02_concepts

June 5, 2024

0.0.1 Bayes' theorem

If Y is a random variable, then $f(y|\theta)$ is a probability distribution representing the sampling model for the observed data $y = (y_1, y_2, ..., y_n)$ given an unknown parameter θ.

The distribution $f(y|\theta)$ is often called the *likelihood* and sometimes written as $L(\theta; y)$.

Often y and θ are vectors, and thus the correct notation would be $\vec{y}$ and $\vec{\theta}$.

We know that $L(\theta; y)$ is not a probability distribution for θ given y. Therefore $\int L(\theta; y)d\theta$ is not necessarily equal to 1 or even finite.

It is possible to find the value of θ that maximises the likelihood function: a *maximum likelihood estimate* (MLE) for θ, as

$$\hat{\theta} = argmax_{\theta} L(\theta; y) \tag{1}$$

In Bayesian statistics, θ is not a fixed (although unknown) parameter but a random quantity.

This is done by adopting a probability distribution, called *prior distribution* , for θ that contains any information we have about θ not related to the data y.

Inferences on θ are based on its *posterior distribution* given by

$$P(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{m(y)} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \tag{2}$$

This formula is known as *Bayes' Theorem.*

The posterior probability is simply the product of the likelihood and the prior, normalised so that integrates to 1. The posterior distribution is therefore a proper (or legitimate) probability distribution.

We will now prove this theorem using an example. One can even use Venn diagrams which are useful but not rigorous.

The greatest loss of vertebrate biodiversity we observed in the past 30 years is due to a chytrids fungus which is responsible for the extinction of over a hundred species of amphibians.

Let's assume we have a sample space S with all the possible outcomes of an experiment and we are interested in a subset of S, representing only some events. In our example we are interested in detecting which samples of frogs are infected or not by the fungus. For doing so, during our fieldwork, we take some samples and test whether they are infected or not.

Let's consider S to consist of all the samples collected in a particular area.

We can split our S in two events: the event "samples with infection" (designated as set A), and "samples with no infection" (complement of set A, or A^c).

What is the probability that a randomly chosen sample is infected?

It is the number of elements in A divided by the number of elements of S.

We can denote the number of elements of A as $|A|$, called the cardinality of A.

The probability of A, $P(A)$, is

$$P(A) = \frac{|A|}{|S|} \tag{3}$$

with $0 \leq P(A) \leq 1$.

Assume that we use a molecular screening test which takes a biological sample (e.g. piece of skin) from a frog and tests for the presence of the fungus. The test will be "positive" for some samples, and "negative" for some other samples.

Let's denote with event B the collection of "samples for which the test is positive".

What is the probability that the test will be "positive" for a randomly selected sample?

It is

$$P(B) = \frac{|B|}{|S|} \tag{4}$$

with $0 \leq P(B) \leq 1$.

We are now dealing with the entire sample space S (all samples), the event A (samples with infection), and the event B (samples with a positive test).

So far we have treated the two events separately, in isolation. What happens if we put them together?

We can calculate the probability of both events occurring ($A \cap B$) simultaneously.

$$P(A \cap B) = \frac{|A \cap B|}{|S|} \tag{5}$$

with $0 \leq P(A \cap B) \leq 1$.

The event $A \cap B$ represents "samples with infection and with a positive test".

Note that sometimes AB is used as shorthand notation for $A \cap B$.

There is also the event $(B - AB)$ or "samples without infection and with a positive test", and the event $(A - AB)$ or "samples with infection and with a negative test".

Given that the test is positive for a randomly selected sample, what is the probability that said sample is infected?

In terms of a Venn diagram, this question translates to "given that we are in region B, what is the probability that we are in region $A \cap B$?". This is equivalent of saying that "if we make region B our new Universe, what is the probability of A?". The notation for the latter conditional probability is $P(A|B)$, called "the probability of A given B".

This probability is equal to

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|S|}{|B|/|S|} = \frac{P(A \cap B)}{P(B)} \tag{6}$$

Given that a randomly selected sample is infected (event A), what is the probability that the test is positive for that sample (event $A \cap B$)?

The conditional probability is

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{7}$$

If we put together these last two equations (by $P(B \cap A)$) we obtain:

$$P(A)P(B|A) = P(B)P(A|B) \tag{8}$$

It follows that

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{9}$$

which is Bayes' theorem.

0.1 Normal/Normal model

Let's assume that we monitored a certain number of frogs in a given pond and want to make some inferences on the infection rate, θ, in the whole area.

If both the prior and the likelihood are Normal (Gaussian) distributions, then

$$f(y|\theta) = N(y|\theta, \sigma^2) \tag{10}$$

$$\pi(\theta) = N(\theta|\mu, \tau^2) \tag{11}$$

μ and τ are known *hyperparameters* while θ is the unknown parameter.

The posterior distribution $p(\theta|y)$ is also a Normal distribution

$$p(\theta|y) = N(\theta|\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}) \tag{12}$$

If

$$B = \frac{\sigma^2}{\sigma^2 + \tau^2} \tag{13}$$

then

$$E(\theta|y) = B\mu + (1 - B)y \tag{14}$$

$$Var(\theta|y) = (1 - B)\sigma^2 \equiv B\tau^2 \tag{15}$$

B is called the *shrinking factor*.

It is called like that because it gives the proportion for how much the posterior mean is "shrunk back" from the classical frequentist estimate y towards the prior mean μ.

Note that $0 \leq B \leq 1$.

The posterior mean is a weighted average of the prior mean μ and the direct estimate Y. The weight on the prior mean B depends on the relative variability of the prior distribution and the likelihood.

If $\sigma^2 >> \tau^2$

then $B \approx 1$ and our prior knowledge is more precise than the data information.

If $\sigma^2 << \tau^2$

then $B \approx 0$ and our prior knowledge is imprecise and the final estimate will move very little towards the prior mean.

- We have a single observation from one pond of 6 infected frogs ($y = 6$).
- Our likelihood function (Normal distribution) has $\sigma = 1$.
- We expected 2 infected frogs before doing the monitoring (with variance equal to 1).

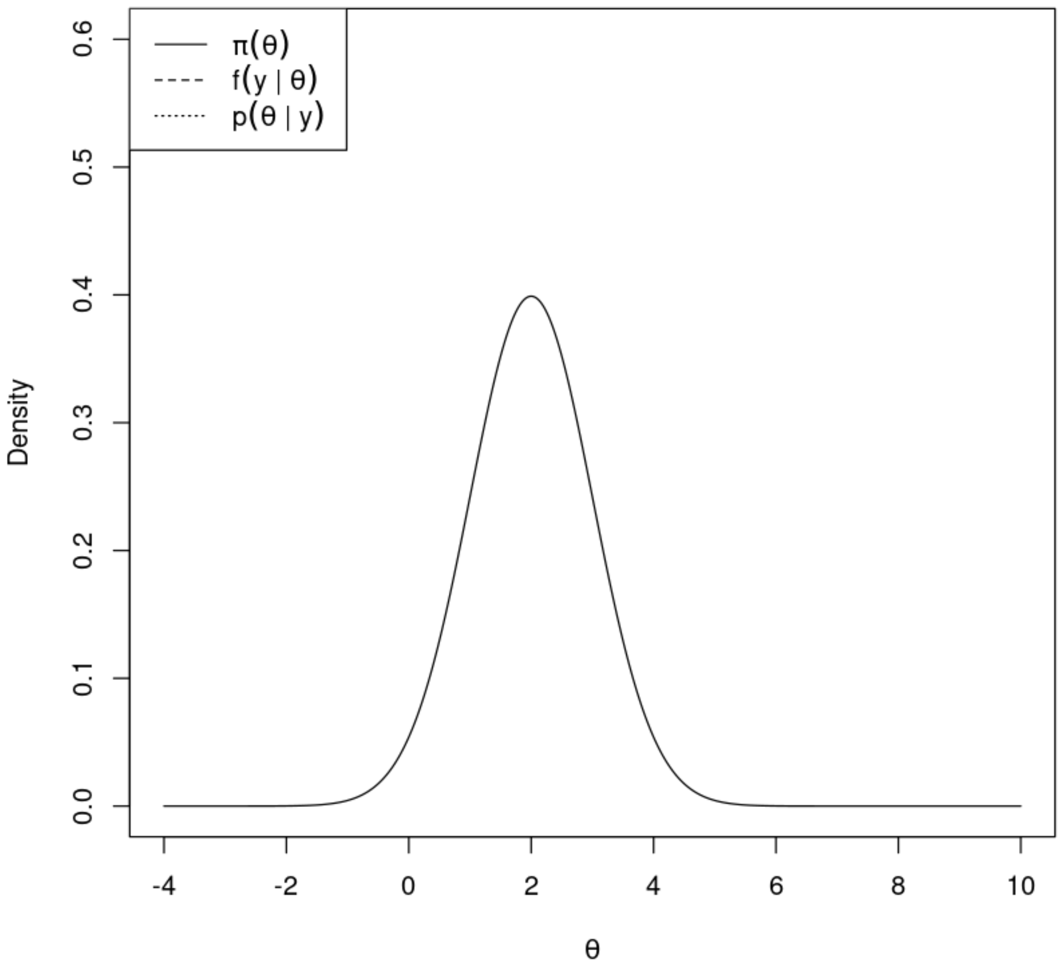
More formally,

$$f(y = 6|\theta) = N(y = 6|\theta, 1) \tag{16}$$

$$\pi(\theta) = N(\theta|2, 1) \tag{17}$$

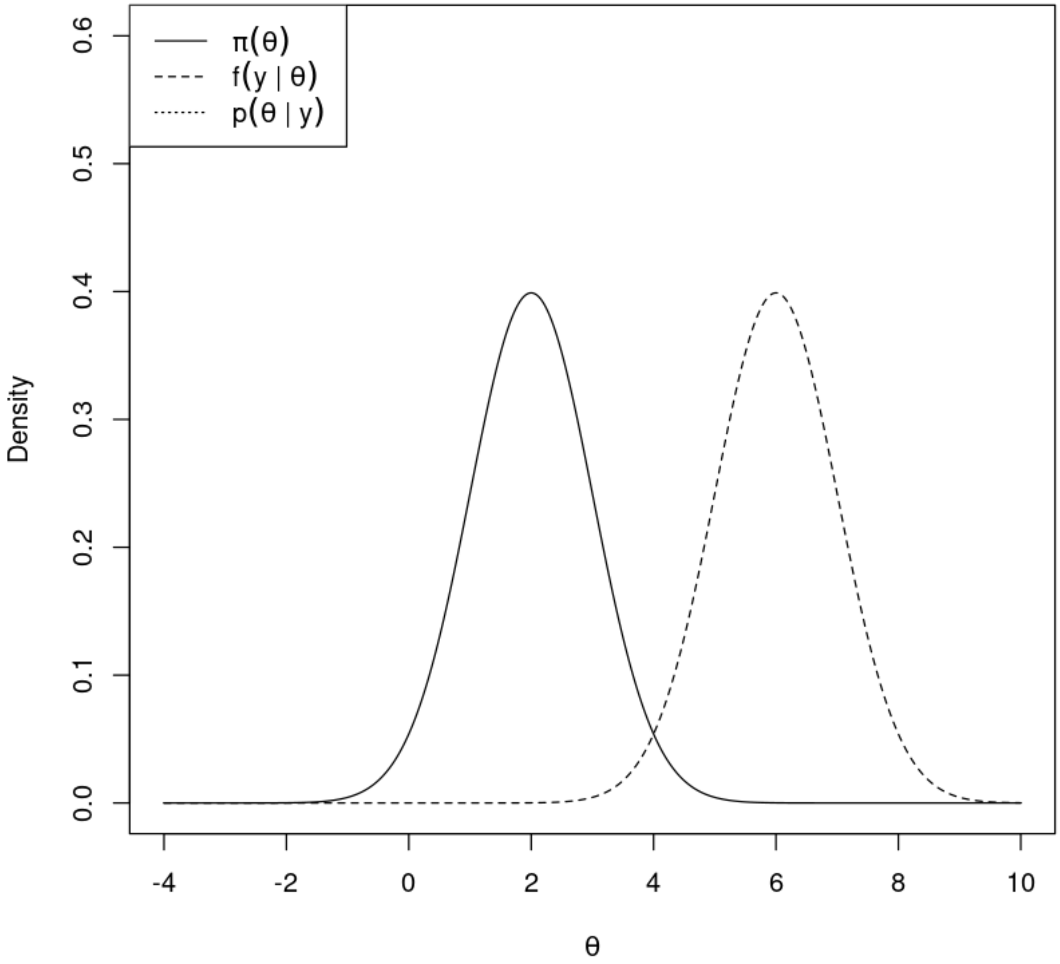
[1]:

```
# prior
mu <- 2
tau <- 1

x <- seq(-4,10,0.01)
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,0.6),
    type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
    legend(x="topleft", legend=c(expression(pi(theta)),
    expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3)
```



[2]:

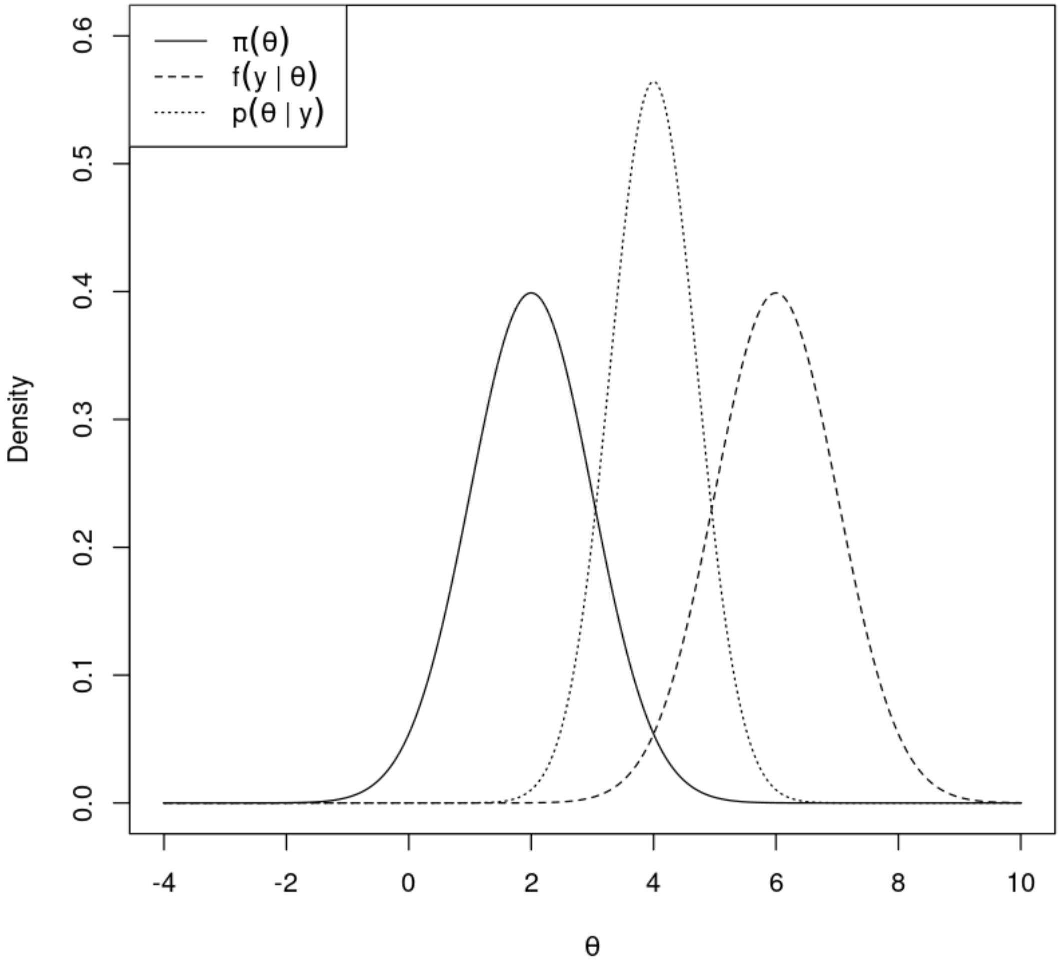
```
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,0.6),
    type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
    legend(x="topleft", legend=c(expression(pi(theta)),
    expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3) # prior

# likelihood
y <- 6
sigma <- 1
points(x=x, y=dnorm(x=y, mean=x, sd=sigma), type="l", lty=2)
```



[3]:

```
# prior
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,0.6),
    type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
    legend(x="topleft", legend=c(expression(pi(theta)),
    expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3) # prior
```

```
# likelihood
points(x=x, y=dnorm(x=y, mean=x, sd=sigma), type="l", lty=2) # likelihood

# posterior
B <- sigma^2/(sigma^2+tau^2)
postMean <- B*mu + (1-B)*y
postVar <- B*tau^2
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)), type="l", lty=3)
```



The prior distribution is centered around 2 (μ). The likelihood function is centered around 6 which is the only observation we have.

The posterior distribution is centred exactly between the prior and the likelihood. In this case $B = 0.5$ and therefore prior and data are equally weighted.

The *maximum a posteriori probability* (MAP) estimate is 4, as it is equal to the mode of the posterior distribution.

[4]:
```
x[which.max(dnorm(x=x, mean=postMean, sd=sqrt(postVar)))]
```

4

The posterior distribution is more skewed than the prior and the likelihood, despite these two distribution having the same variance. The posterior variance is smaller than the variance of either the prior or the likelihood.

The *precision* (the reciprocal of the variance) is the sum of the precisions in the prior and likelihood. The combined strength of prior and likelihood tends to increase the precision, or reduce the variance, in our inference of θ.

In this example, the precision is $1 + 1 = 2$, hence the variance is $1/2$. Therefore the posterior roughly covers $4 \pm 3(\sqrt{1/2}) \approx (1.88, 6.12)$.

ACTIVITY

What happens if we use a skewer (sharper) or wider prior? What is the shape of the posterior distribution if the variance of the prior is smaller (stronger belief) or larger (weaker belief)?

Assume that the prior distribution has $\mu = 2$ and $\tau = 0.5$.

Calculate and plot prior and posterior distributions and evaluate the MAP. Is the posterior mean closer or more distant from the prior mean? What is the shrinking factor?

[ ]:
```
# ...
```

ACTIVITY

Assume that now the prior distribution has $\mu = 2$ and $\tau = 2$.

Calculate and plot prior and posterior distributions and evaluate the MAP. Is the posterior mean closer or more distant from the prior mean? What is the shrinking factor?

[5]:
```
# ...
```

If we have more observations of infected frogs in multiple ponds, our data may look like $\vec{y} = \{6, 5, 4, 5, 6\}$.

Given a sample of n independent observations, then

$$f(\vec{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \tag{18}$$

We can also use a transformation if we can find a statistic $S(\vec{y})$ that is sufficient, meaning that $p(\theta|\vec{y}) = p(\theta|S(\vec{y}))$.

We can use the sufficient statistic $S(\vec{y}) = \bar{y}$, where $\bar{y}$ is the mean of $\vec{y}$.

The likelihood function has the form $f(\bar{y}|\theta) = N(\theta, \sigma^2/n)$ and the posterior distributions is

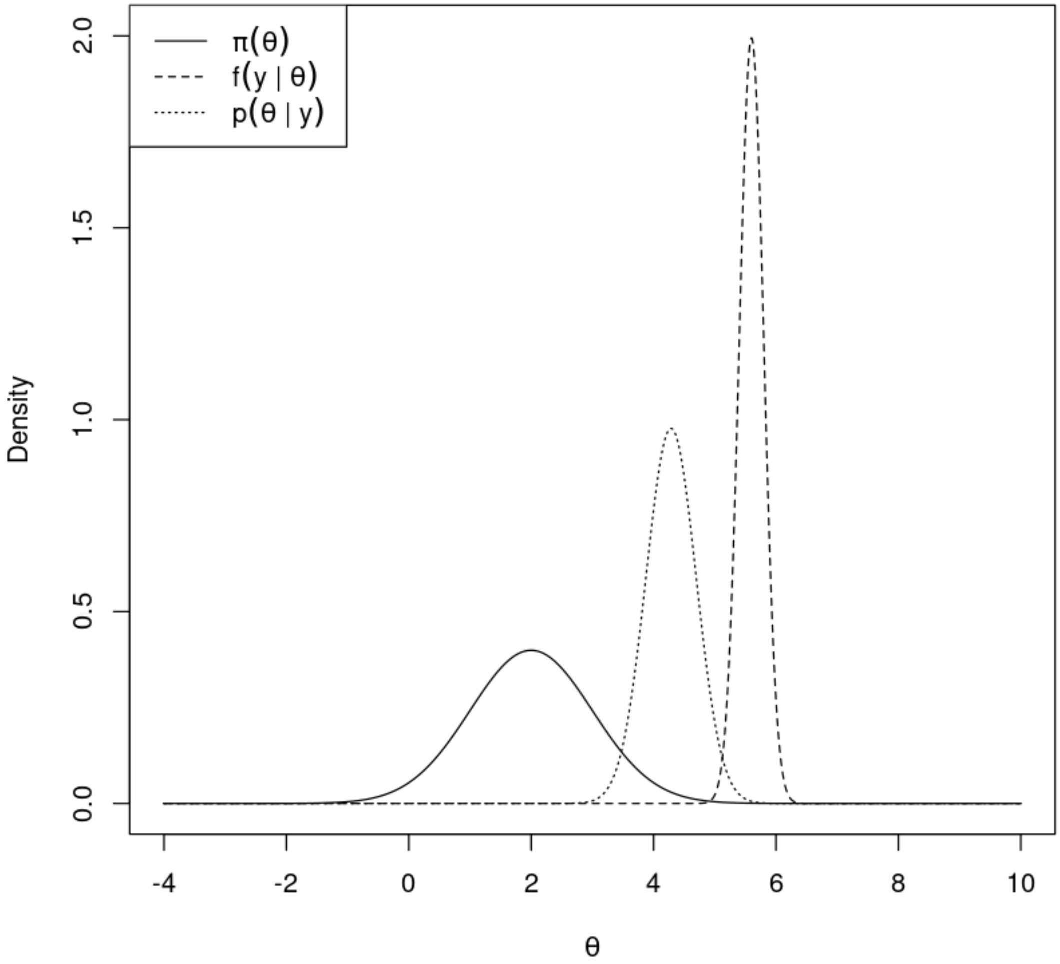
$$p(\theta|\bar{y}) = N(\theta|\frac{(\sigma^2/n)\mu + \tau^2\bar{y}}{(\sigma^2/n) + \tau^2}, \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2}) = N(\theta|\frac{\sigma^2\mu + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}) \tag{19}$$

Suppose we have $\mu = 2$, $\sigma = \tau = 1$ and set $\bar{y} = 5.8$ and $n = 5$. In this case $\theta_{MAP} = 4.43$ with a range of $(3.21, 5.65)$. The MAP has been shifted towards the MLE as we have more data information.

[6]:

```
# prior (obviously it does not change)
mu <- 2
tau <- 1
x <- seq(-4,10,0.01)
plot(x=x, dnorm(x=x, mean=mu, sd=tau), ylim=c(0,2),
    type="l", lty=1, ylab="Density", xlab=expression(theta), main="")
    legend(x="topleft", legend=c(expression(pi(theta)),
    expression(f(y~"|"~theta)), expression(p(theta~"|"~y))), lty=1:3)

# likelihood with more observations
y <- c(6, 5, 6, 4, 7)
n <- length(y)
sigma <- 1
points(x=x, y=dnorm(x=x, mean=mean(y), sd=sigma/n), type="l", lty=2)

# posterior with more observations
postMean <- ( (sigma^2/n)*mu + tau^2*mean(y) ) / ( (sigma^2/n)*mu + tau^2 )
postVar <- ( (sigma^2/n)*tau^2 ) / ( (sigma^2/n) + tau^2 )
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)), type="l", lty=3)

# MAP with more observations
map <-  x[which.max(dnorm(x=x, mean=postMean, sd=sqrt(postVar)))]
cat("MAP:", map, "(", map-3*sqrt(postVar),",",map+3*sqrt(postVar),")\n")
```

```
MAP: 4.29 ( 3.065255 , 5.514745 )
```



0.1.1 Monte Carlo sampling

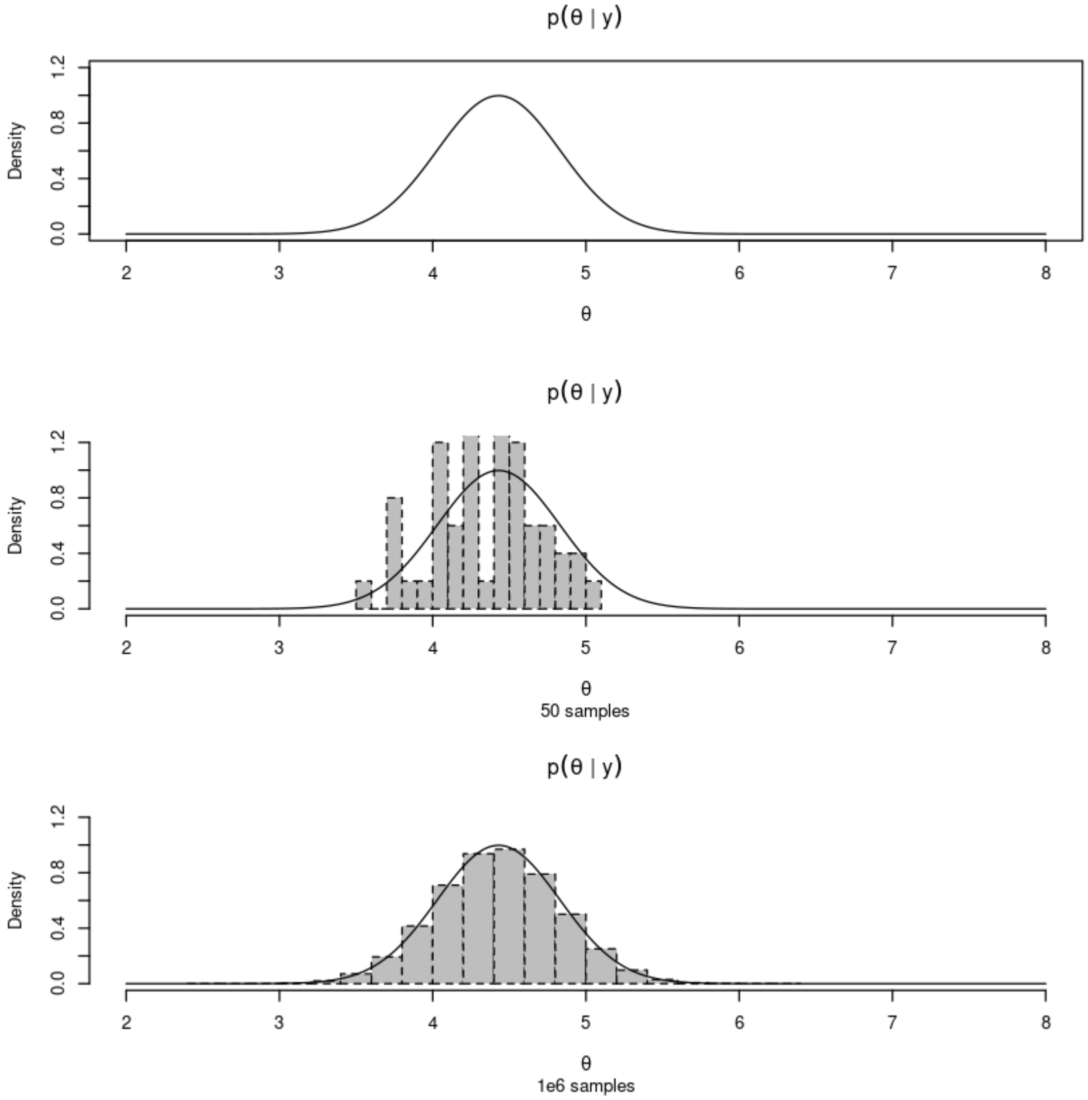
To derive the posterior distribution we can also draw random samples from it instead of directly calculating its parameters.

This procedure is often called *Monte Carlo sampling* after the city famous for its casinos.

In the previous example of the Normal/Normal model with multiple observations, we were able to calculate the posterior mean (4.43) and posterior variance (0.17). From these parameters, we were able to derive (and plot) the density function, the posterior probability itself. Alternatively, we can randomly sample directly from the posterior distribution.

[7]:
```
## Monte Carlo sampling
par(mfrow=c(3,1))
```

```r
# posterior
x <- seq(2,8,0.01)
postMean <- 4.43
postVar <- 0.16
plot(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)), type="l", lty=1,
    ylab="Density", xlab=expression(theta), main=expression(p(theta~"|"~y)),
    ylim=c(0,1.2), xlim=c(2,8))

# sampling
y_sampled_1 <- rnorm(n=50, mean=postMean, sd=sqrt(postVar))
hist(y_sampled_1, breaks=20, freq=F, lty=2, col="grey", ylim=c(0,1.2), xlim=c(2,8),
    sub="50 samples", main=expression(p(theta~"|"~y)), xlab=expression(theta))
    points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)), type="l", lty=1)

# more sampling
y_sampled_2 <- rnorm(n=1e6, mean=postMean, sd=sqrt(postVar))
hist(y_sampled_2, breaks=20, freq=F, lty=2, col="grey", ylim=c(0,1.2), xlim=c(2,8),
    sub="1e6 samples", main=expression(p(theta~"|"~y)), xlab=expression(theta))
points(x=x, y=dnorm(x=x, mean=postMean, sd=sqrt(postVar)), type="l", lty=1, ylab="Density",
    xlab=expression(theta), main=expression(p(theta~"|"~y)), sub="1e6 samples")
```



The more sampling we do, the closer our sampled distribution will be to the "true" posterior distribution. With 50 samples the empirical posterior mean is 4.444 while with $1e6$ samples we have an empirical posterior mean of 4.429 which is very close to our direct estimate of 4.43. With 50 samples the empirical posterior variance is 0.105 while with $1e6$ samples we have an empirical posterior variance of 0.159 which is very close to our direct estimate of 0.16.

In this simple Normal/Normal case, Monte Carlo methods are not strictly necessary since the integral in the denominator of Bayes' theorem can be evaluated in closed form. In these cases, it is preferable to derive a smooth curve rather an a histogram of sampled values, and have the corresponding exact values for the posterior parameters.

However, there are cases where, given the choice for the likelihood and prior functions, this integral cannot be evaluated. In these conditions, Monte Carlo methods are to be preferred for estimating, or rather approximating, the posterior distribution.

As any sample can be drawn from any posterior regardless of the number of unknown parameters θ, we have the ability to work on problems with (theoretically) unlimited complexity, at the price of not obtaining an exact form for the posterior and performing a large number of samplings.

0.1.2 Intended Learning Outcomes

At the end of this part you are now be able to: * appreciate the use of Bayesian statistics in life sciences, * formulate and explain Bayes' theorem, * describe a Normal-Normal model and implement it in R with or without Monte Carlo sampling,

```
[ ]:
```