Session 2: Posterior, and posterior predictive distributions

> **Exercise 2.1**
>
> Sèpia Verda: Estimating a (Poisson) frequency. There is a cultural association called *La Sèpia Verda*, and its members don't know the expected number of weekly visitors to their web page. For this purpose, they register the number weekly visitors in the last 10 weeks. This data can be found in the file *sepiaverda.txt*. If the members of the association believe that the number of visitors will rarely fall under 5 and above 40:

> a. Choose the parameters of a conjugate prior distribution, and explain why you choose them (it might be useful to draw the prior predictive distribution to back your choice up).

Statistical Model : $Y \sim Poisson(\lambda),\ P(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y\,!}$

Prior : $\lambda \sim \Gamma(\alpha, \beta),\ \Pi(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda},\ \lambda > 0$

$$E[\lambda] = \alpha/\beta$$

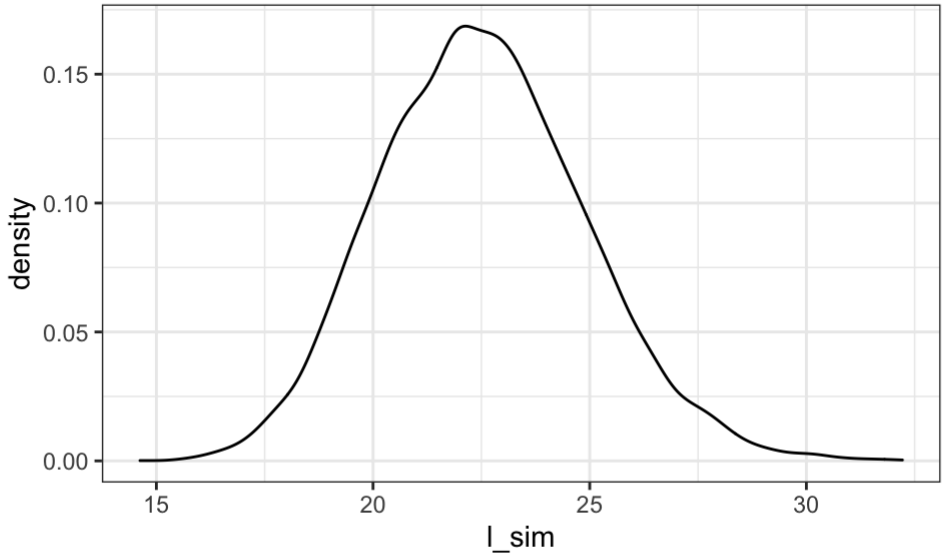
$$Var[\lambda] = \alpha/\beta^2$$

Prior information: *the number of visitors will rarely fall under 5 and above 40*

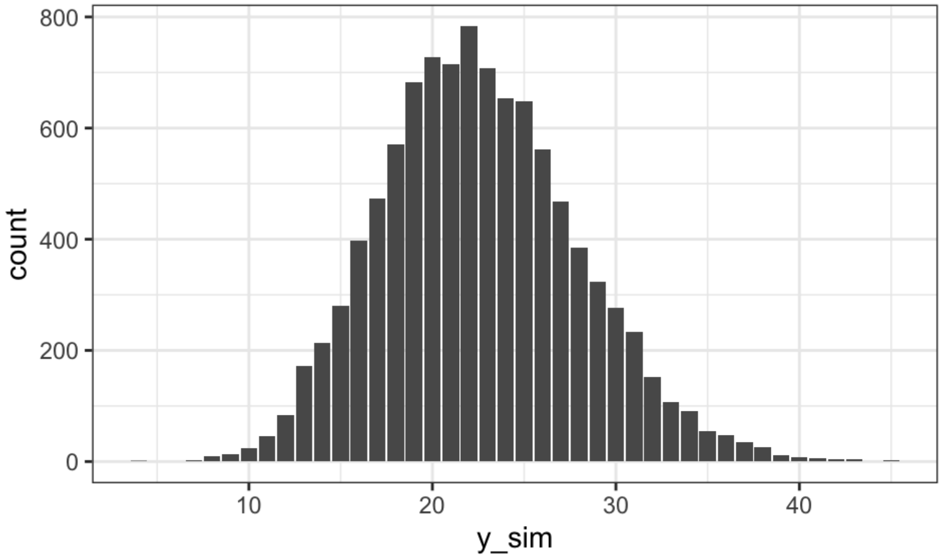
$$5 \leq y \leq 40$$

- The **prior distribution**, $\Pi(\theta)$, summarises the information about the λ parameter, before data
- The **prior predictive distribution**, $P_{\Pi}(\tilde{y})$, summarises the information about the Y variable, before data

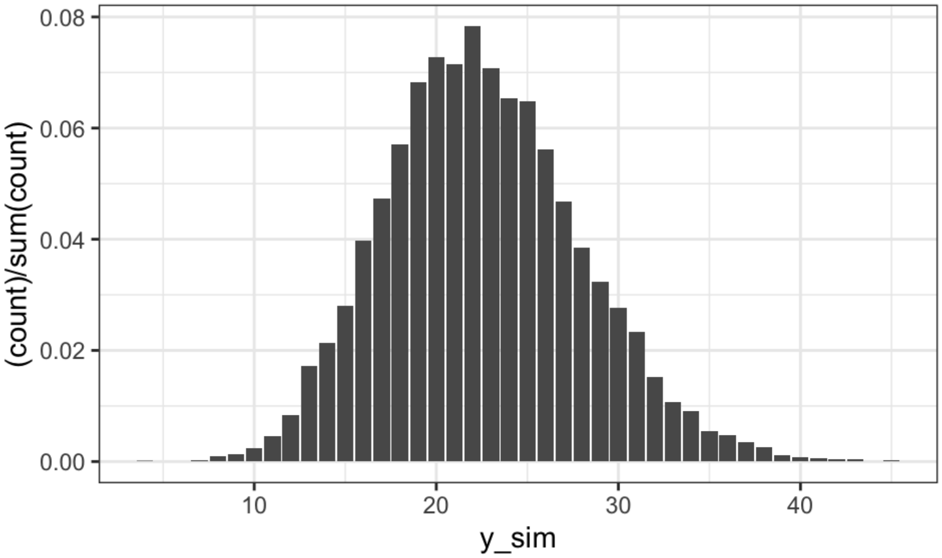
```r
library(ggplot2)
library(dplyr)
library(tidyr)

theme_set(theme_bw())

## a) Prior definition
set.seed(123456)

# Simulate 10000 draws from prior density: l_sim
prior <- c(90, 4)
l_sim <- rgamma(10000, shape = prior[1], rate = prior[2])
ggplot(tibble(l_sim), aes(l_sim)) +
  geom_density()
```



```r
# Simulate 10000 draws from the predictive density: y_sim
y_sim <- rpois(10000, l_sim)

# Plot the prior predictive density
ggplot(tibble(y_sim), aes(y_sim)) +
  geom_bar()
```



```
# Prior predictive with frequency in y axis:
prior_pred_sim <- ggplot(tibble(y_sim), aes(y_sim)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))
prior_pred_sim
```



```
# Compute the probability of less than 5
mean(y_sim < 5)
```

```
[1] 1e-04
```

```
# Compute the probability of more than 40
mean(y_sim > 40)
```

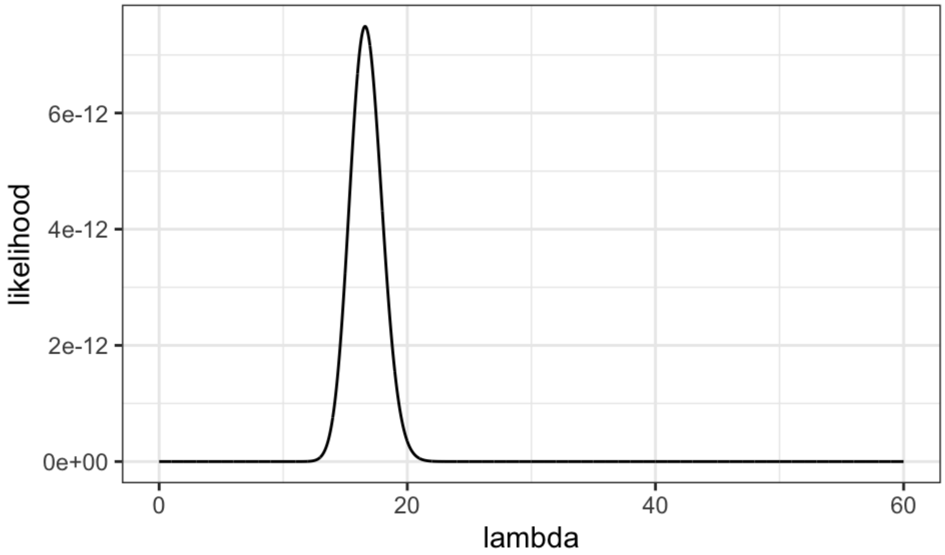
```
[1] 0.0018
```

> b. Draw in the same graph the prior distribution and the likelihood function.

Given $Y = y_1, \ldots, y_n$ i.i.d.:

Likelihood : $P(y|\lambda) = \prod_{i=1}^{n} P(y_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda_i^y}{y_i\,!} = \frac{e^{-\lambda n}\lambda^{\Sigma\, y_i}}{\prod(y_i\,!)} \propto e^{-\lambda n}\lambda^{\Sigma\, y_i}$

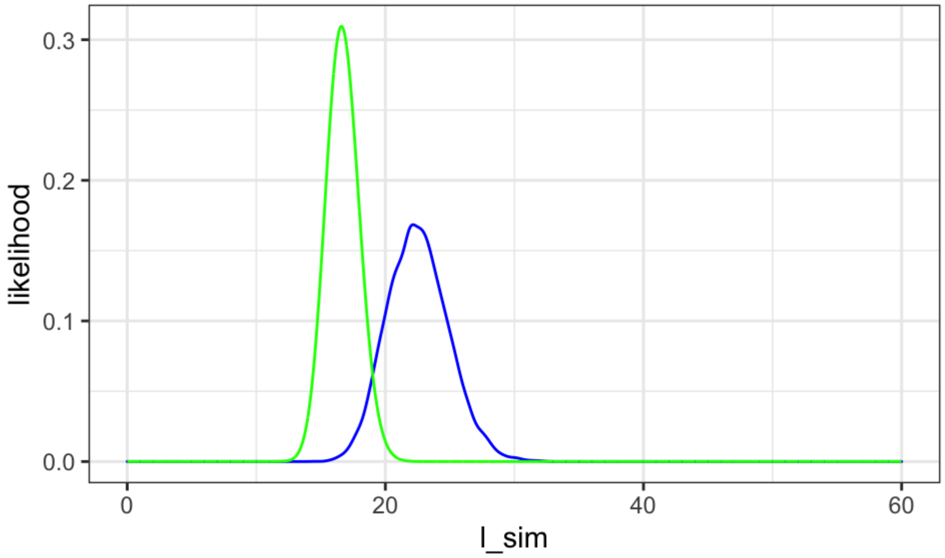
```
## b) Likelihood calculation

library(purrr)

# Likelihood
y <- c(21, 17, 17, 19, 16, 18, 15, 10, 17, 16)

delta_l <- 0.01

lambda <- seq(0, 60, delta_l)
likelihood <- map_dbl(lambda, ~ prod(dpois(y, .x)))
df <- tibble(lambda, likelihood)

# Likelihood graph
ggplot(df) +
  geom_line(aes(x = lambda, y = likelihood))
```



The likelihood function is not a density function (its integral is not 1).

To be able to represent this function together with the *priori* distribution, it is convenient to **standardize the likelihood**.

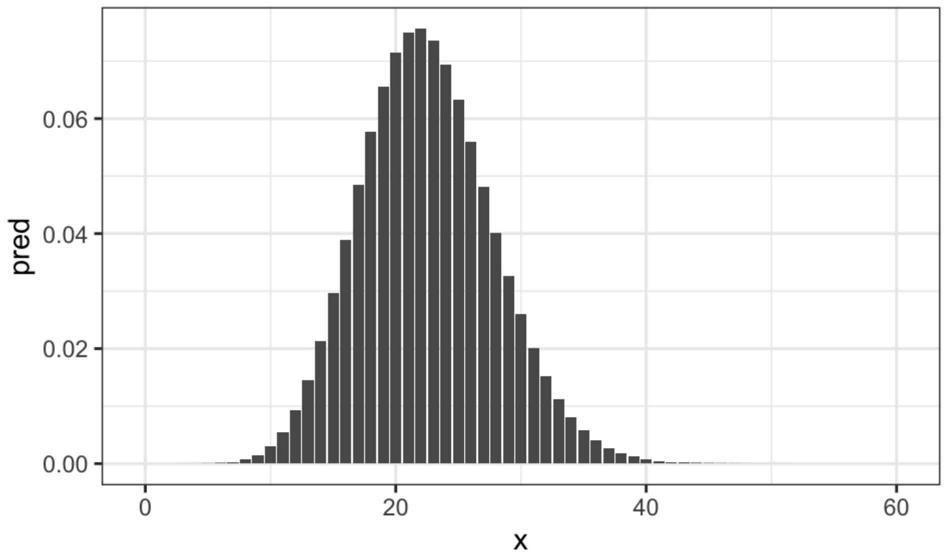
```
# Standardized likelihood
coef <- sum(likelihood)*delta_l
df$likelihood <- df$likelihood/ coef

# Prior + likelihood graph
ggplot(tibble(l_sim), aes(l_sim)) +
  geom_density(col="blue") +
  geom_line(data = df,
    aes(x = lambda, y = likelihood), col="green")
```



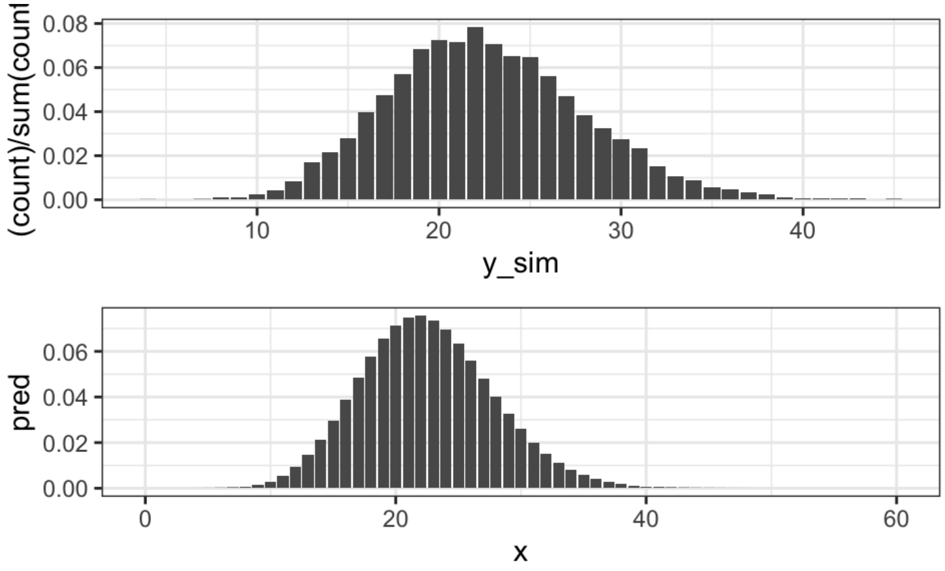
> c. Draw the prior predictive distribution.

Prior Predictive Distribution (conjugate distributions): $\tilde{y} \sim NB(\alpha, \beta/(1+\beta))$

```
# b) By definition
x <- 0:60
pred <- dnbinom(x, prior[1], prior[2]/(1 + prior[2]))
df2 <- tibble(x, pred)
prior_pred_def <- ggplot(df2) +
  geom_col(aes(x, pred))

prior_pred_def
```



```r
library(gridExtra)

grid.arrange(prior_pred_sim, prior_pred_def, ncol = 1)
```



1 Posterior distribution $\Pi(\theta|y)$

We have:
- a statistical model $M = \{p(y|\theta), \theta \in \Omega\}$ and
- a prior distribution $\pi(\theta)$

That's equivalent to $\pi(\theta)p(y|\theta) = f_\Omega(\theta, y)$ (joint distribution).

The **Bayes theorem** states that to compute the posterior distribution:

$\pi(\theta|y) = \frac{f(y,\theta)}{P_\pi(y)} = \frac{\pi(\theta)P(y|\theta)}{P_\pi(y)}$

Where $P_\pi(y)$ (prior predictive distribution) is a constant needed to that the posterior $\pi(\theta|y)$ integrates to 1.

$P_\pi(y) = \int_\Omega P(y|\theta)\pi(\theta)d\theta$

Sometimes you don’t need to integrate. If you compute $\pi(\theta)P(y|\theta)$ and you recognize the distribution, then $\frac{1}{P_\pi(y)}$ will be the constant that gets $\pi(\theta|y)$ to integrate to one.

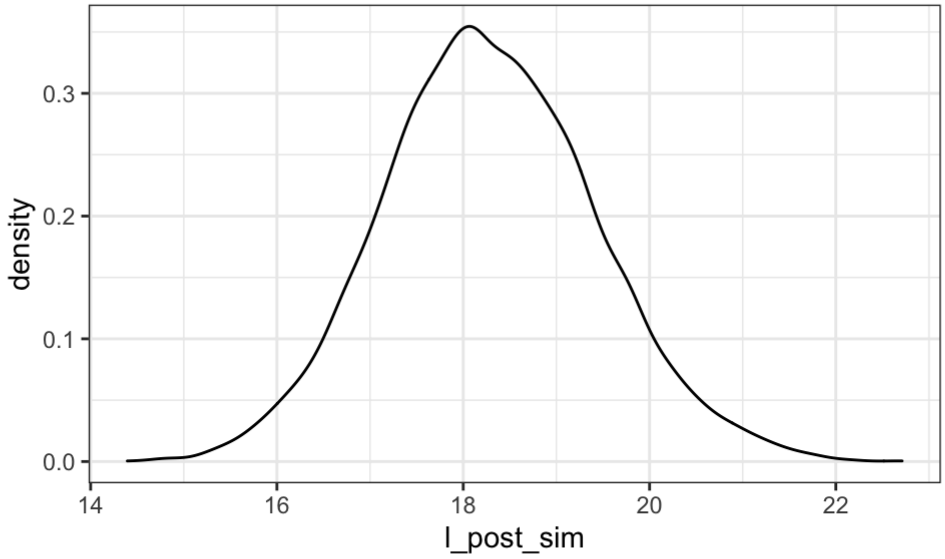
$\pi(\theta|y) = \pi(\theta)\frac{1}{P_\pi(y)}P(y|\theta) \;\propto\; \pi(\theta)\ell_y(\theta)$ where the second term is a version of the likelihood function and $P(y|\theta)$ your statistical model.

Exercise (continued)

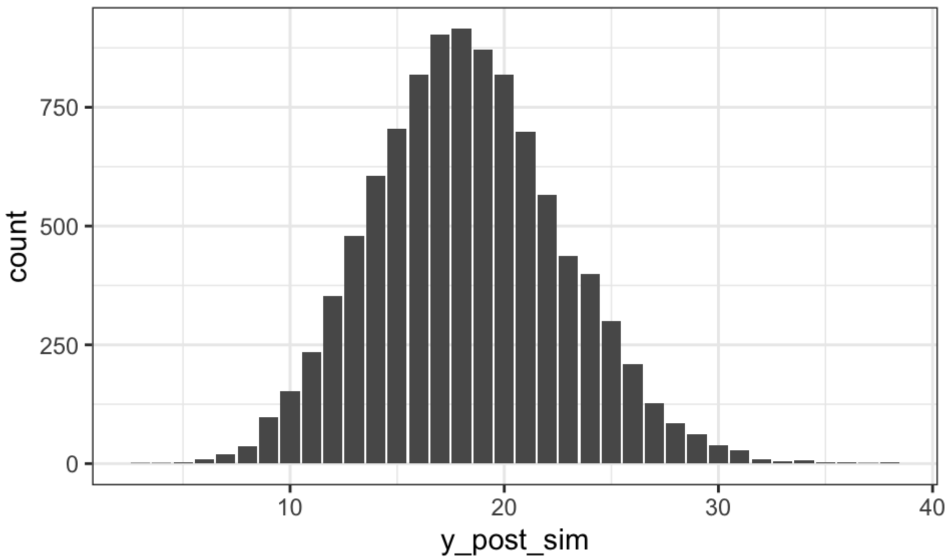
Posterior : $\lambda|y_1 \dots y_n \sim \Gamma(\alpha + \sum y_i, \beta + n)$

d. Draw the posterior distribution, and give a point estimate and a 95% credible interval for the expected value of the number of weekly visitors.

```
## d) Posterior calculation (conjugated)

# Simulate 10000 draws from posterior density: l_post_sim
n <- length(y)

posterior <- c(prior[1] + sum(y), prior[2] + n)
l_post_sim <- rgamma(10000, shape = posterior[1], rate = posterior[2])
ggplot(tibble(l_post_sim), aes(l_post_sim)) +
  geom_density()
```



```
# Simulate 10000 draws from the posterior predictive density: y_sim
y_post_sim <- rpois(10000, l_post_sim)

# Plot the posterior predictive density
ggplot(tibble(y_post_sim), aes(y_post_sim)) +
  geom_bar()
```



```
# Point estimate of the number of weekly visitors
mean(y_post_sim)
```

```
[1] 18.3089
```

```
median(y_post_sim)
```

```
[1] 18
```

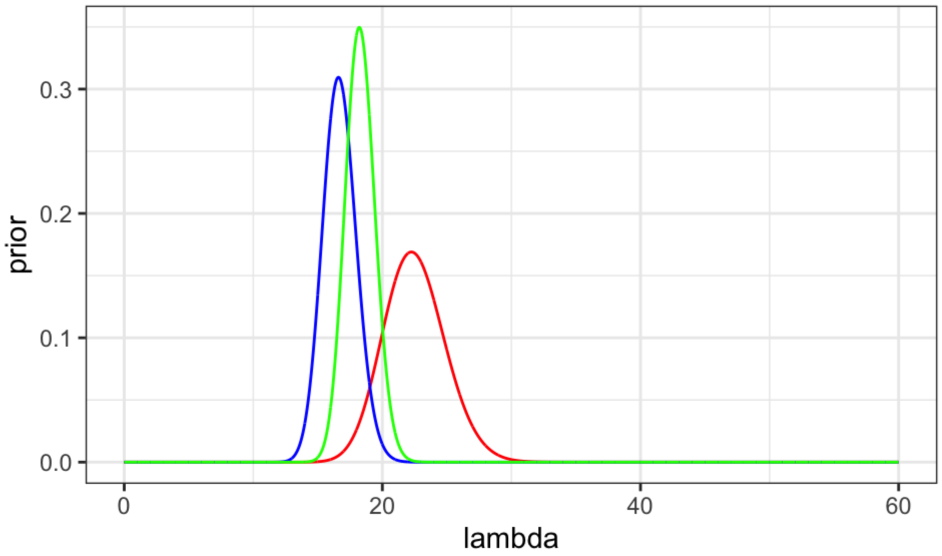
```
# 95% credible interval
quantile(y_post_sim, c(0.025, 0.975))
```

```
 2.5% 97.5% 
   10    27 
```

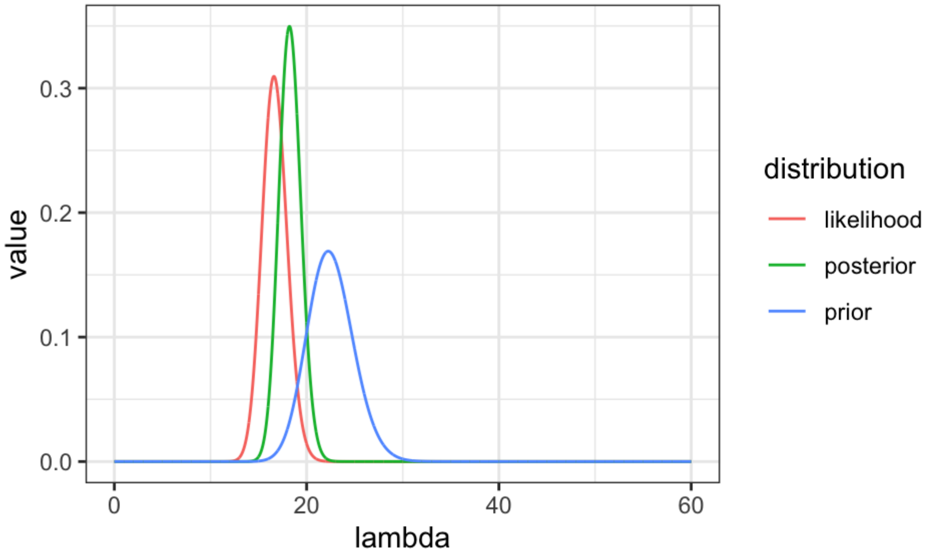
```
## d) Posterior calculation (revisited)

#lambda <- seq(0, 60, 0.01)
df <- df %>%
    mutate(
      prior = dgamma(lambda, shape = 90, rate = 4),
      product = prior * likelihood,
      posterior = product / (sum(product)*0.01))

df
```

```
# A tibble: 6,001 × 5
   lambda likelihood     prior product posterior
    <dbl>      <dbl>     <dbl>   <dbl>     <dbl>
 1   0     0         0               0         0
 2   0.01  0         8.92e-261       0         0
 3   0.02  0         5.30e-234       0         0
 4   0.03  0         2.40e-218       0         0
 5   0.04  0         3.03e-207       0         0
 6   0.05  0         1.23e-198       0         0
 7   0.06  0         1.32e-191       0         0
 8   0.07  0         1.15e-185       0         0
 9   0.08  0         1.60e-180       0         0
10   0.09  1.15e-305 5.48e-176       0         0
# … with 5,991 more rows
```

```
# Option 1
ggplot(df) +
  geom_line(aes(lambda, prior), col="red") +
  geom_line(aes(lambda, likelihood), col = "blue") +
  geom_line(aes(lambda, posterior), col = "green")
```



```
# Option 2:
df %>%
    select(-product) %>%
    gather(key = distribution, value = value, -lambda) %>%
    ggplot() +
    geom_line(aes(x = lambda, y = value, col = distribution))
```



> e. Compute the probability that the number of visitors next week will be lower than 10.

$$P_\Pi(\tilde{y} < 10|y) = \sum_{i=0}^{9} P_\Pi(\tilde{y} = i|y)$$

$P_\Pi(\tilde{y} < 10|y) =$ 0.0156347

```
## e) Probability that the number of visitors next week will be lower than 10
# Option 1) posterior predictive distribution (analytical expression)
sum(dnbinom(0:9, posterior[1], posterior[2]/(1+posterior[2])))
```

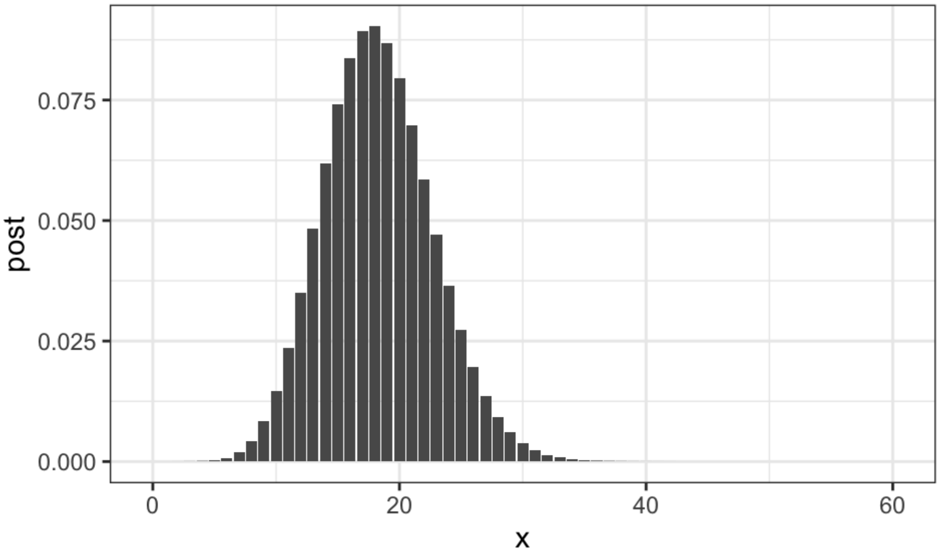
```
[1] 0.01563471
```

```
# Option 2) simulated y
mean(y_post_sim < 10)
```

```
[1] 0.0166
```

> f. Draw the posterior predictive distribution.

```
# b) By definition
x <- 0:60
post <- dnbinom(x, posterior[1], posterior[2]/(1 + posterior[2]))
df <- tibble(x, post)
ggplot(df) +
  geom_col(aes(x, post))
```



> g. Imagine that the members of the cultural association want to do a bet about the number of visitors next week. For what option will you bet?

$$H_1 : \tilde{y} < 15$$

$$H_2 : 15 \leq \tilde{y} < 25$$

$$H_3 : \tilde{y} \geq 25$$

The probabilities of each of the hypotheses can be calculated:

$P(H_1|y) = P_\Pi(\tilde{y} < 15|y) =$ 0.199059
$P(H_2|y) = P_\Pi(15 \leq \tilde{y} < 25|y) =$ 0.7156834
$P(H_3|y) = P_\Pi(\tilde{y} \geq 25|y) =$ 0.0852576

Therefore, we would be left with H_2, which has a higher probability.

```
## g) Option 1: conjugate (analytical expression)
sum(dnbinom(0:14, posterior[1], posterior[2]/(1+posterior[2])))
```

```
[1] 0.199059
```

```
sum(dnbinom(15:24, posterior[1], posterior[2]/(1+posterior[2])))
```

```
[1] 0.7156834
```

```
1 - sum(dnbinom(0:24, posterior[1], posterior[2]/(1+posterior[2])))
```

```
[1] 0.0852576
```

```
## g) Option 2: simulation
mean(y_post_sim < 15)
```

```
[1] 0.199
```

```
mean(y_post_sim < 25) - mean(y_post_sim< 15)
```

```
[1] 0.7132
```

```
mean(y_post_sim >= 25)
```

```
[1] 0.0878
```

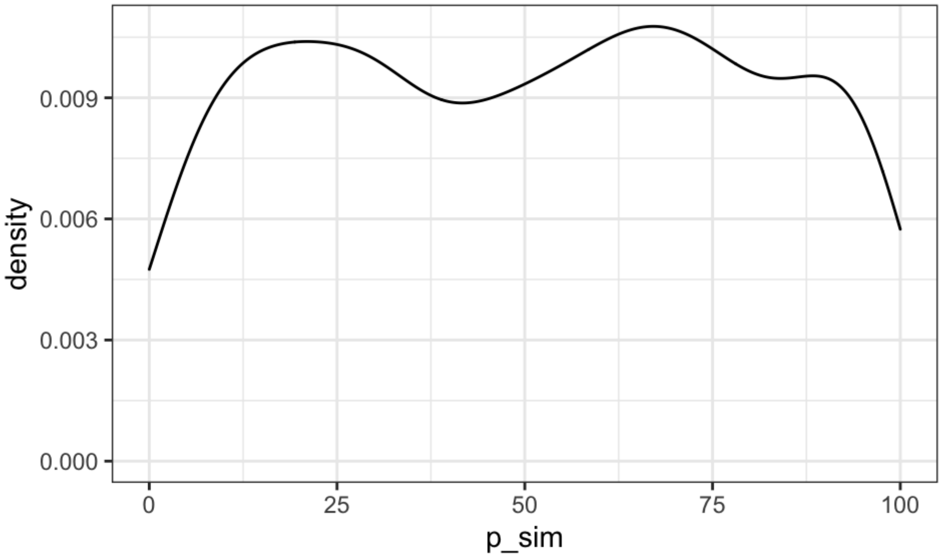
> Now, assume that the members of the association know nothing about the number of weekly visitors. They assume that all the possible values for the poisson’s parameter, λ, are equally likely and hence use a flat prior, π(λ)=1 for λ>0, which is improper.
>
> h. Compute and draw in the same graph the prior distribution, the likelihood function and the posterior distribution.

$\Pi(\theta) = 1$. This prior is not a density function, but it can be used to calculate the posterior distribution:

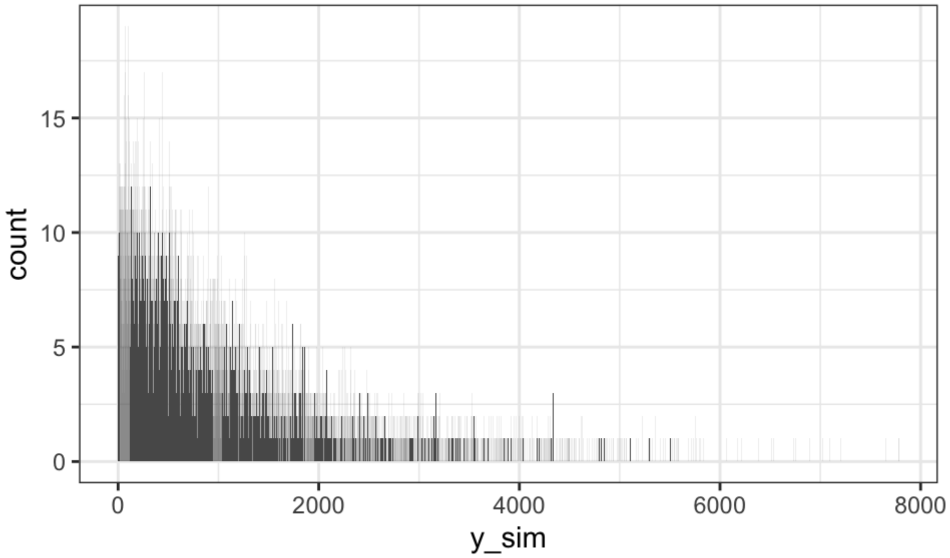
$$\Pi(\theta|y) = \frac{\ell_y(\theta)\Pi(\theta)}{\int \ell_y(\theta)\Pi(\theta)d\theta} = \ell_y^{std}(\theta)$$

Other option should be define a prior distribution with a large variance:

```
## Predictive simulation

# Simulate 10000 draws from prior density: p_sim
p_sim <- rgamma(10000, shape = 1, rate = 0.001)
ggplot(tibble(p_sim), aes(p_sim)) +
  geom_density() +
  scale_x_continuous(limits = c(0,100))
```



```
# Simulate 10000 draws from the predictive density: y_sim
y_sim <- rpois(10000, p_sim)

# Plot the prior predictive density
ggplot(tibble(y_sim), aes(y_sim)) +
  geom_bar()
```


count
y_sim