Inferencia Bayesiana. Distribuciones conjugadas

Félix Belzunce María del Carmen Bueso, Pilar Sanmartín

Introducción

En esta práctica veremos el uso de `R` para realizar inferencia bayesiana en algunos de los modelos de familias conjugadas que hemos visto en clase. Para ello utilzaremos el paquete `bayesrules`.

Modelo binomial-beta

Vemos en primer lugar la inferencia usando como distribución conjugada de la distribución Bernoulli (binomial) la distribución beta. Recordamos este caso.

- **Distribución de la variable** X: X sigue una distribución Bernoulli de parametro p ($X \sim B(p)$).
- **Distribución a priori del parámetro** p: p sigue una distribución beta de parámetros α y β ($X \sim Be(\alpha, \beta)$), es decir,
$$\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}.$$
- **Distribución a posteriori:**
$$\pi(p|\mathbf{x}) \propto p^{\alpha+\sum_{i=1}^n x_i-1}(1-p)^{\beta+n-\sum_{i=1}^n x_i-1},$$
es decir sigue una distribución $Be(\alpha+\sum_{i=1}^n x_i, \beta+n-\sum_{i=1}^n x_i)$.

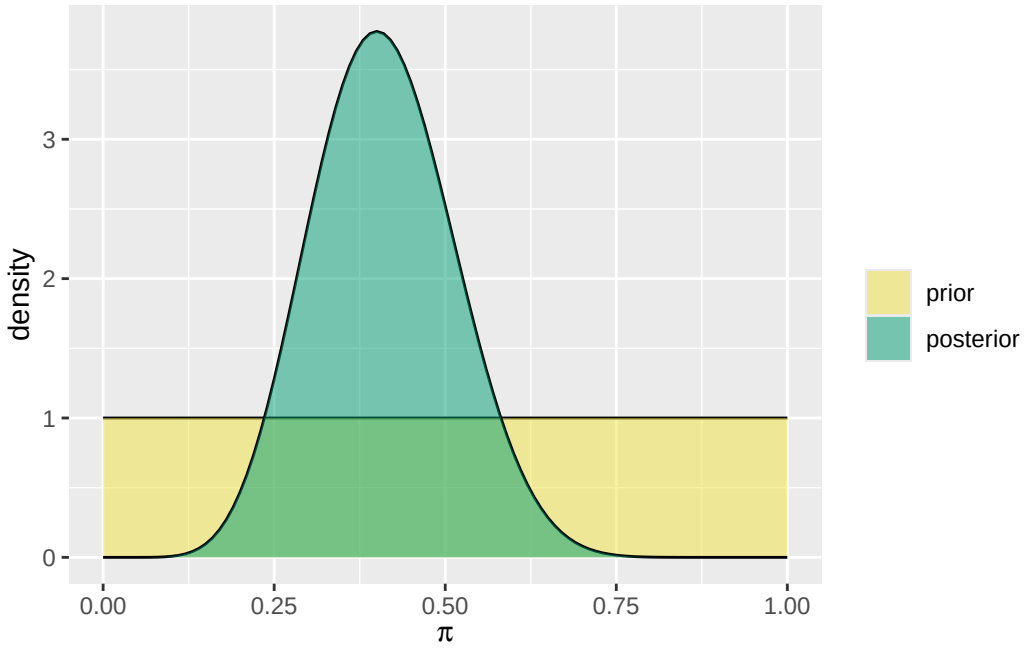
Vamos a utilizar el ejemplo visto en clase.

> Ejemplo:
>
> Consideramos el experimento de lanzar un dado y observar si el resultado es par o impar. Tenemos una variable X con distribución Bernoulli, que toma los valores $X = 1$ si el resultado es par y $X = 0$ si el resultado es impar. Denotaremos $p = P(X = 1)$ y $1 - p = P(X = 0)$. Vamos a considerar que el valor p puede ser cualquier valor, sin ninguna preferencia. Esa información se traduce en asumir una distribución uniforme en el intervalo $(0, 1)$ y por lo tanto $\pi(p) = 1$. Supongamos ahora que observamos un conjunto de observaciones del experimento, en concreto tenemos la siguiente muestra de tamaño 20:
>
> $$\mathbf{x} = (1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0).$$

Veamos como obtener la distribución a posteriori de p a partir de la muestra anterior.

En primer lugar vemos la obtención de la gráfica donde vemos el paso de la distribución a priori a la posteriori a partir de la muestra anterior.

```
library(bayesrules)
library(tidyverse)
plot_beta_binomial(
    alpha = 1,
    beta = 1,
    y = 8,
    n = 20,
    prior = TRUE,
    likelihood = FALSE,
    posterior = TRUE
)
```



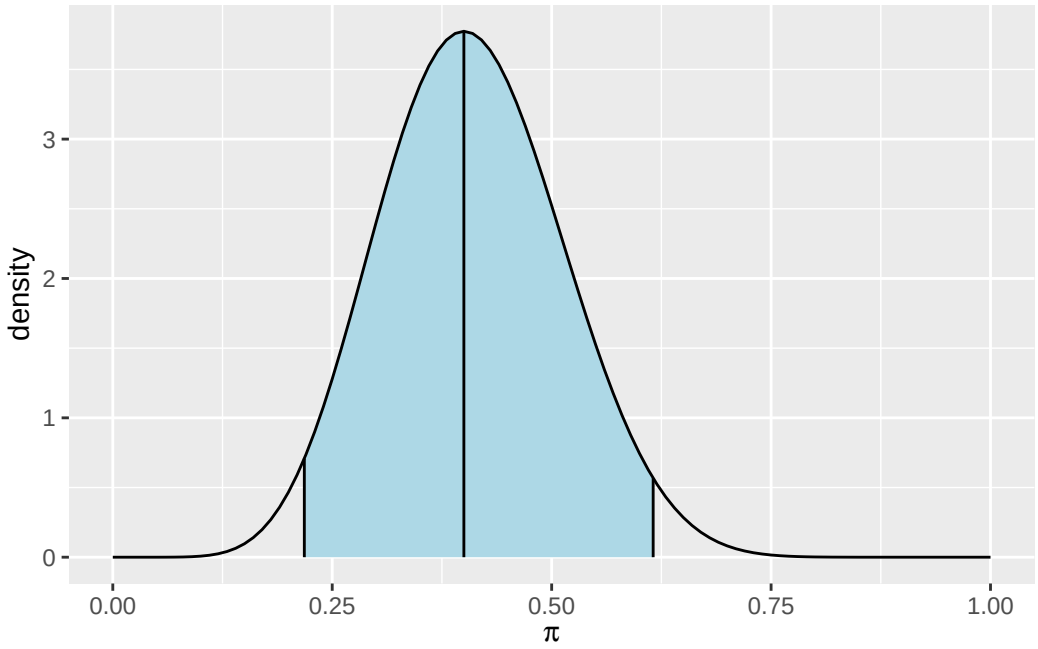
Para la distribución a posteriori podemos obtener los valores que la caracaterizan, así como su media, moda, varianza y desviación típica a posteriori:

```
summarize_beta_binomial(alpha = 1, beta = 1, y = 8, n = 20)
```

```
      model alpha beta      mean mode        var        sd
1     prior     1    1 0.5000000  NaN 0.08333333 0.2886751
2 posterior     9   13 0.4090909  0.4 0.01051024 0.1025195
```

Podemos obtener también un intervalo de credibilidad para p con un invel del 95%, tanto gráficamente

```
plot_beta_ci(9, 13, ci_level = 0.95)
```



como numéricamente

```
qbeta(0.025, 9, 13) #extremo inferior
```

```
[1] 0.2181969
```

```
qbeta(0.975, 9, 13) #extremo superior
```

```
[1] 0.6156456
```

Realizar el siguiente caso.

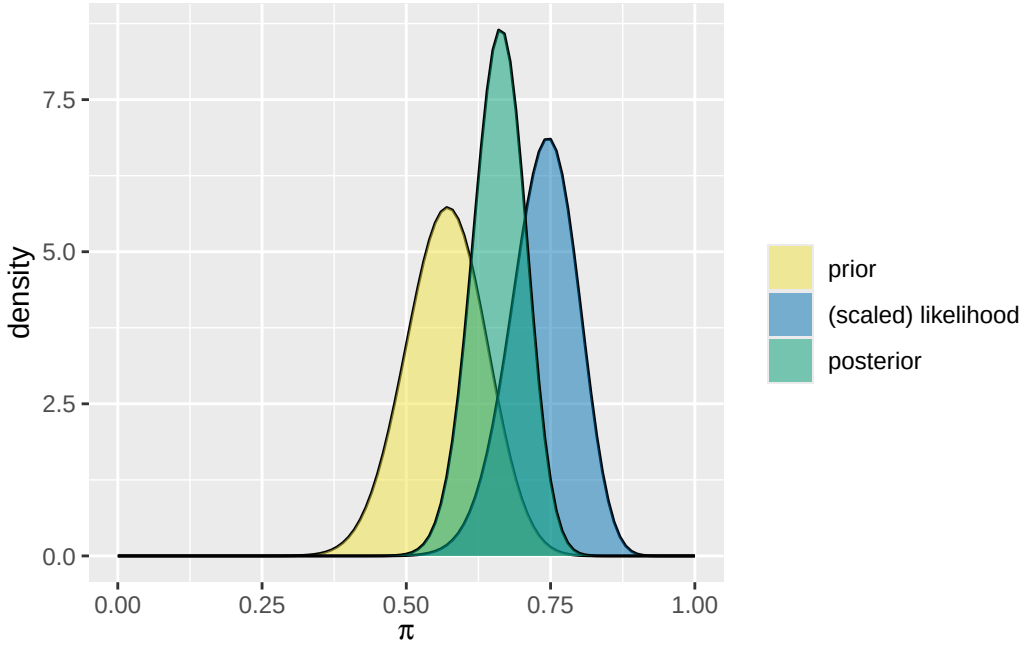
> Ejercicio:
>
> En lo que llevamos de año (2023) de las 55 victimas de género, 41 no presentaron denuncia. Por la experiencia previa se considera que la distribución de la proporción de mujeres que han sido victimas de género y que no han presentado denuncia sigue una distribución $Be(29, 22)$. Obtener a partir de los datos obtenidos en 2023 la información correspondiente de la distribución a posteriori.

Comenta los resultados obtenidos.

```r
# completar aquí
# Valores de los parámetros
alpha_prior <- 29
beta_prior <- 22
n <- 55
y <- 41

# Gráfica de la transición de la distribución previa a la posterior

plot_beta_binomial(
  alpha = alpha_prior,
  beta = beta_prior,
  y = y,
  n = n,
  prior = TRUE,
  likelihood = TRUE,
  posterior = TRUE
)
```




```r
# fin completar aquí
```

- En la distribución previa, se refleja lo que pensábamos antes de ver los datos. Como estaba basada en información previa, era un poco más amplia y centrada alrededor de su media (0.57).
- En la distribución posterior, combinamos lo que sabía antes con los nuevos datos. Está más concentrada, lo que quiere decir que ahora estámos más seguros, y se centra en valores alrededor de su media (0.66), porque los datos muestran que la mayoría de las mujeres no presentó una denuncia.

```
# Resumen de la distribunción posterior
summarize_beta_binomial(alpha = alpha_prior, beta = beta_prior, y = y, n=n)
```

```
      model alpha beta      mean      mode         var         sd
1     prior    29   22 0.5686275 0.5714286 0.004717121 0.06868130
2 posterior    70   36 0.6603774 0.6634615 0.002096066 0.04578282
```

En la tabla se comparan las estadísticas antes y después de ver los datos:

- **Previo:** Antes pensábamos que p estaba cerca de 0.57, pero con algo de incertidumbre.
- **Posterior:** Ahora creemos que p es más alto, alrededor de 0.66, y estamos más seguros porque los datos redujeron la incertidumbre.

```
# Intervalo de credibilidad del 95%
qbeta(0.025, alpha_prior + y, beta_prior + n - y)
```

```
[1] 0.5679935
```

```
qbeta(0.975, alpha_prior + y, beta_prior + n - y)
```

```
[1] 0.7470282
```

El intervalo nos dice que, según el modelo y los datos, p probablemente está entre 0.57 y 0.75 con un 95% de confianza. Esto significa que estamos bastante seguros de que la proporción de mujeres que no denuncian está en ese rango.

Modelo Poisson-gamma

En este modelo la distribución conjugada de la distribución de Poisson es la distribución gamma. Las distribuciones que tenemos son las siguientes.

- **Distribución de la variable** X: X sigue una distribución Poisson parametro λ ($X \sim P(\lambda)$).
- **Distribución a priori del parámetro** λ: λ sigue una distribución gamma de parámetros α y β ($X \sim G(\alpha, \beta)$). Se tiene que

$$\pi(\lambda) \propto \lambda^{\alpha-1} \exp(-\frac{1}{\beta}\lambda).$$

- **Distribución a posteriori:**

$$\pi(\lambda|\mathbf{x}) \propto \lambda^{\alpha+\sum_{i=1}^{n} x_i - 1} \exp(-(\frac{1}{\beta} + n)\lambda),$$

es decir sigue una distribución $G(\alpha + \sum_{i=1}^{n} x_i, \frac{1}{\beta} + n)$

Veamos un ejemplo.

Ejemplo:

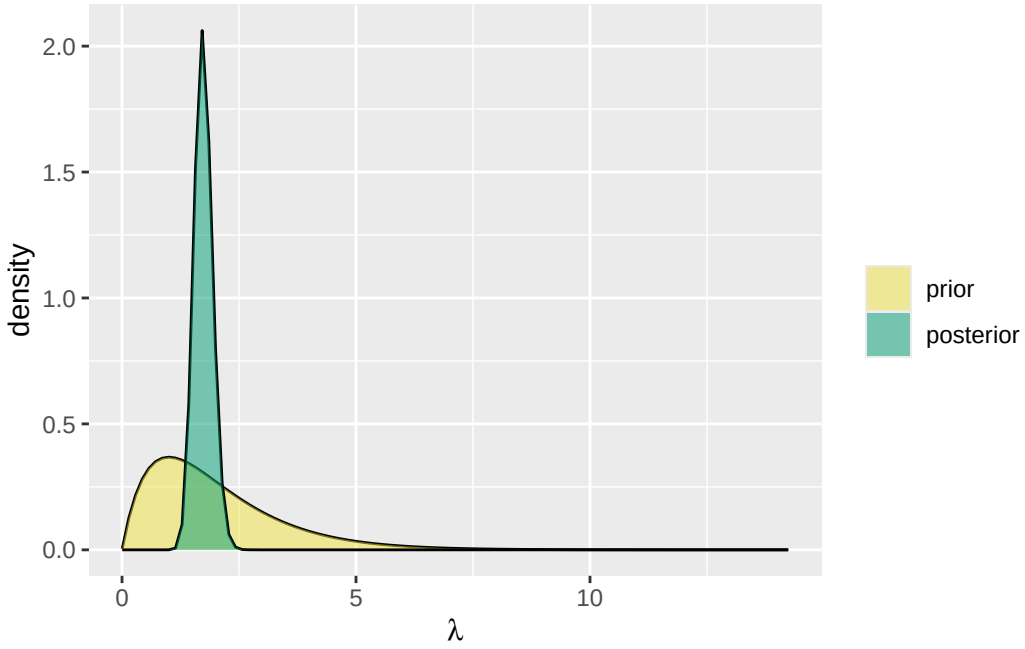
Los datos siguientes se corresponden con el número de llegadas al cajero de un supermercado

```
llegadas <- c(0, 0, 0, 0, 0,
0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2,
2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5)
```

Se considera que la variable X = "Número de llegadas" sigue una distribución de Poisson donde la distribución a priori del parámetro λ sigue una distribución $G(2, 1)$.

Para la obtención de las distribuciones a priori y a posteriori tenemos:

```
plot_gamma_poisson(
    shape = 2,
    rate = 1,
    sum_y = sum(llegadas),
    n = 45,
    prior = TRUE,
    likelihood = FALSE,
    posterior = TRUE
)
```



Para las medidas a posteriori usamos:

```
summarize_gamma_poisson(shape = 2, rate = 1, sum_y = sum(llegadas), n = 45)
```

```
      model shape rate    mean     mode        var        sd
1     prior     2    1 2.00000 1.000000 2.00000000 1.4142136
2 posterior    80   46 1.73913 1.717391 0.03780718 0.1944407
```

Los extremos de un intervalo de credibilidad para λ a un nivel del 95% vienen dados por:

```
qgamma(0.025, shape = 80, rate = 46)
```

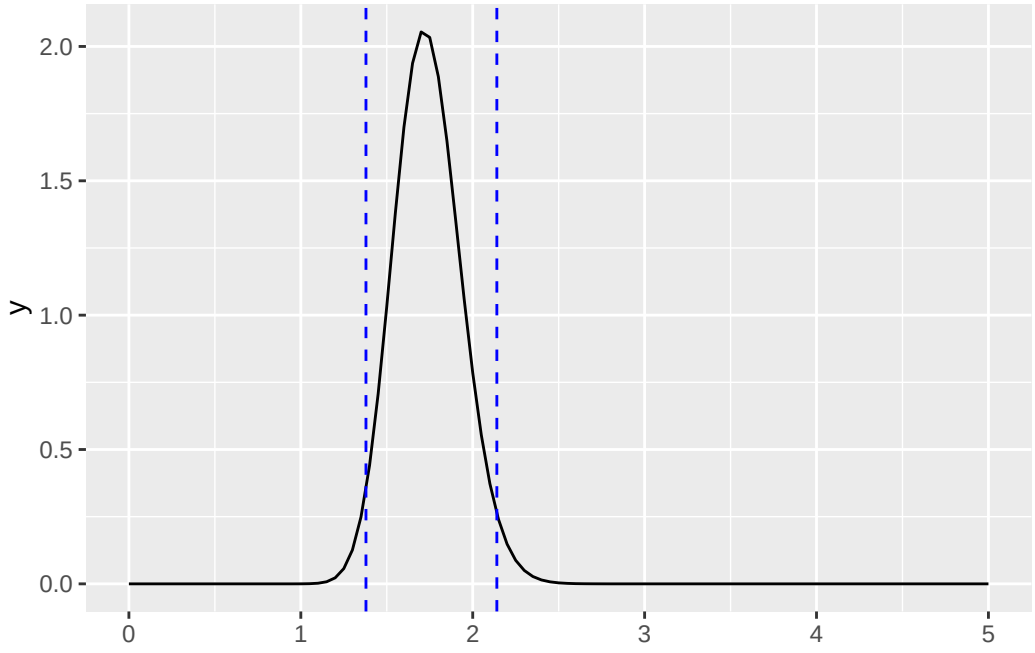
```
[1] 1.379022
```

```
qgamma(0.975, shape = 80, rate = 46)
```

```
[1] 2.140382
```

y gráficamente tenemos:

```r
ggplot() +
    xlim(c(0, 5)) +
    geom_function(
      fun = dgamma,
      args = list(shape = 80, rate = 46)
    ) +
      geom_vline(
          xintercept = qgamma(0.025, shape = 80, rate = 46),
          linetype = 2,
          color = "blue"
      ) +
      geom_vline(
      xintercept = qgamma(0.975, shape = 80, rate = 46),
      linetype = 2,
      color = "blue"
 )
```



Relizar el siguiente ejercicio.

Ejercicio:

Una compañía aseguradora asume que el número de reclamaciones en un año tiene distribución de Poisson, de media λ, y que este número es independiente de un año a otro. Con los datos que tienen hasta el modelo se considera que la distribución a priori de λ sigue una distribución

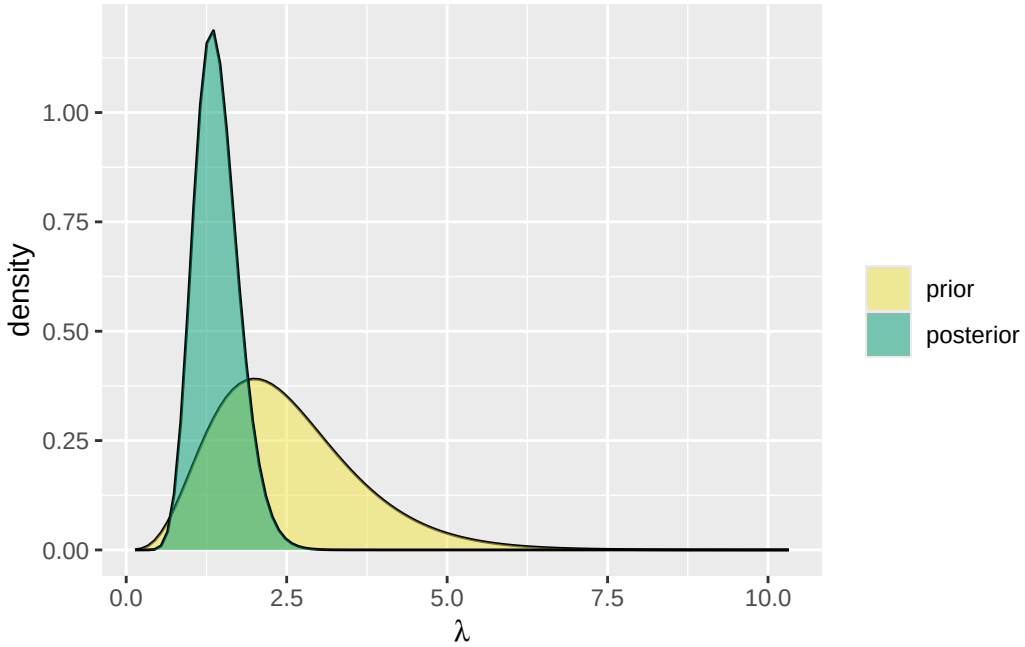
$G(5, 1/2)$. Analizar la distribución a posteriori de un asegurado sabiendo que, en los últimos diez años, el asegurado ha presentado el siguiente número de reclamaciones:

```
nclaims <- c(4, 0, 1, 2, 3, 0, 0, 1, 1, 0)
```

```
# completar aquí
# Parámetros de la distribución previo
shape_prior <- 5
rate_prior <- 1 / 0.5

# Parámetros de la distribución posterior
shape_post <- shape_prior + sum(nclaims)
rate_post <- rate_prior + length(nclaims)

# Gráfica de las distribuciones
plot_gamma_poisson(
  shape = shape_prior,
  rate = rate_prior,
  sum_y = sum(nclaims),
  n = length(nclaims),
  prior = TRUE,
  likelihood = FALSE,
  posterior = TRUE
)
```



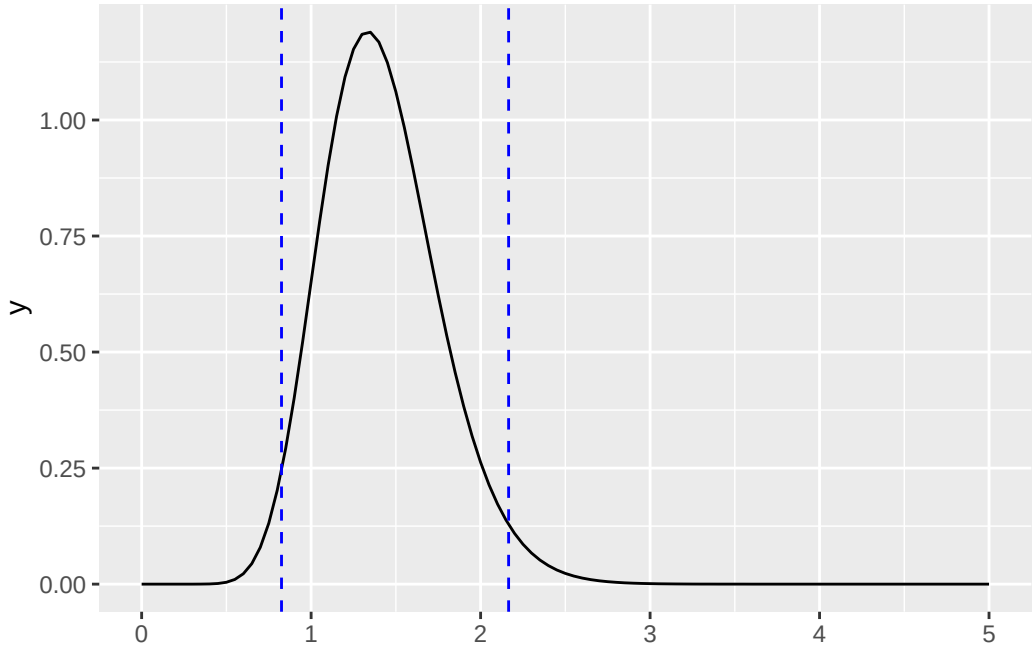
```r
# Intervalo de credibilidad del 95%
qgamma(0.025, shape_post, rate_post)
```

```
[1] 0.8252605
```

```r
qgamma(0.975, shape_post, rate_post)
```

```
[1] 2.16525
```

```r
# Gráfica del intervalo de credibilidad
ggplot() +
  xlim(c(0, 5)) +
  geom_function(
    fun = dgamma,
    args = list(shape = shape_post, rate = rate_post)
  ) +
  geom_vline(
    xintercept = qgamma(0.025, shape_post, rate_post),
    linetype = 2,
    color = "blue"
  ) +
  geom_vline(
    xintercept = qgamma(0.975, shape_post, rate_post),
    linetype = 2,
    color = "blue"
  )
```




```
# fin completar aquí
```

Modelo normal-normal

En este modelo la distribución conjugada de la distribución de normal es la distribución normal para el parámetro μ. Las distribuciones que tenemos son las siguientes.

- **Distribución de la variable** X: $X \sim N(\mu, \sigma^2)$, con σ^2 conocida.
- **Distribución a priori de** μ:

$$\pi(\mu) \propto \exp\left(-\frac{\tau}{2}(\mu - \mu_0)^2\right).$$

- **Distribución a posteriori:**

$$\pi(\mu|\mathbf{x}) \propto \exp\left(-\frac{\tau + n\sigma^2}{2}\left(\mu - \frac{\tau\mu_0 + \sigma^2\sum_{i=1}^n x_i}{\tau + n\sigma^2}\right)^2\right),$$

  es decir sigue una distribución $N\left(\frac{\tau\mu_0+\sigma^2\sum_{i=1}^n x_i}{\tau+n\sigma^2}, \frac{1}{\tau+n\sigma^2}\right)$

Veamos un ejemplo.

Ejemplo:

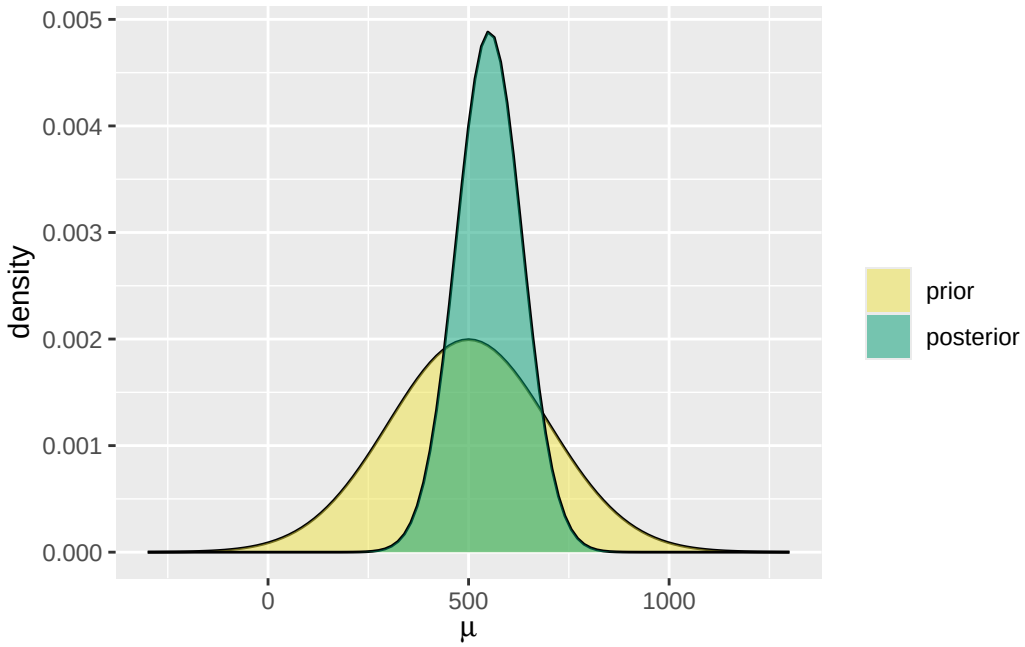
Una de las cantidades de interés en la estadística actuarial, junto con el número de reclamaciones que se reciben, es la cantidad reclamada. Se quiere estudiar el promedio de las reclamaciones de un asegurado. Se asume las cantidades reclamadas siguen una distribución normal con desviación típica conocida e igual a 200 euros. Suponiendo que la distribución a priori de la media de reclamaciones es $N(500, 200^2)$ y conocidas las ultimas reclamaciones de ese cliente:

```
claim <- c(450, 500, 650, 660, 550)
```

Vamos a obtener la distribución a posteriori del promedio de reclamaciones.

Para la obtención de las distribuciones a priori y a posteriori tenemos:

```
plot_normal_normal(
    mean = 500,
    sd = 200,
    sigma = 200,
    y_bar = mean(claim),
    n = length(claim),
    prior = TRUE,
    likelihood = FALSE,
    posterior = TRUE
)
```



Para las medidas a posteriori usamos:

```
l <- summarize_normal_normal(
    mean = 500,
    sd = 200,
    sigma = 200,
    y_bar = mean(claim),
    n = length(claim)
)
l
```

```
      model     mean     mode       var        sd
1     prior 500.0000 500.0000 40000.000 200.00000
2 posterior 551.6667 551.6667  6666.667  81.64966
```

Y para un intervalo de credibilidad al 99

```
qnorm(0.005, mean = l[2, 2], sd = l[2, 5])
```

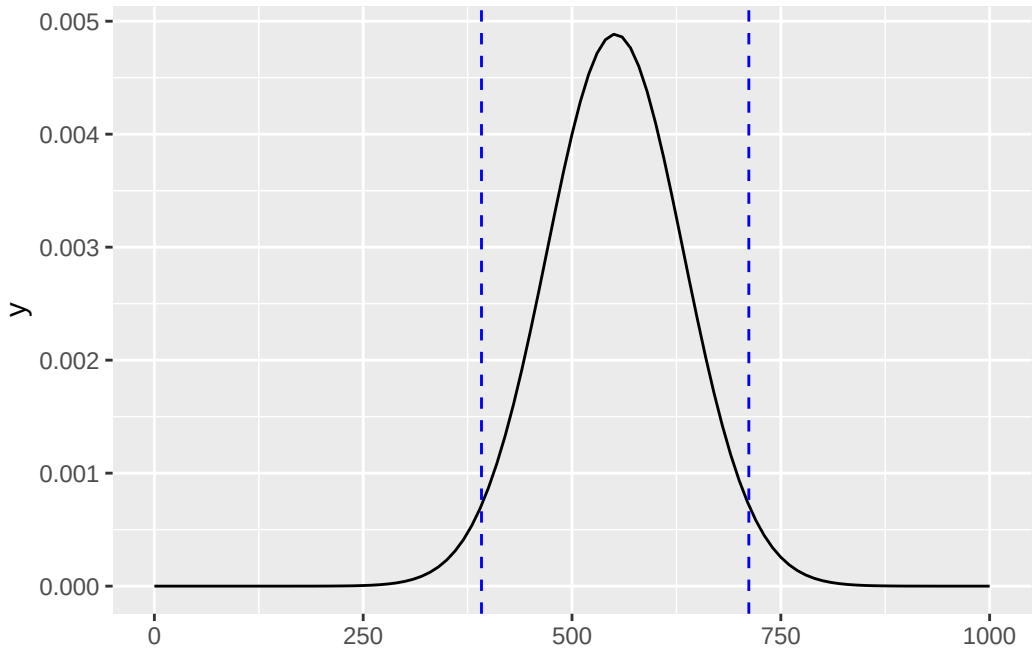
```
[1] 341.3511
```

```
qnorm(0.995, mean = l[2, 2], sd = l[2, 5])
```

```
[1] 761.9822
```

y gráficamente tenemos

```
ggplot() +
    xlim(c(0, 1000)) +
    geom_function(
      fun = dnorm,
      args = list(mean = l[2,2], sd = l[2,5])
    ) +
    geom_vline(
      xintercept = qnorm(0.025, mean = l[2,2], sd = l[2,5]),
      linetype = 2,
      color = "blue"
    ) +
    geom_vline(
      xintercept = qnorm(0.975, mean = l[2,2], sd = l[2,5]),
      linetype = 2,
      color = "blue"
  )
```



Realizar el siguiente ejercicio.

Ejercicio:

Se propone comparar dos métodos de producción de cierta sustancia. Se realizan 16 mediciones y se calcula la diferencia de rendimiento entre ambos métodos (el del primero menos el del segundo). Los datos obtenidos son los siguientes:

```
diferencia <- c(3.7, -6.7,
-10.5, -6.1, -17.6, 2.3, -7.9, -8.9, -4.5, -7.7, -9.4, -10.4, -10.9,
-9.3, -16.7, -7.2)
```

Se considera que la diferencia de mediciones sigue una distribución normal con varianza conocida e igual a 25, donde la distribución a priori de μ es $N(0, 100)$. Analizar la distribución a posteriori y comentar los resultados.

```
# completar aquí
# Parámetros conocidos
mu_0 <- 0                 # Media de la distribución a priori
sd_prior <- 10            # Desviación estándar de la distribución a priori
sigma <- 5                # Desviación estándar conocida de la distribución normal
y_bar <- mean(diferencia) # Media de los datos observados
n <- length(diferencia)   # Número de observaciones

# Parámetros de la distribución a posteriori
```

```r
l <- summarize_normal_normal(
  mean = mu_0,
  sd = sd_prior,
  sigma = sigma,
  y_bar = y_bar,
  n = n
)

print(l)
```

```
      model      mean      mode        var        sd
1     prior  0.000000  0.000000 100.000000 10.000000
2 posterior -7.864615 -7.864615   1.538462  1.240347
```

```r
# Cálculo del intervalo de credibilidad al 99%
lower_bound <- qnorm(0.005, mean = l[2, 2], sd = l[2, 5])
upper_bound <- qnorm(0.995, mean = l[2, 2], sd = l[2, 5])

# Mostrar los resultados del intervalo
print(lower_bound)
```

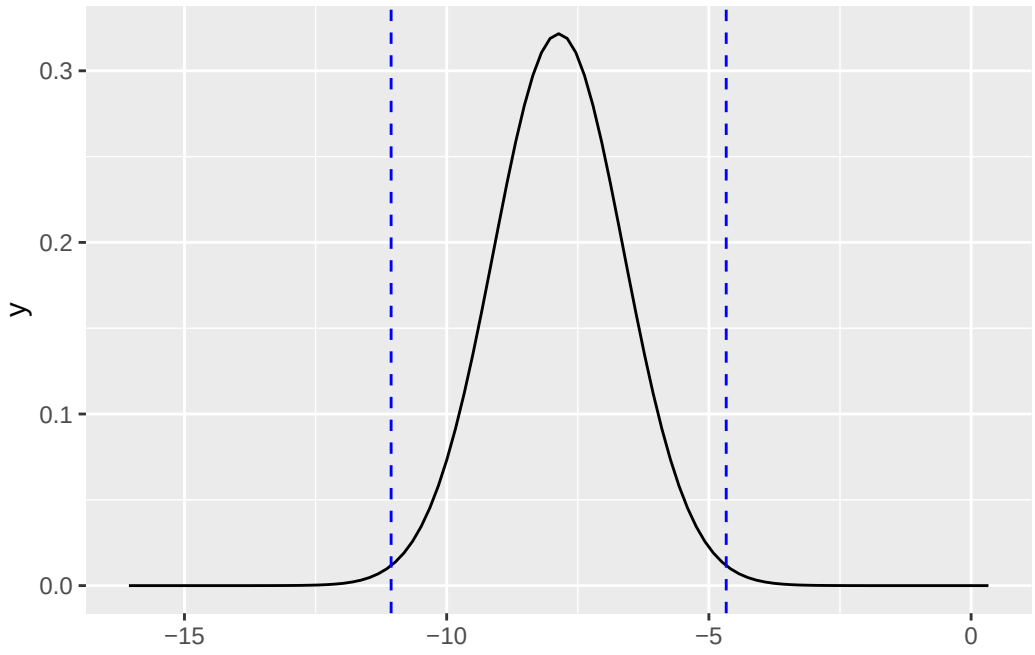
```
[1] -11.05954
```

```r
print(upper_bound)
```

```
[1] -4.669692
```

```r
# Gráfica de la distribución posterior
ggplot() +
  xlim(c(lower_bound - 5, upper_bound + 5)) +
  geom_function(
    fun = dnorm,
    args = list(mean = l[2, 2], sd = l[2, 5])
  ) +
  geom_vline(
    xintercept = lower_bound,
    linetype = 2,
    color = "blue"
  ) +
  geom_vline(
    xintercept = upper_bound,
    linetype = 2,
```

```
    color = "blue"
  )
```




```
# fin completar aquí
```