R Lab. - Exercise 7

Michele Guadagnini - Mt. 1230663

June 3, 2020

Exercise 1 - Normal distribution

A researcher has collected n = 15 observations that are supposed to come from a Normal distribution with known variance σ^2= 16. Assuming a normal prior for μ, Norm(m = 20, s^2 = 25):

A) determine the posterior distribution and find the posterior mean and standard deviation

```
obs <- c(26.8, 31.9, 28.0, 26.3, 28.5, 18.6, 28.3, 27.2,
         22.3, 28.5, 20.9, 25.0, 16.3, 27.5, 31.5)
m <- 20
var <- 16
ymean <- mean(obs)
s2 <- 25
N <- length(obs)

nsamples <- 2000
x <- seq(12, 32, len=nsamples)
delta.x <- 20/nsamples
prior <- dnorm(x, mean=m, sd=sqrt(s2))

likelihood.f <- function(data) {
    Like <- 1
    for (p in data) {
      Like <- Like*dnorm(p, x, sqrt(var))
    }
    return(Like)
}

likelihood <- likelihood.f(obs)
likelihood <- likelihood/(sum(likelihood)*delta.x)

mu.post20 <- (m/s2 + N*ymean/var)/(N/var + 1/s2)
sd.post20 <- (var*s2)/(var + N*s2)
posterior <- dnorm(x, mu.post20, sd.post20)
posterior <- posterior/(sum(posterior)*delta.x)
```

The mean and standard deviation of the Posterior distribution are:

- mean: 25.601
- std : 1.023

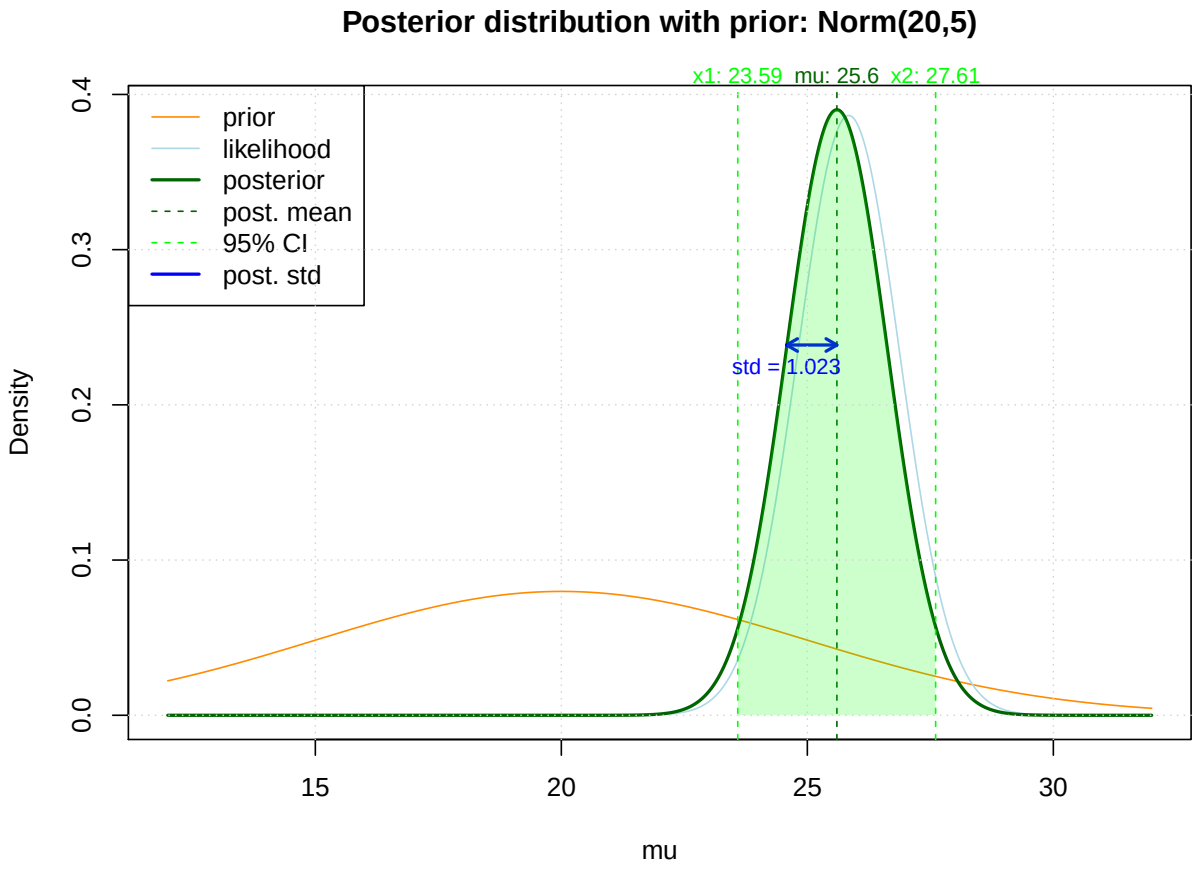
B) find the 95% credibility interval for μ

```
low.20 <- x[max(which(cumsum(posterior)<=0.025*sum(posterior)))]
upp.20 <- x[min(which(cumsum(posterior)>=0.975*sum(posterior)))]
```

The 95% credibility interval for μ is: [23.586, 27.608]

C) plot the posterior distribution, indicating on the same plot: the mean value, the standard deviation, and the 95% credibility interval

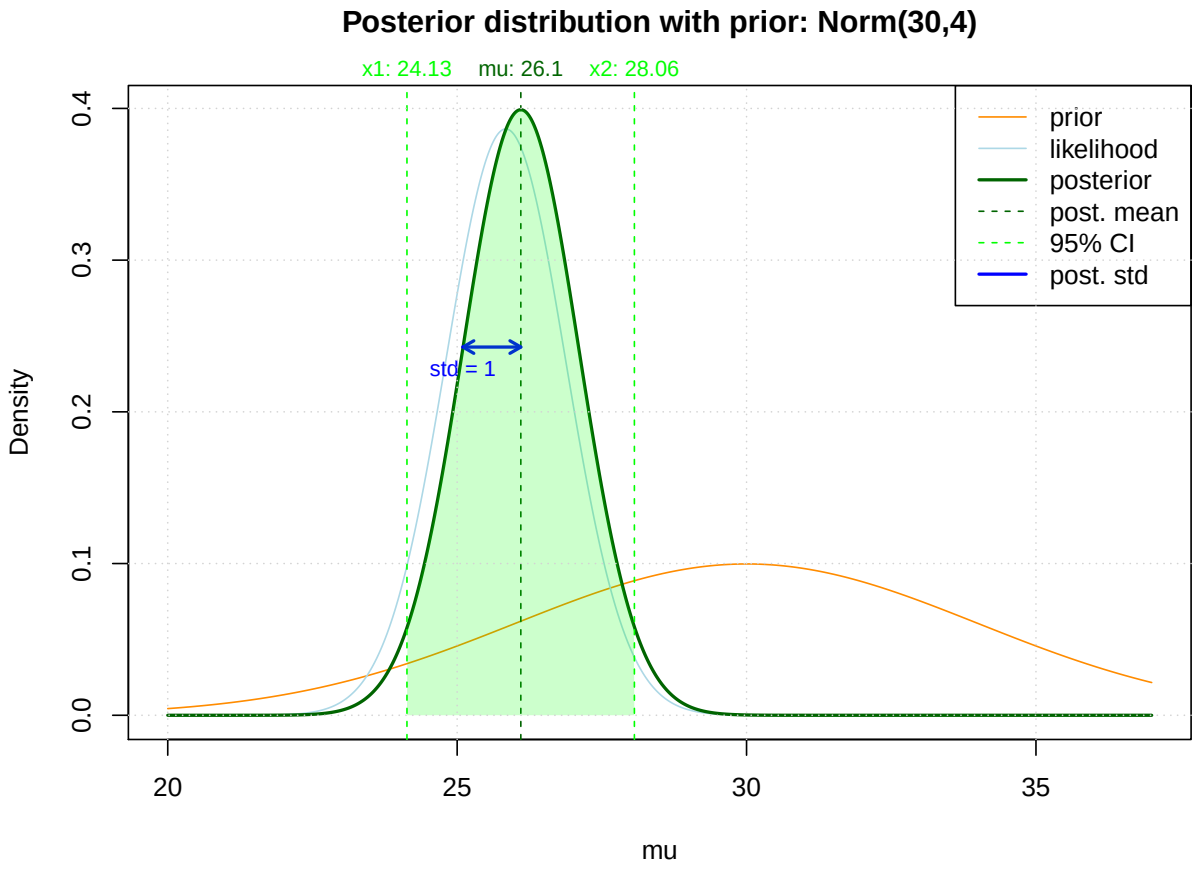
```
ymax <- max(c(prior,likelihood,posterior))
plot(x, prior, type="l", col='darkorange', ylim=c(0,ymax), ylab="Density",
     main="Posterior distribution with prior: Norm(20,5)", xlab="mu")
lines(x, likelihood, col='lightblue')
lines(x, posterior, col='darkgreen', lwd=2)
abline(v=mu.post20, lty=2, col='darkgreen')
ysig <- posterior[min(which(x >= (mu.post20-sd.post20)))]
arrows((mu.post20-sd.post20), ysig, mu.post20, ysig,
       lwd=2, col="blue", code=3, len=0.08)

abline(v=c(low.20,upp.20), lty=2, col='green')
x1 <- low.20; x2 <- upp.20
poly.x <- c(x1, x[x>x1 & x<x2], x2)
poly.y <- c(0, posterior[which(x>x1 & x<x2)],0)
polygon(poly.x, poly.y, col = rgb(0,1,0,alpha=0.2), border=FALSE)

text(x1, par("usr")[4]+0.02, labels = paste("x1:",round(x1, 2)),
     pos=1, cex=0.8, col="green", xpd=TRUE)
text(x2, par("usr")[4]+0.02, labels = paste("x2:",round(x2, 2)),
     pos=1, cex=0.8, col="green", xpd=TRUE)
text(mu.post20, par("usr")[4]+0.02, labels = paste("mu:",round(mu.post20, 2)),
     pos=1, cex=0.8, col="darkgreen", xpd=TRUE)
text((mu.post20-sd.post20), ysig, pos=1, cex=0.8, col="blue",
     labels = paste("std =", round(sd.post20, 3)))

legend("topleft", c("prior","likelihood","posterior","post. mean","95% CI","post. std"),
       col=c("darkorange","lightblue","darkgreen","darkgreen","green","blue"),
       lwd=c(1,1,2,1,1,2), lty=c(1,1,1,2,2,1))
grid()
```



D) repeat the analysis using a different prior Norm(m = 30, s^2 = 16) and plot, on the same graph the likelihood, the prior and the posterior.

```
m <- 30
s2 <- 16
var <- 16
ymean <- mean(obs)
N <- length(obs)

nsamples <- 2000
x <- seq(20, 37, len=nsamples)
delta.x <- 17/nsamples
prior <- dnorm(x, mean=m, sd=sqrt(s2))

likelihood <- likelihood.f(obs)
likelihood <- likelihood/(sum(likelihood)*delta.x)

mu.post30 <- (m/s2 + N*ymean/var)/(N/var + 1/s2)
sd.post30 <- (var*s2)/(var + N*s2)
posterior <- dnorm(x, mu.post30, sd.post30)
posterior <- posterior/(sum(posterior)*delta.x)

low.30 <- x[max(which(cumsum(posterior)<=0.025*sum(posterior)))]
upp.30 <- x[min(which(cumsum(posterior)>=0.975*sum(posterior)))]

ymax <- max(c(prior,likelihood,posterior))
```

```r
plot(x, prior, type="l", col='darkorange', ylim=c(0,ymax), ylab="Density",
     main="Posterior distribution with prior: Norm(30,4)", xlab="mu")
lines(x, likelihood, col='lightblue')
lines(x, posterior, col='darkgreen', lwd=2)
abline(v=mu.post30, lty=2, col='darkgreen')

ysig <- posterior[min(which(x >= (mu.post30-sd.post30)))]
arrows((mu.post30-sd.post30), ysig, mu.post30, ysig,
       lwd=2, col="blue", code=3, len=0.08)

abline(v=c(low.30,upp.30), lty=2, col='green')
x1 <- low.30; x2 <- upp.30
poly.x <- c(x1, x[x>x1 & x<x2], x2)
poly.y <- c(0, posterior[which(x>x1 & x<x2)],0)
polygon(poly.x, poly.y, col = rgb(0,1,0,alpha=0.2), border=FALSE)

text(x1, par("usr")[4]+0.025, labels = paste("x1:",round(x1, 2)),
     pos=1, cex=0.8, col="green", xpd=TRUE)
text(x2, par("usr")[4]+0.025, labels = paste("x2:",round(x2, 2)),
     pos=1, cex=0.8, col="green", xpd=TRUE)
text(mu.post30, par("usr")[4]+0.025, labels = paste("mu:",round(mu.post30, 2)),
     pos=1, cex=0.8, col="darkgreen", xpd=TRUE)
text((mu.post30-sd.post30), ysig, pos=1, cex=0.8, col="blue",
     labels = paste("std =", round(sd.post30, 3)))

legend("topright", c("prior","likelihood","posterior","post. mean","95% CI","post. std"),
       col=c("darkorange","lightblue","darkgreen","darkgreen","green","blue"),
       lwd=c(1,1,2,1,1,2), lty=c(1,1,1,2,2,1))
grid()
```



E) Compare the credibility intervals obtained with the two priors

The 95% credibility interval obtained using $Norm(20, 5)$ as prior is: [23.586, 27.608]

The 95% credibility interval obtained using $Norm(30, 4)$ as prior is: [24.133, 28.062]

Exercise 2 - Normal distribution (2)

A researcher has collected n = 16 observations that are supposed to come from a Normal distribution with known variance $\sigma^2 = 4$. Assuming the prior is a step function:

A) find the posterior distribution, the posterior mean and standard deviation

```
obs <- c(4.09, 4.79, 4.68, 4.49, 1.87, 5.85, 2.62, 5.09,
         5.58, 2.40, 8.68, 6.27, 4.07, 6.30, 4.78, 4.47)

prior.func <- function(p) {
  ifelse(p>0 & p<=3, p,
         ifelse(p>3 & p<=5, 3,
                ifelse(p>5 & p<=8, 8-p, 0)
         )  )
}
Iprior <- integrate(prior.func, 0, 10)$value

var <- 4
N <- length(obs)
nsamples <- 2000
x <- seq(0, 8, len=nsamples)
delta.x <- 8/nsamples
prior <- prior.func(x)/Iprior

likelihood.f <- function(data) {
    Like <- 1
    for (p in data) { Like <- Like*dnorm(p, x, sqrt(var)) }
    return(Like)
}
likelihood <- likelihood.f(obs)
likelihood <- likelihood/(sum(likelihood)*delta.x)

posterior <- prior*likelihood
posterior <- posterior/(sum(posterior)*delta.x)
mu.post <- sum(x*posterior)*delta.x
sd.post <- sqrt(sum(posterior*(x-mu.post)**2)*delta.x)
```

The mean and standard deviation of the Posterior distribution are:

- mean: 4.725
- std : 0.484

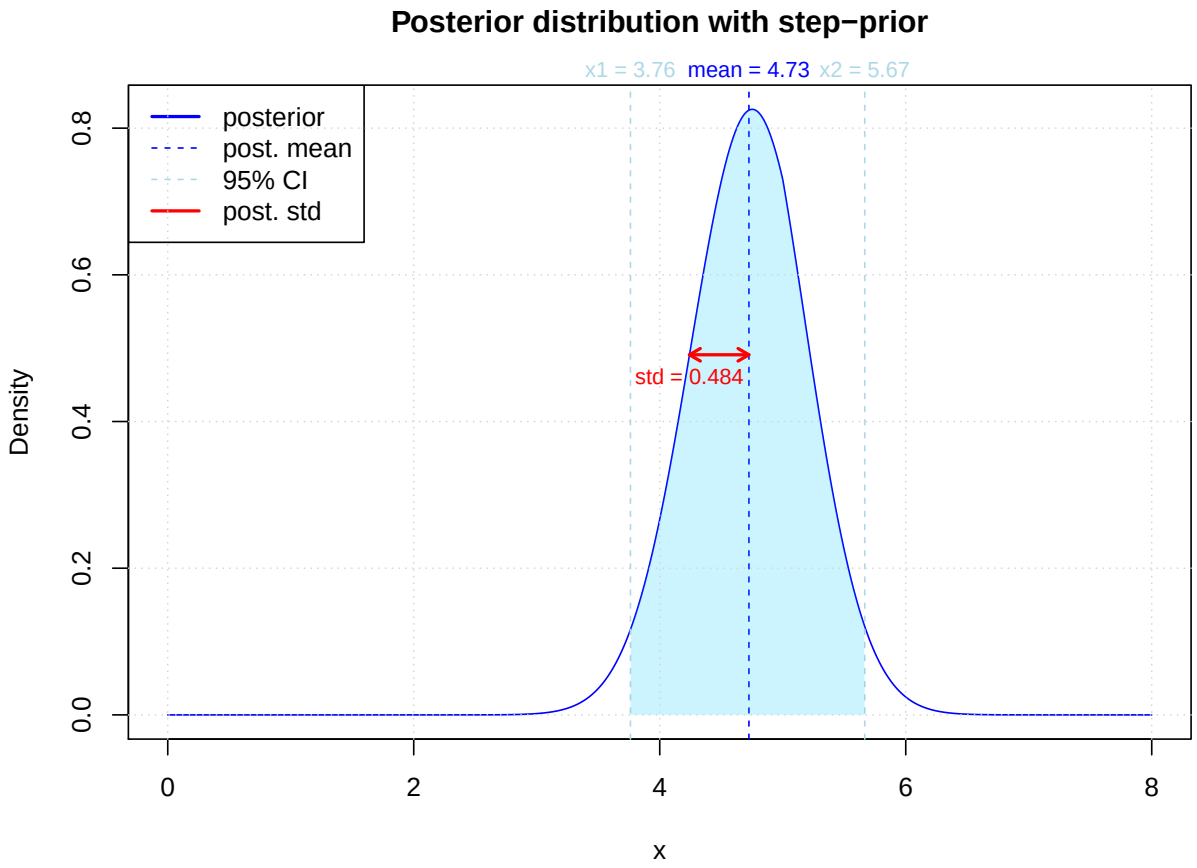
B) find the 95% credibility interval for μ

```
low <- x[max(which(cumsum(posterior)<=0.025*sum(posterior)))]
upp <- x[min(which(cumsum(posterior)>=0.975*sum(posterior)))]
```

The 95% credibility interval for μ is: [3.762, 5.667]

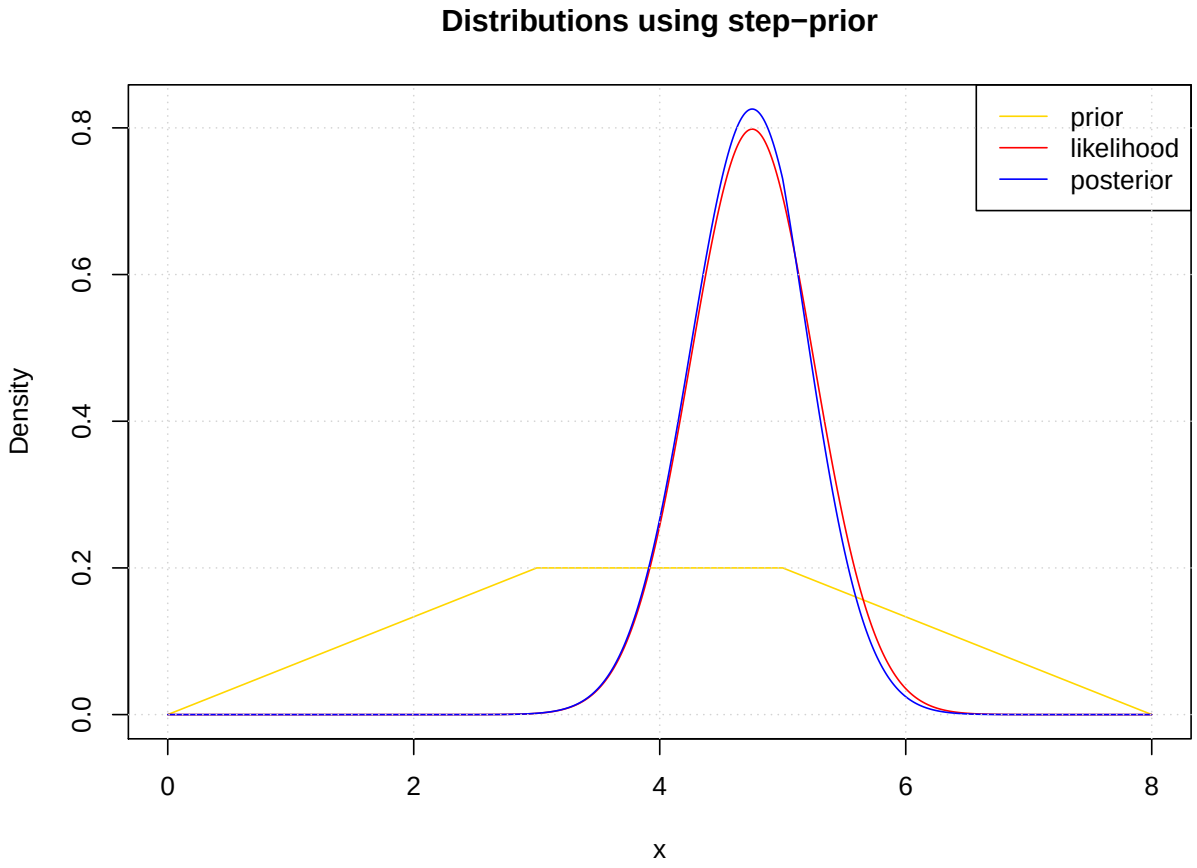
C) plot the posterior distribution, indicating on the same plot: the mean value, the standard deviation, and the 95% credibility interval

```r
ymax <- max(posterior)
plot(x, posterior, type="l", col='blue', ylim=c(0,ymax), ylab="Density",
     main="Posterior distribution with step-prior")
abline(v=mu.post, lty=2, col='blue')
abline(v=c(low, upp), lty=2, col='lightblue')

x1 <- low; x2 <- upp; ysig <- posterior[min(which(x >= (mu.post-sd.post)))]
poly.x <- c(x1, x[x>x1 & x<x2], x2)
poly.y <- c(0, posterior[which(x>x1 & x<x2)],0)
polygon(poly.x, poly.y, col = rgb(0,0.8,1,alpha=0.2), border=FALSE)
arrows((mu.post-sd.post), ysig, mu.post, ysig,
       lwd=2, col="red", code=3, len=0.08)
text(x1, par("usr")[4]+0.05, labels = paste("x1 =",round(x1, 2)),
     pos=1, cex=0.8, col="lightblue", xpd=TRUE)
text(x2, par("usr")[4]+0.05, labels = paste("x2 =",round(x2, 2)),
     pos=1, cex=0.8, col="lightblue", xpd=TRUE)
text(mu.post, par("usr")[4]+0.05, labels = paste("mean =",round(mu.post, 2)),
     pos=1, cex=0.8, col="blue", xpd=TRUE)
text((mu.post-sd.post), ysig, pos=1, cex=0.8, col="red",
     labels = paste("std =", round(sd.post, 3)))
legend("topleft", c("posterior","post. mean","95% CI","post. std"),
       col=c("blue","blue","lightblue","red"),
       lwd=c(2,1,1,2), lty=c(1,2,2,1))
grid()
```



D) plot, on the same graph, the prior, the likelihood and the posterior distribution

```
ymax <- max(c(prior,likelihood,posterior))
plot(x, prior, type="l", col='gold', ylim=c(0,ymax), ylab="Density",
     main="Distributions using step-prior")
lines(x, likelihood, col='red')
lines(x, posterior, col='blue', lwd=1)

legend("topright", c("prior","likelihood","posterior"),
       col=c("gold","red","blue"),
       lwd=c(1,1,1), lty=c(1,1,1))
grid()
```



Exercise 3 - Water quality estimation

A study on water quality of streams, a high level of bacter X was defined as a level greater than 100 per 100 ml of stream water. n = 116 samples were taken from streams having a high environmental impact on pandas. Out of these, y = 11 had a high bacter X level.

Indicating with p the probability that a sample of water taken from the stream has a high bacter X level,

A) find the frequentist estimator for p

```
N <- 116
y <- 11

E.freq <- y/N
var.freq <- E.freq*(1-E.freq)/N
```

The frequentist estimator for p is: $p_F = 0.095$.

B) using a Beta(1, 10) prior for p, calculate the posterior distribution P(p|y)

```
nsamples <- 2000
p <- seq(0, 1, len=nsamples)
alpha <- 1
beta <- 10
prior <- dbeta(p, alpha, beta)
# Beta is a conjugate prior for a binomial likelihood
posterior <- dbeta(p, alpha+y, beta+N-y)
posterior <- posterior/(sum(posterior)/2000)
```

C) find the bayesian estimator for p, the posterior mean and variance, and a 95% credible interval

```
post.mean <- (alpha+y)/(alpha+beta+N)
post.var  <- (alpha*beta)/((alpha+beta+1)*(alpha+beta)**2)

E.bayes <- post.mean

post.low <- p[max(which(cumsum(posterior)<=0.025*sum(posterior)))]
post.upp <- p[min(which(cumsum(posterior)>=0.975*sum(posterior)))]
```

The bayesian estimator for p is: $p_B = 0.094$.
The posterior mean and variance are:

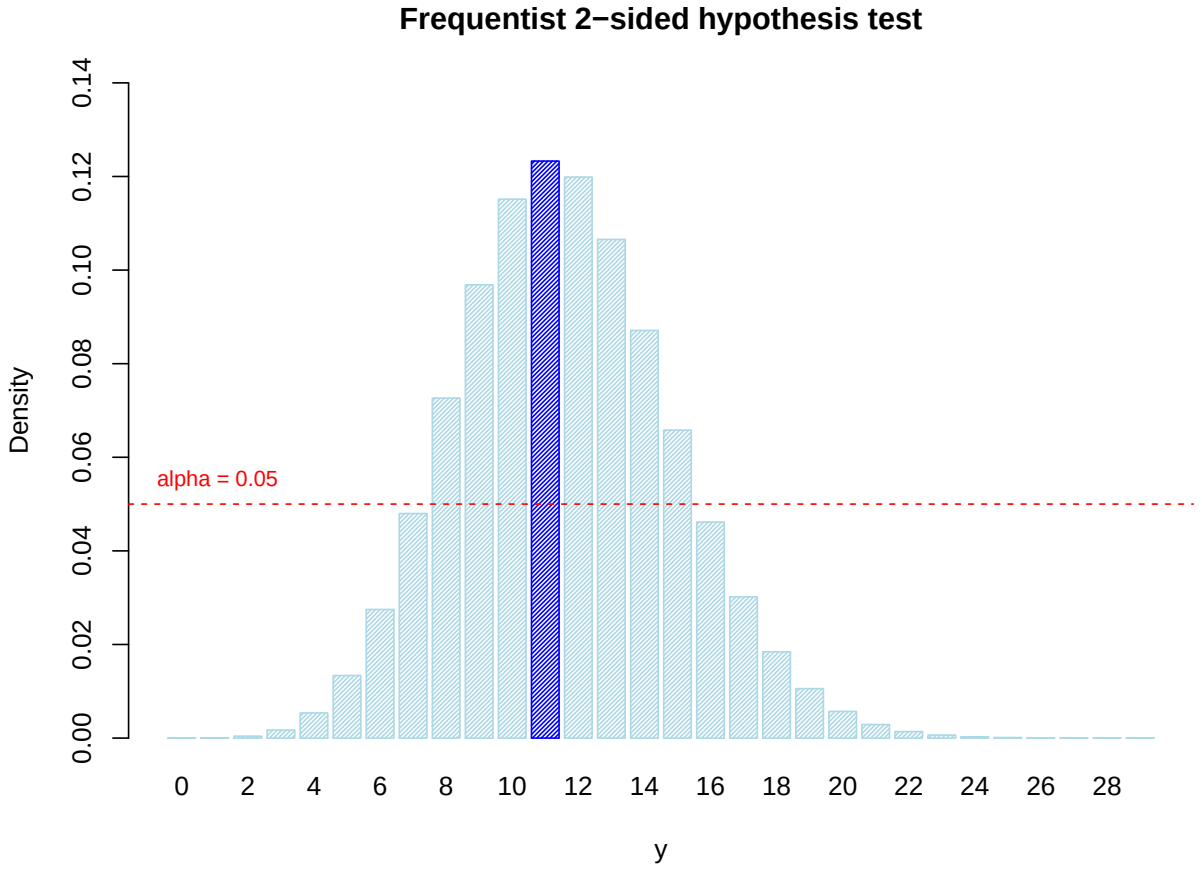
- mean: 0.094
- var : 0.007

The 95% credible interval for p is : [0.05, 0.151].

D) test the hypothesis H_0: p = 0.1 vs H_1: p $\neq$ 0.1 at 5% level of significance with both the frequentist and bayesian approach

```
# Frequentist
p.null <- 0.1
x <- 0:N
d.null <- dbinom(x, size=N, prob=p.null)
```

```r
l.o.s = 0.05

barplot(d.null[1:30], names=x[1:30], ylim=c(0,0.14), density=50, border=TRUE,
        xlab='y',ylab="Density",main="Frequentist 2-sided hypothesis test",
        col=ifelse(x==11, "blue", "lightblue"))
abline(h=l.o.s, lty=2, col='red')
text(2, l.o.s+0.01, labels = paste("alpha =",l.o.s), pos=1, cex=0.8, col="red")
```



In the plot above it can be seen that P(y=11) is greater than $\alpha = 0.05$, so the null hypothesis is accepted.

```r
# Bayesian
post.low <- p[max(which(cumsum(posterior)<=0.025*sum(posterior)))]
post.upp <- p[min(which(cumsum(posterior)>=0.975*sum(posterior)))]
```

The probality $p_0 = 0.1$ is inside the 95% credible interval of the posterior distribution:

- Lower bound: 0.05
- Upper bound: 0.151

So the null hypothesis is accepted.

A new measurement, performed one month later on n = 165 water samples, gives y = 9 high bacter X level.

E) find the frequentist estimator for p

```
N <- 165
y <- 9

E.freq <- y/N
var.freq <- E.freq*(1-E.freq)/N
```

The frequentist estimator for p with the new measuremnet is: $p_F = 0.055$.

F) find the posterior distribution, assuming both a Beta(1, 10) prior for p, and assuming the posterior probability of the older measurement as the prior for the new one.

```
# same prior as before
nsamples <- 2000
p <- seq(0, 1, len=nsamples)

alpha <- 1
beta <- 10
prior <- dbeta(p, alpha, beta)
posterior.beta <- dbeta(p, alpha+y, beta+N-y)
posterior.beta <- posterior.beta/(sum(posterior.beta)/nsamples)

#old posterior as prior
prior <- posterior
likelihood <- dbinom(y, size=N, prob=p)
likelihood <- likelihood/(sum(likelihood)/nsamples)
posterior.new <- prior*likelihood
posterior.new <- posterior.new/(sum(posterior.new)/nsamples)
```

G) find the bayesian estimator for p, the posterior mean and variance, and a 95% credible interval

```
# same prior as before
post.mean.beta <- (alpha+y)/(alpha+beta+N)
post.var.beta  <- (alpha*beta)/((alpha+beta+1)*(alpha+beta)**2)

E.bayes.beta <- post.mean.beta

post.low.beta <- p[max(which(cumsum(posterior.beta)<=0.025*sum(posterior.beta)))]
post.upp.beta <- p[min(which(cumsum(posterior.beta)>=0.975*sum(posterior.beta)))]

# old posterior as new prior
post.mean.new <- sum(p*posterior.new)/nsamples
post.var.new  <- sum(posterior.new*(p-post.mean)**2)/nsamples

E.bayes.new <- post.mean.new

post.low.new <- p[max(which(cumsum(posterior.new)<=0.025*sum(posterior.new)))]
post.upp.new <- p[min(which(cumsum(posterior.new)>=0.975*sum(posterior.new)))]
```

The bayesian estimator for p using $Beta(1, 10)$ as prior is: $p_B = 0.057$.

The posterior mean and variance are:

- mean: 0.057
- var : 0.007

The 95% credible interval for p is : [0.027, 0.096].

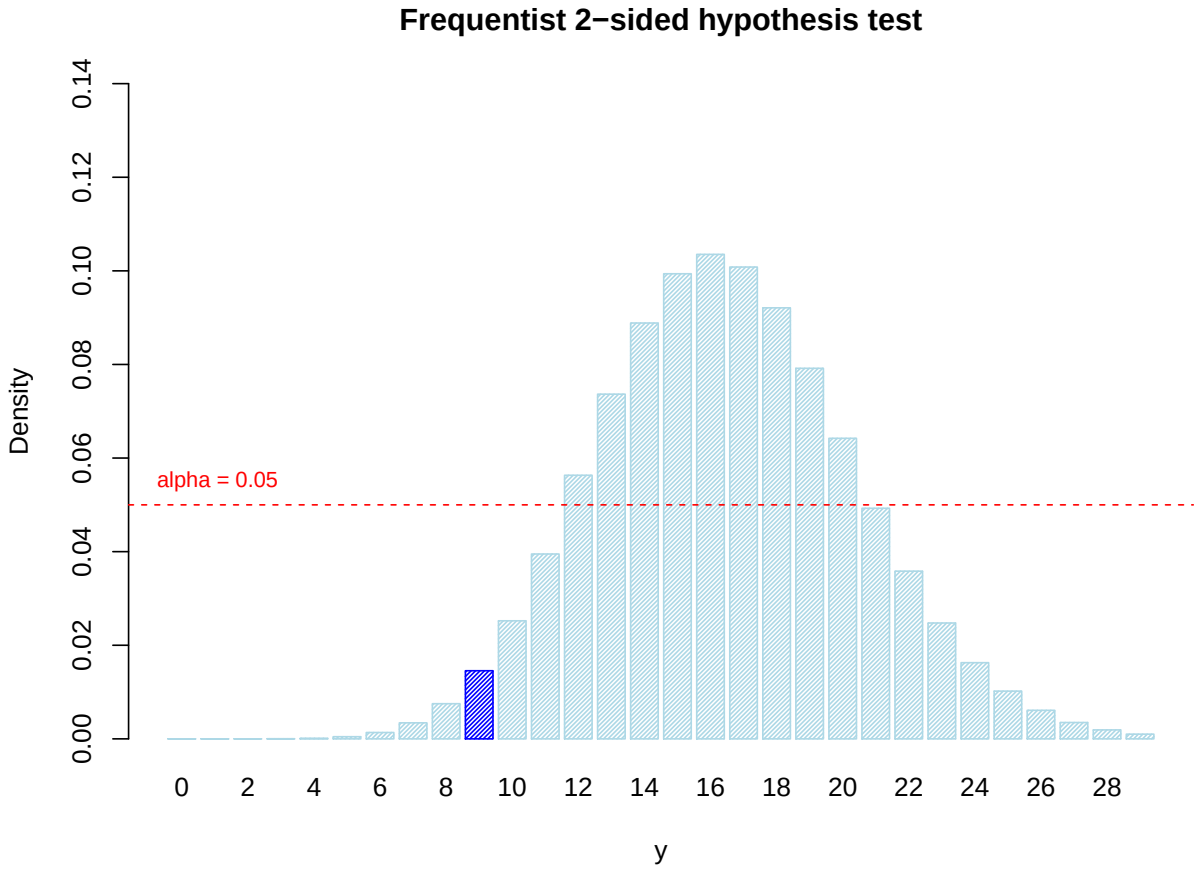
The bayesian estimator for p using the posterior of the old measurement as prior is: $p_B = 0.072$.
The posterior mean and variance are:

- mean: 0.072
- var : 0.001

The 95% credible interval for p is : [0.045, 0.104].

H) test the hypothesis H_0: p = 0.1 vs H_1: p $\neq$ 0.1 at 5% level of significance with both the frequentist and bayesian approach

```
### Frequentist
p.null <- 0.1
x <- 0:N
d.null <- dbinom(x, size=N, prob=p.null)

l.o.s = 0.05

barplot(d.null[1:30], names=x[1:30], ylim=c(0,0.14), density=50, border=TRUE,
        xlab='y',ylab="Density",main="Frequentist 2-sided hypothesis test",
        col=ifelse(x==9, "blue", "lightblue"))
abline(h=l.o.s, lty=2, col='red')
text(2, l.o.s+0.01, labels = paste("alpha =",l.o.s), pos=1, cex=0.8, col="red")
```



In the plot above it can be seen that P(y=9) is smaller than $\alpha = 0.05$, so the null hypothesis is rejected.

```
### Bayesian
# same prior as before
post.low.beta <- p[max(which(cumsum(posterior.beta)<=0.025*sum(posterior.beta)))]
post.upp.beta <- p[min(which(cumsum(posterior.beta)>=0.975*sum(posterior.beta)))]

# old posterior as new prior
post.low.new <- p[max(which(cumsum(posterior.new)<=0.025*sum(posterior.new)))]
post.upp.new <- p[min(which(cumsum(posterior.new)>=0.975*sum(posterior.new)))]
```

The probality $p_0 = 0.1$ is outside the 95% credible interval of the posterior distribution obtained using as prior a $Beta(1, 10)$ distribution:

- Lower bound: 0.027
- Upper bound: 0.096

The probality $p_0 = 0.1$ is inside the 95% credible interval of the posterior distribution obtained using as prior the posterior distribution of the previous measures:

- Lower bound: 0.045
- Upper bound: 0.104

So in the first case the null hypothesis is rejected, while in the second one it is accepted.