10 Bayesian analysis II: Solutions to Practical

We will look at the posterior distribution for the mean of the Normal distribution and do some simple Bayesian linear regression in R.

> **Intended learning objectives:** By the end of this session you will be able to: * Conduct a conjugate analysis for the mean of a Normal distribution; * Summarize the posterior in multiple ways (by plots, computing credible and HPD intervals); * Find Bayesian predictive distributions for Normal data.

Question 1. Posterior for the mean of the Normal distribution

We will use a dataset on 1,174 mother-newborn pairs where we have information on the birth weight, gestational days and several variables on the mother such as her age and height. We load the data and look at the first few rows:

In [3]:

```
library(readr)
baby <- read_csv("https://raw.githubusercontent.com/data-8/textbook/gh-pages/data/baby.csv")
head(baby)
```

```
Parsed with column specification:
cols(
  `Birth Weight` = col_double(),
  `Gestational Days` = col_double(),
  `Maternal Age` = col_double(),
  `Maternal Height` = col_double(),
  `Maternal Pregnancy Weight` = col_double(),
  `Maternal Smoker` = col_logical()
)
```

Birth Weight	Gestational Days	Maternal Age	Maternal Height	Maternal Pregnancy Weight	Maternal Smoker
120	284	27	62	100	FALSE
113	282	33	64	135	FALSE
128	279	28	64	115	TRUE
108	282	23	67	125	TRUE
136	286	25	62	93	FALSE
138	244	33	62	178	FALSE

Note that the birth weight is in ounces, maternal age is in years, maternal height is in inches, and maternal pregnancy weight is in pounds.

We change the column names of the dataset so they do not contain spaces:

In [4]:

```
#change column names so they do not contain spaces:
colnames(baby) <- c("bweight", "gestd", "age", "height", "weight", "smoke")
```

Suppose that we are interested in the number of gestational days.

> 1: Plot a histogram of the number of gestational days, and obtain some summary statistics.

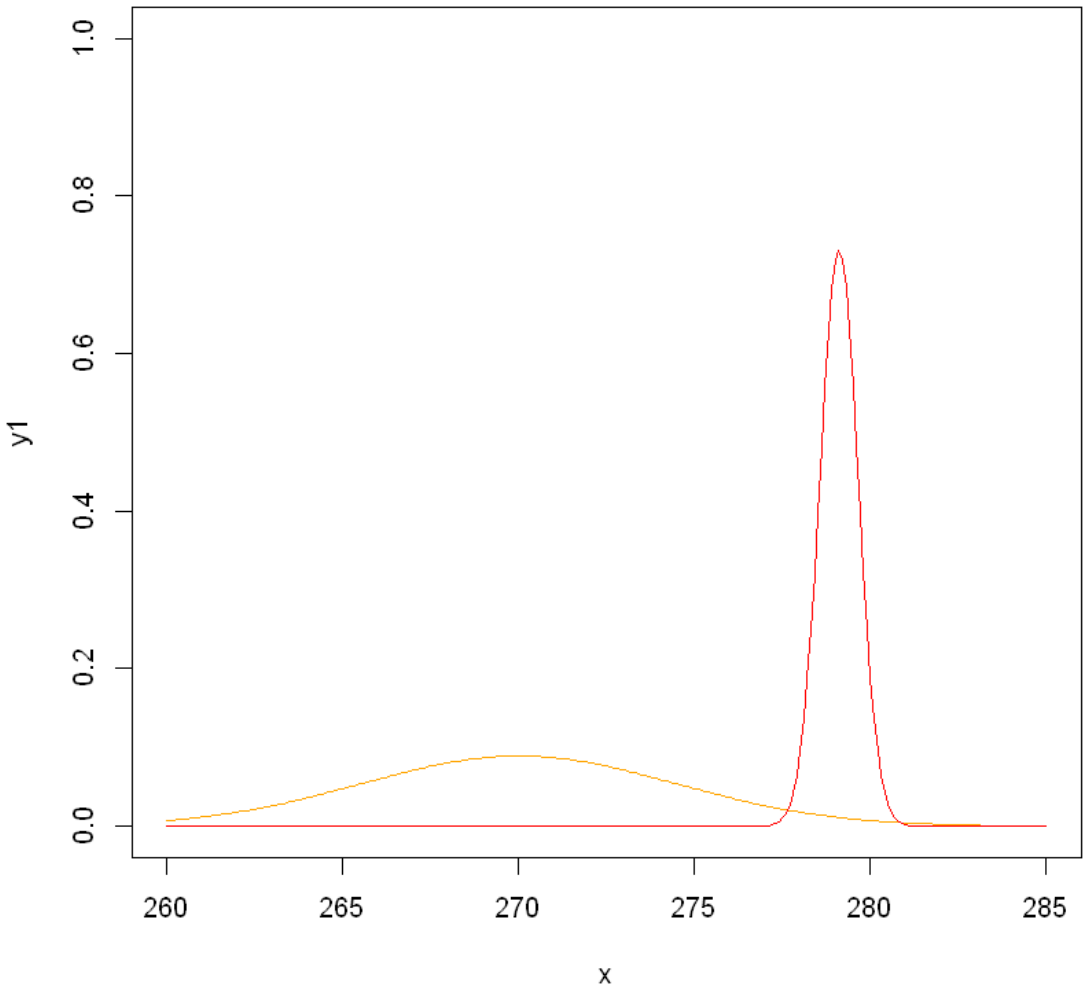
We wish to obtain the posterior distribution of the mean number of gestational days μ for this population. For now, we assume that we know the variance of gestational days in the population to be 350.

We choose a vague prior for the mean: $\mu \sim N(270, 20)$.

> 1. Work out the posterior distribution for μ. The following code plots the prior and the distribution of the data on the same scale. Once you have worked out the posterior distribution, adapt the code below so that the posterior distribution is plotted as well. Try to use a dashed line for the posterior (you can use the option lty=4).

In [5]:

```
#Choose a vague prior for the mean:
prior_mean <- 270
prior_var <- 20

x<-seq(260, 285, 0.1)
#plot the prior
y1 <- dnorm(x, mean=prior_mean, sd=sqrt(prior_var))
plot(x, y1, type="l", lwd=1, col="orange", ylim=c(0,1))
#plot the observed distribution
y2 <- dnorm(x, mean=mean(baby$gestd), sd=sqrt(350/1174))
lines(x, y2, type="l", lwd=1, col="red")
```


y1
0.0
0.2
0.4
0.6
0.8
1.0
260
265
270
275
280
285
x

ANSWER

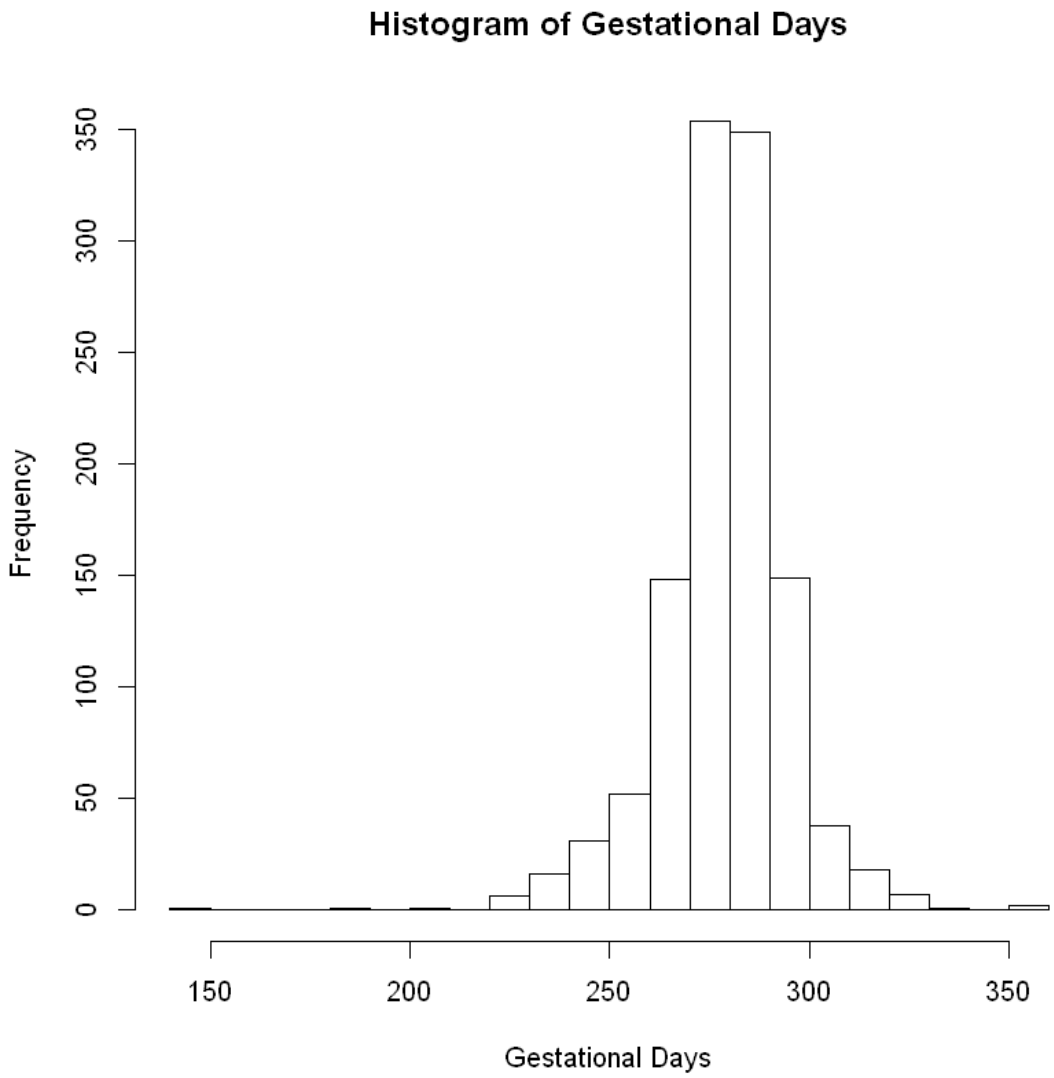
1) Plotting the histogram and obtaining summary statistics:

In [18]:

```
hist(baby$gestd, xlab="Gestational Days", main="Histogram of Gestational Days", breaks=
20)

summary(baby$gestd)

var(baby$gestd)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  148.0   272.0   280.0   279.1   288.0   353.0 
```

256.329870263786



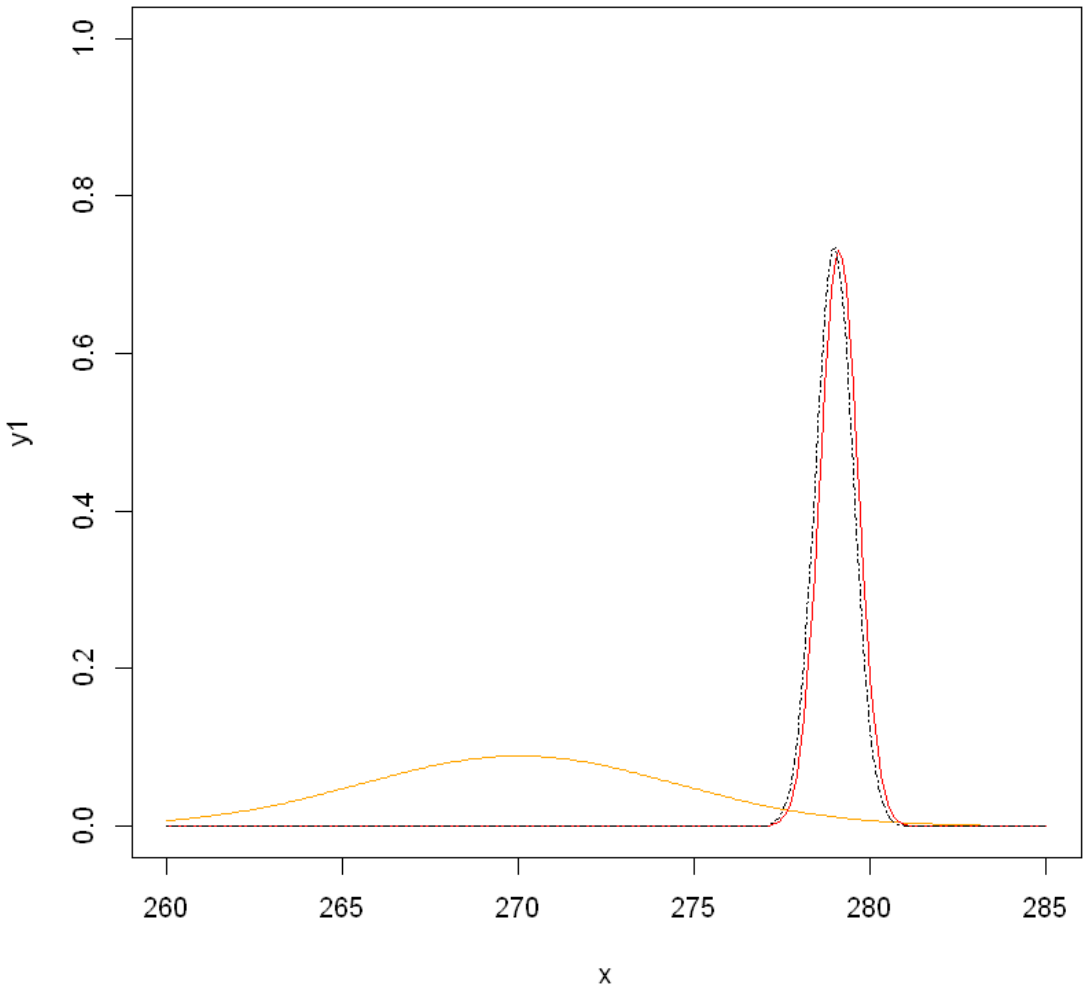
1. Finding the posterior distribution and adding the plot:

In [8]:

```
prior_mean <- 270
prior_var <-20

like_mean <- mean(baby$gestd)
like_var <- 350
n<-nrow(baby)

post_mean <- (n*prior_var*like_mean + prior_mean*like_var)/(n*prior_var + like_var)  ;
post_mean

post_var <- like_var*prior_var/(n*prior_var + like_var)  ; post_var


x<-seq(260, 285, 0.1)

y1<-dnorm(x, mean=prior_mean, sd=sqrt(prior_var))
plot(x, y1, type="l", lwd=1, col="orange", ylim=c(0,1))

y2<-dnorm(x, mean=like_mean, sd=sqrt(like_var/n))
lines(x, y2, type="l", lwd=1, col="red")

y3<-dnorm(x, mean=post_mean, sd=sqrt(post_var))
lines(x, y3, type="l", lwd=1, col="black", lty=4)
```

278.967687788502

0.29374737725556



Question 2

1: Given a normal posterior distribution with mean ψ and variance γ^2, write down the 95% HPD interval.
2: For the mother-baby dataset in Question 1, obtain the 95% HPD interval for μ, given the data.
3: A fetus born before the 37th week of gestation (259 days) is considered to be preterm. What is the **prior** predictive probability that the next mother will have a preterm birth? You may find the pnorm() function helpful.

In [ ]:

```
pnorm(q, mean, sd) ###remember that R uses standard deviation as an argument, not the v
ariance!
```

> 4: What is the posterior predictive probability that the next mother will have a preterm birth, given that we have observed the data?

ANSWER

1. The 95\% HPD interval is given by $\psi \pm 1.96 \times \gamma$.
2. The code below computes the 95% HPD:

In [10]:

```
post_mean-1.96*sqrt(post_var)  ;  post_mean+1.96*sqrt(post_var)
# 95%HPD interval 277.9, 280.0
```

277.905397844812

280.029977732192

1. Using notation from the lectures, we have that $y \sim N(\phi, \tau^2 + \sigma^2)$ and we wish to compute $p(y < 259)$:

In [19]:

```
pnorm(259, prior_mean, sqrt(prior_var+like_var))
```

0.283707474267189

There is approximately a 28\% chance that the next baby will be pre-term.

1. From the lectures, we know that the posterior predictive distribution has the form $N\left\{\frac{\tau^2 n\bar{y}+\sigma^2\phi}{\tau^2 n+\sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n+\sigma^2} + \sigma^2\right\}$; the mean of the distribution is the posterior mean, and the variance of the distribution is the posterior variance plus the variance of the likelihood:

In [20]:

```
pnorm(259, post_mean, sqrt(post_var+like_var))
```

0.143015226406651

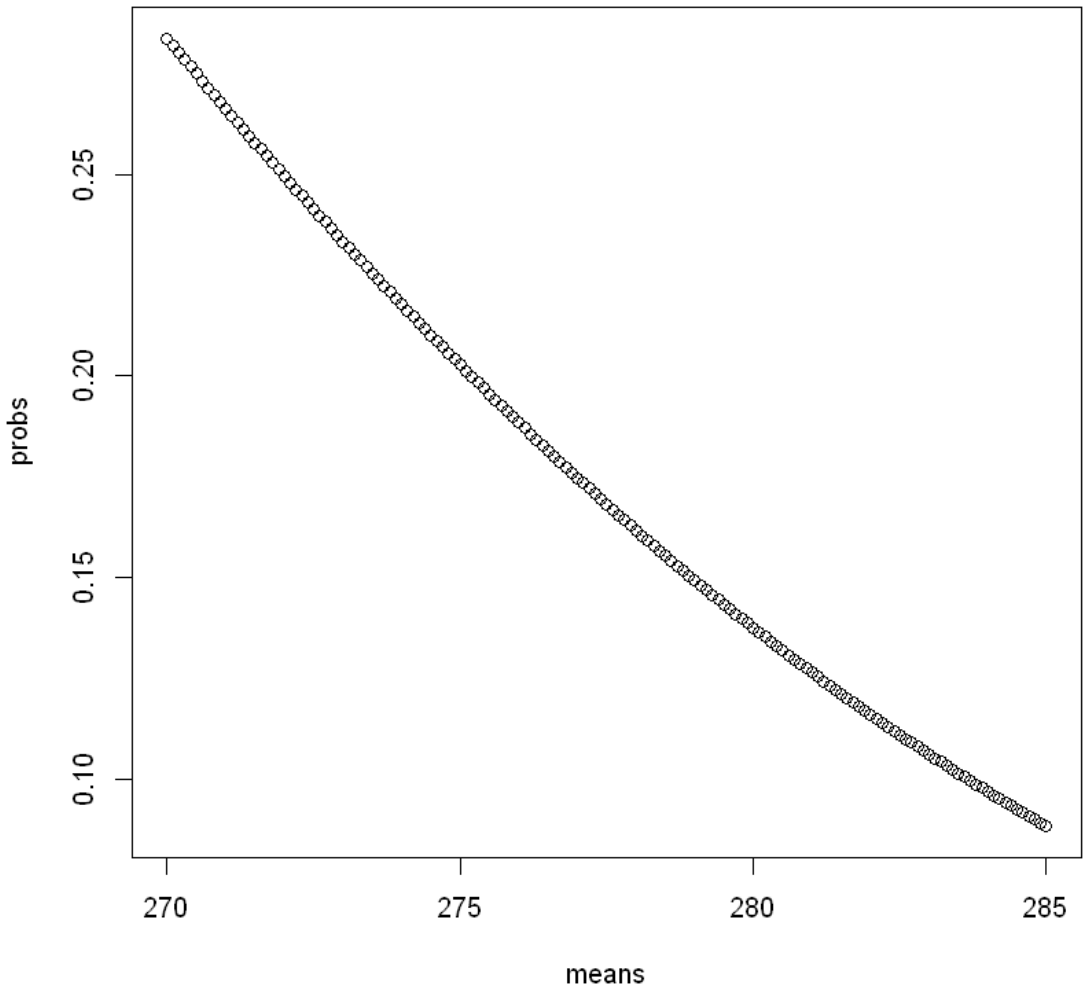
There is approximately a 14% chance that the next baby is pre-term.

Question 3 (Optional)

1. Create an R function that:
   - Takes the following parameters as arguments: ϕ, τ^2 (parameters of the prior), σ^2 (population variance);
   - Computes the prior predictive probability that the mean gestation is less than 259 (pre-term) for the next mother in the study;
   - Provides this probability as the output.
2. Suppose, as before, that the population variance of gestational days is 350. Compute the prior predictive probability that a the next mother will have a preterm birth for a range of values for the prior mean (keeping prior variance fixed). Create a plot of the prior predictive probability vs prior mean.

In [12]:

```
pre_term_prob <- function(prior_mean, prior_var, like_var){

    prob <- pnorm(259, prior_mean, sqrt(prior_var+like_var))
    return(prob)
}

#we keep the prior variance fixed at 20.
#we create a vector of possible prior means between 270 and 285
means <- seq(270, 285, 0.1)

probs <- pre_term_prob(means, prior_var=20, like_var=350)

plot(means, probs)
```


probs
0.10
0.15
0.20
0.25
270
275
280
285
means