Homework #01

Statistical Methods in Data Science II & Lab

2021/2022

April 22th, 2022

Francesco Pinto 1871045

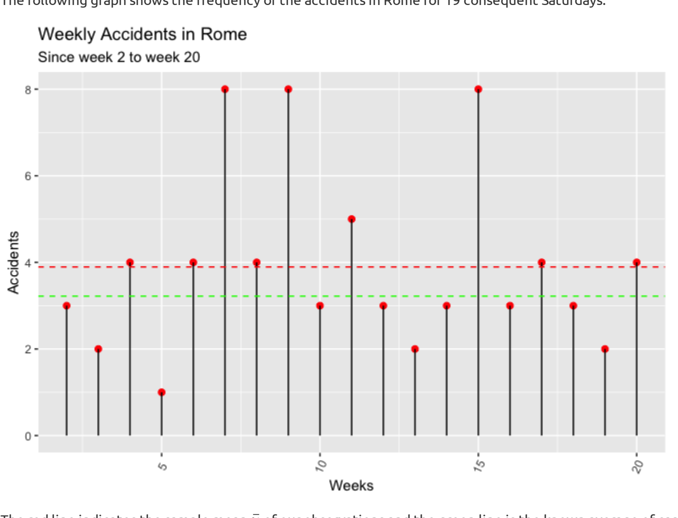
Fully Bayesian conjugate analysis of Rome car accidents

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

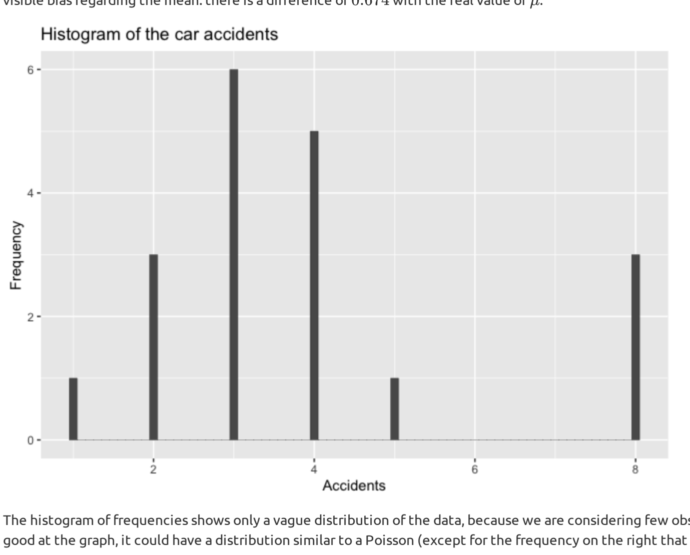
```
mydata <- subset(roma,subset=sign_up_number==104)
y_prior = mydata$car_accidents
```

1. The data

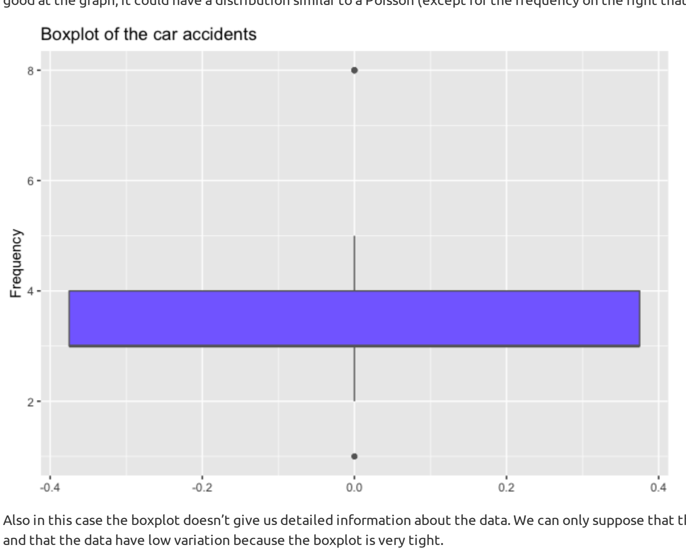
The following graph shows the frequency of the accidents in Rome for 19 consequent Saturdays.



The red line indicates the sample mean $\bar{y}$ of our observations and the green line is the known average of car accidents. In our sample there is a visible bias regarding the mean: there is a difference of 0.674 with the real value of μ.



The histogram of frequencies shows only a vague distribution of the data, because we are considering few observations. By the way, looking good at the graph, it could have a distribution similar to a Poisson (except for the frequency on the right that could distort this hypothesis).



Also in this case the boxplot doesn't give us detailed information about the data. We can only suppose that the median of the car accidents is 3 and that the data have low variation because the boxplot is very tight.

2. The Ingredients

Statistical model: $Y_i|\theta \sim Poi(\theta)$

The best choice for the statistical model, in this case, is a Poisson model. The Poisson distribution expresses the probability for the number of events that occur successively and independently in a given time interval, knowing that in average a number occurs θ times (exactly like the number of car accidents during a day).

Prior distribution: $\theta \sim Gamma(s, r)$

In order to do a Conjugate Bayesian analysis, the prior distribution chosen for θ is a Gamma, let's explain why…

Given a random sample $y_1, \dots, y_n$ from a $Pois(\theta)$, the likelihood function is

$$L(\theta; y_n) = \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!} \propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta}$$

that can be interpreted as a Gamma with parameters $s = \sum_{i=1}^n y_i + 1$ and $r = n$. In fact, the Gamma probability distribution for θ is

$$\pi(\theta) = \frac{r^s}{\Gamma(s)} \theta^{(s-1)} e^{-r\theta} \propto \theta^{(s-1)} e^{-r\theta}$$

where s is the *shape* parameter and r is the *rate* parameter.

We can now merge the Poisson likelihood with the Gamma distribution:

So the posterior distribution is

$$\pi(\theta|y_n) \propto \pi(\theta) L(\theta|y_n) \propto \theta^{(s+\sum_{i=1}^n y_i - 1)} e^{-(r+n)\theta} = \theta^{(s+y_n)-1} e^{-(r+n)\theta}$$

That corresponds to a gamma with parameters $\theta|y_n \sim Gamma(\bar{s}, \bar{r})$, where $\bar{s} = s + y_n$ and $\bar{r} = r + n$.

By using the function *manipulate*, the best parameters I found are $s = 3.22$ and $r = 1$ because their mean and their variance correspond to the respective values of our statistical model (in our Poisson, $\mu = \sigma^2 = 3.22$).

```
s = 3.22
r = 1
```

Furthermore, Gamma distribution and Poisson distribution (and also Exponential distribution) are models that pattern different aspects of the same process, which concerns the waiting time.

3.a Point estimating

Let's start with the estimation of the **Mean**: $E[\theta|y_n] = \frac{\bar{s}}{\bar{r}} = \frac{s_{prior} + \sum_{i=1}^n y_i}{r_{prior} + n}$

```
s_hat = s + sum(y_prior)
r_hat = r + length(y_prior)
mu_post = s_hat/r_hat
mu_post
```

```
## [1] 3.861
```

Another estimate could be the **Mode** estimate: $Mode_{post} = \frac{\bar{s}-1}{\bar{r}}$

```
mode_post = (s_hat - 1)/(r_hat)
mode_post
```

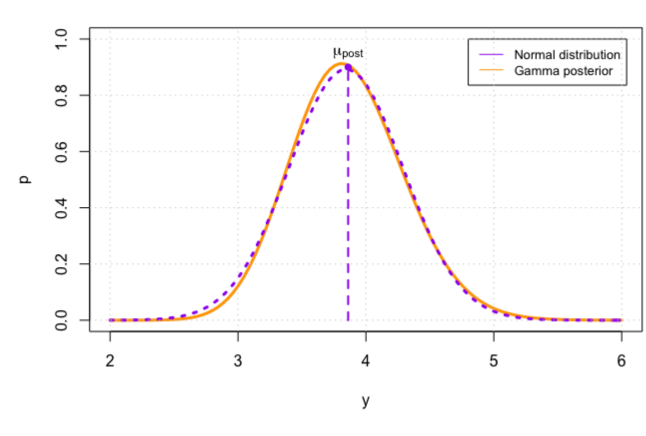
```
## [1] 3.811
```

And finally, it's the **Median** time…

```
median_post = qgamma(0.5, rate = r_hat, shape = s_hat)
median_post
```

```
## [1] 3.844346
```

The three results are very close, so we can suppose that the posterior distribution is simil-Normal. In fact, looking at the plot below, the posterior distribution looks similar to a normal distribution with $\mu = 3.844$ (corresponding to the mode of the Gamma) and $\sigma^2 = 0.2$:



In fact, the posterior distribution is proportional to the likelihood function that, as the sample number increases, tends to concentrate around its maximum point and to take the form of a Gaussian.

Consequently it's demonstrable that, under certain conditions and for a sufficiently elevated sample numerosity, also the posterior density tends to a normal density.

3.b The posterior uncertainty

To measure the posterior uncertainty for a new sample, a good indicator could be the predictive variance. The uncertainty of y_n in fact is sometimes contained into the variability of the data that distorts a good prediction for the parameters.

In this case, the posterior variance is $\sigma^2_{post} = \frac{\bar{s}}{\bar{r}^2}$:

```
sigma2_post = s_hat/r_hat^2
round(sigma2_post,3)
```

```
## [1] 0.193
```

The variance is very low, so we can conclude that the values in the posterior distribution are mainly concentrated around the mean. It's a good fact because the expected value can be considered "plausible".

3.c Interval estimating

Now it's time to estimate credible intervals. The credible intervals are Bayesian confidence intervals, with the difference that in the Bayesian case a 95% credible interval actually contains a true parameter value with 95% probability.

The first interval to estimate is the **Equal-Tail interval (ET)** at level $\alpha = 0.05$

$$I_\alpha = [q_{\alpha/2}, q_{1-\alpha/2}]$$

Before estimating the ETI, I define the quantile function as

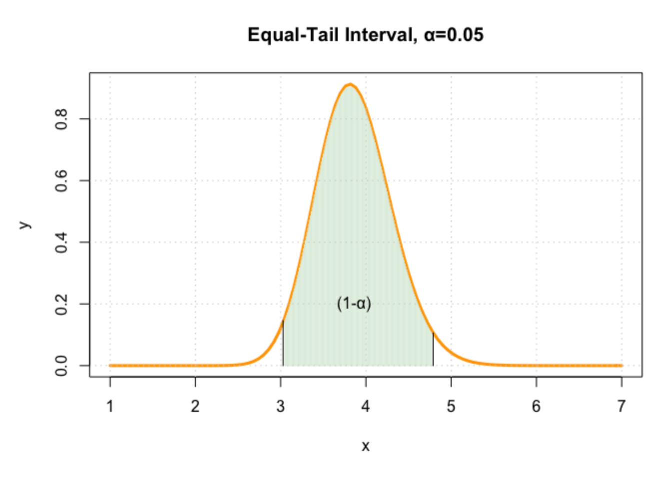
```
posterior_qf = function(x){
  qgamma(x,rate = r_hat, shape = s_hat)
}
```

and then

```
alpha_lev=0.05

ET_interval = c(posterior_qf(alpha_lev/2),posterior_qf(1-alpha_lev/2))
round(ET_interval, 3)
```

```
## [1] 3.048 4.768
```



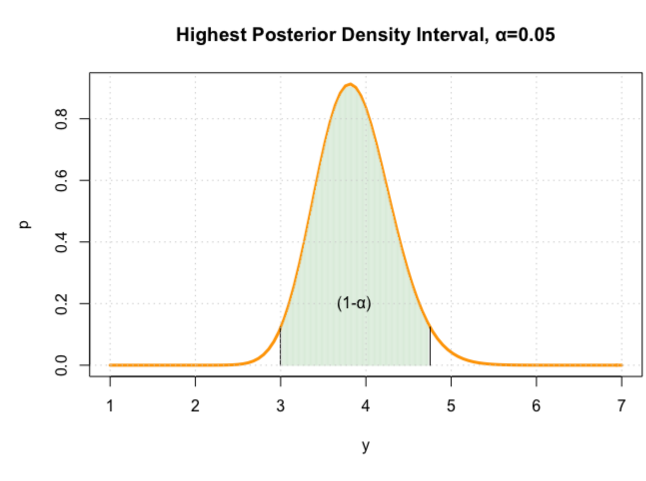
By the way, if the posterior distribution is not unimodal and symmetric (Gamma is not symmetric) there could be points points out of the ET interval having a higher posterior density than some points of the interval. So in this case it's better to study the **Highest Posterior Density interval (HPD)**.

The particularity of the HPD interval is that the posterior density for every point in the confidence region I_α is higher than the posterior density for any point outside of this set.

I will find the HPD interval at level $\alpha = 0.05$ with tolerance 0.0000001 for optimize with the *hpd* function from the *TeachingDemo* package:

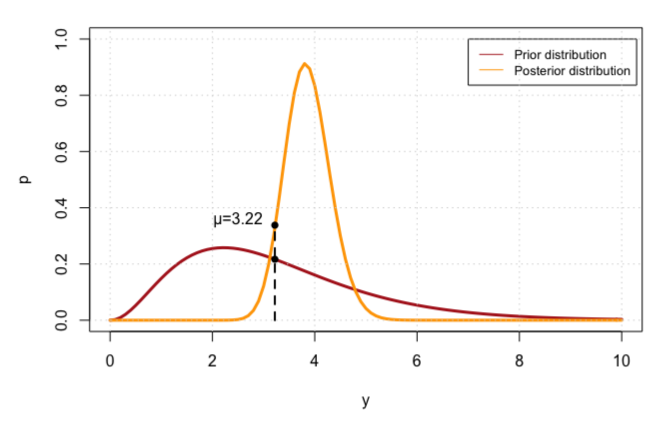
```
HPD_interval = hpd(posterior.icdf=posterior_qf, conf=1-alpha_lev, tol=0.0000001)
round(HPD_interval, 3)
```

```
## [1] 3.017 4.733
```



3.d Differences

The red curve in the plot is the prior distribution, that represents the original distribution before the introduction of the observed data ($Gamma(s, r)$). The orange curve instead is the posterior distribution that takes into account the new information ($Gamma(\bar{s}, \bar{r})$).



The prior distribution is more squeezed down than the posterior: it means that the distribution gives similar probabilities to different values of y, and it isn't an accurate way to predict data. Also observing the value of the known average ($\mu_{known} = 3.22$), it hasn't an high probability like other values.

The posterior distribution instead is taller and narrower because, increasing the data, the information about the distribution becomes more clear and the curve becomes thinner around its maximum (maximum likelihood estimate). In this case, we are assigning strongest probabilities to more plausible values. The distribution becomes also more similar to a normal distribution due to the reason explained in the point **3.a**.

However, the real value of the average is not particularly considered by both the probability distributions.

3.e Posterior predictive distribution

Before the data are considered, the distribution of the unknown Y is

$$m(y) = \int_\Theta f(y|\theta) d\theta$$

It's the marginal distribution of y and it's called *prior predictive distribution*. With this distribution we can only try to do predictions with the information given by the single parameter.

After observing the data $Y = (Y_1, \dots, Y_n) = y = (y_1, \dots, y_n)$, the **posterior predictive distribution** is given by

$$m(y_{new}|y_1, \dots, y_n) = \int_\Theta f(y_{new}|\theta)\pi(\theta|\mathbf{y})$$

It is the distribution of possible unobserved values conditional on the observed values.

In the case of a conjugate Poisson-Gamma, the posterior predictive becomes

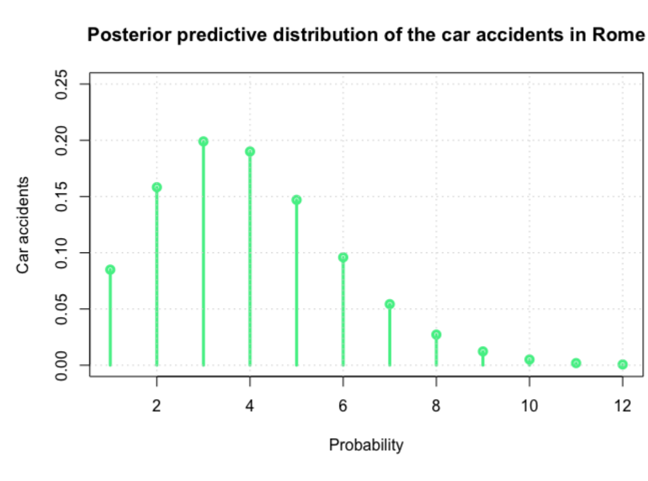
$$m(y_{new}|y_1, \dots, y_n) = \int_\Theta \theta^{y_{new}} \frac{e^{-\theta}}{y_{new}!} \frac{\bar{r}^{\bar{s}}}{\Gamma(\bar{s})} \theta^{\bar{s}-1} e^{-\bar{r}\theta} \, d\theta = \frac{\bar{r}^{\bar{s}}}{\Gamma(\bar{s}) y_{new}!} \int_\Theta \theta^{y_{new}+\bar{s}-1} e^{-(\bar{r}+1)\theta} \, d\theta = \dots$$

$$\dots = \frac{\bar{r}^{\bar{s}}}{\Gamma(\bar{s}) y_{new}!} \cdot \left(\frac{1}{\bar{r}+1}\right)^{y_{new}+\bar{s}} \Gamma(y_{new}+\bar{s}) = \frac{\Gamma(y_{new}+\bar{s})}{\Gamma(\bar{s}) y_{new}!} \left(1 - \frac{\bar{r}}{\bar{r}+1}\right)^{y_{new}} \left(\frac{\bar{r}}{\bar{r}+1}\right)^{\bar{s}}$$

That is the density function of a Negative Binomial distribution as follows:

$$Y_{new}|y_1, \dots, y_n \sim NegBin(\bar{s}, \frac{\bar{r}}{\bar{r}+1})$$

and has the following pattern



Also in this case, the most plausible number of accidents on Saturday is 3 or 4! In fact, also in the posterior distribution the mean and the mode were between these two numbers.

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.

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.

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

1. The ingredients

Statistical model: $Y_i|\theta \sim Exp(\theta)$

The exponential distribution describes the lifetime of a given phenomena, so it's the best model we can choose in this case.

The characteristic of the exponential distribution is the *memory loss*: it means that our phenomena is independent by its past and "doesn't get old"! So it fits perfectly with our probability case, where a bulb has lifetime independent by the previous time (it can die in every moment). Mathematically speaking, the memory loss is that property such that

$$P(X > x + y | X > y) = P(X > x)$$

The Exponential distribution has probability density function:

$$f(y|\theta) = \theta e^{-\theta y}$$

for $y > 0$

Prior distribution: $\theta \sim Gamma(s, r)$

The chosen prior distribution for θ is Gamma, but in order to study the lifetime of the bulb, we are interested in $\psi = \frac{1}{\theta}$, so the purpose is to find the parameters for the Gamma distribution and then study the distribution of ψ with the Inverse Gamma function.

The Gamma distribution has following pdf:

$$\pi(\theta) = \frac{r^s}{\Gamma(s)} \theta^{(s-1)} e^{-r\theta}$$

2. The parametrized conjugate prior distribution

Knowing that

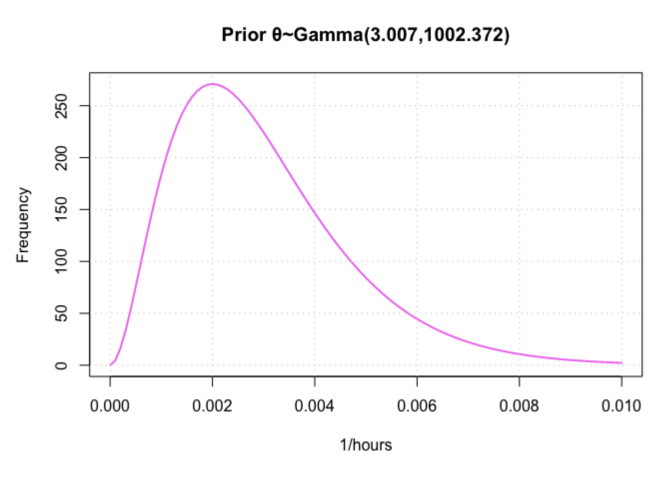
$$\begin{cases} \frac{s}{r} = 0.003 \\ \sqrt{(\frac{s}{r^2})} = 0.00173 \end{cases}$$

after solving the system we obtain $s = 3.007$ and $r = 1002.372$.

Due to the fact that the parameter of the Exponential distribution we are looking for is $\theta = \frac{1}{\psi}$, the objective is to study an inverted random variable $\frac{1}{X}$ and in order to do it, we will use an Inverse Gamma distribution (from the *invgamma* package).

```
s = 3.007
r = 1002.372
```

So the prior distribution for θ is $\pi(\theta) \sim Gamma(3.007, 1002.372)$



3. Vague prior opinions

Of course the distribution of θ does not give us important information regarding the lifetime of the bulb: the values are very small and, even if we try to estimate the mean lifetime by doing $\frac{1}{\mu_\theta}$, it's very far from the sample mean (that reflects a little part of the real distribution).

```
mu_theta = 1/(s/r)
mu_theta
```

```
## [1] 333.3462
```

The bias is very high!

```
BIAS = mean(y_prior) - mu_theta
BIAS
```

```
## [1] 146.6038
```

The best way to predict the representative values of the mean lifetime is with the **Inverse Gamma Distribution**.In fact, given a random variable $X \sim Gamma(s, r)$, then $Y = \frac{1}{X} \sim InvGamma(s, r)$.

For example, the expected value if we use an Inverse Gamma is $\frac{r}{s-1} = 498.69$ that is very close to the sample mean. So from now the analysis will shift on the mean $\psi = \frac{1}{\theta} \sim InvGamma(3.007, 1002.372)$ and no more on θ

4. Fitting into the Conjugate Bayesian Analysis

As mentioned above, (point 2 of exercise 1), Exponential distribution and Gamma distribution models are different aspects of the same process. So, a good idea would be merge these two distribution into one and get a Gamma posterior distribution.

In this case-study, the values s and r are studied as Gamma parameters, and then inserted into the InverseGamma.

By the way, mathematically speaking…

Above I said that the ingredients for the CBA are $Y_i|\theta \sim Exp(\theta)$ as statistical model and $\theta \sim Gamma(s, r)$ as prior distribution… but why are they the right ingredients for the analysis?

Knowing that the data have an Exponential distribution $Y_i|\theta \sim Exp(\theta)$, the likelihood is

$$L(\theta|y_n) = \theta^n e^{-\theta \sum_{i=1}^n y_i}$$

that is a Gamma with parameters $s = n$ and $r = \sum_{i=1}^n y_i$.

Multiplying it to the following Gamma density function

$$\pi(\theta) = \frac{r^s}{\Gamma(s)} \theta^{(s-1)} e^{-r\theta} \propto \theta^{(s-1)} e^{-r\theta}$$

we obtain the posterior density as follows

$$\pi(\theta|y_n) \propto \theta^n e^{-\theta \sum_{i=1}^n y_i} \theta^{s-1} e^{-r\theta} = \theta^{n+s-1} e^{-\theta(\sum_{i=1}^n y_i + r)}$$

that can be interpreted as a Gamma distribution: $\theta|y_n \sim Gamma(s + n, r + \sum_{i=1}^n y_i)$

So, in a posterior point of view, the Gamma prior distribution is in line with our hypothesis for conjugate models and, inverting it, we will use the Inverse Gamma distribution.

5. Characteristic of the lifetime of my bulb

About the parameter θ we have useless information for our analysis (almost all the values of the distribution are less than 0.01). In fact, we are interested in knowing something about the ψ parameter and we don't care about θ.

Instead, let's study the distribution of $\frac{1}{\theta}$

The mean lifetime in our experiment (the sample mean) is $\bar{y} = 479.95$, but if we dive more deeply into the Bayesian Model, we have a new mean given by $\frac{\bar{r}}{\bar{s}-1}$ that is

```
s_post = s + length(y_prior)
r_post = r + sum(y_prior)
mu_post = (r_post)/(s_post-1)

round(mu_post, 2)
```

```
## [1] 481.73
```

This value is accurate and very close to the sample mean, so the posterior distribution seems to give good results!

Now let's look at the mode $Mode = \frac{r}{s+1}$

```
mode_post = r_post/(s_post+1)

round(mode_post, 2)
```

```
## [1] 441.6
```

And finally the median (with the *quantile function*)

```
median_post = qinvgamma(0.5, s_post, r_post)

round(median_post, 2)
```

```
## [1] 467.55
```

The three values are relatively near, but the distribution remains positively asymmetric.

And lastly, I'm interested in knowing approximately how much time will last my bulb and with which probability. My distribution is not symmetric, so I'll use the HPD interval to study it.

I will study how many hours will my bulb persist with high probability (95%)

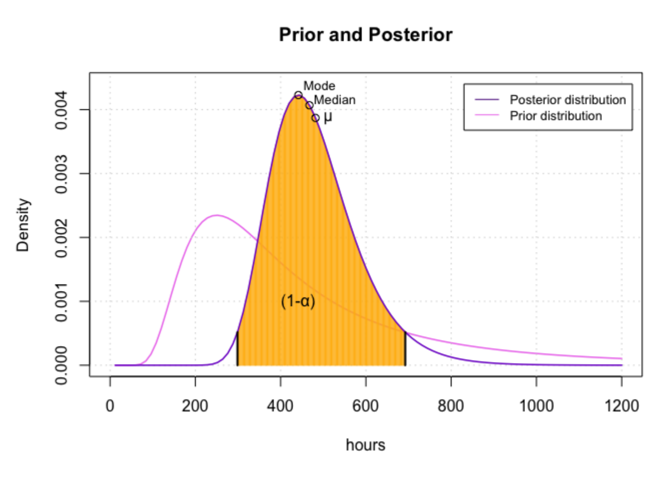
```
posterior_qf = function(x){
  qinvgamma(x,rate = r_post, shape = s_post)
}

alpha_lev = 0.05

HPD_interval = hpd(posterior.icdf=posterior_qf, conf=1-alpha_lev, tol=0.0000001)
round(HPD_interval)
```

```
## [1] 299 692
```

Now let's compare prior and posterior distributions graphically as follows.



What does the posterior distribution tell us?

After considering the data on the bulbs, we obtain the posterior distribution with the relative indexes that can help us to make previsions on the future bulbs I predict.

Thanks to the credible interval found ($I_\alpha = [299, 692]$) a first supposition is that my innovative bulb will have a duration between 12 and 26 days with probability of 95%, and that in general it will be around 481 hours (around 20 days).

6. Previsions

My boss is asking me what's the probability that the mean lifetime of a bulb exceeds 550 hours. To satisfy her, after observing the data, I will use the basic probability laws of random variables.

The boss is asking me to find $P(\psi > 550|y_1, \dots, y_n)$ and I have to use the **Cumulative density function (CDF)** to find probabilities:

$$P(\psi > 550|y_1, \dots, y_n) = 1 - P(\psi \le 550|y_1, \dots, y_n) = 1 - \int_0^{550} f(\psi|y_n) d\psi$$

Instead of writing the integral, I can use the *cdf* function of R for an Inverse Gamma distribution:

```
P = 1-pinvgamma(550, s_post, r_post)

round(P,4)*100
```

```
## [1] 22.54
```

The probability of having the mean lifetime of the bulbs over 550 is 22.54%

In a first moment, without a Bayesian point of view, we could have believed (ignorantly) that the probability of having mean more than 550 would be something like $\sum_{i=1}^{20} I(y_i > 550)$.

But thanks to the Bayesian inference and to the statistical model used, we modeled the distribution in a way such that we could study also the average of the lifetime of the bulbs. The possibility to "continuize" discrete values and model the distribution in a way to obtain more precise values was fundamental to give a satisfying answer to my boss!