04_priors

June 5, 2024

0.0.1 Intended Learning Outcomes

At the end of this part you will be able to: * describe the pros and cons of using different priors (e.g. elicited, conjugate, …), * evaluate the interplay between prior and posterior distributions, * calculate several quantities of interest from posterior distributions, * apply Bayesian inference to estimate population variation data.

One of the main feature of Bayesian statistics is that we assign probability distributions not only to data variables y but also to parameters θ. We quantify whatever feelings or believes we have about θ before observing y.

Using Bayes' theorem, we obtain a posterior distribution of θ, a blend of the information between the data and the prior.

How can we decide which prior distribution is more appropriate in our study?

Prior distributions can * be derived from past information or personal opinions from experts; * be distributed as familiar distribution functions; * bear little information.

0.0.2 Elicited priors

The simplest approach to specify $\pi(\theta)$ is to define the collection of θ which are possible.

Then one can assign some probability to each one of these cases and make sure that they sum to 1.

If θ is discrete, this looks like a natural approach.

Imagine that your prior distribution describes the number of kits a mother rabbit will have in the next litter.

Perhaps, you want to make some inference on the biological mechanisms for the number of kits. In this case you may have a likelihood function relating some observations $\vec{y}$ (e.g. genetic or environmental markers) to the number of kits θ.

θ is clearly discrete and you may have some past information on its distribution from the literature.

"Rabbits can have anywhere from one to 14 babies, also called kits, in one litter. An average litter size is 6. Hereditary and environmental factors play a role in the number of kits born in a litter."

$$\pi(\theta = 0) = \pi(\theta > 14) = 0 \tag{1}$$

If it is more probable that a mother will have 6 kits, as this is the average litter size based on past information, then

$$\pi(\theta = 2) < \pi(\theta = 6) > \pi(\theta = 10) \tag{2}$$

We must ensure that

$$\sum_{i=1}^{14} \pi(\theta = i) = 1 \tag{3}$$

On the other hand, it θ is continuous, a simple solution would be to discretise the prior distribution by assigning masses to intervals. In other words, you create a histogram prior for θ.

Imagine that your prior distribution specifies the recorded temperature in hot springs at Lassen Volcanic National Park. Specifically, you are interested in relating the temperature of different pools at Bumpass Hell with the occurrence of certain extremophile micro-organisms, capable of surviving in extremely hot environments.

You want to assign a prior distribution for the pool temperature, θ. Clearly θ is continuous.

From past observations, we know that pool temperatures, θ have a range of $(80, 110)$ with an average of 88, in Celsius degrees.

A simple solution would be to derive a prior histogram of θ, as

$$\pi(80 \leq \theta < 85) < \pi(85 \leq \theta < 90) > \pi(90 \leq \theta < 95) \tag{4}$$

Again, you have to make sure that all these probabilities sum to 1.

Furthermore, it is important that the histogram is sufficiently wide, as the posterior will have support only for values that are included in the prior.

Alternatively, we may assume that the prior distribution for θ belongs to a parametric distributional family $\pi(\theta|\nu)$.

Here we choose ν so that $\pi(\theta|\nu)$ closely matches our elicited beliefs.

This approach has several advantages:

- it reduces the effort to the elicitee (you don't have to decide a probability for each value θ can have);
- it overcomes the finite support problem (as in the case of the histogram);
- it may lead to simplifications in the computation of the posterior (as we will see later on).

A limitation of this approach is that it would be impossible to find a distribution that perfectly matches the elicitee's beliefs.

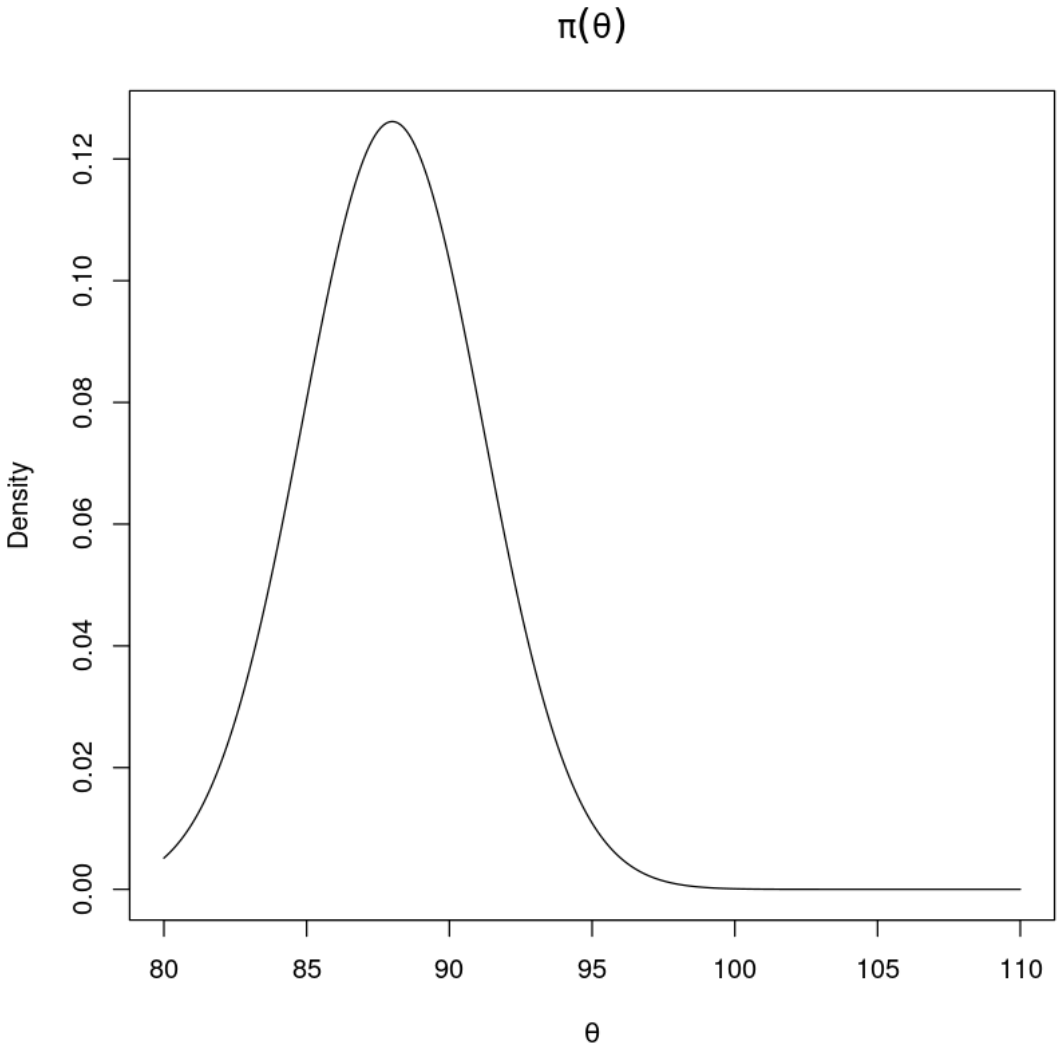
For instance, the prior for temperatures could be Normally distributed as $N(\mu, \sigma^2)$ bounded at $(80, 110)$, that is

$$\pi(\theta) = 0 \qquad \text{for } \theta < 80 \text{ or } \theta > 110 \tag{5}$$

$$\pi(\theta) = N(\mu, \sigma^2) \qquad \text{for } 80 \leq \theta \leq 110 \tag{6}$$

with $\mu = 88$ and $\sigma^2 = 10$.

[1]:
```
## elicited prior
mu <- 88
sigma2 <- 10
x <- seq(80,110,0.1)
plot(x=x, dnorm(x=x, mean=mu, sd=sqrt(sigma2)), type="l", lty=1, ylab="Density",
    xlab=expression(theta), main=expression(pi(theta)))
```



Note that this distribution is not defined outside the interval $(80, 110)$. As such, the posterior distribution will not have mass outside this interval. Overconfidence may result into failing to condition on events outside the range of personal experience or previous observations. For instance, the fact that a temperature lower than 80 has never been observed may be better represented by setting a small (but greater than 0) probability of occurring.

As a rule of thumb, for elicited priors, it is recommended to focus on quantiles close to the middle of the distribution (e.g. the 50^{th}, 25^{th} and 75^{th}) rather than extreme quantiles (e.g. the 95^{th} and 5^{th}). You should also assess the symmetry of your prior.

Elicited priors can be updated and reassessed as new information is available.

They are very useful for experimental design where some ideas on the nature of the studied system is given in input.

0.0.3 Conjugate priors

When choosing a prior distribution $\pi(\theta|\nu)$ some family distributions will make the calculation of posterior distributions more convenient than others will do.

It is possible to select a member of that family that is *conjugate* with the likelihood $f(y|\theta)$, so that the posterior distribution $p(\theta|y)$ belongs to the same distributional family as the prior.

Suppose we are interested in modelling the arrival of herds of elephants to a specific water pond in the savannah in a day during the migratory season. We may be interested in this estimate for tracking migratory routes or assessing population sizes.

Y is the count of distinct elephant herds or groups (not the total number of elephants!) arriving at the pool in a day during the migration season (not during the whole year!).

A Poisson distribution has a natural interpretation to model arrival rates for discrete variables.

Indeed, the Poisson distribution is a discrete probability distribution that gives the probability of a given number of events occurring in a fixed interval of time (or space) when such events occur independently and with a known average rate.

The Poisson distribution is an appropriate model under certain assumptions: 1. Y is the number of times an event occurs in an interval and it can take values any positive integer values including 0; 2. the occurrence of one event does not affect the probability that a second event will occur (i.e. events occur independently); 3. the rate at which events occur is constant (it cannot be higher in some intervals and lower in other intervals); 4. two events cannot occur at exactly the same instant; 5. the probability of an event in an interval is proportional to the length of the interval.

Condition 1 is clearly met in our case. Conditions 2 and 4 assumes that different herds do not follow each other (perhaps by taking different routes). For the sake of illustrating this distribution, we will assume this to be true. You can see that if Y were the number of elephants (not the herds) then condition 2 is not met as elephants tend to migrate in group. Condition 3 is met when we focus our analysis on the annual period where we expect to see herds, not during the whole year. Condition 5 is easily met, as the number of herds arriving in a week is likely to be higher than the number in a day. If we assume that all these conditions are true, then Y, the number of elepehant herds arriving at a pool, is a Poisson random variable.

The event Y can occur $0, 1, 2, ...$ times in the interval. The average number of events in an interval is the parameter θ.

The probability of observing y events in an interval is given by

$$f(y|\theta) = \frac{e^{-\theta}\theta^y}{y!},\ y \in \{0, 1, 2, ...\},\ \theta > 0 \tag{7}$$

Here θ is the event rate, or the rate parameter.

Note that the equation above is a probability mass function (pmf), as it is defined only for discrete values of y.

Note that the parameter of the Poisson distribution is typically written as λ.

This is our **likelihood** distribution and, once we know θ, then the whole distribution is defined.

θ has to be positive (not necessarily an integer) and y is a positive integer.

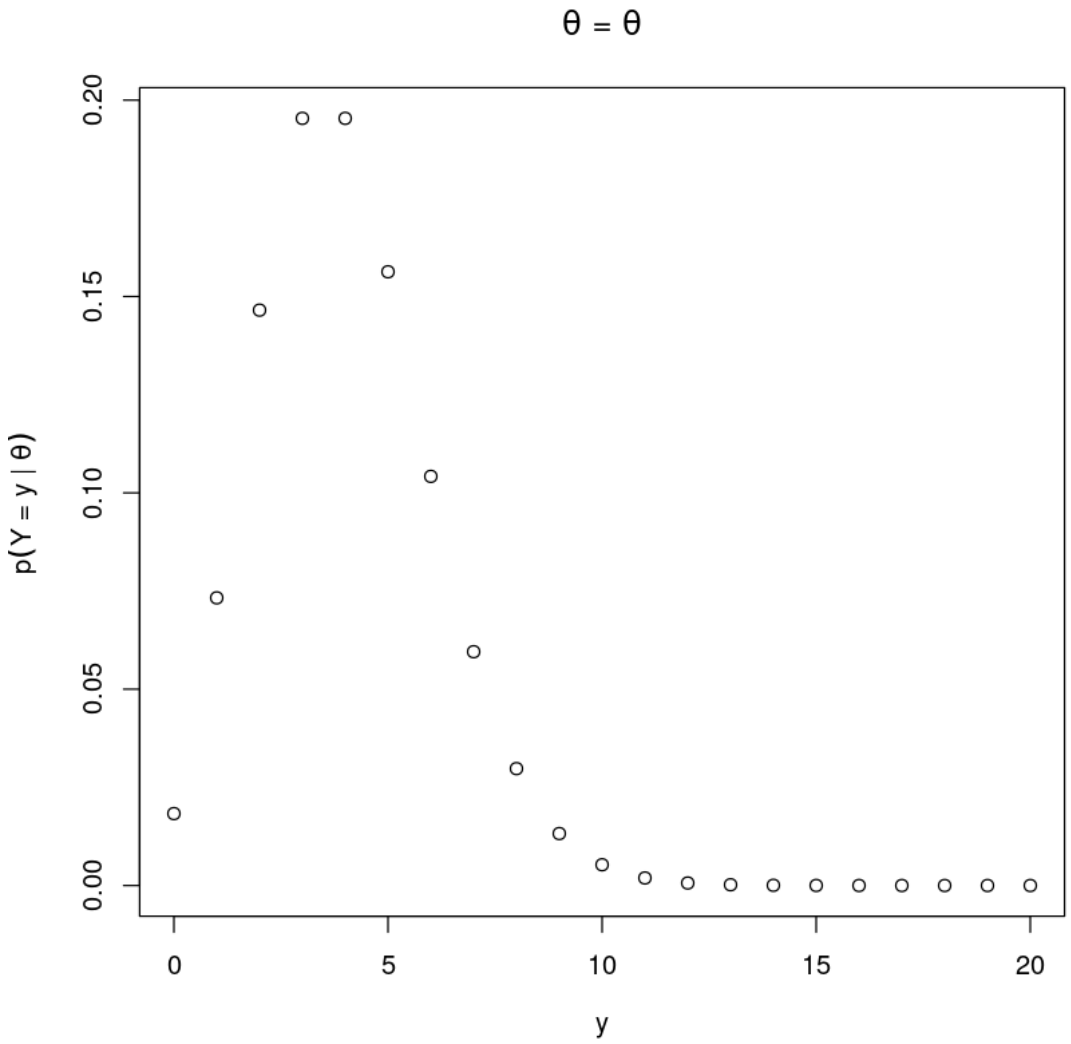
Assuming that the rate θ is set to 4 (4 herds per day during migration season), then:

$$P(y=0) = \frac{e^{-4}4^0}{0!} = e^{-4} = 0.0183 \tag{8}$$

$$P(y=1) = \frac{e^{-4}4^1}{1!} = ... = 0.0733 \tag{9}$$

$$P(y=2) = \frac{e^{-4}4^2}{2!} = ... = 0.147 \tag{10}$$

[2]:
```
## Poisson distribution
theta <- 4
y <- seq(0, 20, 1)
plot(x=y, dpois(x=y, lambda=theta), type="p", lty=1, xlab=expression(y),
    main=expression(theta~"="~theta), ylab=expression(p(Y~"="~y~"|"~theta)))
```



As you can see, the highest mass is towards 4 and above 12 the probability is very close to 0. You may recall that a Poisson distribution has expected value and variance equal to the rate parameter. Note that we have some non-zero probability of observing 0 events.

We now need to define a **prior** distribution for θ having support for positive values.

A reasonable choice is given by the Gamma distribution

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \theta > 0, \alpha > 0, \beta > 0 \tag{11}$$

α is the shape parameter and β is the rate parameter.

$$E(G(\alpha,\beta)) = \alpha\beta \tag{12}$$
$$Var(G(\alpha,\beta)) = \alpha\beta^2 \tag{13}$$

The Gamma distribution is the prior distribution, that is $\theta \sim G(\alpha,\beta)$. The Gamma distribution is a two-parameter family of continuous probability distributions. Please note that the common exponential distribution and chi-squared distribution are special cases of the Gamma distribution.

We have also suppressed the dependency of π to hyperparameters $\nu = (\alpha,\beta)$ since we assume them to be known.

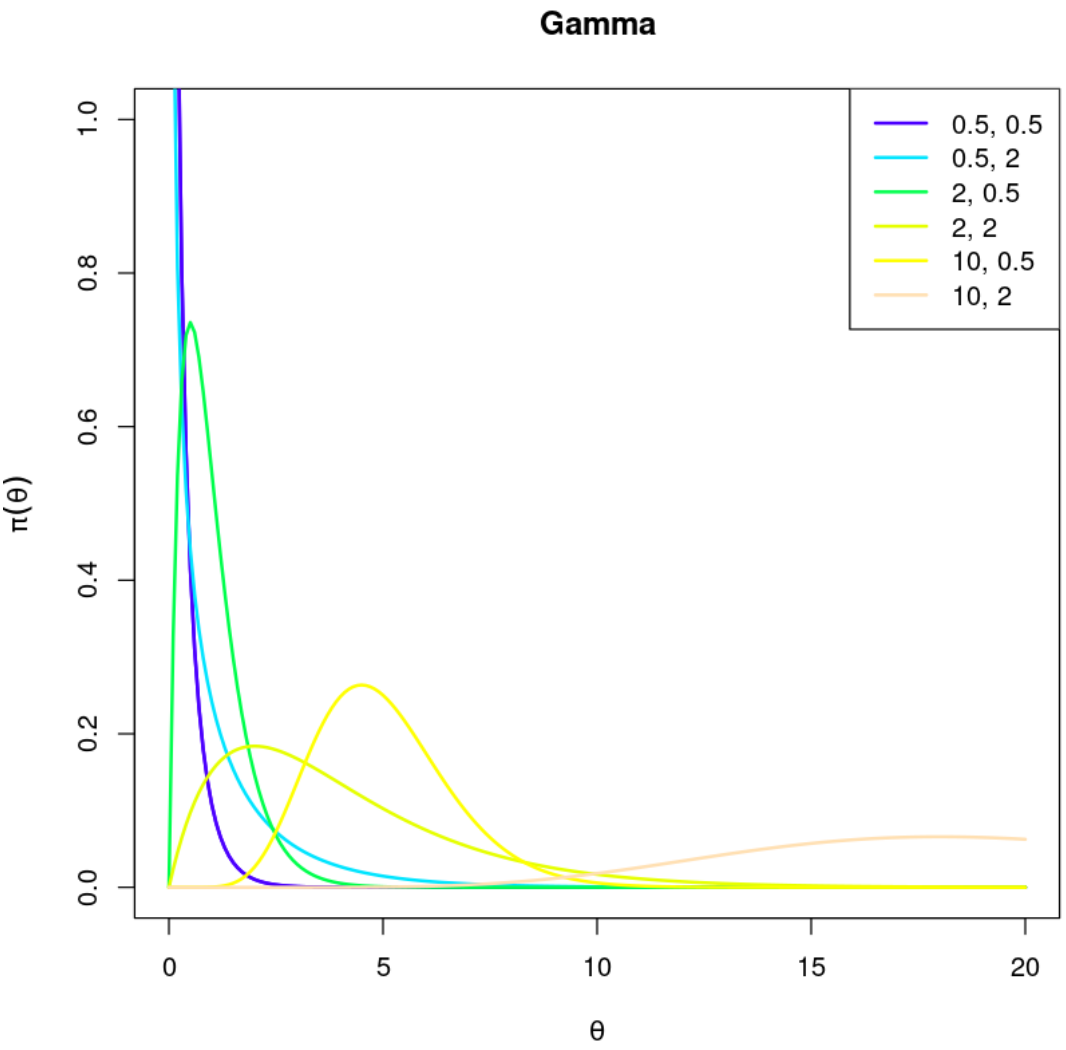
[3]:
```
## Gamma distribution
alpha <- c(0.5,2,10)
beta <- c(0.5,2)
thetas <- seq(0, 20, 0.1)

mycolors <- topo.colors(6)
plot(x=thetas, dgamma(x=thetas, shape=alpha[1], scale=beta[1]), type="l",
    lty=1, xlab=expression(theta), main="Gamma", ylab=expression(pi(theta)),
    ylim=c(0,1.0), col=mycolors[1], lwd=2)

index <- 0
for (i in alpha) {
    for (j in beta) {
        index <- index+1
        points(x=thetas, dgamma(x=thetas, shape=i,␣
 ↪scale=j),col=mycolors[index], ty="l", lwd=2)
    }
}

names <- cbind(rep(alpha,each=2),rep(beta))

legend(x="topright", legend=apply(FUN=paste, MAR=1, X=names, sep="",␣
 ↪collapse=", "), col=mycolors, lty=1, lwd=2)
```



The Gamma distribution is very flexible and it can have one tail ($\alpha \leq 1$) or two tails ($\alpha > 1$). For very large values of α the Gamma distribution resembles a Normal distribution. The β parameter shrinks or stretches the distribution relative to 0 but it doesn't change its shape.

Using the Bayes' theorem, we can now obtain the posterior probability

$$P(\theta|y) \approx f(y|\theta)\pi(\theta) \tag{14}$$
$$\approx (e^{-\theta}\theta^{y})(\theta^{\alpha-1}e^{-\theta/\beta}) \tag{15}$$
$$= \theta^{y+\alpha-1}e^{-\theta(1+1/\beta)} \tag{16}$$

Since we are only interested in a normalised function of θ, we drop all functions that do not depend on θ.

The posterior distribution is a Gamma distribution $G(\alpha', \beta')$ with $\alpha' = y+\alpha$ and $\beta' = (1+1/\beta)^{-1}$.

We were able to do this operation because the Gamma distribution (prior) is the conjugate family for the Poisson distribution (likelihood).

ACTIVITY

Suppose that, before looking at the actual data, we have some intuition that we expect to see 3 herds per day. Let's also assume that we are not extremely confident and we assign a moderate variance to it.

We then observe 4 herds.

Derive and plot the posterior distribution using both the exact solution and Monte Carlo sampling.

0.0.4 Noninformative priors

If no reliable prior information on θ is available, can we still employ a Bayesian approach?

It is still appropriate if we find a distribution $\pi(\theta)$ that contains "no information" about θ, in the sense that it does not favour one value over another.

We refer to such a distribution as a *noninformative prior* for θ.

All the information in the posterior will arise from the data.

If the parameter space is discrete and finite, that is, $\vec{\theta} = \{\theta_1, \theta_2, ..., \theta_n\}$, a possible noninformative prior is

$$p(\theta_i) = \frac{1}{n},\ i = 1, 2, ..., n \tag{17}$$

The prior distribution must be a proper (legitimate) probability distribution, meaning that

$$\sum_{1}^{n} \frac{1}{n} = 1 \tag{18}$$

If $\vec{\theta}$ is continuous and bounded, as $\vec{\theta} = [a, b]$ with $-\infty < a < b < +\infty$, a uniform prior in the form

$$P(\theta) = \frac{1}{b-a},\ a < \theta < b \tag{19}$$

is a noninformative prior distribution.

This assertion is less clear to be true than in the discrete case.

If $\vec{\theta}$ being continuous and unbounded, so that $\vec{\theta} = (-\infty, +\infty)$, a noninformative prior could be set as

$$P(\theta) = c, \text{ any } c > 0 \tag{20}$$

However, this distribution is clearly *improper* as

$$\int_{-\infty}^{+\infty} p(\theta) d\theta = +\infty \tag{21}$$

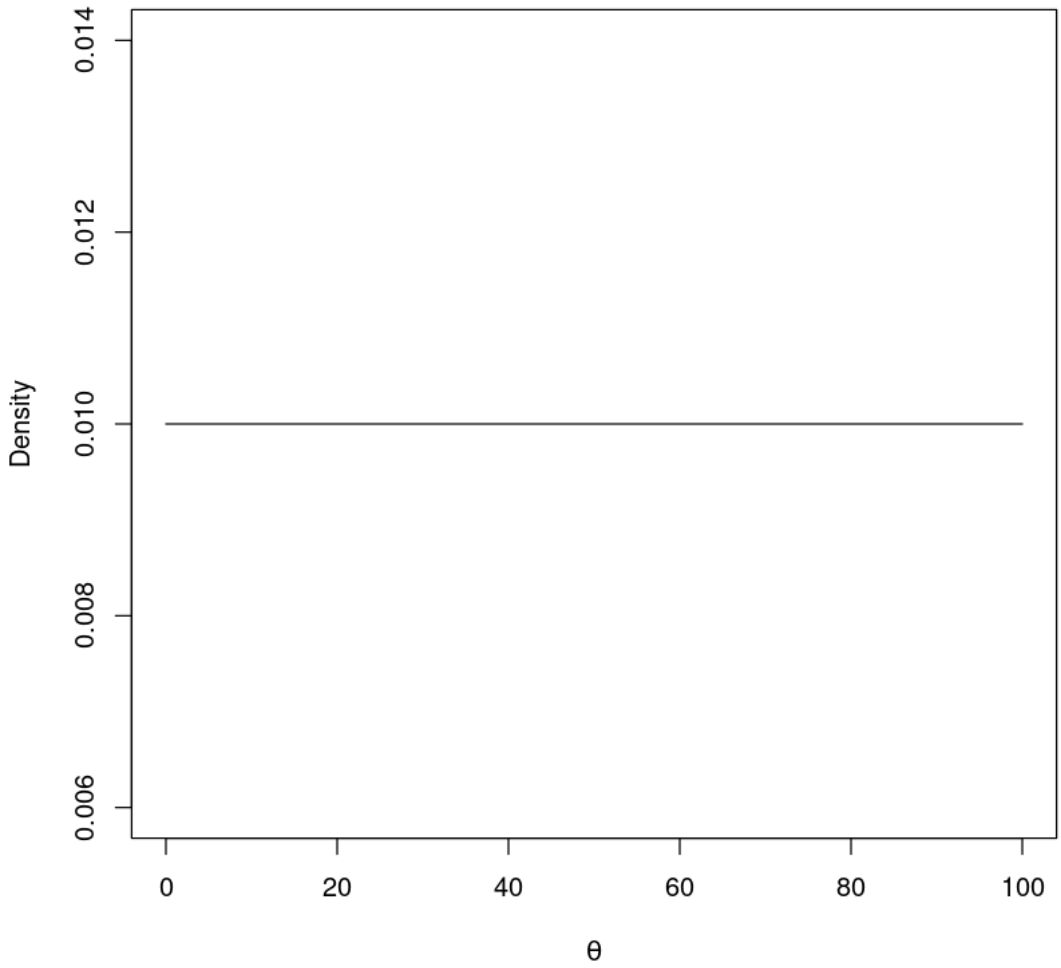
This may suggest that in this case a Bayesian approach cannot be used.

However, a Bayesian inference is still possible if the integral of of the likelihood $f(y|\theta)$ with respect to θ equals some finite value K, since

$$\int \frac{f(y|\theta) \cdot c}{\int f(y|\theta) \cdot c d\theta} d\theta = 1 \tag{22}$$

If use a uniform prior U for the mean arrival θ, our parameter, it should be $[0, +\infty)$. However, we can limit the range define a uniform prior distribution as $U(0, 100)$.

[4]:
```
## Uniform distribution
theta <- seq(0, 100, 0.1)
prior <- dunif(x=theta, min=0, max=100)
plot(x=theta, y=prior, xlab=expression(theta), ylab="Density", type="l")
```



This choice rules out scenarios that are impossible in real life. This means that the posterior will be truncated at 0 and 100. As we lack a conjugate model, we can sample from the posterior to obtain its distribution.

Noninformative priors are related to the notion of *reference* priors. These are not necessarily noninformative but a convenient, default choice for prior distributions.

0.0.5 Hierarchical modelling

A posterior distribution is typically obtained with two stages, one for $f(y, \theta)$, the likelihood of the data, and one for $\pi(\theta, \nu)$, the prior distribution of θ given a vector of *hyperparameters* ν.

Note that we drop the notation of all these parameters being vectors for simplicity.

If we are uncertain about the values for ν, we need an additional stage, a *hyperprior* , defining the density distribution of hyperparameters.

If we denote this distribution as $h(\nu)$, then the posterior distribution is

$$P(\theta|y) = \frac{\int f(y|\theta)\pi(\theta|\nu)h(\nu)d\nu}{\int\int f(y|\theta)\pi(\theta|\nu)h(\nu)d\nu d\theta} \tag{23}$$

Another possibility is to replace ν with an estimate $\hat{\nu}$ obtained by maximising the marginal distribution $m(y|\nu)$.

Inferences are made on the *estimated posterior* $P(\theta|y, \hat{\nu})$, by inserting $\hat{\nu}$ in the Bayes' theorem equation.

This approach is called *Empirical Bayesian* analysis as we are using the data to estimate the hyperparameter.

The empirical estimation of the prior seems a violation of Bayesian principles. Indeed, the update of the prior based on the data would use the data twice, both for the likelihood and the prior. Inferences from such modelling tend to be "overconfident" and methods that ignore this fact are called *naive Empirical Bayesian* approaches.

ν can depend on a collection of unknown parameters λ, with $h(\nu|\lambda)$ and a third-stage prior $g(\lambda)$.

This procedure of specifying a model over several layers is called *hierarchical modelling*.

This framework is very much used in graphical modelling (e.g. Bayesian networks). As we add extra layers and levels of randomness, subtle changes at the top levels (hyperpriors) will not have a strong effect at the bottom level (the data).