Homework #01

Statistical Methods in Data Science II & Lab

Michelangelo Saveriano

April 22th, 2022

Michelangelo Saveriano 1823326

*Your answers for each data analysis question should discuss the problem, data, model, method, conclusions.

Fully Bayesian conjugate analysis of Rome car accidents

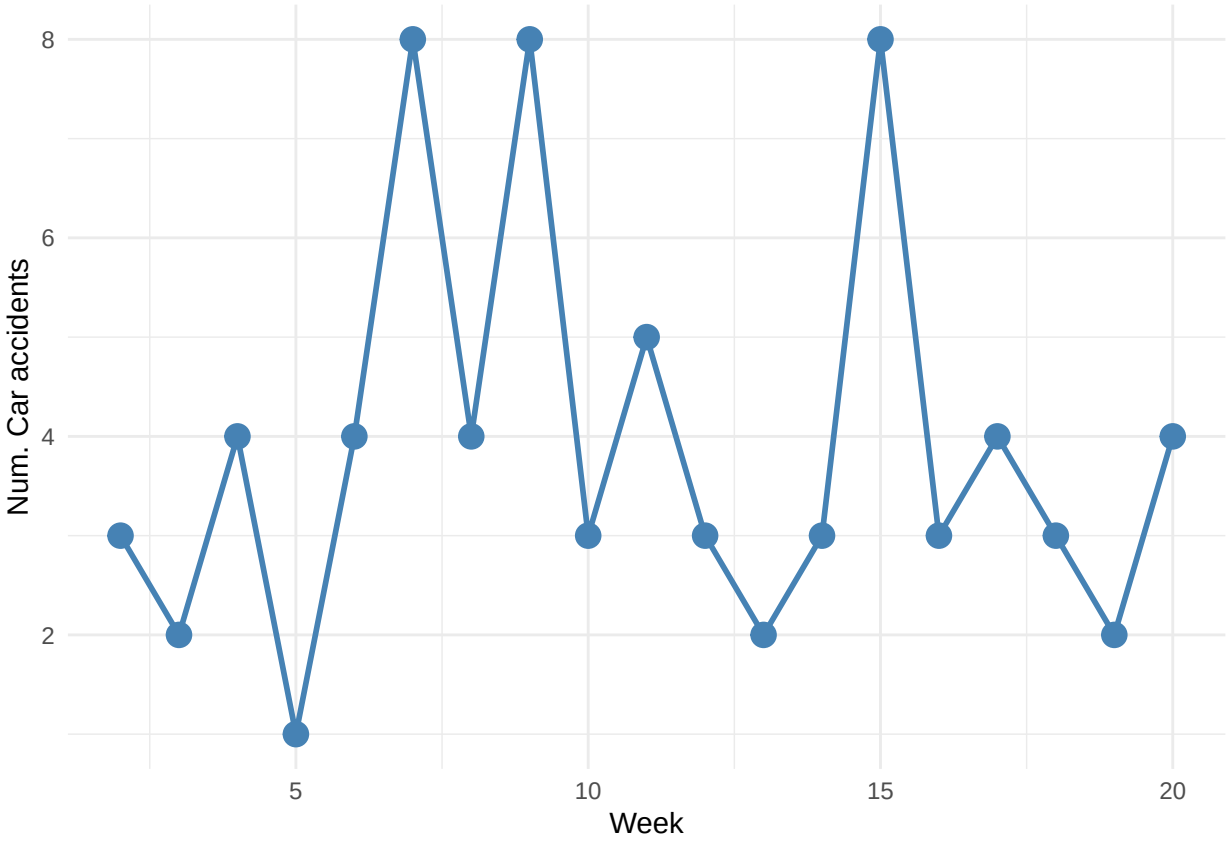
Consider the car accident in Rome (year 2016) contained in the `data.frame` named `roma`. Select your data using the following code

```
mydata <- subset(roma,subset=sign_up_number==104)
str(mydata)
```

```
## 'data.frame':    19 obs. of  5 variables:
##  $ week          : int  2 3 4 5 6 7 8 9 10 11 ...
##  $ weekday       : chr  "Saturday" "Saturday" "Saturday" "Saturday" ...
##  $ hour          : int  9 9 9 9 9 9 9 9 9 9 ...
##  $ car_accidents : int  3 2 4 1 4 8 4 8 3 5 ...
##  $ sign_up_number: int  104 104 104 104 104 104 104 104 104 104 ...
```

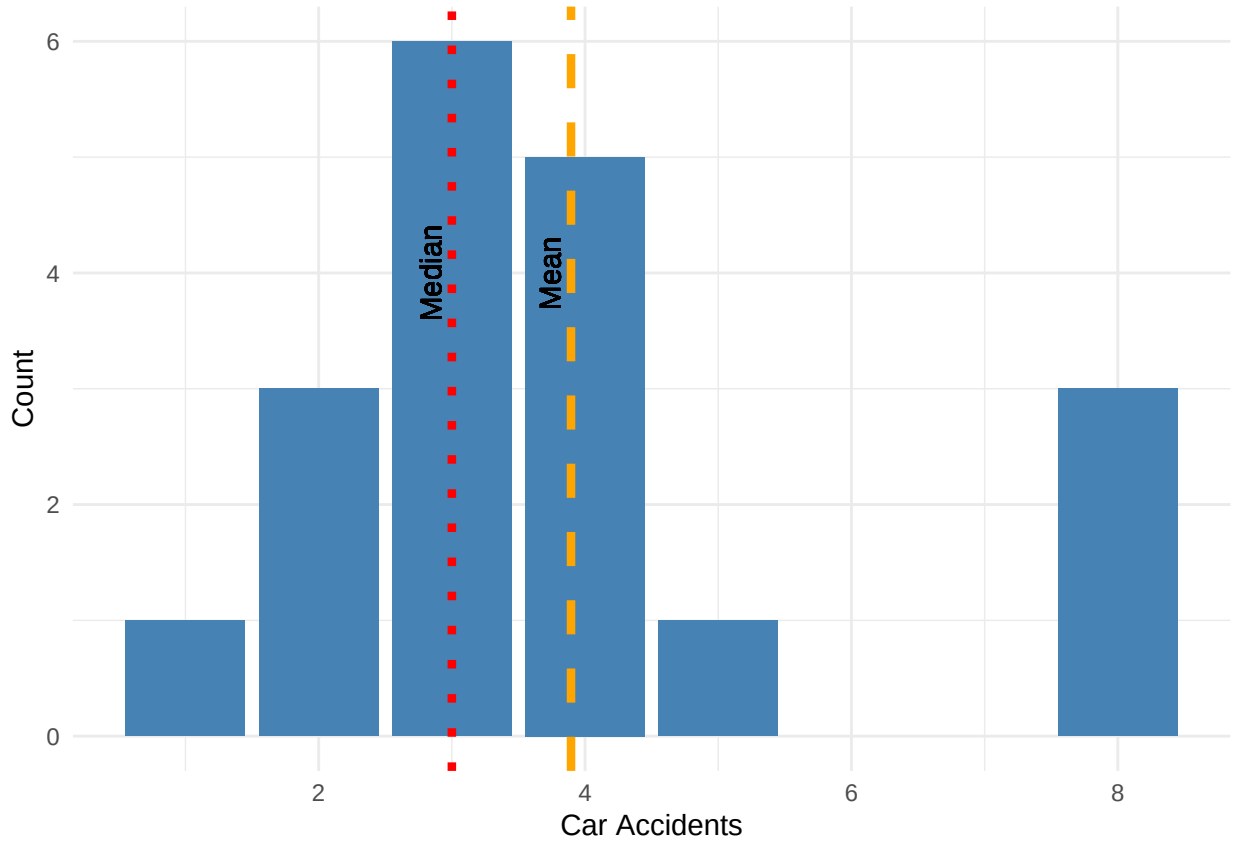
The column `car_accidents` contains the number of car accidents $Y_i = y_i$ occurred in a specific weekday during a specific hour of the day in some of the weeks of 2016. Using the observed outcomes of the number of car accidents do a fully Bayesian analysis using as a statistical model a conditionally i.i.d. Poisson distribution with unknown parameter. Take into account that it is known that the average number of hourly car accidents occurring in Rome during the day is **3.22**. In particular do the following:

Describe your observed data The dataset consists of 19 observations recording the number of car accidents occurred in a specific weekday during a specific hour of the day. If we assume that the number of accidents depends only on the weekday and the hour of the day we can also assume that the sample are identically distributed. To assess that we can look at the number of accidents over the weeks and we observe that the week number doesn't seem to have an effect on the number of accidents.



Below we can see a little summary containing the main statistics for the observed data as well as the distribution itself shown using a barplot of counts:

- mean number of car accidents: 3.89
- variance of the distribution: 4.21
- median number of car accidents: 3



Justify as best you can your choices for the ingredients of your Bayesian model especially for the choices you make for the prior distribution In this setting we can assume that the samples can be modeled as conditionally iid random variables drawn from a Poisson distribution with parameter θ.

We decided to model the unknown parameter θ as a Gamma distribution with parameter rate r and shape s

$$\theta \sim Gamma(r, s)$$

A crucial step in the creation of the Bayesian model is the choice of the prior distribution where we embed our knowledge and beliefs into the distribution's parameters, mainly:

- as we are told it is known that the average number of hourly car accidents occurring in Rome during the day is 3.22 therefore we can set $\mathbb{E}[\theta] = \mu = \frac{s}{r} = 3.22$
- we want also impose an additional constraint requiring that a certain amount of probability density falls within a given range, in particular we want that $\mathbb{P}(2 \leq \theta \leq 5) = 0.70$. To satisfy this constraint we let $\mathbb{V}ar[\theta] = \sigma^2 = \frac{s}{r^2}$ vary.

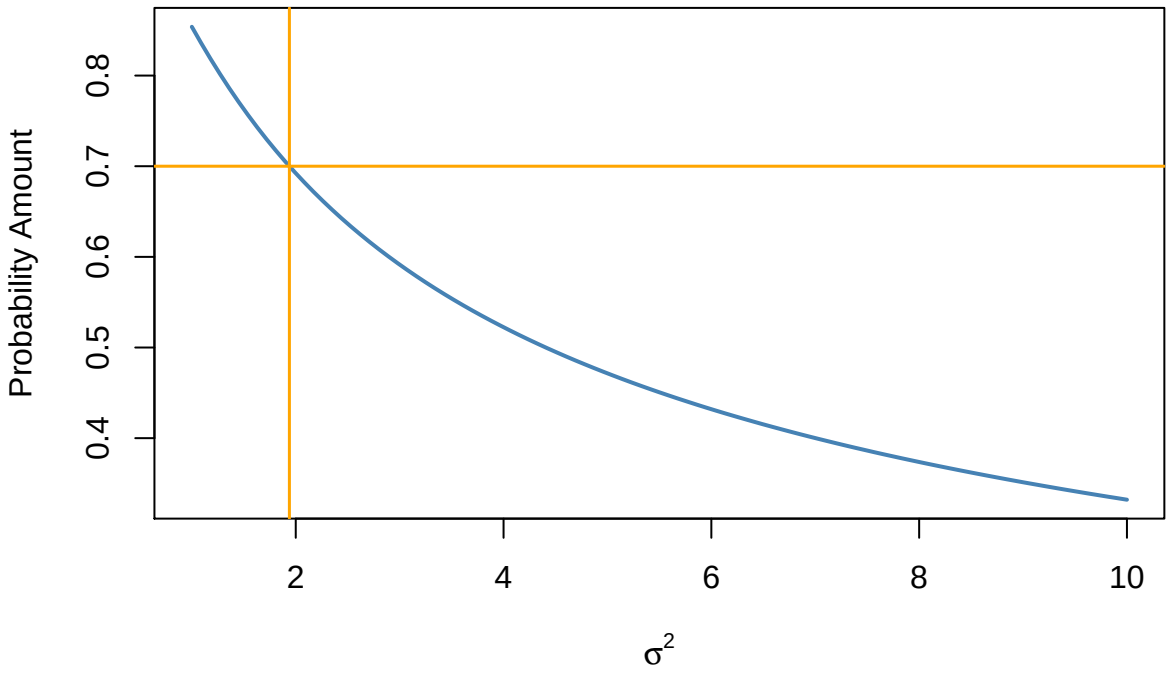
To impose this constraints we first have to define the map from the (μ, σ^2) parametrisation to the (s, r) parametrisation, which can be easily proven be this one:

$$r = \frac{\mu}{\sigma^2}, \; s = \frac{\mu^2}{\sigma^2}$$

Then we can run a simulation to find the value of σ^2 that satisfy such constraint.

```r
prob_amount <- function (mean_, var_, from=2, to=5){
  rate <- mean_ / var_
  shape <- mean_ ^ 2  / var_
  pgamma(to, rate = rate, shape = shape) - pgamma(from, rate = rate, shape = shape)
}

p_thresh <- 0.7
mean_theta_prior <- 3.22
var_theta_prior <- uniroot(function(x) prob_amount(mean_theta_prior, x) - p_thresh, interval = c(1, 10))
rate_prior <- mean_theta_prior / var_theta_prior
shape_prior <- mean_theta_prior ^ 2  / var_theta_prior
median_theta_prior <- qgamma(.5, shape_prior, rate_prior)
```



In the end we obtain that the prior distribution is characterised by:

- $\mu = 3.22$
- $\sigma^2 = 1.94$
- $r = 1.66$
- $s = 5.35$
- $median = 3.02$

Report your main inferential findings using your posterior distribution of the unknown parameter in terms of

3a) possible alternative point estimates with comments on how similar they are and, in case, why + 3b) posterior uncertainty First of all we have to update the parameter according to the following rules:

$$r^{\text{posterior}} = r^{\text{prior}} + ns^{\text{posterior}} = s^{\text{prior}} + \sum_{i=1}^{n} y_i$$

```
# parameter update
rate_posterior <- rate_prior + length(mydata$car_accidents)
shape_posterior <- shape_prior + sum(mydata$car_accidents)
mean_theta_posterior <- shape_posterior / rate_posterior
var_theta_posterior <- shape_posterior / rate_posterior ^ 2
median_theta_posterior <- qgamma(.5, shape_posterior, rate_posterior)
```

The updated values for rate and shape are $r^{\text{posterior}} = 20.66$ and $s^{\text{posterior}} = 79.35$.

Therefore our posterior distribution is:

$$\theta^{\text{posterior}} \sim Gamma(20.66, 79.35)$$

We can then compute the updated point estimates like mean and variance using the formulas we saw above:

- $\mu^{\text{posterior}} = 3.84$
- $\sigma^{2^{\text{posterior}}} = 0.19$

As well as the median via the quantile function:

- $median = 3.82$

From the point estimates we can see that mean and median are close to each other and the variance has shrunk drastically going from ~ 2 in the prior to almost 0.2 in the posterior.

3c) interval estimates justifying your (possibly best) choices The two main type of credible intervals are *equal-tailed interval* and *highest posterior density interval* (HPD). The first can be evaluated using the quantile function while to construct the latter we need to find the point with the higher posterior density.

```
# Confidence level
alpha <- 0.05

# HPD interval
hpd_ci <- hpd(qgamma, rate = rate_posterior, shape = shape_posterior, conf = 1-alpha)
print(paste0('HPD -> Lower: ', round(hpd_ci[1], 3), ' - Upper: ', round(hpd_ci[2], 3), ' - Width: ', rou
```

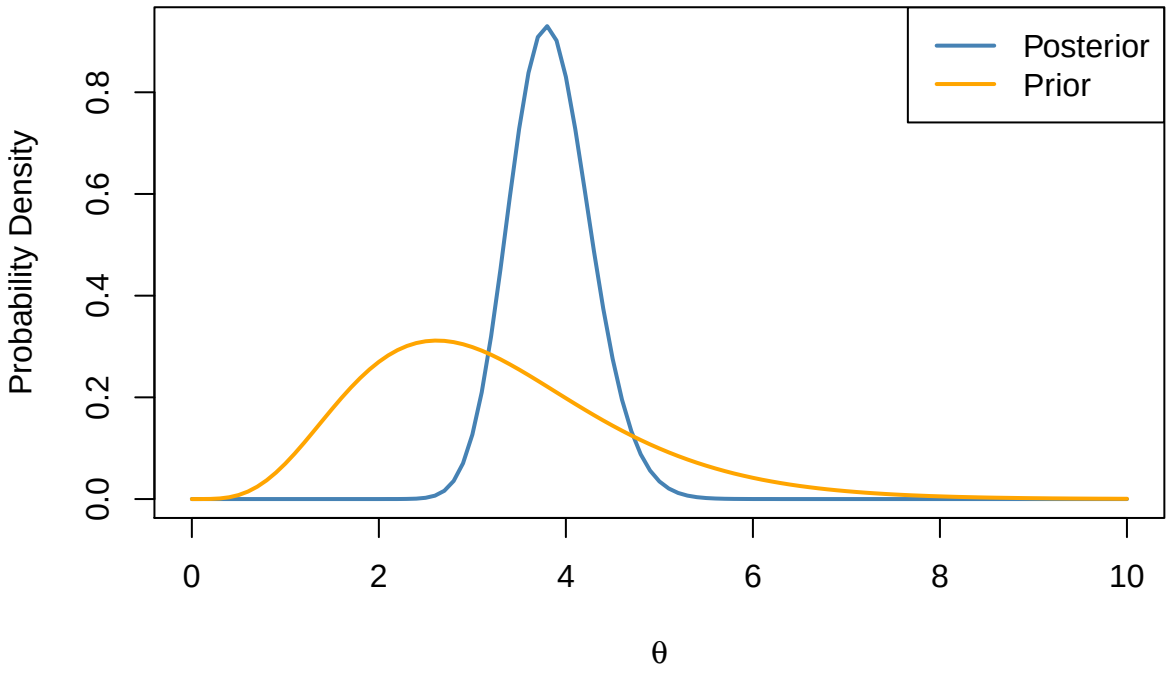
```
## [1] "HPD -> Lower: 3.012 - Upper: 4.696 - Width: 1.683"
```

```
#Equal-Tailed interval
eqt_ci <- qgamma(c(alpha/2, 1 - alpha/2), shape_posterior, rate_posterior)
print(paste0('Equal-Tailed -> Lower: ', round(eqt_ci[1], 3), ' - Upper: ', round(eqt_ci[2], 3), ' - Wid
```

```
## [1] "Equal-Tailed -> Lower: 3.042 - Upper: 4.73 - Width: 1.688"
```

The HPD intervals are often preferred because they convey narrower intervals than Equal-tailed. In our case though the intervals' width are almost equal.

3d) suitable comments on the differences between the prior and the posterior We already pointed out few differences between prior and posterior, mainly regarding mean, median and variance. Now we take a look at the difference in the prior and posterior gamma distributions visually by plotting their densities.



As already shown previously we notice that the posterior is shifted to the right and heavily concentrated. The posterior is also less skewed appearing almost symmetrical, this aspect also explains the closeness of the updated mean and median.

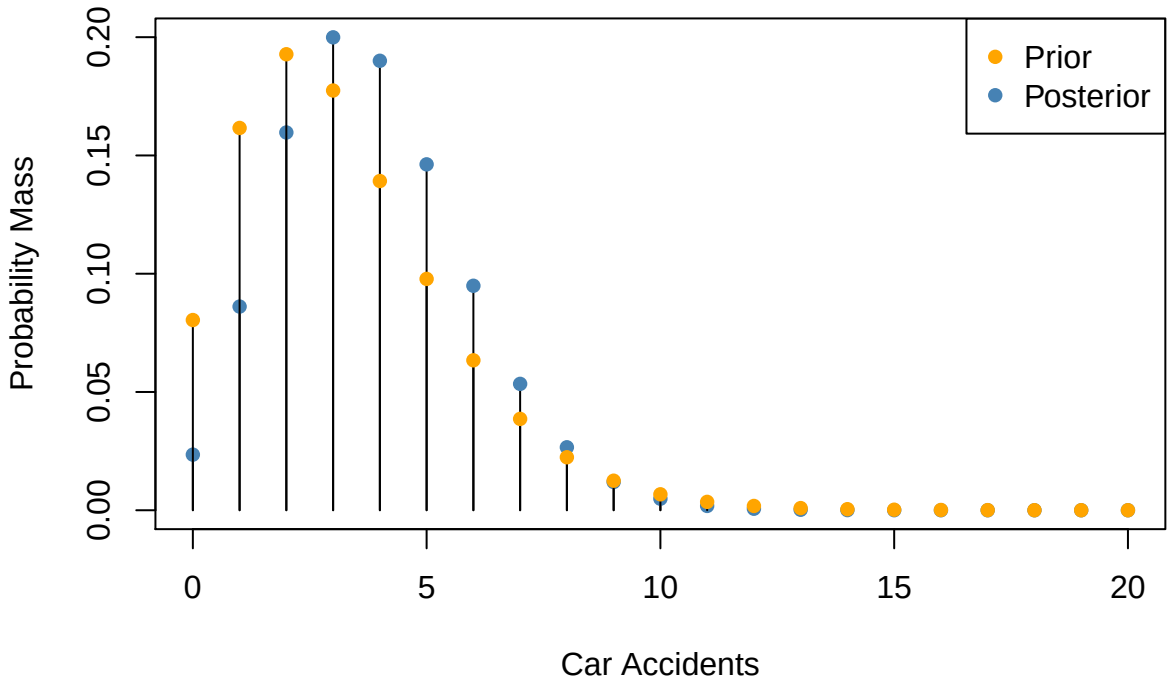
3e) (optional) Provide a formal definition of the posterior predictive distribution of $Y_{next}|y_1, ..., y_n$ and try to compare the posterior predictive distribution for a future observable with the actually observed data If we want to compute the posterior predictive distribution we have to marginalise the join distribution over the parameter θ

$$m(y_{\text{next}}|y_1, .., y_n) = \int_0^{+\infty} f(y_{\text{next}}|\theta, y_1, \dots, y_n)\pi(\theta|y_1, \dots, y_n)d\theta$$

It can be proven that this distribution is indeed a *Negative Binomial* distribution with the following parameters:

$$Y_{\text{next}} \sim NegBin\Big(p = \frac{r^{\text{prior}} + n}{r^{\text{prior}} + n + 1}, size = s^{\text{prior}} + \sum_{i=0}^{n} y_i\Big)$$

Below we can see the **pmf**s of both the prior and the posterior.


Prior
Posterior
Probability Mass
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
Car Accidents

Bulb lifetime

You work for Light Bulbs International. You have developed an innovative bulb, and you are interested in characterizing it statistically. You test 20 innovative bulbs to determine their lifetimes, and you observe the following data (in hours), which have been sorted from smallest to largest.

```
bulbs_lifetime <- c(1, 13, 27, 43, 73, 75, 154, 196, 220, 297,
           344, 610, 734, 783, 796, 845, 859, 992, 1066, 1471)
```

Based on your experience with light bulbs, you believe that their lifetimes Y_i can be modeled using an exponential distribution conditionally on θ where $\psi = 1/\theta$ is the average bulb lifetime.

Write the main ingredients of the Bayesian model. The main ingredients in Bayesian modeling are:

$$\pi(\theta|y) = \frac{f_Y(y|\theta)\pi(\theta)}{m(y)} = \frac{f_Y(y|\theta)\pi(\theta)}{\int f_Y(y|\theta)\pi(\theta)d\theta} \propto f_Y(y|\theta)\pi(\theta)$$

Where:

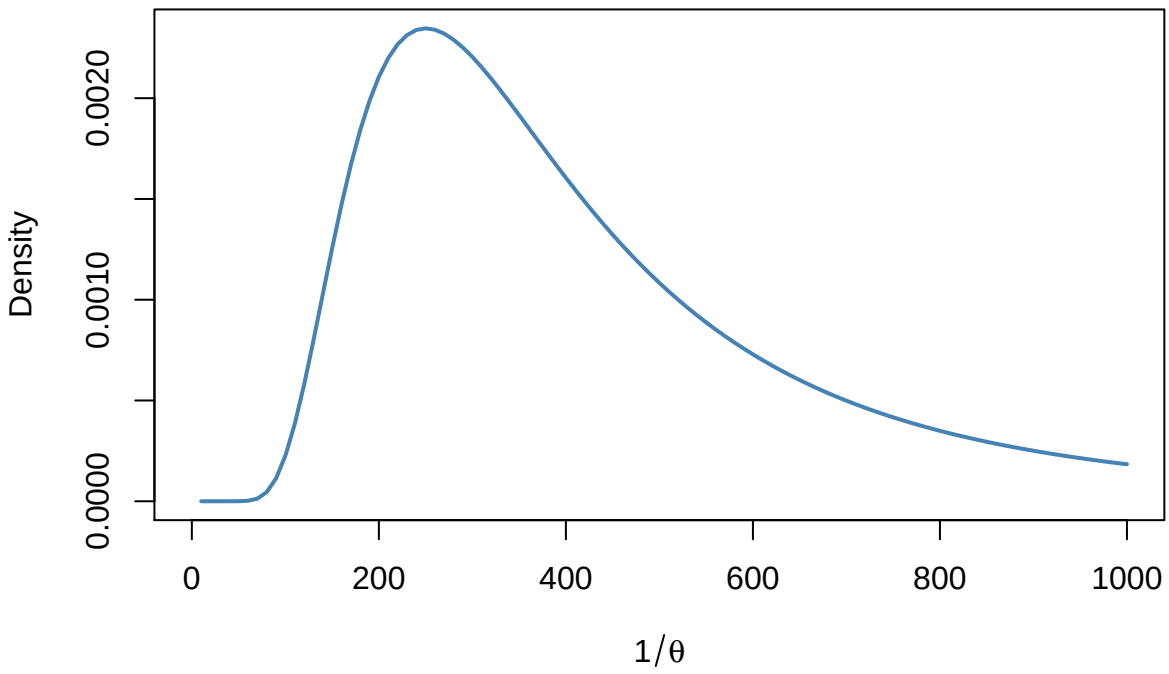
- $\pi(\theta)$ is the prior distribution of the parameter θ
- $\pi(\theta|y)$ is the posterior distribution of the parameter θ
- $f_Y(y|\theta)$ is the likelihood and represents how likely are the observed data given the parameter θ, in our case it can be computed as the product of probabilities drawn from an exponential distribution $f_Y(y|\theta) = \prod_i \theta e^{-\theta y_i} = \theta^n e^{-\theta \sum_i y_i}$
- $m(y)$ is the marginal predictive distribution of y which can be computed as the integral of the product between the likelihood and the prior over the θ parameter's space

Choose a conjugate prior distribution $\pi(\theta)$ with mean equal to 0.003 and standard deviation 0.00173. As prior distribution we choose a gamma distribution with density $\pi(\theta) \propto \theta^{s-1}e^{-r\theta}$. We can use the same transformation as before to find the distribution's rate and shape values.

```
mean_prior <- 0.003
std_prior <- 0.00173
var_prior <- std_prior ^ 2
rate_prior <- mean_prior / var_prior
shape_prior <- mean_prior ^ 2  / var_prior
```

so our conjugate prior distribution is $\theta \sim Gamma(r = 1002.37\ , s = 3.007\ )$

Argue why with this choice you are providing only a vague prior opinion on the average lifetime of the bulb. Since the average bulb lifetime can be expressed as $\psi = \frac{1}{\theta}$ this means that it follows an *Inverse Gamma* distribution, we can therefore give a look at its density.



As we can see the distribution is widespread on the real line therefore our prior opinion on the bulb lifetime is really vague.

Show that this setup fits into the framework of the conjugate Bayesian analysis We already shown that $f_Y(y|\theta) = \theta^n e^{-\theta \sum_i y_i}$, we can then use to change the posterior formula:

$$\pi(\theta|y_1,\dots,y_n) \propto f_Y(y|\theta)\pi(\theta) = \theta^n e^{-\theta \sum_i y_i} \theta^{s-1} e^{-r\theta} = \theta^{\,(s+n)-1} e^{-(\,r+\sum y_i)\,\theta}$$

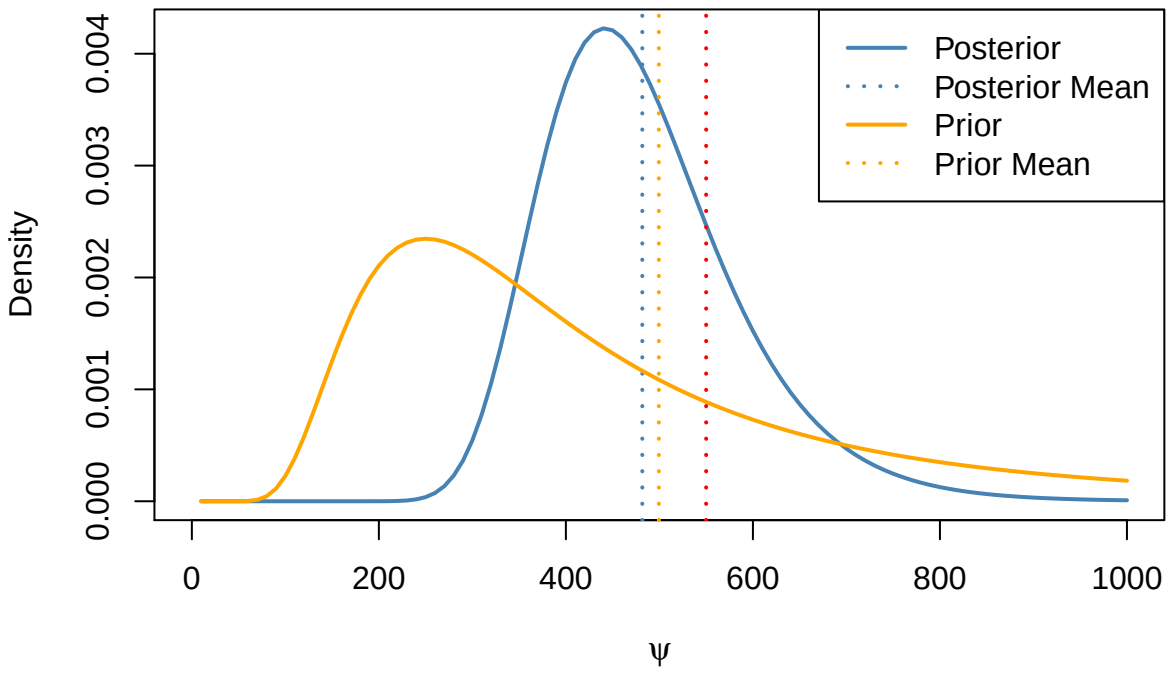
This shows that the posterior is indeed a Gamma distribution with updated parameters $r^{\text{posterior}}, s^{\text{posterior}}$:

$$\theta|y_1,\dots,y_n \sim Gamma(r^{\text{posterior}} = r^{\text{prior}} + \sum_{i=1}^{n} y_i,\ s^{\text{posterior}} = s^{\text{prior}} + n)$$

Based on the information gathered on the 20 bulbs, what can you say about the main characteristics of the lifetime of your innovative bulb? Argue that we have learnt some relevant information about the θ parameter and this can be converted into relevant information about $1/\theta$ Firts of all we have to update the parameters according to the rules we just derived.

```
rate_posterior <- rate_prior + sum(bulbs_lifetime)
shape_posterior <- shape_prior + length(bulbs_lifetime)
```

We can now give a look at the prior and posterior distributions along with the point estimates for their mean, which in the case of an Inverse Gamma can be computed using the formula $\mathbb{E}[\psi] = \frac{r}{s-1}$.



After observing the data the mean decreased from the 499.4 hours in the prior to 481.7 according to the posterior therefore we argue we gained information.

However, your boss would be interested in the probability that the average bulb lifetime $1/\theta$ exceeds 550 hours. What can you say about that after observing the data? Provide her with a meaningful Bayesian answer. In order to answer our boss' question we have to evaluate

$$\mathbb{P}(\psi \geq 550) = 1 - \mathbb{P}(\psi < 550)$$

which can be easily done using the `pinvgamma` function.

```
p_exceed = 1 - pinvgamma(550, shape_posterior, rate_posterior)
```

The probability that an average bulb lifetime exceeds 550 is **0.2254117**