BDA - Assignment 3

Anonymous

Table of Contents

Exercise 1: Inference for normal mean and deviation 1
Exercise 2: Inference for the difference between proportions 6
Exercise 3: Inference for the difference between normal means 10

Exercise 1: Inference for normal mean and deviation

A factory has a production line for manufacturing car windshields. A sample of windshields has been taken for testing hardness. The observed hardness values y_1 can be found in file windshieldy1.txt. The data can also be accessed from the aaltobda R package as follows:

```
library(aaltobda)
data("windshieldy1")
head(windshieldy1)

## [1] 13.357 14.928 14.896 15.297 14.820 12.067
```

We may assume that the observations follow a normal distribution with an unknown standard deviation σ. We wish to obtain information about the unknown average hardness μ. For simplicity we assume standard noninformative prior discussed in the book, that is, $p(\mu, \sigma) \propto \sigma^{-1}$. It is not necessary to derive the posterior distribution in the report, as it has already been done in the book.

```
windshieldy_test <- c(13.357, 14.928, 14.896, 14.820)
n = length(windshieldy1)
y_mean = mean(windshieldy1)
nu = length(windshieldy1) - 1
s = sd(windshieldy1)
cat("Number of datapoints:", n,"\n")

## Number of datapoints: 9

cat("Data mean:", y_mean,"\n")

## Data mean: 14.61122

cat("Degree of freedom:", nu,"\n")

## Degree of freedom: 8
```

```
cat("standard deviation:",s,"\n")
```

```
## standard deviation: 1.474162
```

In the report, formulate (1) model likelihood, (2) the prior, and (3) the resulting posterior. Because the prior is noninformative:

(1) The model likelihood is the Guassian distribution: $p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)}$ or $p(y|\mu) = \frac{1}{1.47\sqrt{2\pi}} e^{\left(-\frac{1}{4.3459}(y-\mu)^2\right)}$

(2) The prior is noninformative:

$$p(\mu) = \sum_1^n (y_i - \overline{y})^2 + n(\overline{y} - \mu)^2 = \sum_1^n (y_i - 14.611)^2 + 9(14.611 - \mu)^2$$

(3) Under noninformative prior with unknown mean, the posterior distribution of μ follows the Student t-distribution, which has the form of: $p(\mu|y) = t_{n-1}(\mu|\overline{y}, s^2/n)$ or $t = \frac{\overline{y}-\mu}{s/\sqrt{n}}$. The scaling factor is $scale = \frac{s}{\sqrt{n}}$.
In this case, it is $p(\mu|y) = t_8(\mu|14.6112,0.2414)$

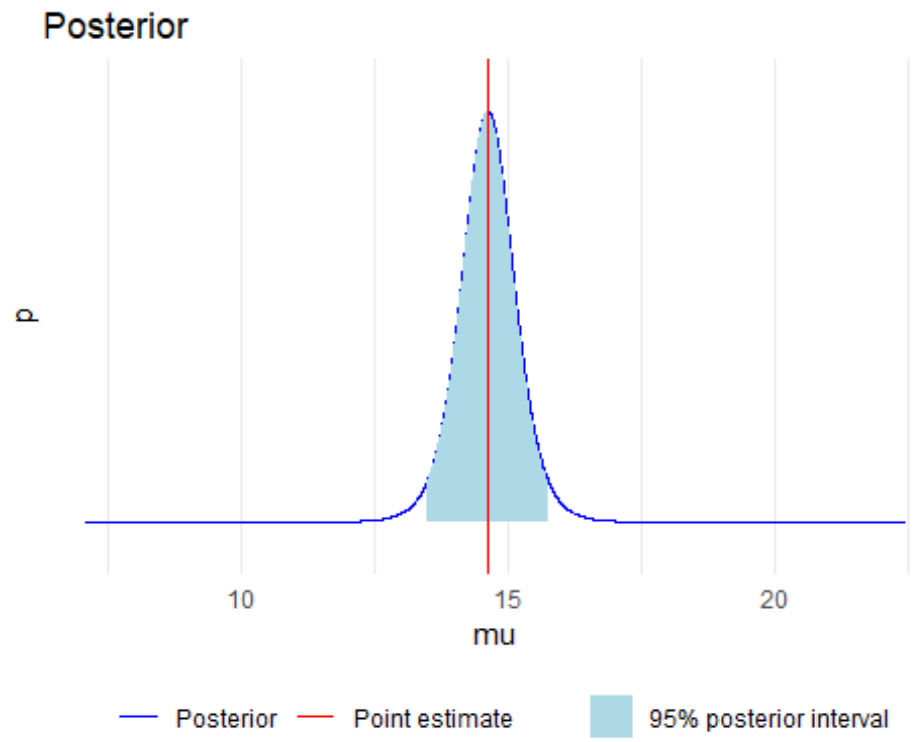
a) What can you say about the unknown μ? Summarize your results using Bayesian point estimate (i.e. $E(\mu|y)$), a posterior interval (95%), and plot the density. A test example can be found below for an uninformative prior. Note! Posterior intervals are also called credible intervals and are different from confidence intervals.

```
mu_point_est <- function(data){
  nu <- length(data) - 1
  n <- length(data)
  mu <- mean(data)
  scale <- sd(data)/sqrt(n)
  x <- rtnew(10000000, nu, mu, scale)
  return(mean(x))
}

mu_interval <- function(data, prob){
  probStart <- (1 - prob) /2
  probEnd <- 1 - (1 - prob) /2
  nu <- length(data) - 1
  n <- length(data)
  mu <- mean(data)
  scale <- sd(data)/sqrt(n)
  x <- rtnew(10000000, nu, mu, scale)
  return(quantile(x, c(probStart, probEnd)))
}

# mu_point_est(data = windshieldy_test)
# mu_interval(data = windshieldy_test, prob = 0.95)
print("The point estimate for the hardness mean of windshield is")
```

```
## [1] "The point estimate for the hardness mean of windshield is"

mu_point_est(data = windshieldy1)

## [1] 14.61114

print("The 95% posterior interval for the hardness mean of windshield is")

## [1] "The 95% posterior interval for the hardness mean of windshield is"

mu_interval(data = windshieldy1, prob = 0.95)

##     2.5%    97.5%
## 13.47738 15.74421

# mark_my_assignment()
```

The point estimate for the hardness mean of windshield is 14.61114 and the 95% posterior interval for the hardness mean of windshield is (2.5%, 97.5%) = (13.47738, 15.74421)

```
plotPosterior <- function(data){
nu <- length(data) - 1
n <- length(data)
mu <- mean(data)
scale <- sd(data)/sqrt(n)
start <- min(data) - 5
end <- max(data) + 5
x <- rtnew(10000000, nu, mu, scale)
expectedValue <- mean(x)
df1 <- data.frame(theta = seq(start, end, 0.001))
df1$p <- dtnew(df1$theta, nu, mu, scale)
df2 <- data.frame(theta = seq(qtnew(0.025, nu, mu, scale), qtnew(0.975, nu, m
u, scale), length.out = 100))
# compute the posterior density
df2$p <- dtnew(df2$theta, nu, mu, scale)
ggplot(mapping = aes(theta, p)) +
  geom_line(data = df1, aes(colour = "Posterior")) +
  # Add a layer of colorized 95% posterior interval
  geom_area(data = df2, aes(fill='1')) +
  # Decorate the plot a little
  # Add the expected value
  geom_vline(xintercept = expectedValue, linetype='solid', colour = "red") +
  labs(title='Posterior', y = 'p', x = 'mu') +
  scale_y_continuous(expand = c(0, 0.1), breaks = NULL) +
  scale_fill_manual(values = 'lightblue', labels = '95% posterior interval')
+
  scale_color_manual(name = "Type", values = c("Posterior" = "blue", "Point e
stimate" = "red")) +
  theme(legend.position = 'bottom', legend.title = element_blank())
}

plotPosterior(windshieldy1)
```



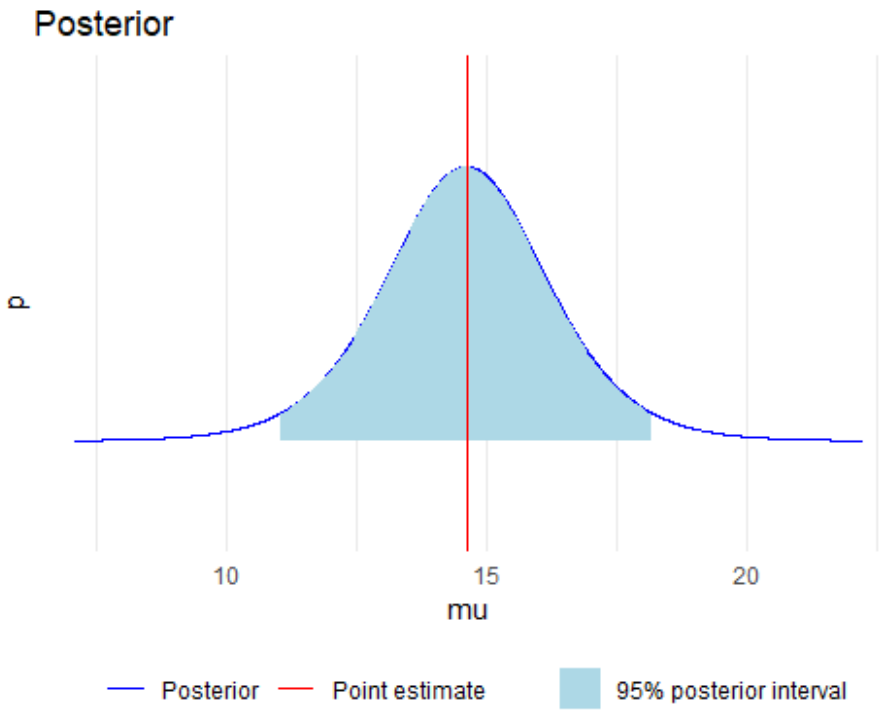
b) What can you say about the hardness of the next windshield coming from the production line before actually measuring the hardness? Summarize your results using Bayesian point estimate, a predictive interval (95%), and plot the density.
The predictive posterior distribution for the average hardness is:
$p(\tilde{\mu}|y) = t_{n-1}\left(\tilde{\mu}|\overline{y}, \left(1+\frac{1}{n}\right)s^2\right)$. In this case, $scale = \sqrt{\left(1+\frac{1}{n}\right)s^2}$

```
mu_pred_point_est <- function(data){
  nu <- length(data) - 1
  n <- length(data)
  mu <- mean(data)
  scale <- sqrt((sd(data)^2) * (1 + (1/n)))
  x <- rtnew(10000000, nu, mu, scale)
  return(mean(x))
}

mu_pred_interval <- function(data, prob){
  probStart <- (1 - prob) /2
  probEnd <- 1 - (1 - prob) /2
  nu <- length(data) - 1
  n <- length(data)
  mu <- mean(data)
  scale <- sqrt((sd(data)^2) * (1 + (1/n)))
  x <- rtnew(10000000, nu, mu, scale)
  return(quantile(x, c(probStart, probEnd)))
}
```

```r
# mu_pred_point_est(data = windshieldy_test)
# mu_pred_interval(data = windshieldy_test, prob = 0.95)
print("The predicted point estimate for the hardness mean of windshield is")
## [1] "The predicted point estimate for the hardness mean of windshield is"
mu_pred_point_est(data = windshieldy1)
## [1] 14.61159
print("The 95% predictive interval for the hardness mean of windshield is")
## [1] "The 95% predictive interval for the hardness mean of windshield is"
mu_pred_interval(data = windshieldy1, prob = 0.95)
##     2.5%    97.5%
## 11.03042 18.19367

# mark_my_assignment()
```

The predictive point estimate for the hardness mean of windshield is 14.61159 and the 95% predictive posterior interval for the hardness mean of windshield is (2.5%, 97.5%) = (11.03042, 18.19367)

```r
plotPredictivePosterior <- function(data){
nu <- length(windshieldy1) - 1
n <- length(windshieldy1)
mu <- mean(windshieldy1)
scale <- sqrt((sd(windshieldy1)^2) * (1 + (1/n)))
start <- min(windshieldy1) - 5
end <- max(windshieldy2) + 5
x <- rtnew(10000000, nu, mu, scale)
expectedValue <- mean(x)
df1 <- data.frame(theta = seq(start, end, 0.001))
df1$p <- dtnew(df1$theta, nu, mu, scale)
df2 <- data.frame(theta = seq(qtnew(0.025, nu, mu, scale), qtnew(0.975, nu, m
u, scale), length.out = 100))
# compute the posterior density
df2$p <- dtnew(df2$theta, nu, mu, scale)
ggplot(mapping = aes(theta, p)) +
  geom_line(data = df1, aes(colour = "Posterior")) +
  # Add a layer of colorized 95% posterior interval
  geom_area(data = df2, aes(fill='1')) +
  # Decorate the plot a little
  # Add the expected value
  geom_vline(xintercept = expectedValue, linetype='solid', colour = "red") +
  labs(title='Posterior', y = 'p', x = 'mu') +
  scale_y_continuous(expand = c(0, 0.1), breaks = NULL) +
  scale_fill_manual(values = 'lightblue', labels = '95% posterior interval')
+
  scale_color_manual(name = "Type", values = c("Posterior" = "blue", "Point e
stimate" = "red")) +
  theme(legend.position = 'bottom', legend.title = element_blank())
}
plotPredictivePosterior(windshieldy1)
```



Exercise 2: Inference for the difference between proportions

An experiment was performed to estimate the effect of beta-blockers on mortality of cardiac patients. A group of patients was randomly assigned to treatment and control groups: out of 674 patients receiving the control, 39 died, and out of 680 receiving the treatment, 22 died. Assume that the outcomes are independent and binomially distributed, with probabilities of death of p_0 and p_1 under the control and treatment, respectively. Set up a noninformative or weakly informative prior distribution on $(p_0; p_1)$.

In the report, formulate (1) model likelihood, (2) the prior, and (3) the resulting posterior. Because the two groups treatment and control are independent, this means there are two sets of prior, likelihood, and posteriors – one for each group.

(1) The model likelihood:

For p_0: $p(y|p_0) = \binom{n}{y} p_0^y (1-p_0)^{n-y} = \binom{674}{39} p_0^{39} (1-p_0)^{635}$

For p_1: $p(y|p_1) = \binom{n}{y} p_1^y (1-p_1)^{n-y} = \binom{680}{22} p_1^{22} (1-p_1)^{658}$

(2) The prior: prior $p(p_0)$ and $p(p_1)$ is Beta(1,1), which is a uniform or noninformative prior for binomial model.

(3) The resulting posterior:

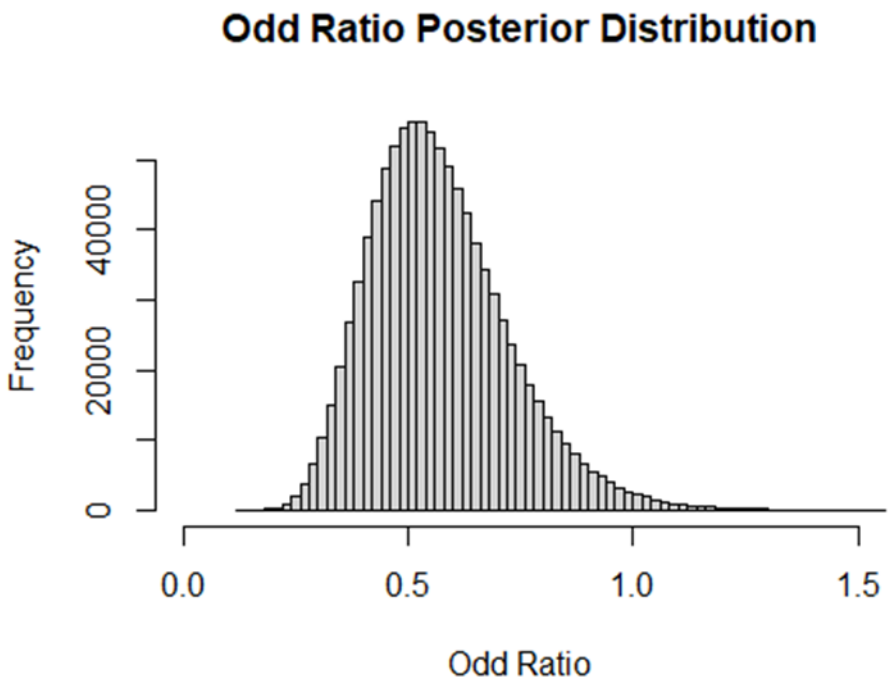
For p_0: $p(p_0|y) = Beta(1+39, 1+635) = Beta(40,636)$

For p_1: $p(p_1|y) = Beta(1+22, 1+658) = Beta(23,659)$

a) Summarize the posterior distribution for the odds ratio, $\frac{p_1/(1-p_1)}{p_0/(1-p_0)}$. Compute the point estimate, a posterior interval (95%), and plot the histogram. Use Frank Harrell's recommendations how to state results in Bayesian two group comparison.

```
set.seed(4711)
p0_test <- rbeta(1000000, 5, 95)
p1_test <- rbeta(1000000, 10, 90)
p0 <- rbeta(1000000, 40, 636)
p1 <- rbeta(1000000, 23, 659)
posterior_odds_ratio_point_est <- function(p0, p1){
  nominator = p1/(1 - p1)
  denominator = p0/(1 - p0)
  odd_ratio_posterior_distribution <- nominator/denominator
  posterior_mean <- mean(odd_ratio_posterior_distribution)
  hist(odd_ratio_posterior_distribution, breaks = 100, xlim=c(0,1.5), xlab="O
dd Ratio", main="Odd Ratio Posterior Distribution")
  return(posterior_mean)
}
posterior_odds_ratio_interval <- function(p0, p1, prob){
  probStart <- (1 - prob) /2
  probEnd <- 1 - (1 - prob) /2
  nominator = p1/(1 - p1)
  denominator = p0/(1 - p0)
  odd_ratio_posterior_distribution <- nominator/denominator
  posterior_quantile <- quantile(odd_ratio_posterior_distribution, probs=c(pr
obStart, probEnd))
  return(posterior_quantile)
}
# posterior_odds_ratio_point_est(p0 = p0_test, p1 = p1_test)
# posterior_odds_ratio_interval(p0 = p0_test, p1 = p1_test, prob = 0.9)
print("The point estimate for the odd ratio is")

## [1] "The point estimate for the odd ratio is"

posterior_odds_ratio_point_est(p0 = p0, p1 = p1)

## [1] 0.5702356

print("The 95% posterior interval for the odd ratio is")

## [1] "The 95% posterior interval for the odd ratio is"

posterior_odds_ratio_interval(p0 = p0, p1 = p1, prob = 0.95)

##      2.5%     97.5%
## 0.3210311 0.9260771
```

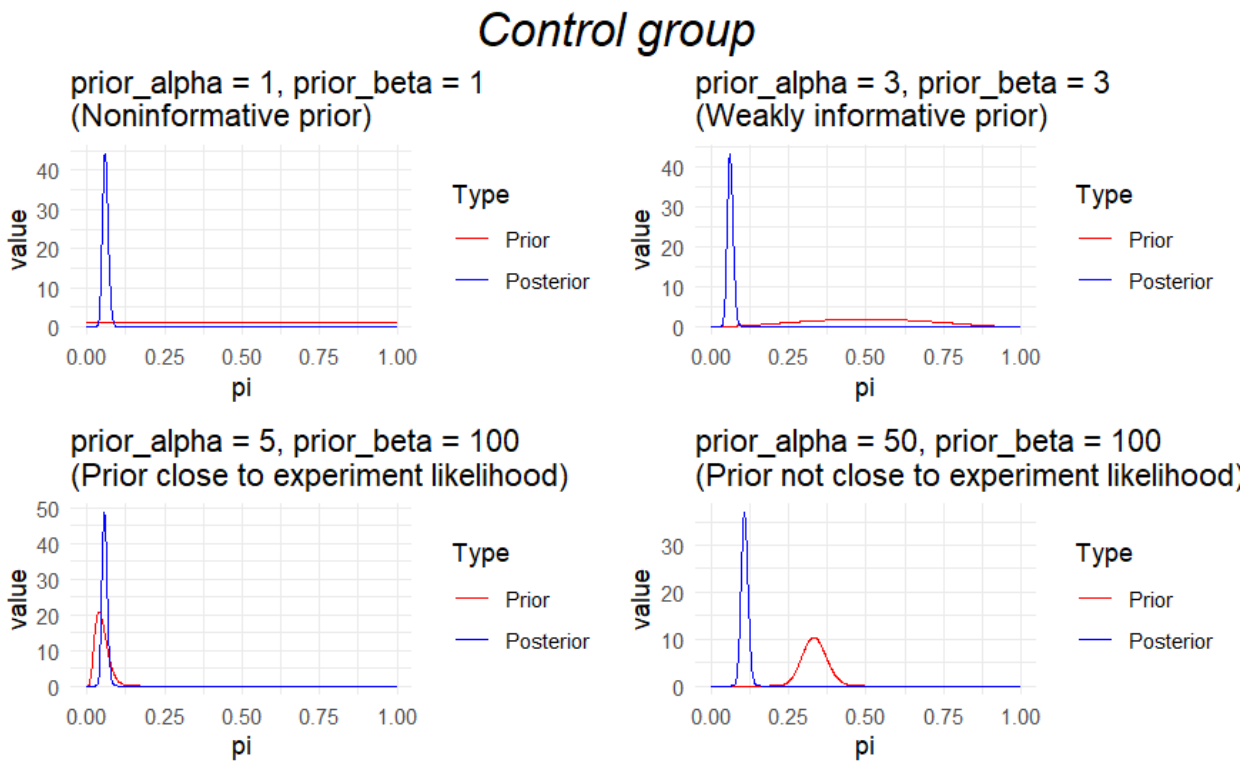
The point estimate for the odd ratio is 0.5702356 and the 95% posterior interval for odd ratio is (2.5%, 97.5%) = (0.3210311, 0.9260771)



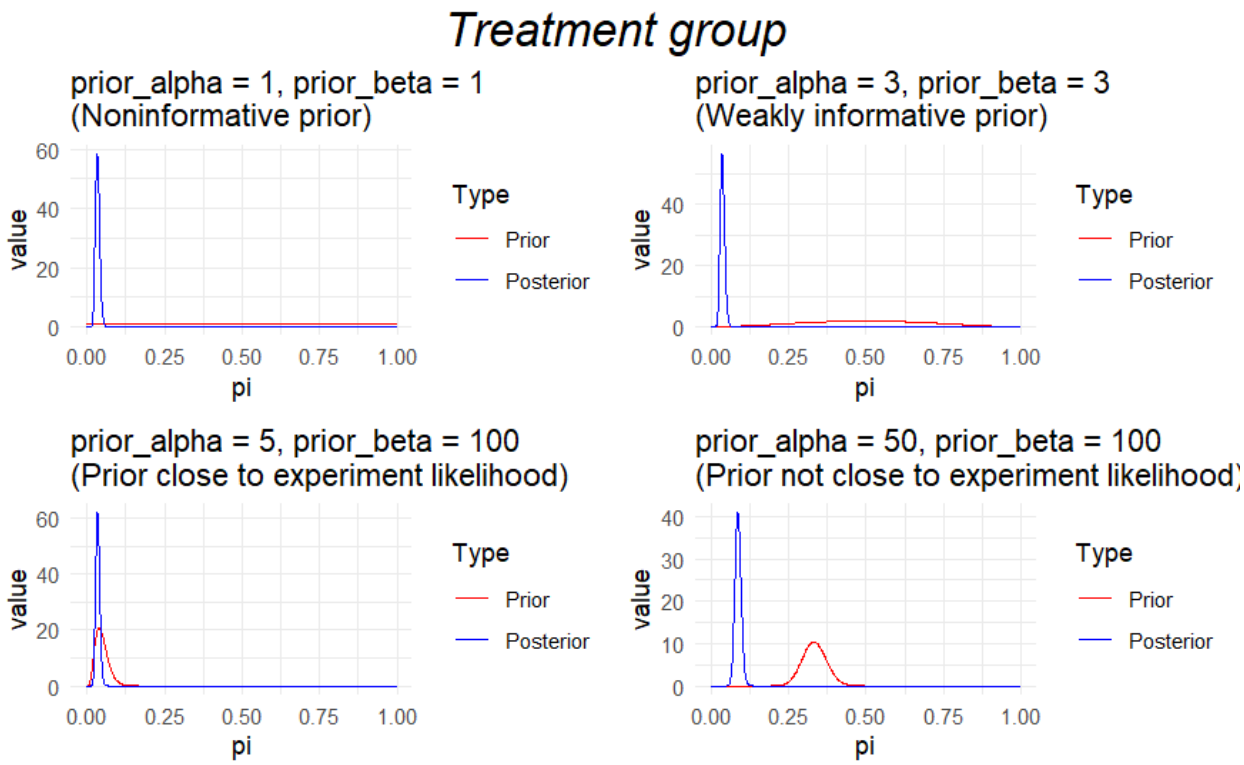
b) Discuss the sensitivity of your inference to your choice of prior density with a couple of sentences.

```
library(gridExtra)
library(grid)
prior_sensitivity_analysis <- function(prior_alpha, prior_beta, failure, succ
ess, description){
df <- data.frame(pi = seq(0, 1.0, 0.001))
posterior_alpha <- prior_alpha + failure
posterior_beta <- prior_beta + success
df$prior <- dbeta(df$pi, prior_alpha, prior_beta)
df$posterior <- dbeta(df$pi, posterior_alpha, posterior_beta)
plt<- (ggplot(df)
+geom_line(aes(x = pi, y = prior, colour="Prior"))
+geom_line(aes(x = pi, y = posterior, colour="Posterior"))
+labs(title=paste0('prior_alpha = ', prior_alpha, ', prior_beta = ', prior_be
ta, " \n(", description, ")"), y = 'value')
+scale_color_manual(name = "Type", values = c("Prior" = "red", "Posterior" =
"blue"))
)
return(plt)
}
# For the control group
death = 39
survive = 635
p1 <- prior_sensitivity_analysis(prior_alpha = 1, prior_beta = 1, death, surv
```

```
ive, "Noninformative prior")
p2 <- prior_sensitivity_analysis(prior_alpha = 3, prior_beta = 3, death, surv
ive, "Weakly informative prior")
p3 <- prior_sensitivity_analysis(prior_alpha = 5, prior_beta = 100, death, su
rvive, "Prior close to experiment likelihood")
p4 <- prior_sensitivity_analysis(prior_alpha = 50, prior_beta = 100, death, s
urvive, "Prior not close to experiment likelihood")
grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2, top = textGrob("Control grou
p", gp=gpar(fontsize=20,font=3)))
```




```
# For the treatment group
death = 22
survive = 658
p1 <- prior_sensitivity_analysis(prior_alpha = 1, prior_beta = 1, death, surv
ive, "Noninformative prior")
p2 <- prior_sensitivity_analysis(prior_alpha = 3, prior_beta = 3, death, surv
ive, "Weakly informative prior")
p3 <- prior_sensitivity_analysis(prior_alpha = 5, prior_beta = 100, death, su
rvive, "Prior close to experiment likelihood")
p4 <- prior_sensitivity_analysis(prior_alpha = 50, prior_beta = 100, death, s
urvive, "Prior not close to experiment likelihood")
grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2, top = textGrob("Treatment gr
oup", gp=gpar(fontsize=20,font=3)))
```



Sensitivity of posterior distribution based on the prior:

As we can see for both control and treament group, the noninformative or weakly informative prior has no or miniscule effects on the posterior distribution. When the prior is close to the likelihood, the shape of the posterior bell shape becomes much sharper due to stronger belief in the true distribution. However, if the prior is far away from the experiment likelihood, the posterior bell shape becomes a little flatter and shifted slightly towards the prior location, suggesting that the posterior has less belief in the data and tends towards the given prior.

Exercise 3: Inference for the difference between normal means

Consider a case where the same factory has two production lines for manufacturing car windshields. Independent samples from the two production lines were tested for hardness. The hardness measurements for the two samples y1 and y2 are given in the files windshieldy1.txt and windshieldy2.txt. These can be accessed directly with

```
data("windshieldy1")
n = length(windshieldy1)
y_mean = mean(windshieldy1)
df = length(windshieldy1) - 1
s = sd(windshieldy1)
print("windshieldy1 stats")
```

```
## [1] "windshieldy1 stats"

cat("Number of datapoints:", n,"\n")

## Number of datapoints: 9

cat("Data mean:", y_mean,"\n")

## Data mean: 14.61122

cat("Degree of freedom:", df,"\n")

## Degree of freedom: 8

cat("standard deviation:",s,"\n")

## standard deviation: 1.474162

data("windshieldy2")
n = length(windshieldy2)
y_mean = mean(windshieldy2)
df = length(windshieldy2) - 1
s = sd(windshieldy2)
print("windshieldy2 stats")

## [1] "windshieldy2 stats"

cat("Number of datapoints:", n,"\n")

## Number of datapoints: 13

cat("Data mean:", y_mean,"\n")

## Data mean: 15.82108

cat("Degree of freedom:", df,"\n")

## Degree of freedom: 12

cat("standard deviation:",s,"\n")

## standard deviation: 0.8726099
```

We assume that the samples have unknown standard deviations σ_1 and σ_2. In the report, formulate (1) model likelihood, (2) the prior, and (3) the resulting posterior.
Because the two production lines are independent, this means there are two sets of prior, likelihood, and posteriors – one for each production line.

For windshieldy1 data:

(1) Likelihood: $p(y|\mu_1,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(y-\mu_1)^2\right)} = \frac{1}{1.47\sqrt{2\pi}} e^{\left(-\frac{1}{4.3459}(y-\mu_1)^2\right)}$

(2) The noninformative prior: $p(\mu_1) = \sum_1^n (y_i - \overline{y})^2 + n(\overline{y} - \mu_1)^2 = \sum_1^n (y_i - 14.611)^2 + 9(14.611 - \mu_1)^2$

(3) The posterior distribution of μ_1 follows the Student t-distribution: $p(\mu_1|y) = t_{n-1}(\mu_1|\overline{y}, s^2/n)$ or $t = \frac{\overline{y}-\mu_1}{s/\sqrt{n}}/$. The scaling factor is $scale = \frac{s}{\sqrt{n}}$

In this case, it is $p(\mu_1|y) = t_5(\mu_1|14.6112,0.2414)$

For windshieldy2 data:

(1) Likelihood: $p(y|\mu_2, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(y-\mu_2)^2\right)} = \frac{1}{0.87\sqrt{2\pi}} e^{\left(-\frac{1}{1.5228}(y-\mu_2)^2\right)}$

(2) The noninformative prior:

$p(\mu_2) = \sum_1^n (y_i - \overline{y})^2 + n(\overline{y} - \mu_2)^2 = \sum_1^n (y_i - 15.821)^2 + 13(15.821 - \mu_2)^2$

(3) The posterior distribution of μ_1 follows the Student t-distribution:

$p(\mu_2|y) = t_{n-1}(\mu_2|\overline{y}, s^2/n)$ or $t = \frac{\overline{y}-\mu_2}{s/\sqrt{n}}/$. The scaling factor is $scale = \frac{s}{\sqrt{n}}$

In this case, it is $p(\mu_2|y) = t_{12}(\mu_2|15.821,0.0585)$

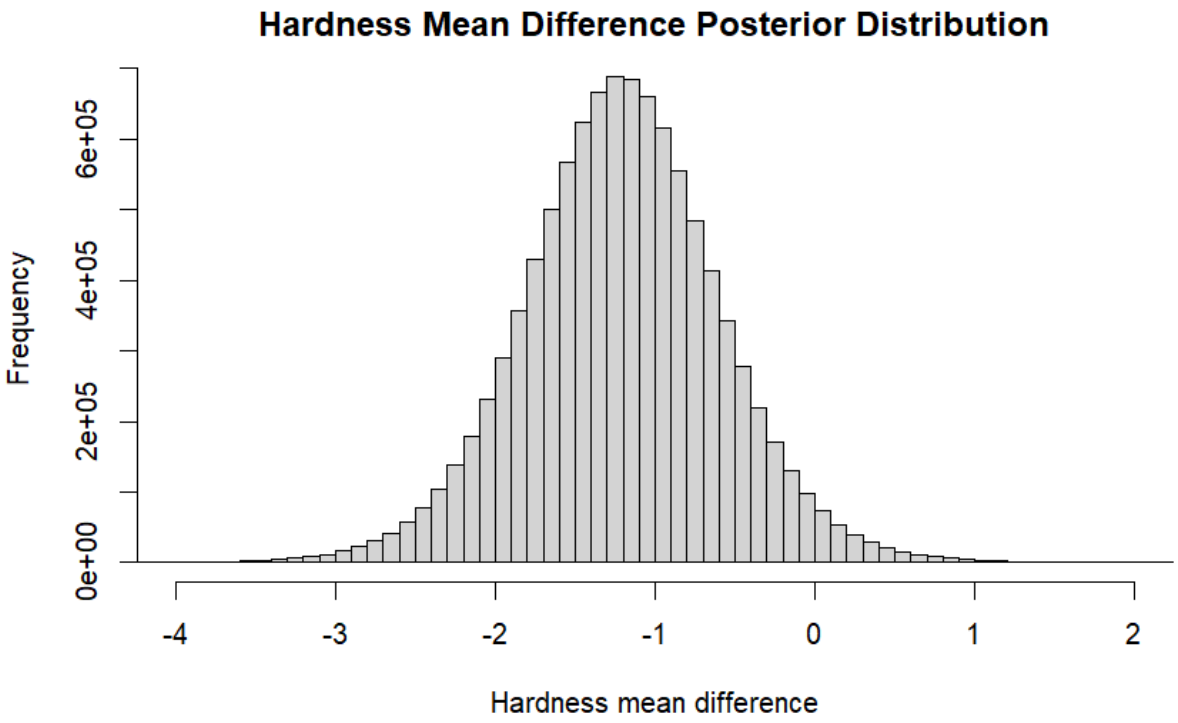
Use uninformative or weakly informative priors and answer the following questions:

a) What can you say about $\mu_d = \mu_1 - \mu_2$? Summarize your results using a Bayesian point estimate, a posterior interval (95%), and plot the histogram. Use Frank Harrell's recommendations how to state results in Bayesian two group comparison.

```
mu_difference_point_est <- function(data1, data2){
  df1 <- length(data1) - 1
  n1 <- length(data1)
  mu1 <- mean(data1)
  scale1 <- sd(data1)/sqrt(n1)
  x1 <- rtnew(10000000, df1, mu1, scale1)
  df2 <- length(data2) - 1
  n2 <- length(data2)
  mu2 <- mean(data2)
  scale2 <- sd(data2)/sqrt(n2)
  x2 <- rtnew(10000000, df2, mu2, scale2)

  mu_difference_posterior_distribution = x1 - x2
  hist(mu_difference_posterior_distribution, breaks = 200, xlim=c(-4,2), xlab
="Hardness mean difference", main="Hardness Mean Difference Posterior Distrib
ution")
  return(mean(mu_difference_posterior_distribution))
}
```

```r
mu_difference_interval <- function(data1, data2, prob){
  probStart <- (1 - prob) /2
  probEnd <- 1 - (1 - prob) /2
  df1 <- length(data1) - 1
  n1 <- length(data1)
  mu1 <- mean(data1)
  scale1 <- sd(data1)/sqrt(n1)
  x1 <- rtnew(10000000, df1, mu1, scale1)

  df2 <- length(data2) - 1
  n2 <- length(data2)
  mu2 <- mean(data2)
  scale2 <- sd(data2)/sqrt(n2)
  x2 <- rtnew(10000000, df2, mu2, scale2)
  mu_difference_posterior_distribution = x1 - x2
  return(quantile(mu_difference_posterior_distribution, c(probStart, probEnd)
))
}

print("The point estimate for the hardness difference mean of windshieldy1 an
d windshieldy2 is")

## [1] "The point estimate for the hardness difference mean of windshieldy1 a
nd windshieldy2 is"

mu_difference_point_est(data1 = windshieldy1, data2 = windshieldy2)

## [1] -1.209878

print("The 95% posterior interval for the hardness difference mean of windshi
eldy1 and windshieldy2 is")

## [1] "The 95% posterior interval for the hardness difference mean of windsh
ieldy1 and windshieldy2 is"

mu_difference_interval(data1 = windshieldy1, data2 = windshieldy2, prob = 0.9
5)

##        2.5%       97.5%
## -2.45307530  0.03300713
```

The point estimate for the odd ratio is -1.209878 and the 95% posterior interval for odd ratio is (2.5%, 97.5%) = (-2.45307530, 0.03300713)



b) Given this specific model, what is the probability that the means are exactly the same ($\mu_1 = \mu_2$)? Explain your reasoning.

Because the posterior distributions of μ_1 and μ_2 are continuous, the posterior distribution of the difference $\mu_d = \mu_1 - \mu_2$ is also continuous. In order words, the probability that the means are exactly the same, or $\mu_d = 0$ is complete 0 because continuous distributions are undefined, which is 0, for single point estimate. Continuous distribution only has defined probability for ranges of values.