammans med priorfördelningen.

Det ses i figuren att priorfördelningen och posteriorfördelning är väldigt olika varandra. Där priorfördelningen har betydligt mycket mer varians än posteriorfördelningen samt högre medelvärde. Anledningen till att prior och posteriorfördelningen blir så olika är just för att vi har så många datapunkter (1612), således är det mest data som kommer styra fördelningen av posteriorn.



Figur 1.2

1.3 c)

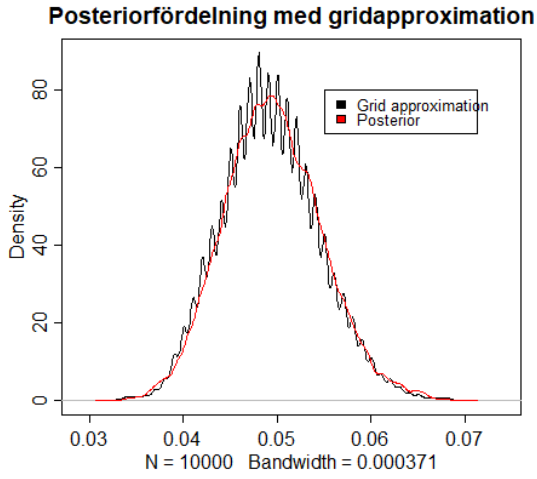
```
#grid approximation

p_grid <- seq( from=0 , to=1 , length.out=1000 ) # grid av v?rden

# Allm?n Beta prior, BW, se f?rel?sningsslides
prior <- dbeta(p_grid, alpha, beta, ncp = 0, log = FALSE)

likelihood <- dbinom( round(sumxi) , size=n , prob=p_grid )
posterior <- likelihood * prior # posterior prop mot likelihood*prior
posterior <- posterior / sum(posterior) # g?r posteriorn till en t?thet

# R code 3.3
samples_grid <- sample( p_grid , prob=posterior , size=1e4 , replace=TRUE ) # dra v?rden fr?n
    posteriorn

# R code 3.4
plot( samples_grid ) # plotta v?rdena

dens(samples_grid, main = "Posteriorfördelning med grid approximation")

#kvadratisk approximation

# R code 2.6, tag fram posteriorf?rdelningen med kvadratisk approximation
KA <- map(
  alist(
    w ~ dbinom(1612,p) , # binomial likelihood
    p ~ dbeta(8.389154,112.6174) # beta prior
```

```
  ) ,
  data=list(w=77) )

# display summary of quadratic approximation
precis( KA , digits=5)

# Plotta kvadratiska approximationen mot grid approximationen. Anv?nd medelv?rde och st.avv.
    fr?n funktion precis.
samplesQ <- rnorm(1e4,0.04875,0.00517)
#Black = gridapprox, Red = kvadrat approx
dens(samples_grid,type="l",col="black", main = "Posteriorfördelning med approximation")
dens(samplesQ,type="l",col="red",add=TRUE)
legend(0.055, 80, legend = c("Grid approximation", "Kvadrat approximation"), fill=c("black",
    "red"), cex = 0.8)
```

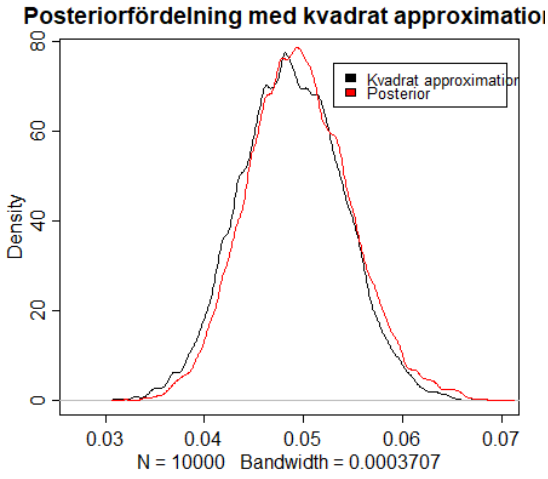
Återigen ska data från *Demoskops* väljarundersökning användas från september 2018 för att simulera fram 10 000 värden från posteriorfördelningen med hjälp av gridapproximation samt kvadratisk approximation.

Resultatet från de två approximeringarna redovisas i figur 1.3 respektive 1.4 tillsammans med posteriorn från föregående uppgift.

De två approximeringarna ser följa den exakta posteriorn godtyckligt. De har dock helt olika utseende där gridapproxmiationen får ett hackigt utseende medans kvadratiska approximationen får ett mer likt utseende till den exakta posteriorn. Det är relativt lite skillnad på de två approximationerna, detta beror på att då vi har tillgång till mycket data, 1612 observationer i detta fall så blir båda approximationerna bra. Gridapproximationen förväntas även ha ett mer jämt utseende om vi har fler lyckade utfall i data, i detta fall var endast 77 av 1612 av utfallen lyckade.



Figur 1.3



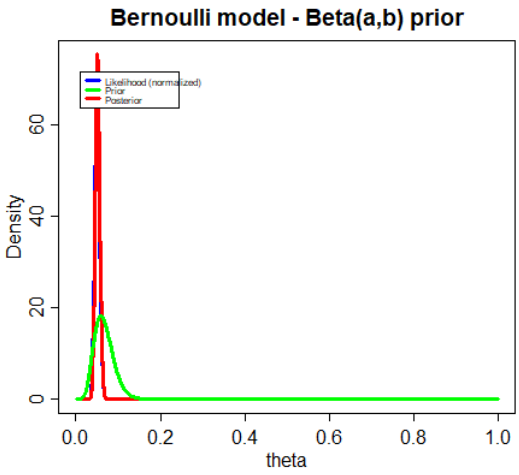
Figur 1.4

1.4 d)

En känslighetsnaalys ska först utföras i den valda konjugerade prior, sedan med en uniform prior och tillsist med en betydligt mer informativ prior en den tidigare vald konjugerade prior.

1.4.1 i.

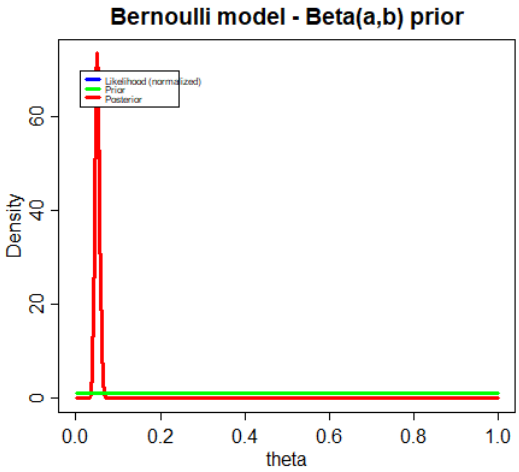
I figur 1.5 presenteras den valda konjugerade prior tillsammans med den normaliserade likelihooden samt posteriorfördelningen. Det blir väldigt tydligt i denna figur hur lite den valda priorn faktiskt påverkar posteriorfördelningen, detta eftersom likelihooden och posteriorn nästan ligger på varandra. Detta är pågrund av den stora mängden data som fanns att tillgå i detta exempel där $n = 1612$.



Figur 1.5

1.4.2 ii.

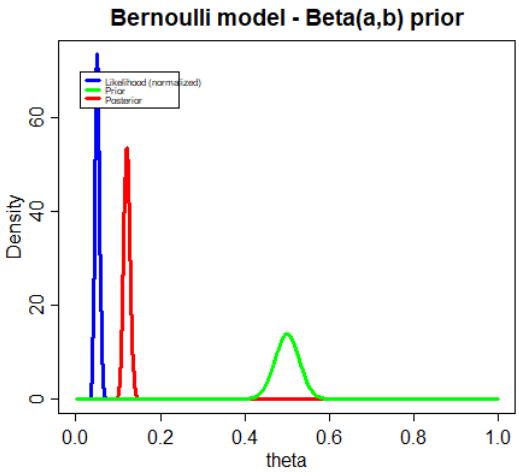
I figur 1.6 ses en uniform prior tillsammans med den normaliserade likelihooden och posteriorfördelningen. Eftersom den valda priorn är av den uniforma fördelningen så kommer inte den ha nån påverkan på posteriorfördelningen. Detta ses tydligt eftersom posteriorn och likelihooden ligger exakt på varandra.



Figur 1.6

1.4.3 iii.

I figur 1.7 ses den normaliserade likelihooden, posteriorn och en betydligt mer informativ prior än de tidgiare redovisade prior fördelningarna. Detta ger en tydlig bild av att posterior fördelningen dras bort från likelihooden trots det stora antalet observationer, detta pågrund av att priorn är så informativ.



Figur 1.7

1.5 e)

```
mean(samples_grid)
sd(samples_grid)
```

Posteriorfördelningens medelvärde och standardavvikelse ska beräknas med hjälp av informationen från gridapproximationen i uppgift (c). Medelvärdet för posteriorfördelningen framtagen med gridapproxmiation ses vara 0.049, således förväntas miljöpartiet få ungefär 4.9 procent av rösterna med denna modell. Standardavikelsen beräknas till 0.005 vilket ger att andelen röster på miljöpartiet kommer avvika med ungefär 0.5 procent.

1.6 f)

```
posterior_PI <- rbeta(1e4,alpha,beta,ncp = 0)
PI( posterior_PI , prob=0.909 )
```

Ett 90.9 procent kredibilitetsintervall för andelen väljare av Miljöpartiet ska presenteras genom att använda 10000 samplade värden från den exakta posterior fördelningen från uppgift (b).

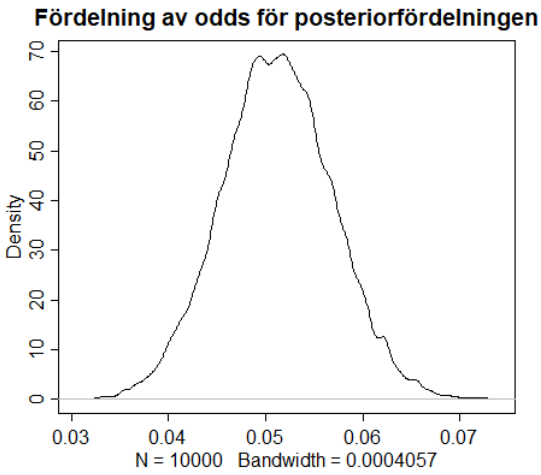
Intervallet produceras och säger att sannolikheten att den okända andelen röster Miljöpartiet kommer få ligger mellan 3.56 procent och 11.21 procent med 90.9 procent sannolikhet.

1.7 g)

```
odds <- samplesQ / (1 - samplesQ)
dens(odds, main="Fördelning av odds för posteriorfördelningen")
```

Posteriorfördelningen för oddset ska beräknas. Posteriorfördelningen från den kvadratiska approximationen som producerades i uppgift (c) ska användas.

Oddset är förhållandet mellan sannolikheten att ett event sker gentemot att det inte sker, således förhållandet mellan att en person röstar på Miljöpartiet gentemot att denne inte gör det. Eftersom sannolikheten att rösta på Miljöpartiet är så låg blir posteriorfördelningen för oddset väldigt lik den allmäna posteriorfördelningen. Posteriorfördelningen för oddset visualiseras i figur 1.8


Fördelning av odds för posteriorfördelningen
Density
0
10
20
30
40
50
60
70
0.03
0.04
0.05
0.06
0.07
N = 10000 Bandwidth = 0.0004057

Figur 1.8

2. Uppgift 2

Uppgiften behandlar data om vikter för kycklingar efter 6 veckors föda.

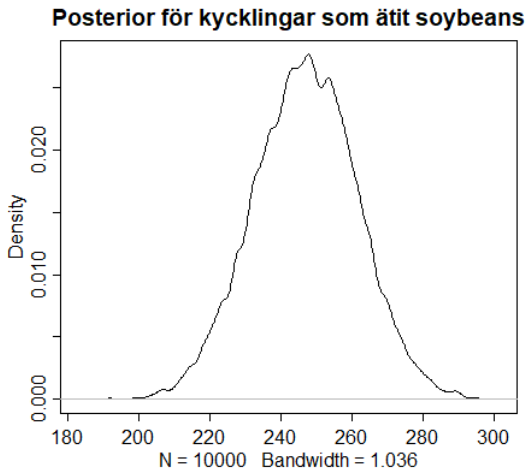
2.1 a)

```
data(chickwts)

soybeans_index <- chickwts$feed == "soybean"
meatmeal_index <- chickwts$feed == "meatmeal"

soybean_data <- chickwts[soybeans_index,1:2]
meatmeal_data <- chickwts[meatmeal_index,1:2]

ejsoy_ejmeat <-chickwts[!soybeans_index&!meatmeal_index,1:2]


#a)
Stavv <- sd(soybean_data[,1])

PriorMu <- mean(ejsoy_ejmeat[,1])
PriorStavv <- sd(ejsoy_ejmeat[,1])

flist <- alist(
  weight ~ dnorm(mu, Stavv) , # likelihood fr?n normalf?rdelning
  mu ~ dnorm( PriorMu , PriorStavv ) # normalf?rdelad prior
)

resNormal <- map(flist, data=soybean_data)
precis(resNormal)

samples_2a <- rnorm(1e4, 246.85, 14.27)
dens(samples_2a, type="l",col="black", main = "Fördelning av vikten på kyckling som ätit
    soybeans")
```

Kvadratapproximation ska användas för att simulera fram 10000 värden från följande posterior för en genomisnittsvikt på kycklingarna:

Först ska viktdata för kycklingar som fått födan 'soybean' användas och antagande ska göras om att vikterna följer en normalfördelning med känd standardavvikelse som standardavvikelsen av vikterna. Sedan ska en normalfördelad prior för medelvikten användas där medelvärdet och standardavvikelsen i priorn är lika med medelvärdet och standardavvikelsen för ikterna på kycklingar som inte fått varken 'soybean' eller 'meatmeal'.

Den framtagna posteriorfördelningen redovisas i figur 2.1 där det ses att medelvärdes vikten på kycklingar som ätit soybeans är ungefär 245 gram med en avvikelse på ungefär 14 gram.



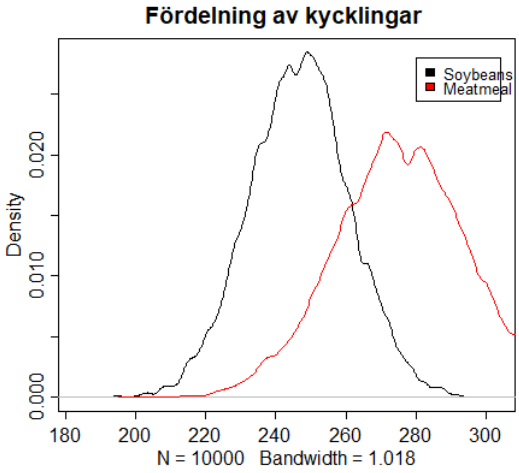
Figur 2.1

2.2 b)

```
Stavv_b <- sd(meatmeal_data[,1])

PriorMu_b <- mean(ejsoy_ejmeat[,1])
PriorStavv_b <- sd(ejsoy_ejmeat[,1])

flist_b <- alist(
  weight ~ dnorm(mu, Stavv_b) , # likelihood fr?n normalf?rdelning
  mu ~ dnorm( PriorMu_b , PriorStavv_b ) # normalf?rdelad prior
)

resNormal_b <- map(flist_b, data=meatmeal_data)
precis(resNormal_b)

samples_2b <- rnorm(1e4, 276.2, 19.09)
dens(samples_2b, type="l",col="black")

dens(samples_2a, type="l",col="black", main = "Fördelning av kycklingar ")
dens(samples_2b, type="l",col="red",add=TRUE)
legend(280, 0.028, legend = c("Soybeans", "Meatmeal"), fill=c("black", "red"), cex = 0.8)
```

På samma sätt som i uppgift (a) ska 10000 värden simuleras från posteriorn för en genomsnittsvikt μ, men i detta fall ska viktdata för kycklingar som fått födan 'meatmeal' användas.

I figur 2.2 ses posteriorfördelningen för kycklingar som ätit 'soybeans' samt 'meatmeal' i samma graf. Det ses tydligt att kycklingar som ätit 'meatmeal' väger mer i genomsnitt men har också högre varians. Mer exakt en genomsnittsvikt på 276 och en standardavvikelse på 19 för kycklingar som ätit födan 'meatmeal'.



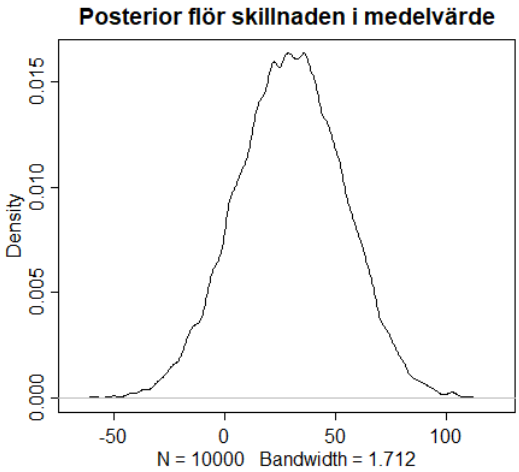
Figur 2.2

2.3 c)

```
posterior_diff <- samples_2b - samples_2a
dens(posterior_diff)

sum( samples_2b > samples_2a) / 1e4
```

Genom att använda de simluerade posteriorvärden i uppgift (a) och (b) ska posteriorfördelningen för skillnaden mellan genomsnittsvikterna för födorna 'meatmeal' och 'soybean' tas fram. Sedan ska posteriorsannolikheten att medelvärdet för kycklingar som ätit 'meatmeal' är större än medelvärdet för kycklingar som ätit 'soybean' beräknas.

Posteriorfördelningen för skillnad i medelvärde mellan kycklingar som ätit de två olika födorna redovisas i figur 2.3. Det syns tydligt att medelskillnaden i vikten mellan de två födorna är ungefär 30 gram med en avvikelse på ungefär 24 gram. Således kan även sannolikheten för att medelvärdes vikten för kycklingar som ätit 'meatmeal' är större än medelvärdes vikten för kycklingar som ätit 'soybeans' beräknas till 0.8927, alltså 89.27 procent.



Figur 2.3