ETC 2420/5242 Lab 10 2016

Souhaib Ben Taieb

Week 10

Purpose

This lab is to compute conditional probabilities and practice Bayesian inference.

Question 1

A situation where Bayesian analysis is routinely used is your spam filter in your mail server. The message is scrutinized for the appearance of key words which make it likely that the message is spam. Let us describe how one one of these filters might work. We imagine that the evidence for spam is that the subject message of the mail contains the sentence "check this out". We define events *spam* (the message is spam) and *check this out* (the subject line contains this sentence).

From previous experience we know that 40% of emails are spam, 1% of spam email have "check this out" in the subject line, and .4% of non-spam emails have this sentence in the subject line.

Explain the different steps to compute the conditional probability *P(spam | check this out)*.

$P(\text{spam}|\text{check this out}) = \frac{P(\text{check this out}|\text{spam})P(\text{spam})}{P(\text{check this out})}$

$P(\text{spam}) = 0.4$

$\text{check this out}|\text{spam} = 0.01$

$$\begin{aligned} P(\text{check this out}) &= P(\text{check this out}|\text{spam})P(\text{spam}) + P(\text{check this out}|\text{not spam})P(\text{not spam}) \\ &= 0.01 \times 0.4 + 0.004 \times 0.6 = 0.0064 \end{aligned}$$

$P(\text{spam}|\text{check this out}) = \frac{0.004}{0.0064} = \frac{5}{8} = 0.625$

Question 2

Let $X_1, \ldots, X_n \sim N(\theta, 9)$.

a. If $\theta \sim N(\mu, \tau^2)$, what is $\pi(\theta|x_1, \ldots, x_n)$?
b. What is the posterior mean $E[\theta|x_1, \ldots, x_n]$?
c. What is the MLE estimate $\hat{\theta}_{\text{MLE}}$?

See the slides of week 9

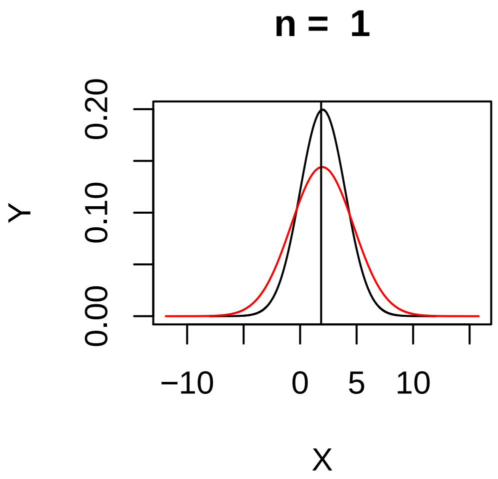
Suppose the "true" value is $\theta = 2$. Consider (1) $\mu = 5$ and $\tau = 1$, and (2) $\mu = 2$ and $\tau = 2$.

For $n \in \{1, 10, 20, 50, 100, 10000\}$:

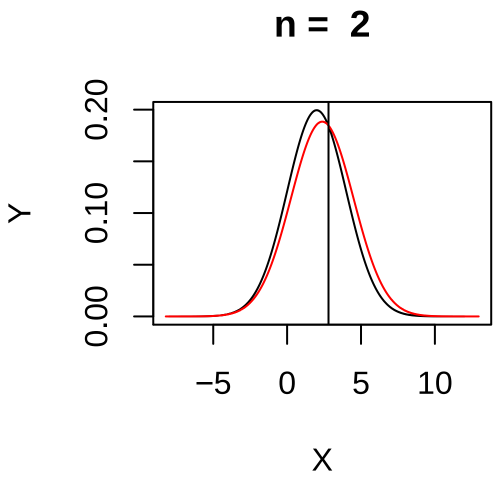
a. Simulate a data set consisting of n observations
b. Plot on the same graphic $\pi(\theta)$, $\pi(\theta|x_1, \ldots, x_n)$ and $\hat{\theta}_{\text{MLE}}$.

Discuss the behavior of $\pi(\theta|x_1, \ldots, x_n)$ as n increases and the impact of the prior distribution.

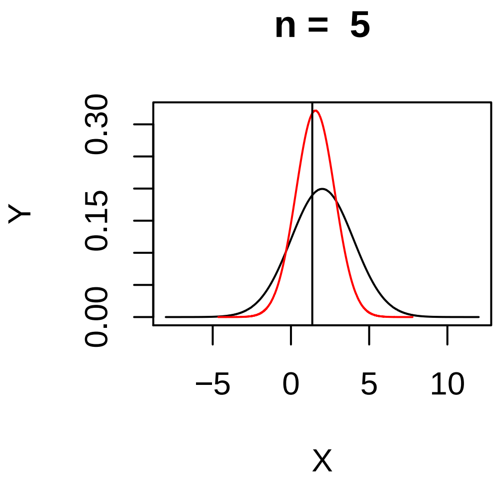
```
set.seed(1986)
theta <- 2
sigma_0   <- 3

alln <- c(1, 2, 5, 10, 100, 10000)
for(case in c(1, 2)){
  if(case == 1){
    prior_mu <- 2
    prior_tau <- 2
  }else if(case == 2){
    prior_mu <- 5
    prior_tau <- 1
  }

  for(n in alln){
    x <- rnorm(n, mean = theta, sd= sigma_0)
    x_bar <- mean(x)

    a <- (n * x_bar)/sigma_0^2 + prior_mu/prior_tau^2
    b <- n/sigma_0^2 + 1/prior_tau^2

    post_mu <- a/b
    print(post_mu)
    post_sigma <- 1/(n/sigma_0^2 + 1/prior_tau^2)

    xx <- seq(-5, 5, by = 0.001)
    xx_prior <- xx * prior_tau + prior_mu
    xx_post <-  xx * post_sigma + post_mu

    Y <- cbind(dnorm(xx_prior, mean = prior_mu, sd= prior_tau), dnorm(xx_post, mean = post_mu, sd = pos
    X <- cbind(xx_prior, xx_post)
    matplot(X, Y, type = 'l', lty = 1, main = paste("n = ", n))
    abline(v = x_bar, lty = 1)
  }
}
# [1] 1.957306
```



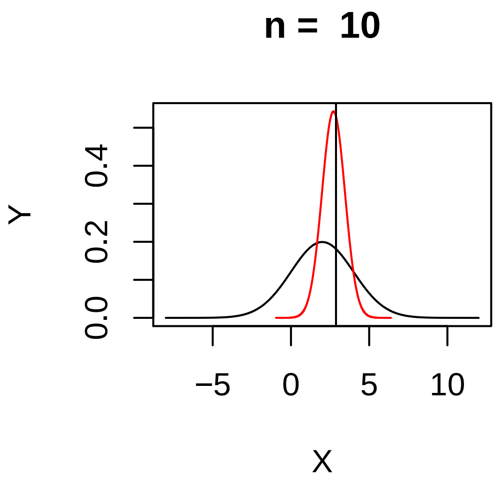
```
# [1] 2.376356
```



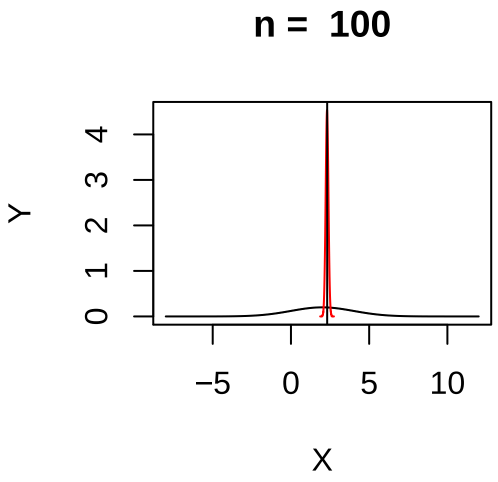
```
# [1] 1.561813
```



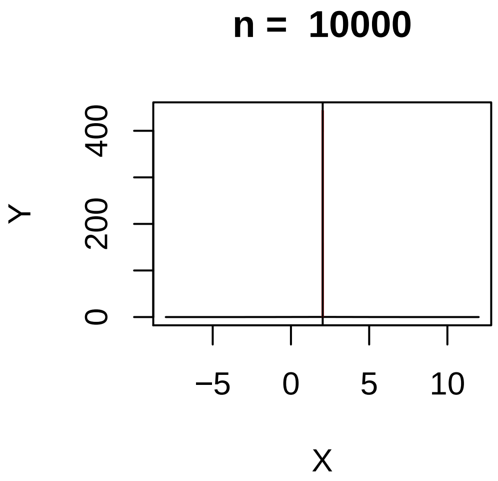
```
# [1] 2.718445
```



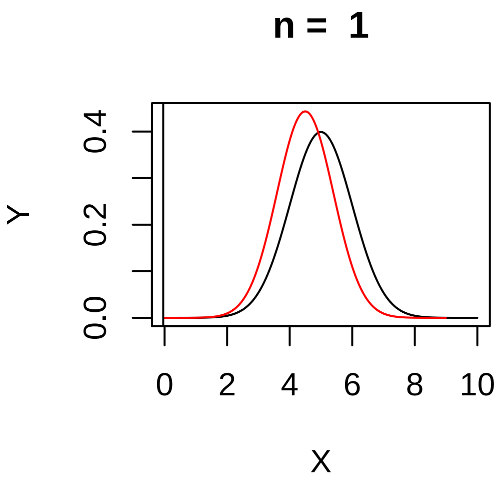
```
# [1] 2.307785
```



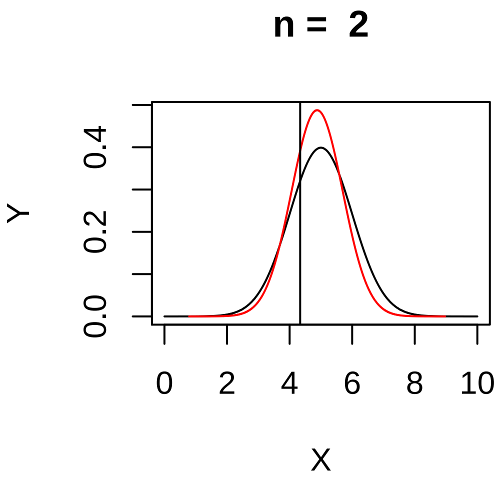
```
# [1] 2.028544
```



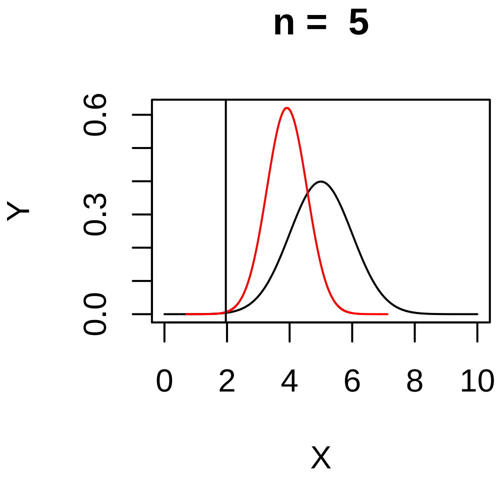
```
# [1] 4.495602
```



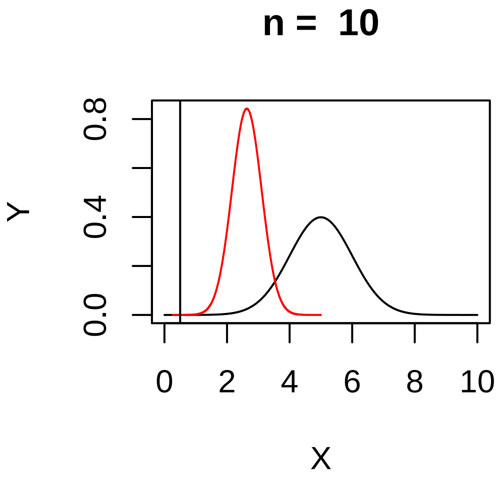
```
# [1] 4.879569
```



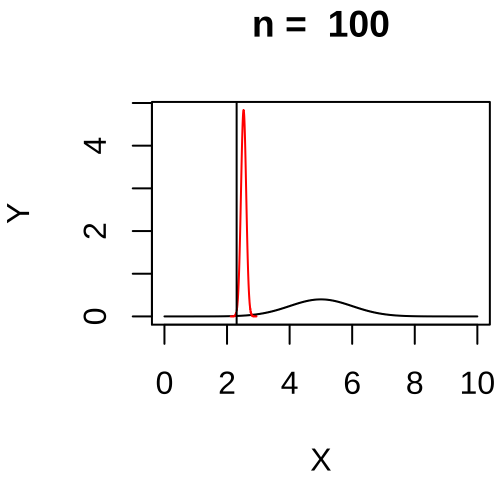
```
# [1] 3.914996
```



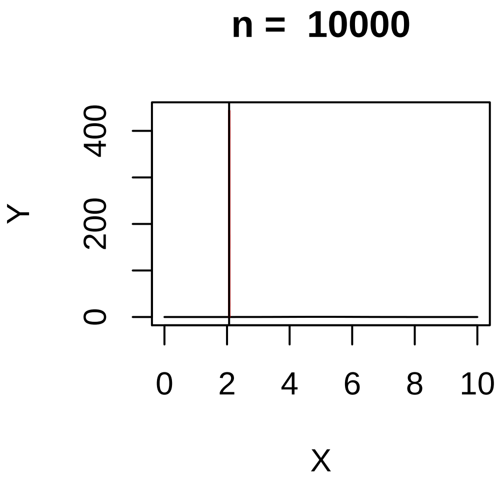
```
# [1] 2.632301
```



```
# [1] 2.529208
```



```
# [1] 2.069093
```



Question 3

Suppose there is a Beta(4, 4) prior distribution on the the probability θ that a coin will yield a “head” when spun in a specified maner. The coin is independently spun ten times, and “heads” appear fewer than 3 times. You are not told how many heads were seen, only that the number is less than 3. Calculate your exact posterior density (up to a proportionality constant) for θ and plot it.

Prior density:

$\pi(\theta) \propto \theta^3(1-\theta)^3$

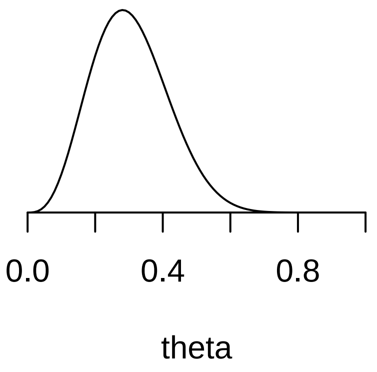
Likelihood:

$$\begin{aligned} f(\text{data}|\theta) &= \binom{10}{0}\theta^0(1-\theta)^10 + \binom{10}{1}\theta^1(1-\theta)^9 + \binom{10}{2}\theta^2(1-\theta)^8 \\ &= (1-\theta)^{10} + 10\theta(1-\theta)^9 + 45\theta^2(1-\theta)^8 \end{aligned}$$

Posterior density:

$\pi(\theta|\text{data}) \propto \theta^3(1-\theta)^{13} + 10\theta^4(1-\theta)^{12} + 45\theta^5(1-\theta)^{11}$

```
theta <- seq(0, 1, .01)
dens <- theta^3 * (1-theta)^13 + 10 * theta^4 * (1-theta)^12 + 45 * theta^5 * (1-theta)^11
plot (theta, dens, ylim=c(0,1.1*max(dens)), type="l", xlab="theta", ylab="", xaxs="i",yaxs="i", yaxt="n
```



Question 4

Suppose your prior distribution for θ, the proportion of Californians who support the deat penalty, is beta with mean 0.6 and standard deviation 0.3.

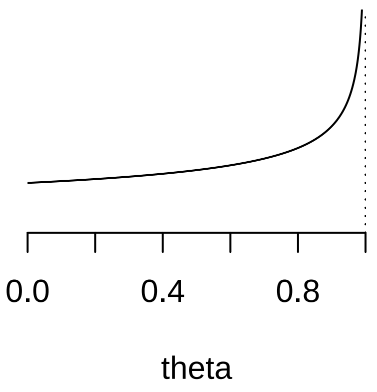
a. Determine the parameters α and β of your prior distribution. Plot the prior density function.
b. A random sample of 1000 Californians is taken, and 65% support the death penalty. What are your posterior mean and variance for θ? Plot the posterior density function.

$\alpha + \beta = \frac{E[\theta](1-E[\theta])}{var(\theta)} - 1 = 1.67$

$\alpha = (\alpha + \beta)E[\theta] = 1$

$\beta = (\alpha + \beta)(1 - E[\theta]) = 0.67$

```
theta <- seq(0,1,.001)
dens <- dbeta(theta,1,.67)
plot (theta, dens, xlim=c(0,1), ylim=c(0,3),
      type="l", xlab="theta", ylab="", xaxs="i",
      yaxs="i", yaxt="n", bty="n", cex=2)
lines (c(1,1),c(0,3),col=0)
lines (c(1,1),c(0,3),lty=3)
```



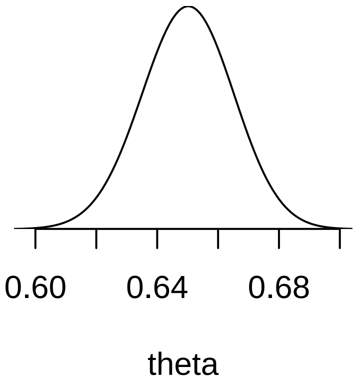
Posterior distribution:

$\pi(\theta|\text{data}) = \text{Beta}(\alpha + 650, \beta + 350) = \text{Beta}(651, 350.67)$

$E(\theta|data) = 0.6499$

$sd(\theta|data) = 0.015$

```r
theta <- seq(0,1,.001)
dens <- dbeta(theta,651,350.67)
cond <- dens/max(dens) > 0.001
plot (theta[cond], dens[cond],
      type="l", xlab="theta", ylab="", xaxs="i",
      yaxs="i", yaxt="n", bty="n", cex=2)
```



Question 5

10 Prussian cavalry corp were monitored for 20 years (200 Corp-Years) and the number of fatalities due to horse kicks were recorded:

x = # Deaths	Number of Corp-Years with x Fatalities
0	109
1	65
2	22
3	3
4	1

Let $x_i, i = 1, \ldots, 200$, be the number of deaths in observation i. Assume that $x_i \overset{i.i.d}{\sim} \text{Poisson}(\theta)$.

a. Compute the MLE estimate $\hat{\theta}_{\text{MLE}}$?

$\hat{\theta}_{\text{MLE}} = \bar{x} = \frac{122}{200} = 0.61$

Suppose $\theta \sim \text{Gamma}(\alpha, \beta)$.

a. What is the prior mean and variance.

$E[\theta] = \frac{\alpha}{\beta}$

$Var[\theta] = \frac{\alpha}{\beta^2}$

b. What is the posterior distribution $\pi(\theta|x)$?

Gamma$(\alpha + n * \bar{x}, \beta + n)$

c. What is the posterior mean and variance.

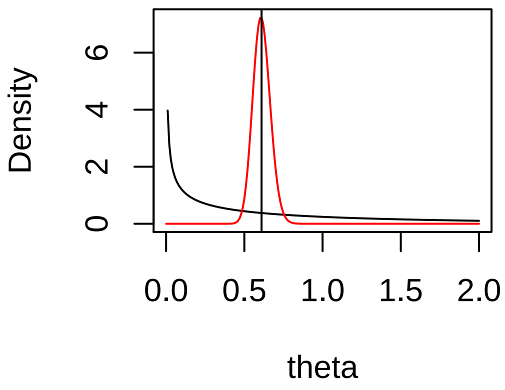
$E[\theta|x] = \frac{\alpha+n*\bar{x}}{\beta+n}$

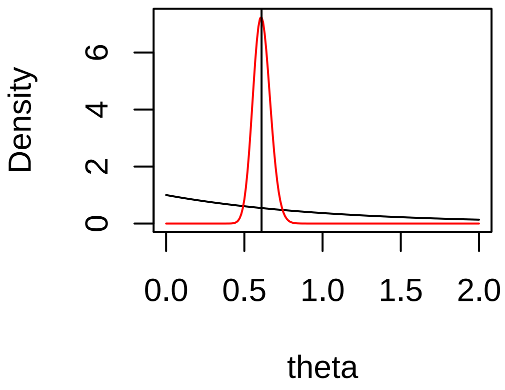
$Var[\theta|x] = \frac{\alpha+n*\bar{x}}{(\beta+n)^2}$

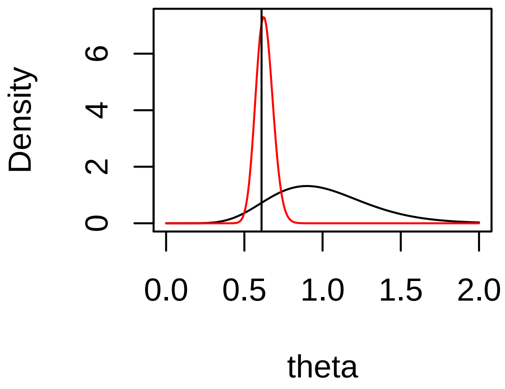
Plot on the same graphic $\pi(\theta)$, $\pi(\theta|x)$ and $\hat{\theta}_{\text{MLE}}$ for

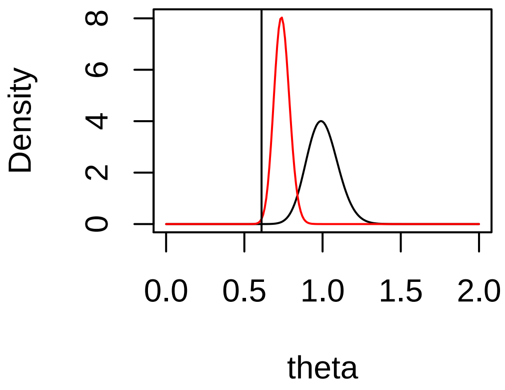
a. $\alpha = \beta = 0.5$
b. $\alpha = \beta = 1$
c. $\alpha = \beta = 10$
d. $\alpha = \beta = 100$

```r
n <- 200
DT <- data.frame(c(0, 1, 2, 3, 4), c(109, 65, 22, 3, 1))
xbar <- sum(DT[, 1] * DT[, 2])/n

x <- seq(0, 2, by = 0.01)

for(case in c(1, 2, 3, 4)){
  if(case == 1){
    alpha <- beta <- 0.5
  }else if(case == 2){
    alpha <- beta <- 1
  }else if(case == 3){
    alpha <- beta <- 10
  }else if(case == 4){
    alpha <- beta <- 100
  }

  dens <- dgamma(x, shape = alpha, rate = beta)

  alpha_posterior <- alpha + n * xbar
  beta_posterior <- beta + n
  dens_posterior <- dgamma(x, shape = alpha_posterior, rate = beta_posterior)

  matplot(x, cbind(dens, dens_posterior), lty = 1, type = 'l', ylab = "Density", xlab = "theta")
  abline(v = xbar)

}
```


Density
0
2
4
6
0.0
0.5
1.0
1.5
2.0
theta


Density
0
2
4
6
0.0
0.5
1.0
1.5
2.0
theta


Density
0
2
4
6
0.0
0.5
1.0
1.5
2.0
theta


Density
0
2
4
6
8
0.0
0.5
1.0
1.5
2.0
theta

TURN IN

- Your `.Rmd` file
- Your Word (or pdf) file that results from knitting the Rmd.
- Make sure your group members are listed as authors, one person per group will turn in the report
- DUE: Wednesday after the lab, by 7am, loaded into moodle

Resources

- Lecture slides on Bayesian reasoning