BSDA: Assignment 3

Anonymous student

Contents

General Information to include

- **Time used for reading and self-study exercises**: ~ 6 hours.
- **Time used for the assignment**: ~11 hours
- **Good with assignment**: The level of the questions were tough but fair. I like that if you go to class and you read the book you basically can solve everything.
- **Things to improve in the assignment**: I feel like in class we discussed too little about sensitivity analysis and how to properly do one. I struggle coming up with some meaningful alternative priors that are informative.

1. Inference for normal mean and deviation

```
library(bsda)
library(ggplot2)
data("windshieldy1")
head(windshieldy1)
```

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

```
length(windshieldy1)
```

```
## [1] 9
```

1. Likelihood:

The sampling distribution with unknown μ and σ^2 is

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

2. Prior:

The uninformative prior is

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

NOTE: We can derive the prior on variance scale if prior is uniform on $(\mu, log\sigma)$ in this way:

$$\begin{aligned} \text{Let Y} &= log(\sigma^2) \\ p(Y) &\propto 1 \\ \frac{dY}{d\sigma^2} &= \frac{1}{\sigma^2} \\ \text{Then if X} &= \sigma^2 \\ p(X) = p(Y)|\frac{dY}{d\sigma^2}| &\propto 1 \times \frac{1}{\sigma^2} \propto (\sigma^2)^{-1} \end{aligned}$$

3. Posterior:

With a sample of independent and normally distributed observations $y = (y_1, \ldots, y_9)$, the posterior density is:

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto (\sigma^2)^{-1} \prod_{i=1}^{9} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ &\propto \sigma^{-9-2} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{9}(y_i-\mu)^2\right) \\ &= \sigma^{-9-2} exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{9}(y_i-\bar{y})^2 + 9(\bar{y}-\mu)^2\right]\right), \\ \text{where } \bar{y} &= \frac{1}{9}\sum_{i=1}^{9} y_i \\ &= \sigma^{-9-2} exp\left(-\frac{1}{2\sigma^2}\left[(9-1)s^2 + 9(\bar{y}-\mu)^2\right]\right), \\ \text{where } s^2 &= \frac{1}{9-1}\sum_{i=1}^{9}(y_i-\bar{y})^2 \end{aligned}$$

The marginal posterior of μ can be obtained by integrating σ^2 out of the joint posterior

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y) d\sigma^2$$

Using integration by substitution

$$x = \frac{A}{2\sigma^2}, \text{where } A = (9-1)s^2 + 9(\mu - \bar{y})$$

Recognizing the result is an unnormalized gamma integral:

$$\begin{aligned} p(\mu|y) &\propto A^{-9/2} \int_0^\infty z^{(9-2)/2} exp(-z) dz \\ &\propto [(9-1)s^2 + 9(\mu-\bar{y})^2]^{-9/2} \\ &\propto \left[1 + \frac{9(\mu-\bar{y})^2}{(9-1)s^2}\right]^{-9/2} \end{aligned}$$

This is the Student's t $p(\mu|y) = t_{9-1}(\mu|\bar{y}, s^2/9)$.

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.

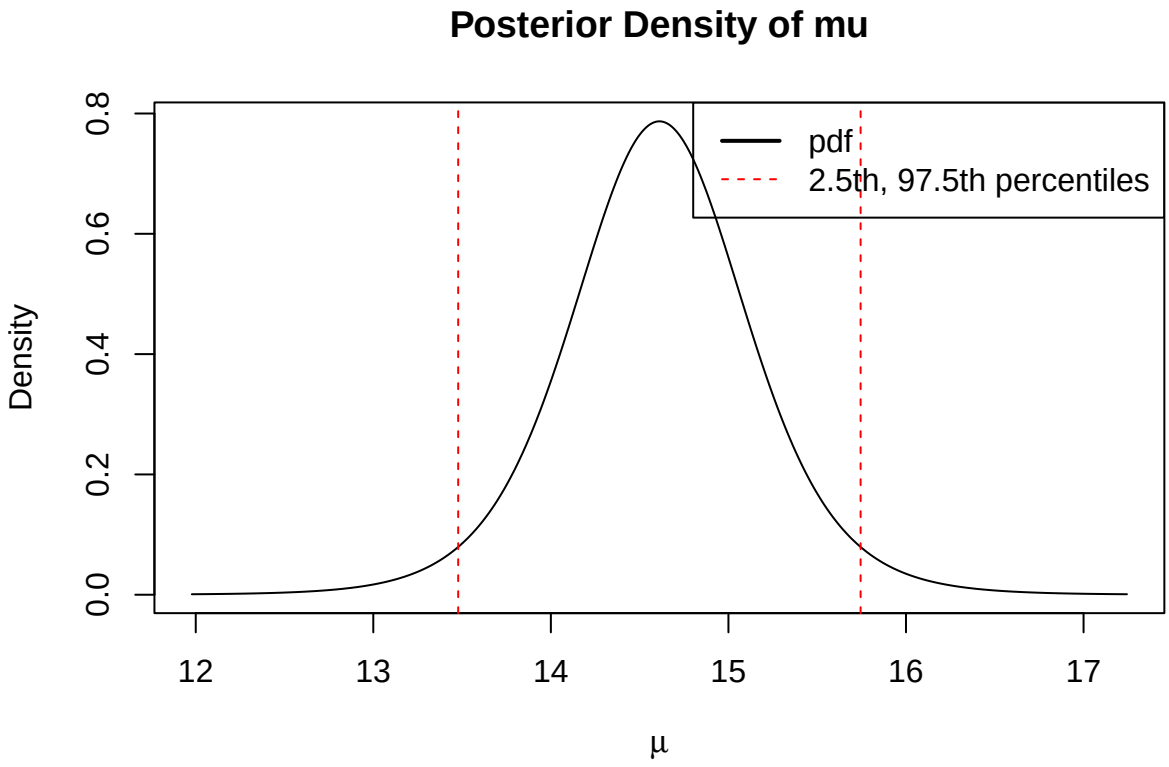
```
#define function for point estimate
mu_point_est <- function(data) {

  y_bar <- mean(data)

  return(y_bar)
}

mu_point_est(windshieldy1)
```

```
## [1] 14.61122
```

```
mu_interval <- function(data, prob) {

  n <- length(data)
  y_bar <- mean(data)
  s2 <- var(data)
  df <- n - 1
  alpha <- 1-prob

  lower_quantile <- qtnew(alpha / 2, df, mean = y_bar, scale = sqrt(s2/n))
  upper_quantile <- qtnew(1 - alpha / 2, df, mean = y_bar, scale = sqrt(s2/n))

  return(c(lower_quantile, upper_quantile))
}

mu_interval(windshieldy1, prob = 0.95)
```

```
## [1] 13.47808 15.74436
```

```
#plot the posterior density for mu using the Student's t-distribution
x <- seq(mu_interval(windshieldy1, prob= 0.95)[1]-1.5,
         mu_interval(windshieldy1, prob= 0.95)[2]+1.5, length.out = 1000)

posterior_density <- dtnew(x, 8, mean = mean(windshieldy1),
                           scale = sqrt(var(windshieldy1)/9))

plot(x, posterior_density, type = 'l', xlab = expression(mu), ylab = 'Density',
     main = 'Posterior Density of mu')

#add vertical lines for the credible interval
abline(v = mu_interval(windshieldy1, prob= 0.95)[1], col = 'red', lty = 2)
abline(v = mu_interval(windshieldy1, prob= 0.95)[2], col = 'red', lty = 2)

legend("topright", legend = c("pdf", "2.5th, 97.5th percentiles"),
       col = c("black", "red"), lwd = c(2, 1), lty = c(1, 2))
```



$$E(\mu|y) = \bar{y} = \frac{1}{9}\sum_{i=1}^{9} y_i = 14.611$$

To find the posterior interval we we can find the 2.5th and 97.5th percentiles of the t-distribution. We get the following interval: There is a 95% probability that μ, average hardness, falls in $[13.478, 15.744]$

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 posterior predictive distribution for $\tilde{y}$ can be written as:

$$p(\tilde{y}|y) = \int\int p(\tilde{y}|\mu, \sigma^2, y)p(\mu, \sigma^2|y)d\mu d\sigma^2$$

Using similar steps as in the derivation of the posterior distribution of μ, we can obtain an analytic form for the posterior predictive distribution of $\tilde{y}$, which is a t distribution.

$$p(\tilde{y}|y) = t_{9-1}\left(\tilde{y} \mid \bar{y}, \left(1 + \frac{1}{9}\right)s^2\right)$$

```
mu_pred_point_est <- function(data) {
```

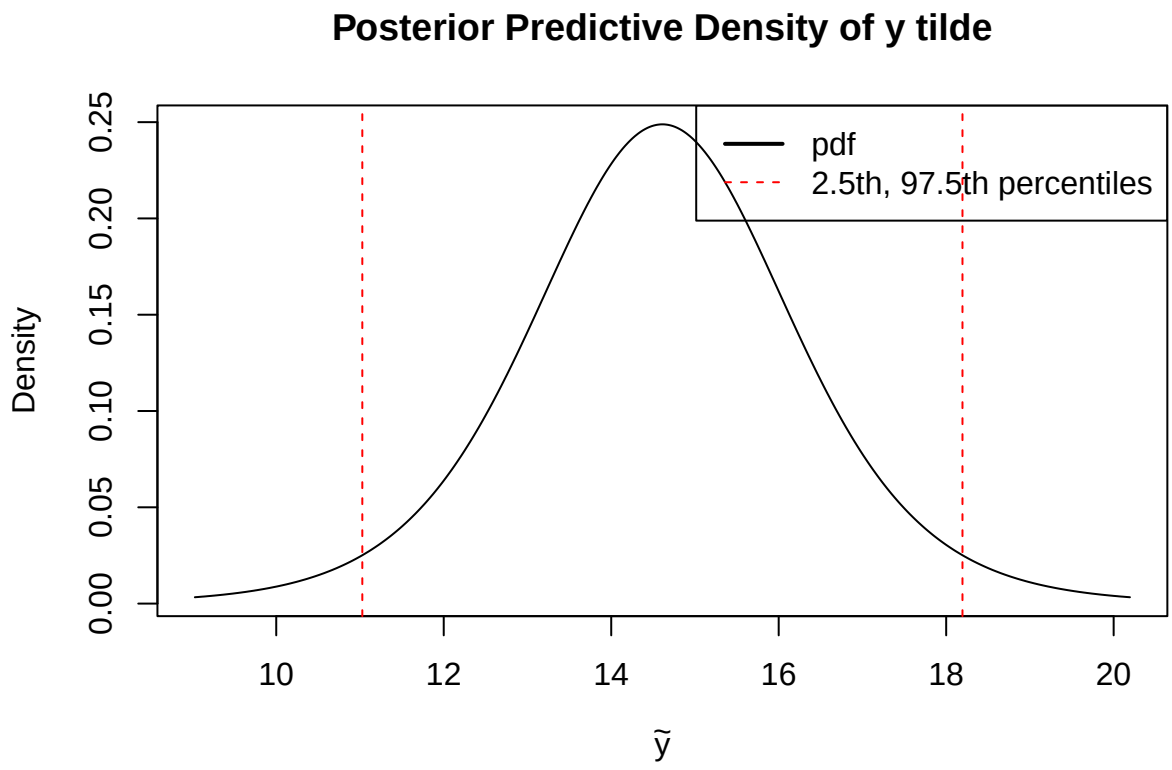
```
  y_bar <- mean(data)

  return(y_bar)
}

mu_pred_point_est(windshieldy1)
```

```
## [1] 14.61122
```

```
mu_pred_interval <- function(data, prob) {

  n <- length(data)
  y_bar <- mean(data)
  s2 <- var(data)
  df <- n - 1
  alpha <- 1-prob

  lower_quantile <- qtnew(alpha / 2, df,
                          mean = y_bar, scale = sqrt(s2*(1+1/n)))
  upper_quantile <- qtnew(1 - alpha / 2, df,
                          mean = y_bar, scale = sqrt(s2*(1+1/n)))

  return(c(lower_quantile, upper_quantile))
}

mu_pred_interval(windshieldy1, prob = 0.95)
```

```
## [1] 11.02792 18.19453
```

```
#plot the posterior density for mu using the Student's t-distribution
x_pred <- seq(mu_pred_interval(windshieldy1, prob= 0.95)[1]-2,
         mu_pred_interval(windshieldy1, prob= 0.95)[2]+2, length.out = 1000)

posterior_pred_density <- dtnew(x_pred, 8, mean = mean(windshieldy1),
                           scale = sqrt(var(windshieldy1)*(1+1/9))
                             )

plot(x_pred, posterior_pred_density, type = 'l', xlab = expression(tilde(y)), ylab = 'Density',
     main = 'Posterior Predictive Density of y tilde')

#add vertical lines for the credible interval
abline(v = mu_pred_interval(windshieldy1, prob= 0.95)[1], col = 'red', lty = 2)
abline(v = mu_pred_interval(windshieldy1, prob= 0.95)[2], col = 'red', lty = 2)

legend("topright", legend = c("pdf", "2.5th, 97.5th percentiles"),
       col = c("black", "red"), lwd = c(2, 1), lty = c(1, 2))
```



Point estimate:

$$E(\tilde{y}|y) = \bar{y} = \frac{1}{9}\sum_{i=1}^{9} y_i = 14.611$$

Predictive interval:

There is a 95% probability that the hardness windshield of the next is between $[11.028, 18.195]$

2. Inference for the difference between proportions

We have data of the form:

$$(x_i, n_i, y_i);\ i = 0, 1$$

where x_i represents the ith of either receiving the treatment or not given to n_i patients, of which y_i died.

1. The likelihood:

The likelihood for each group i (i.e., control and treatment) assuming that the outcomes are independent and binomially distributed is:

$$\begin{aligned} p(y_i|p_i, n_i) = p(y_i|p_i) = Bin(y_i|n_i, p_i) &= \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i} \\ &\propto p_i^{y_i}(1-p_i)^{n_i - y_i} \end{aligned}$$

where $i = 0, 1$

2. Uninformative prior:

independent and locally uniform in the two parameters that is

$$p(p_0, p_1) \propto 1$$

3. Posterior:

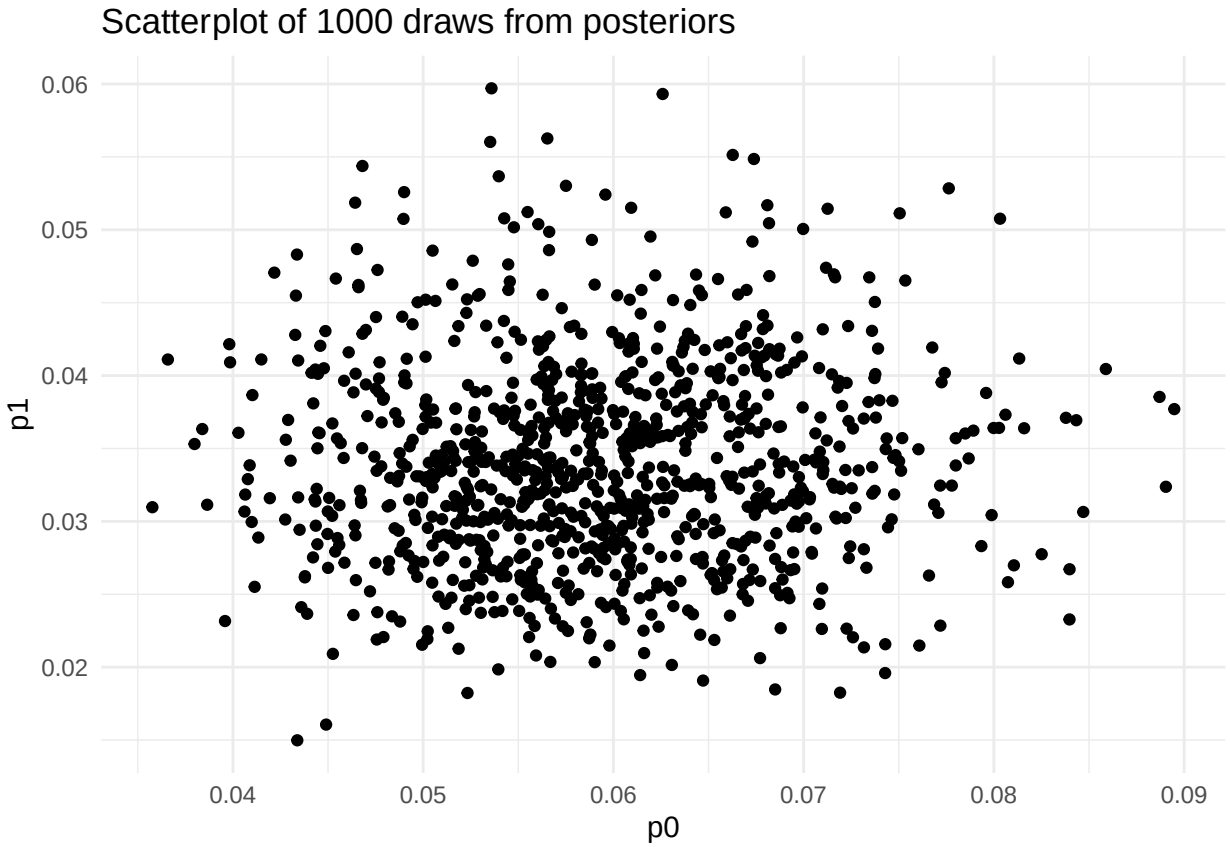
$$\begin{aligned} p(p_0, p_1|y) &\propto p(p_0, p_1)p(y|p_0, p_1) \\ &\propto p(p_0, p_1) \prod_{i=0}^{1} p(y_i|p_0, p_1) \\ &\propto (p_0^{y_0}(1-p_0)^{n_0-y_0})(p_1^{y_1}(1-p_1)^{n_1-y_1}) \end{aligned}$$

The parameters p_0 and p_1 would have independent beta posterior distributions.

$$\begin{aligned} p(p_0|y_0) &= Beta(p_0|\alpha + y_0, \beta + n_0 - y_0) = Beta(p_0|40, 636) \\ p(p_1|y_1) &= Beta(p_1|\alpha + y_1, \beta + n_1 - y_1) = Beta(p_1|23, 659) \end{aligned}$$

a) Summarize the posterior distribution for the odds ratio, $(p_1/(1-p_1))/(p_0/(1-p_0))$. Compute the point estimate, a posterior interval (95%), and plot the histogram.

```
set.seed(4711)
p0 <- rbeta(100000, 40, 636)
p1 <- rbeta(100000, 23, 659)
odds_ratios <- (p1 / (1 - p1)) / (p0 / (1 - p0))

data <- data.frame(rbeta(1000, 40, 636), rbeta(1000, 23, 659))

ggplot(data, aes(x = rbeta(1000, 40, 636), y = rbeta(1000, 23, 659))) +
  geom_point() +
  labs(x = "p0", y = "p1") +
  ggtitle("Scatterplot of 1000 draws from posteriors") +
  theme_minimal()
```



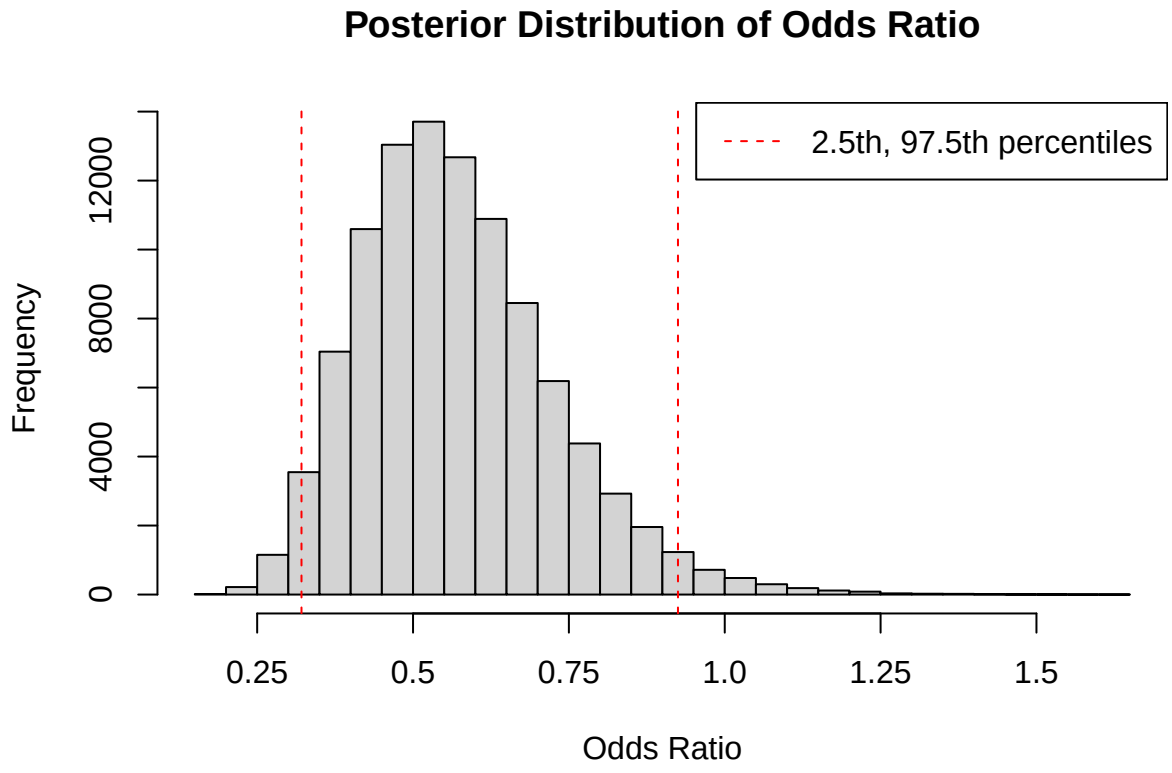

```
posterior_odds_ratio_point_est <- function(p0, p1) {

  odds_ratios <- (p1 / (1 - p1)) / (p0 / (1 - p0))
  odds_ratio_point_estimate <- mean(odds_ratios)

  return(odds_ratio_point_estimate)
}

#point estimate
p_or_est <- posterior_odds_ratio_point_est(p0 = p0, p1 = p1)
posterior_odds_ratio_point_est(p0 = p0, p1 = p1)
```

```
## [1] 0.570978
```

```
posterior_odds_ratio_interval <- function(p0, p1, prob) {

  odds_ratios <- (p1 / (1 - p1)) / (p0 / (1 - p0))
  lower_quantile <- quantile(odds_ratios, (1-prob)/2)
  upper_quantile <- quantile(odds_ratios, 1-(1-prob)/2)
  odds_ratio_interval <- c(lower_quantile, upper_quantile)

  return(odds_ratio_interval)
}

#posterior interval
p_or_int <- posterior_odds_ratio_interval(p0 = p0, p1 = p1, prob = 0.95)
posterior_odds_ratio_interval(p0 = p0, p1 = p1, prob = 0.95)
```

```
##     2.5%    97.5%
## 0.321063 0.924998
```

```r
#plot the histogram of the odds ratio samples
hist(odds_ratios, breaks = 50, main = "Posterior Distribution of Odds Ratio",
     xlab = "Odds Ratio", ylab = "Frequency")

axis(1, at=seq(0.25,1.25,by=0.50), labels=seq(0.25,1.25,by=0.50))

abline(v = posterior_odds_ratio_interval(p0 = p0, p1 = p1, prob = 0.95)[1],
       col = 'red', lty = 2)
abline(v = posterior_odds_ratio_interval(p0 = p0, p1 = p1, prob = 0.95)[2],
       col = 'red', lty = 2)

legend("topright", legend = c("2.5th, 97.5th percentiles"),
       col = c("red"), lwd = c(1), lty = c( 2))
```



Point estimate:

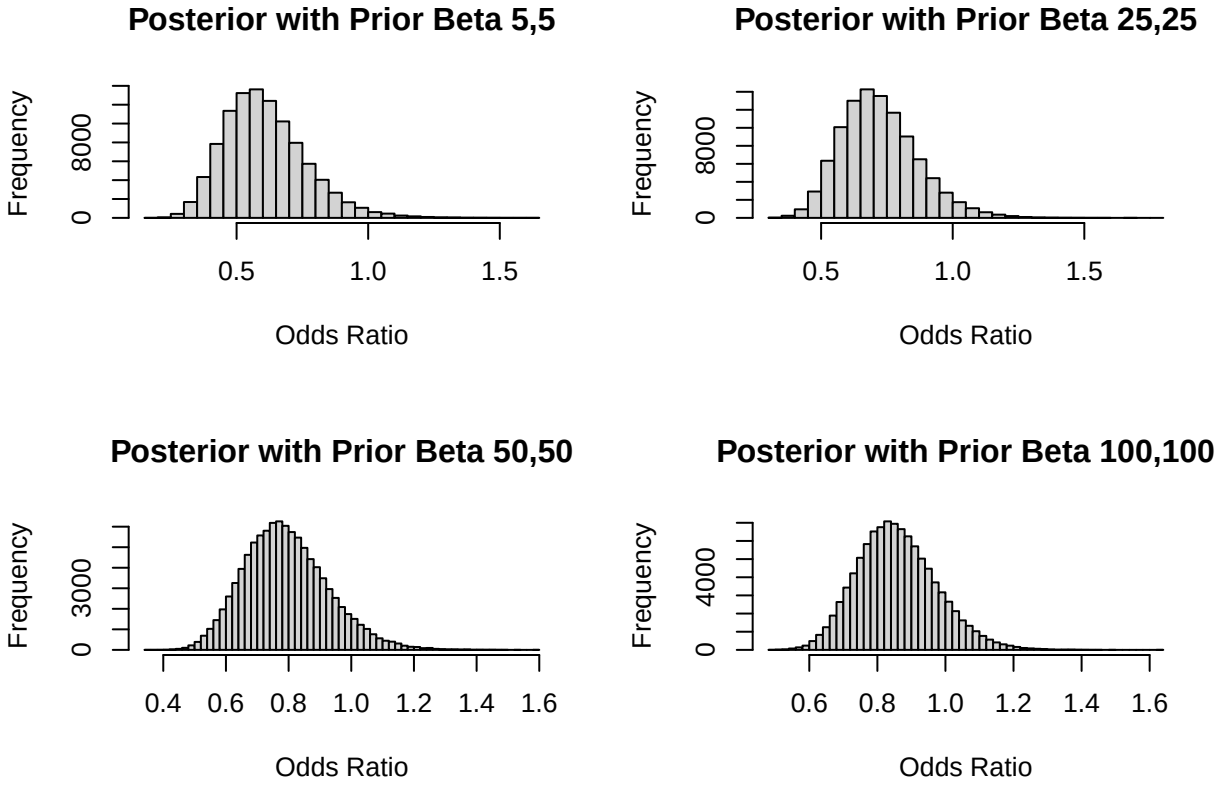
$$E\left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}|y\right) = 0.571$$

Posterior interval:

There is a 95% probability that the odds ratio, falls in $[0.321, 0.925]$ given the evidence provided by the observed data.

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

Using priors increasingly concentrated around 0.5



Prior distribution		Posterior distribution	
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	Posterior odds ratio	95% posterior interval odds ratio
0.5	2	0.570978	[0.321063, 0.924998]
0.5	10	0.6079862	[0.3561132, 0.9606887]
0.5	50	0.7217833	[0.4778566, 1.0477749]
0.5	100	0.7923978	[0.5598988, 1.0830403]
0.5	200	0.8580166	[0.6520223, 1.1074219]

The odds ratio posterior seems to be somewhat sensible to the prior distribution, we can see in the plots that the posterior is moving. When the priors contain information equivalent to more than 100 observations, our posterior distribution and intervals are largely pulled towards 0.8; by looking, at the last intervals seems like prior mean 0.5 is unlikely in the eyes of the posterior.

3. Inference for the difference between normal means

```
data("windshieldy1")
data("windshieldy2")
```

Assuming exchangeability among the two groups.

1. Likelihood:

Assuming measurements were taken at random from normal distributions

$$p(y|\mu_1, \mu_2, \sigma_1, \sigma_2) = \prod_{i=1}^{9} N(y_{1i}|\mu_1, \sigma_1^2) \prod_{i=1}^{13} N(y_{2i}|\mu_2, \sigma_2^2)$$

2. Prior:

Uniform prior on $(\mu_i, log(\sigma_i))$, with $i = 1, 2$

$$p(\mu_i, \sigma_i^2) \propto (\sigma_i^2)^{-1}$$

Then,

$$p(\mu_1, \mu_2, log\sigma_1, log\sigma_2) \propto 1$$

3. Posterior:

$$\begin{aligned} p(\mu_1, \mu_2, log\sigma_1, log\sigma_2|y) &= p(\mu_1, \mu_2, log\sigma_1, log\sigma_2)p(y|\mu_1, \mu_2, log\sigma_1, log\sigma_2) \\ &= \prod_{i=1}^{9} N(y_{1i}|\mu_1, \sigma_1^2) \prod_{i=1}^{13} N(y_{2i}|\mu_2, \sigma_2^2) \end{aligned}$$

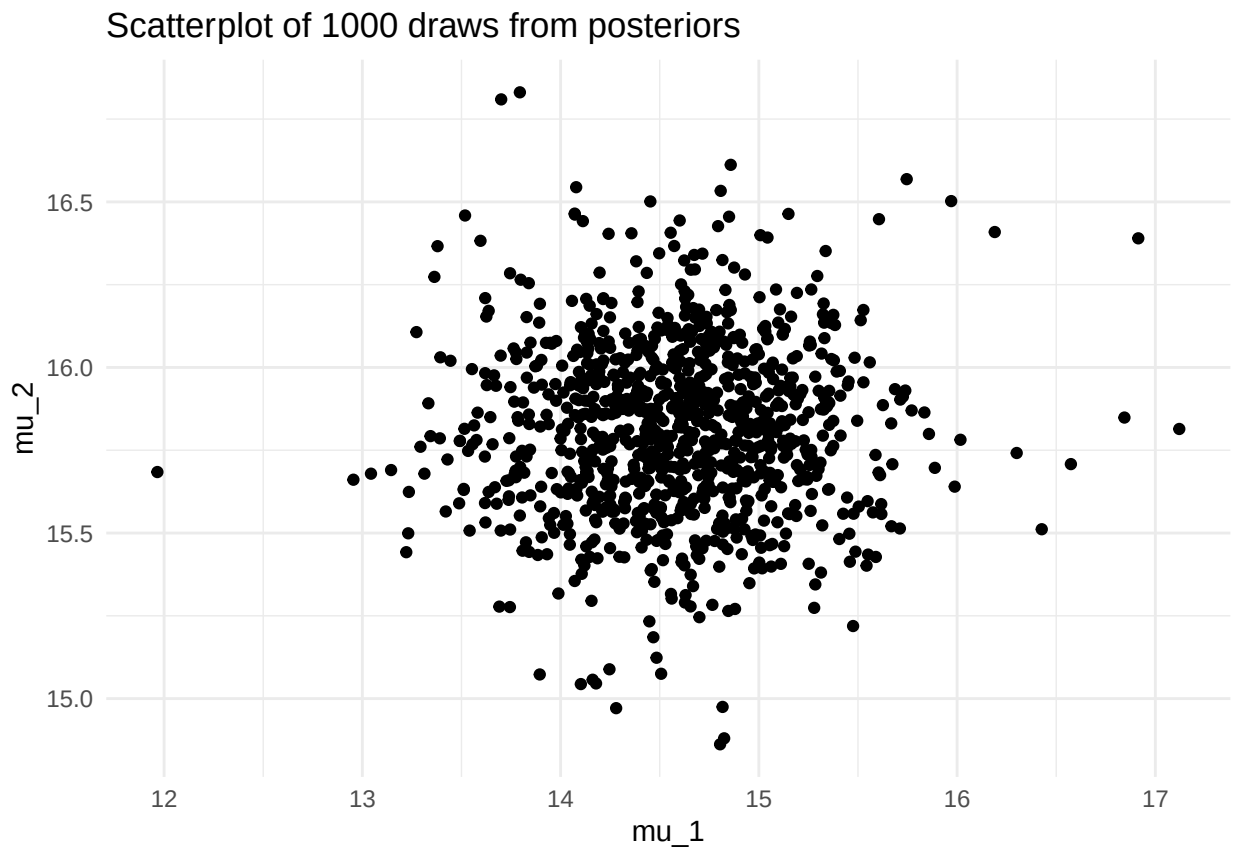
The posterior density factors, (μ_1, σ_1) are independent of (μ_2, σ_2) in the posterior distribution.

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.

The marginal posterior distributions for μ_1 and μ_2 are:

$$\mu_1|y \sim t_8(14.6112222, 0.2414614)$$
$$\mu_2|y \sim t_12(15.8210769, 0.0585729)$$

```
set.seed(4711)
mu_1 <- rtnew(1000, 8,mean(windshieldy1), sqrt(var(windshieldy1)/length(windshieldy1)))
mu_2 <- rtnew(1000,12,mean(windshieldy2), sqrt(var(windshieldy2)/length(windshieldy2)))
mu_d <- mu_1 - mu_2

data_d <- data.frame(rtnew(1000,12,mean(windshieldy2),
                           sqrt(var(windshieldy2)/length(windshieldy2))),
                     rtnew(1000,12,mean(windshieldy2),
                           sqrt(var(windshieldy2)/length(windshieldy2))))

ggplot(data_d, aes(x = mu_1, y = mu_2)) +
  geom_point() +
  labs(x = "mu_1", y = "mu_2") +
  ggtitle("Scatterplot of 1000 draws from posteriors") +
  theme_minimal()
```



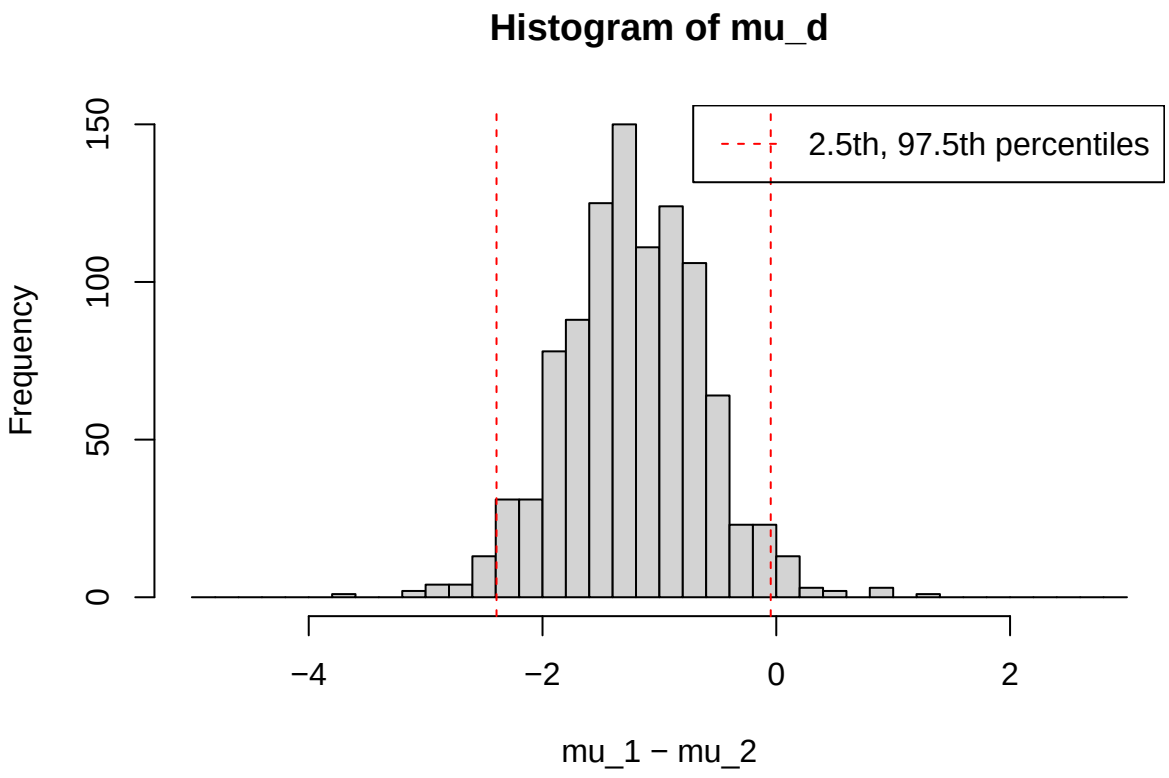

```
#point estimate
mean(mu_d)
```

```
## [1] -1.223479
```

```
lower_quantile_d <- quantile(mu_d, (1-.95)/2)
upper_quantile_d <- quantile(mu_d, 1-(1-.95)/2)

#posterior interval
cat(lower_quantile_d, upper_quantile_d)
```

```
## -2.393461 -0.04749882
```

```
hist (mu_d, xlab="mu_1 - mu_2",
breaks=seq(-5,3,0.20))

abline(v = lower_quantile_d,col ='red', lty = 2)
abline(v = upper_quantile_d,col ='red', lty = 2)

legend("topright", legend = c("2.5th, 97.5th percentiles"),
       col = c("red"), lwd = c(1), lty = c( 2))
```



Point estimate:

$$E\left(\mu_1 - \mu_2 | y\right) = -1.223$$

Posterior interval:

There is a 95% probability that μ_d, the mean difference, falls in $[-2.393, -0.047]$ given the evidence provided by the observed data.

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

In a continuous distribution, the probability that two values are exactly the same, $\mu_1 = \mu_2$, is essentially **zero**. This is because in a continuous distribution, the probability of any specific value is zero.

Instead, we can check the relative probability of different ranges close to zero within the posterior distribution. The smaller we set the tolerance the smaller the probability:

```
tolerance <- 0.20

probab_equal <- sum(abs(mu_d) < tolerance) / length(mu_d)

cat("tolerance 0.20: ", probab_equal)
```

```
## tolerance 0.20:  0.036
```

```
tolerance <- 0.10
```

```r
probab_equal <- sum(abs(mu_d) < tolerance) / length(mu_d)

cat("tolerance 0.10: ", probab_equal)
```

```
## tolerance 0.10:  0.021
```

```r
tolerance <- 0.01

probab_equal <- sum(abs(mu_d) < tolerance) / length(mu_d)

cat("tolerance 0.01: ", probab_equal)
```

```
## tolerance 0.01:  0.001
```

```r
tolerance <- 0.001

probab_equal <- sum(abs(mu_d) < tolerance) / length(mu_d)

cat("tolerance 0.001: ", probab_equal)
```

```
## tolerance 0.001:  0
```

Probability of $\mu_d = 0$ is 0.

Appendix: Code for plots in 2.b)

```
set.seed(4711)
n_0 <- 674
y_0 <- 39
n_1 <- 680
y_1 <- 22

posterior_d_0 <- function(a, b) {
  alpha <- a + y_0
  beta <- b + n_0 - y_0
  rbeta(100000, alpha, beta)
}

posterior_d_1 <- function(a, b) {
  alpha <- a + y_1
  beta <- b + n_1 - y_1
  rbeta(100000, alpha, beta)
}

hyperparameters <- list(c(5, 5), c(25, 25), c(50, 50), c(100, 100))
results <- list()
mean_odds_ratios <- numeric(length(hyperparameters))
posterior_odds_ratios_l <- numeric(length(hyperparameters))
posterior_odds_ratios_u <- numeric(length(hyperparameters))

for (i in 1:length(hyperparameters)) {
  posterior_samples_p0 <- posterior_d_0(hyperparameters[[i]][1], hyperparameters[[i]][2])
  posterior_samples_p1 <- posterior_d_1(hyperparameters[[i]][1], hyperparameters[[i]][2])

  odds_ratio_samples <- (posterior_samples_p1 / (1 - posterior_samples_p1)) /
    (posterior_samples_p0 / (1 - posterior_samples_p0))

  results[[paste("Prior =", paste(hyperparameters[[i]], collapse = ","))]] <- odds_ratio_samples

  mean_odds_ratio <- posterior_odds_ratio_point_est(p0 = posterior_samples_p0,
                                                    p1 = posterior_samples_p1)
  posterior_odds_ratio <- posterior_odds_ratio_interval(p0 = posterior_samples_p0,
                                                        p1 = posterior_samples_p1,
                                                        prob = 0.95)

  mean_odds_ratios[i] <- mean_odds_ratio
  posterior_odds_ratios_l[i] <- posterior_odds_ratio[1]
  posterior_odds_ratios_u[i] <- posterior_odds_ratio[2]
}


par(mfrow = c(2, 2))


for (i in 1:length(hyperparameters)) {
  odds_ratio_samples <- as.numeric(results[[paste("Prior =",
                                                  paste(unlist(hyperparameters[i]),
```

```
                                                    collapse = ","))]])
    hist(odds_ratio_samples,
         breaks = 50,
         main = paste("Posterior with Prior Beta",
                      paste(unlist(hyperparameters[i]), collapse = ",")),
         xlab = "Odds Ratio", ylab = "Frequency")
}
```