# HW1

## 1 Exponential model

We write $Y \sim Exp(\theta)$ to indicate that Y has the Exponential distribution, that is, its p.d.f. is $p(y|\theta) = Exp(y|\theta) = \theta e^{-\theta y} 1_{\{y>0\}}$
The Exponential distribution has some special properties that make it a good model for certain applications. It has been used to model the time between events (such as neuron spikes, website hits, neutrinos captured in a detector), extreme values such as maximum daily rainfall over a period of one year, or the amount of time until a product fails (lightbulbs are a standard example).
Suppose you have data $y_1, ..., y_n$ which you are modeling as iid observations from an Exponential distribution, and suppose that your prior is $\theta \sim Gamma(a, b)$, defined as in the previous question.
(a). Derive the formula for the posterior density, $p(\theta|y_{1:n})$. Give the form of the posterior in terms of one of the following distributions: Bernoulli, Beta, Exponential, and Gamma

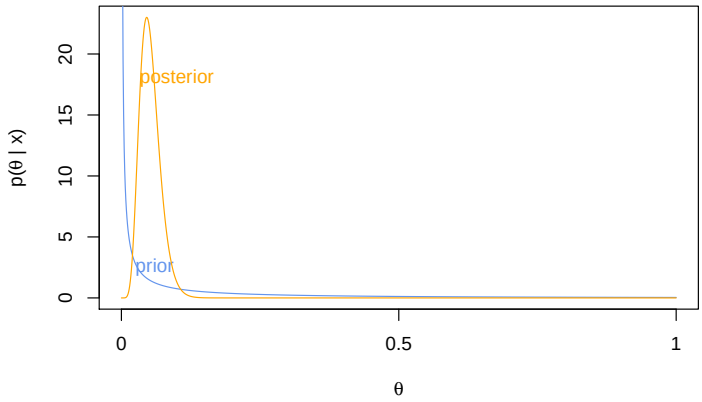
$$\Pr(Y_{1:n} = y_{1:n}|\theta) = p(y_{1:n}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \theta e^{-\theta y_i} 1_{\{y>0\}} = \theta^n e^{-\theta \sum y_i} \left( \prod_{i=1}^{n} 1_{\{y>0\}} \right)$$

$$p(\theta) = \text{Gamma}(\theta|a, b) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} 1_{\{\theta>0\}}$$

$$p(\theta|y_{1:n}) \propto p(y_{1:n}|\theta)p(\theta) = \left( \theta^n e^{-\theta \sum y_i} \right) \left( \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} 1_{\{\theta>0\}} \right) \propto \theta^{a+n-1} e^{-(b+\sum y_i)\theta} 1_{\{\theta>0\}} \propto \text{Gamma}(a+n, b+\sum y_i)$$

(b). Now, suppose you are measuring the number of seconds between lightning strikes during a storm, your prior is Gamma(0.1,1.0), and your data is
$(y_1, ..., y_8) = (20.9, 69.7, 3.6, 21.8, 21.4, 0.4, 6.7, 10.0)$
Plot the prior and posterior pdf's. (Be sure to make your plots on a scale that allows you to clearly see the important features.)

```
y <- cumsum(c(20.9,69.7,3.6,21.8,21.4,0.4,6.7,10.0))
N <- 1:8
k <- 8
ps <- seq(0,1,length.out = 1000)
prior <- dgamma(ps,shape=0.1,rate=1)
post <- dgamma(ps,shape=(0.1+N[k]),rate=(1+y[k]))
```




(c). Give a specific example of an application where an Exponential model would be reasonable. Give an example where an Exponential model would NOT be appropriate, and explain why.
\
In stable traffic, the time between the vehicles entering the highway has an exponential distribution.
The distribution of urban size (human population) within a region have the "rank-size" property called Zipf 's law. Reed et al. (2004) found "the time since foundation follows an exponential distribution, and that at foundation sizes are lognormally distributed, then the current sizes should follow the Double Pareto-Lognormal Distribution".

The p.d.f. of Pareto distribution has a scale parameter of $\alpha$ and its support is dependent on the value of $\alpha$ (lower limit). Thus, the Exponential model is not appropriate for the distribution of urban size.

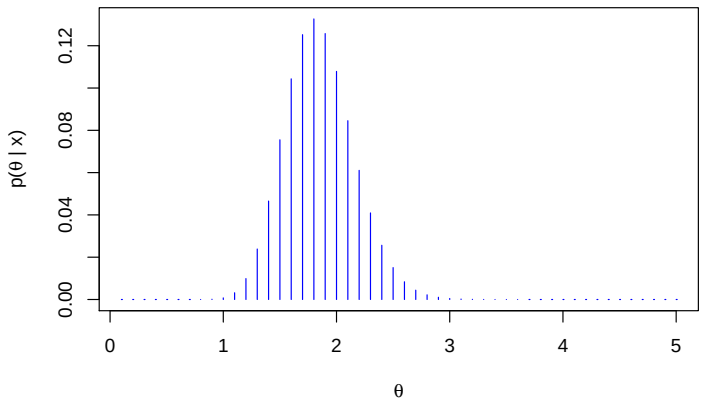
## 2 A Poisson Case by Bayes rule

An ecologist records the number of eggs laid in a sample of sparrow nests of size $n = 20$. Let $Y_i$ be the number of eggs laid in nest i for $i = 1, ..., 20$. Based on this sample, the ecologist is interested in estimating $\theta$, the mean number of eggs per nest in the general population of nests. Assume $Y_1, ..., Y_n|\theta \sim Poisson(\theta)$ iid, and also for now that $\theta \in \Theta = \{0.1, 0.2, ..., 4.9, 5.0\}$ and that $p(\theta) = 1/50$ for each $\theta \in \Theta$.
(a). Let $X = \sum_i Y_i$ denote the random variable that counts the total number of eggs laid in the 20 nests. Using the form of Bayes rule on page 15 of the Hoff book, write down a formula for $p(\theta|x)$ and simplify as much as possible.
$X = \sum_i Y_i \sim \text{Pois}(n\theta)$, $p(x|\theta) = \frac{(20\theta)^x e^{-20\theta}}{x!}\mathbf{1}_{\{\theta\in\Theta\}}$,

$$p(\theta \mid x) = \frac{p(x|\theta_j)p(\theta_j)}{\sum_{k=1}^{K} p(x|\theta_k)p(\theta_k)} = \frac{\frac{(20\theta_j)^x e^{-20\theta_j}}{x!}\mathbf{1}_{\{\theta_j\in\Theta\}}\frac{1}{50}}{\sum_{k=1}^{50}\frac{(20\theta_k)^x e^{-20\theta_k}}{x!}\mathbf{1}_{\{\theta_k\in\Theta\}}\frac{1}{50}} = \frac{\theta_j^x e^{-20\theta_j}\mathbf{1}_{\{\theta_j\in\Theta\}}}{\sum_{k=1}^{50}\theta_k^x e^{-20\theta_k}\mathbf{1}_{\{\theta_k\in\Theta\}}}$$

(b). The ecologist observes that $x = 36$. Make a plot of $p(\theta|x)$ versus $\theta$ for$\theta \in \Theta$

```
x <- 36; n <- 20
Theta <- seq(0.1,5,by = 0.1)
post.theta <- exp(-n*Theta)*(Theta^x)
post.theta <- post.theta/sum(post.theta)
```




(c). Find $E[\theta|x]$, the posterior mean of $\theta$.

```
sum(Theta*post.theta)
```

```
## [1] 1.85
```

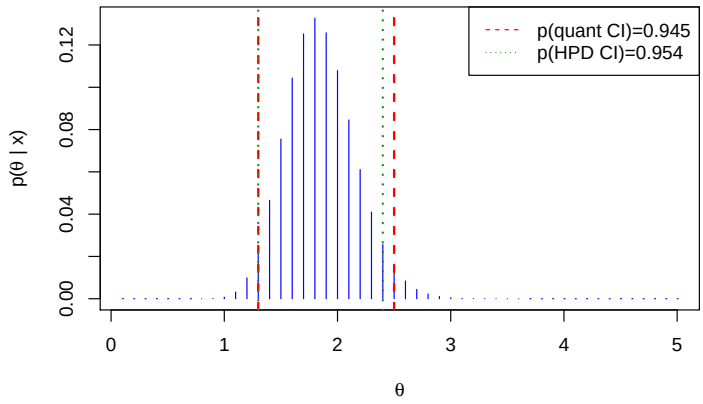
- A method may not be right, but give a same value.

$p(\theta|x) \propto p(x|\theta)p(\theta) = \frac{(20\theta)^x e^{-20\theta}}{x!}\mathbf{1}_{\{\theta\in\Theta\}}\frac{1}{50} \propto \theta^x e^{-20\theta}\mathbf{1}_{\{\theta\in\Theta\}} \propto \text{Gamma}(x+1, 20)$
$E[\theta|X = 36] = \frac{36+1}{20} = 1.85$
(d). Find two numbers, $\theta_l$ and $\theta_h$ such that $Pr(\theta_l \le \theta \le \theta_h|x) \approx 0.95$. This is an approximate 95% posterior confidence interval. Note that there is more than one way of doing this.

```
quant_ci<- c(q2.5=min(Theta[which(cumsum(post.theta)>=0.025)]),
             q97.5=min(Theta[which(cumsum(post.theta)>=0.975)]))
## Get a closer CI by moving the upper bound with one quantile
check_quant <- diff(cumsum(post.theta)[which(Theta%in%quant_ci)])
quant_ci
##  q2.5 q97.5
##   1.3   2.5
check_quant
## [1] 0.9452372
```

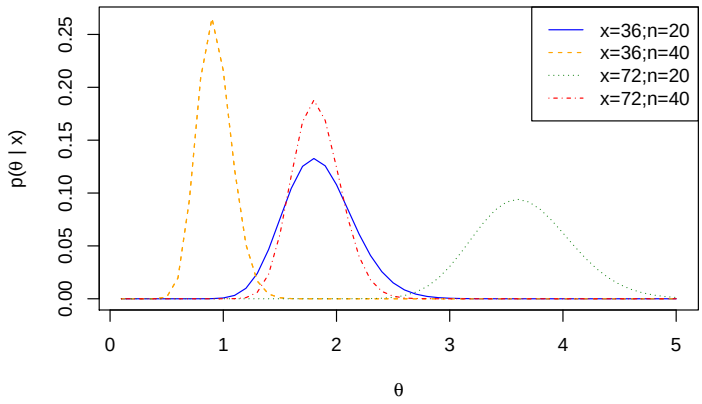
```r
cumprob <- 0
pmf <- sort(unique(post.theta),decreasing = T) #order pmf
hpd_ci <- range(Theta)
th.vals <- NULL
k <- 1
while(round(cumprob,2)<0.95){
  th.vals <- Theta[post.theta>=pmf[k]]
  hpd_ci <- range(th.vals)
  cumprob <- sum(post.theta[post.theta>=pmf[k]])
  k <- k+1
}
names(hpd_ci) <- c("l(x)","h(x)")

pos.lims.hpd <- which(Theta%in%hpd_ci)
#move position one to the left of lower limit since:
# P(theta \in [l(y),h(y)]) = P(theta <= u(y))- P(theta < l(y))
pos.lims.hpd[1] <- pos.lims.hpd[1]-1
#get cumulative probs for each limit
cumprobs.lims.hpd <- cumsum(post.theta)[pos.lims.hpd]
#the difference below gives the actual Bayesian coverage of the HPD
check_hpd <- diff(cumprobs.lims.hpd)
hpd_ci
## l(x) h(x)
##  1.3  2.4
check_hpd
## [1] 0.9540186
```




The quantile CI bound (red) from $l(x) = 1.3$ to $h(x) = 2.5$ is wider than HPD CI (green) from $l(x) = 1.3$ to $h(x) = 2.4$, I feel confused that the p(HPD CI)=0.954>p(quant CI)=0.945.
(e). Remake the plot in part (b) but with $n = 40$ and $x = 72$. Describe and explain the differences you see between this plot and the one in (b).

```r
x <- c(36,36,72,72)
n <- c(20,40,20,40)
#define parameter space Theta (all values that theta can take)
Theta <- seq(0.1,5,by = 0.1)
#calculate the numerator of p(theta | x)
post.theta <- matrix(NA,nrow=50,ncol=4)
for(i in 1:4){
post.theta[,i]=exp(-(n[i])*Theta)*(Theta^(x[i]))
#renormalize (i.e., divide by the sum to get them to add up to 1)
post.theta[,i] <- post.theta[,i]/sum(post.theta[,i])
}
```




The result shows that the ratio of observed value over sample size will decide the shape and location of the probability distribution of $\theta$.A higher ratio will give a sharper and left-shift probability distribution, while a lower ratio will give a blunt and right-shift distribution.

### 3 posterior predictive density

Suppose the data $y_{1:n}|\theta$ is modeled as iid $Exp(\theta)$, and the prior is $p(\theta) = Gamma(\theta|a,b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}1_{\{\theta>0\}}$
From problem 1, we know that the posterior is $p(\theta|y_{1:n}) = Gamma(\theta|a_n,b_n)$, where $a_n = a+n$ and $b_n = b+\sum_{i=1}^n y_i$. What is the posterior predictive density $p(y_{n+1}|y_{1:n})$? Give your answer as a closed-form expression (not an integral).
$p(y_{n+1}|y) = p(y_{n+1}|y_{1:n}) = \int_{\theta\in\Theta} p(y_{n+1},\theta|y_{1:n})d\theta = \int_{\theta\in\Theta} p(y_{n+1}|\theta,y_{1:n})p(\theta|y_{1:n})d\theta = \int_{\theta\in\Theta} p(y_{n+1}|\theta)p(\theta|y_{1:n})d\theta$

$$= \int_{\theta\in\Theta} \theta e^{-\theta y_{n+1}}\mathbf{1}_{\{y>0\}}\frac{(b+\sum y_i)^{a+n}}{\Gamma(a+n)}\theta^{a+n-1}e^{-(b+\sum y_i)\theta}\mathbf{1}_{\{\theta>0\}}d\theta$$

$$= \frac{(a+n)(b+\sum y_i)^{a+n}}{(b+\sum y_i+y_{n+1})^{a+n+1}}\mathbf{1}_{\{y>0\}}\int_{\theta\in\Theta}\frac{(b+\sum y_i+y_{n+1})^{a+n+1}}{\Gamma(a+n+1)}\theta^{a+n}e^{-(b+\sum y_i+y_{n+1})\theta}\mathbf{1}_{\{\theta>0\}}d\theta$$

$$= \frac{(a+n)(b+\sum_{i=1}^n y_i)^{a+n}}{(b+\sum_{i=1}^{n+1} y_i)^{a+n+1}}\mathbf{1}_{\{y>0\}}$$

### 4 Loss function, posterior expected loss, Bayesian decision procedure

Suppose that in the small imaginary city of our class example, where the prevalence of a rare disease was studied (Foundations: Section 4, page 11), public health officials need to decide the amount of resources to allocate towards prevention and treatment of the disease we are concerned with, with the fraction of infected individuals $\theta$ still unknown. They will decide on the resources needed based on a fraction $c$ of the population. If $c$ is chosen too large, there will be wasted resources, while if it is too small, preventable cases may occur and some individuals may go untreated. After some deliberation, they tentatively adopt the following loss function:
$\ell(\theta,c) = (\theta-c+0.5)^2$ for $c\in[0,1]$
(a). Assume that the number of people sampled is again $n=20$, that $y=0$, and use $\theta\sim Beta(2,20)$ as the prior again to derive the posterior expected loss.

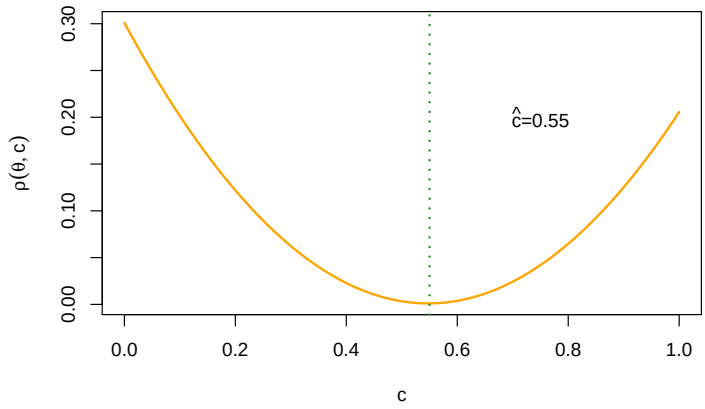
$$p(\theta|y)\propto p(y|\theta)p(\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}\frac{1}{B(2,20)}\theta^{2-1}(1-\theta)^{20-1}\propto\theta^{2+0-1}(1-\theta)^{20+(20-0)-1}\propto \text{Beta}(2,40)$$

$$\rho(c,y) = E(\ell(\boldsymbol{\theta},c)|y) = \int_0^1\ell(\theta,c)\frac{1}{B(a_n,b_n)}\theta^{a_n-1}(1-\theta)^{b_n-1}d\theta = \int_0^1(\theta-c+0.5)^2\frac{1}{B(2,40)}\theta^{2-1}(1-\theta)^{40-1}d\theta$$

(b). Find an optimal value (call it $\hat{c}$) for $c$ following a Bayesian decision procedure using the set up from part (a).

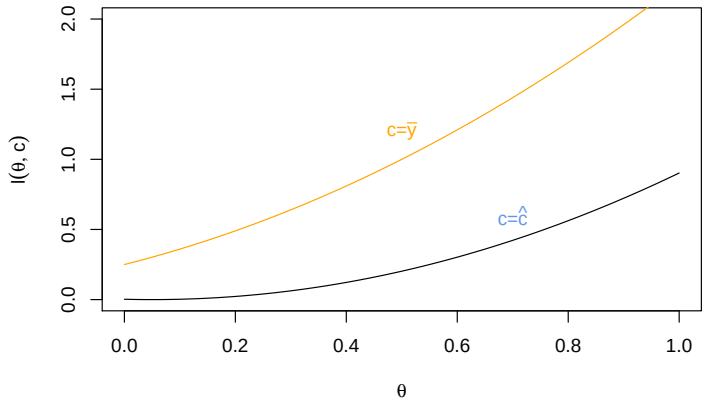
```
n <- 20;a <- 2;b <- 20;y <- 0
rhofn <- function(c){
  ff <- function(theta){
      db <- dbeta(theta,shape1=(a+y),shape2=(b+n-y))
      (theta-c+0.5)^2 * db
    }
    res <- integrate(ff, lower=0, upper=1)
```

```
    res$value
}
vrhofn <- Vectorize(rhofn, vectorize.args = "c")
c.values <- seq(0,1,by=0.005)
rho.values <- vrhofn(c.values)
c.argmin <- c.values[which(rho.values==min(rho.values))]
```




The Bayesian decision procedure gives that $\hat{c} = 0.55$ can minimize the poseriorexpected loss.
Check with $\ell(\theta, c) = \theta^2 - 2\hat{\theta}\theta + \hat{\theta}^2$; $\frac{d}{d\hat{\theta}}\rho(\hat{\theta}, y_{1:n}) = -2E(\boldsymbol{\theta}|y_{1:n}) + 2\hat{\theta} = 0$; $\hat{\theta} = E(\boldsymbol{\theta}|y_{1:n}) = \frac{2}{2+40} = \frac{2}{42}$
Let $\hat{\theta} = c - 0.5$, $\hat{\theta} = \delta(y_{1:n}) = E(\boldsymbol{\theta}|y_{1:n}) = \frac{2}{42} = c - 0.5 \implies \hat{c} = \frac{23}{42} = 0.547619$, which is very close to the result by Bayesian decision procedure.
And, $\rho(\hat{\theta}, y_{1:n}) = E(\ell(\boldsymbol{\theta}, \hat{\theta})|y_{1:n}) = E[(\boldsymbol{\theta} - \hat{\theta})^2|y_{1:n}] = E[(\boldsymbol{\theta} - \mu)^2|y_{1:n}] = Var(\boldsymbol{\theta}|y_{1:n}) = \frac{2*40}{42^2(42+1)} = 0.001054685$
(c). Graphically compare $l(\theta, \hat{c})$ and $l(\theta, \bar{y})$ as $\theta$ ranges from 0 to 1.

```
n <- 20; a <- 2; b <- 20; y <- 0
lossfn <- function(theta,c){
  loss <- (theta-c+0.5)^2
}
theta.values <- seq(0,1,by=0.005)
loss.values_hat_c <- lossfn(theta.values,0.55)
loss.values_bar_y <- lossfn(theta.values,0)
```

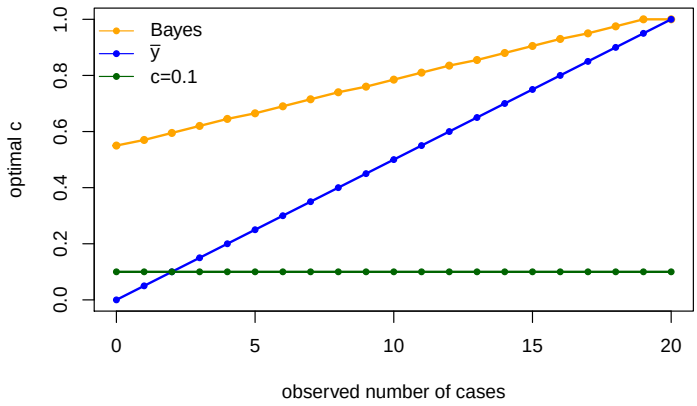



(d). Compare the outcome from the Bayesian decision procedure to the procedures that choose (i) $c = \bar{y}$ and (ii) $c = 0.1$ constant (i.e., setting it to the prior mean). This comparison can be done by observing the optimal quantity $c$ derived from each of the decision procedures considered while varying the value of $y$ (the number of infected cases).

```r
n <- 20;a <- 2;b <- 20
yvec <- 0:20 ; c.fixed <- 0.1; c.argminvec <- NA

for(y in yvec){
  rhofn <- function(c){

    ff <- function(theta){
      db <- dbeta(theta,shape1=(a+y),shape2=(b+n-y))
      (theta-c+0.5)^2 * db
    }
    res <- integrate(ff, lower=0, upper=1)
    res$value
  }
  vrhofn <- Vectorize(rhofn, vectorize.args = "c")
  c.values <- seq(0,1,by=0.005)
  rho.values <- vrhofn(c.values)
  #value of c that minimizes the posterior expected loss
  (c.argminvec[y+1] <- c.values[which(rho.values==min(rho.values))])
}
```




The result shows that, if set $c = 0.1$ constant, the number of infected cases will not change the optimal c; if set $c = \bar{y}$, the more infected cases, the optimal c is closer to Bayesian decision procedure. This may suggest that the set of loss function is more decisive when there is a few infected cases.