Assignment 3

anonymous

1.General information

I used AI to search some related information (Generate a frame to draw the figure in 3.b) in this assignment.

2.(a)

1.Likelihood:

Given Observations y1,y2,y3….yn which following the mean μ and the standard deviation is σ Normal distribution. The likelihood function for single observation value yi is that: $f(y_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$.

We have a number of observed hardness values yi, so we can formulate the joint likelihood, because for n independent observations, joint likelihood is the product of individual likelihood which should be: $L(\mu,\sigma|\mathbf{y}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right)$

2.The prior distribution for μ and σ given is:

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

3.Resulting Posterior:
We can using Bayes' theorem, the posterior distribution is proportional to the product of likelihood and prior:
$p(\mu,\sigma|\mathbf{y}) \propto L(\mu,\sigma|\mathbf{y}) \times p(\mu,\sigma)$ Then we can formulate the Resulting Posterior:
$p(\mu,\sigma|\mathbf{y}) \propto \sigma^{-n-1}\exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right)$

2.(b)

Keep the below name and format for the functions to work with `markmyassignment`:

```
# Useful functions: mean(), length(), sqrt(), sum()
# and qtnew(), dtnew() (from aaltobda)

mu_point_est <- function(data) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
  point_estimate <- mean(data)
    #14.5
```

```r
    return(point_estimate)
}
print(mu_point_est(windshieldy1))
```

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

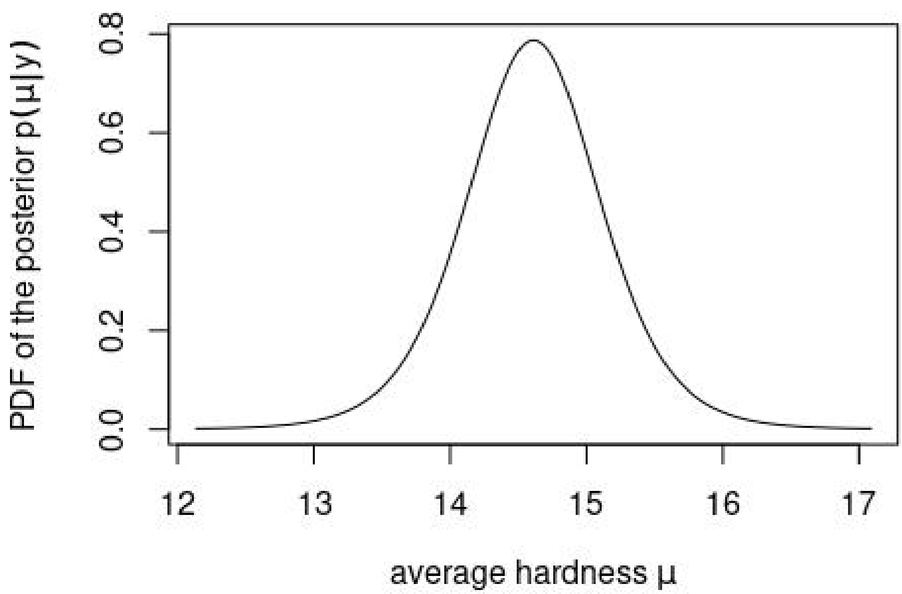
```r
mu_interval <- function(data, prob = 0.95) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    n<-length(data)
    s <- sd(data)
    point_estimate <- mean(data)
    lower_bound <- point_estimate+qt((1-prob)/2,df=n-1)*(s/sqrt(n))
    upper_bound <- point_estimate+qt(1-(1-prob)/2, df=n-1)*(s/sqrt(n))
    credible <- c(lower_bound, upper_bound)
    return(credible)
}
print(mu_interval(windshieldy1,prob=0.95))
```

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

The point estimate is 14.61122.This means that based on our data, the best estimate or average hardness μ The central value of is 14.61122. 95% confidence interval:The results are: 2.5%=13.478, 97.5%=15.744 This means that based on the data we observe, μ There is a 95% probability that the true value of falls between 13.478 and 15.744. In other words, we are interested in μ The uncertainty range of is within this range.

```r
mu_pdf <- function(data, x){
    # Compute necessary parameters here.
    n <- length(data)
    sample_variance <- var(data)
    sample_mean <- mean(data)
    # These are the correct parameters for `windshieldy_test`
    # with the provided uninformative prior.
    df = n-1
    location = sample_mean
    scale = sqrt(sample_variance / n)
    # Use the computed parameters as below to compute the PDF:

    dtnew(x, df, location, scale)
}

x_interval = mu_interval(windshieldy1, .999)
lower_x = x_interval[1]
upper_x = x_interval[2]
x = seq(lower_x, upper_x, length.out=1000)
plot(
    x, mu_pdf(windshieldy1, x), type="l",
    xlab=TeX(r'(average hardness $\mu$)'),
```

```
    ylab=TeX(r'(PDF of the posterior $p(\mu|y)$)')
)
```



From the graph, we can see that the peak of the posterior distribution is at the average hardness μ which is bout 14.6. The shape and width of the distribution reflect our understanding of μ uncertainty: The wider the distribution, the greater the uncertainty, the narrower the distribution, the smaller the uncertainty.

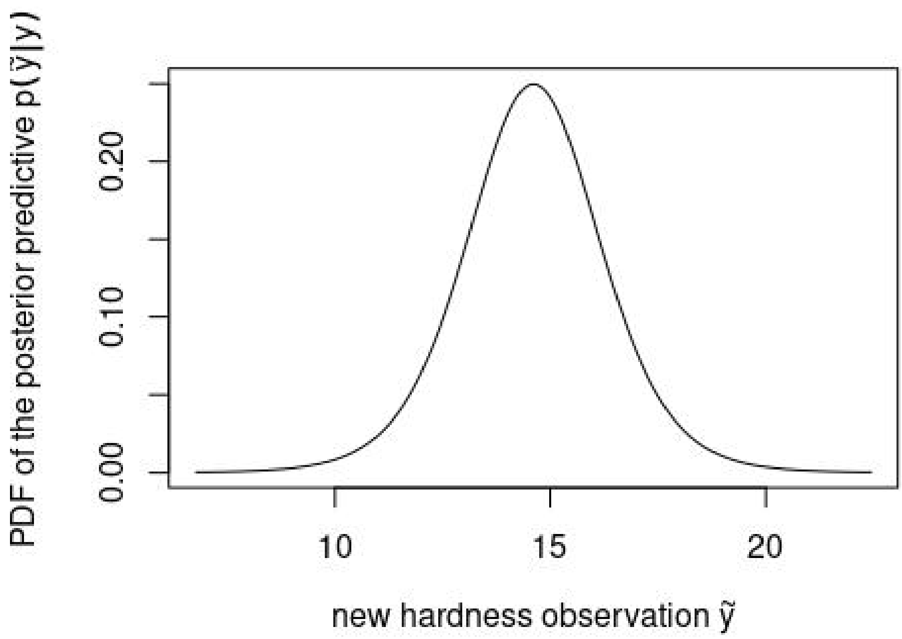
2.(c)

```
# Useful functions: mean(), length(), sqrt(), sum()
# and qtnew(), dtnew() (from aaltobda)

mu_pred_point_est <- function(data) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    #14.5
    pointEstimate <- mean(data)
    return(pointEstimate)
}
mu_pred_point_est(windshieldy1)

## [1] 14.61122

mu_pred_interval <- function(data, prob = 0.95) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    #c(11.8, 17.2)
```

```r
    n <- length(data)
    s2 <- var(data)
    y_bar <- mean(data)
    predictive_variance <- s2 + s2/n
    lower_bound <- y_bar + qt((1-prob)/2,df=n-1)*sqrt(predictive_varian
ce)
    upper_bound <- y_bar + qt(prob+(1-prob)/2, df=n-1)*sqrt(predictive_
variance)
    interval<-c(lower_bound, upper_bound)
    return(interval)
}
mu_pred_interval(windshieldy1)

## [1] 11.02792 18.19453
```

The point estimate provides a single most probable value for the hardness of the next windshield which is about 14.611. 95% confidence interval:The results are: 2.5%=11.02792, 97.5%=18.19453.

```r
mu_pred_pdf <- function(data, x){
    # Compute necessary parameters here.
    # These are the correct parameters for `windshieldy_test`
    # with the provided uninformative prior.
    s2 <- var(data)
    n <- length(data)
    sample_mean <- mean(data)
    predictive_variance <- s2 + s2/n
    df = n
    location = sample_mean
    scale = sqrt(predictive_variance)
    # Use the computed parameters as below to compute the PDF:

    dtnew(x, df, location, scale)
}

x_interval = mu_pred_interval(windshieldy1, .999)
lower_x = x_interval[1]
upper_x = x_interval[2]
x = seq(lower_x, upper_x, length.out=1000)
plot(
    x, mu_pred_pdf(windshieldy1, x), type="l",
    xlab=TeX(r'(new hardness observation $\tilde{y}$)'),
    ylab=TeX(r'(PDF of the posterior predictive $p(\tilde{y}|y)$)')
)
```



The peak value of the curve is approximately 14.6 on the x-axis and 0.24 on the y-axis. This indicates that the most likely new hardness observation is approximately 15, corresponding to a probability density of approximately 0.24. Overall, this chart shows the distribution of new hardness observations and their probabilities. The most likely new hardness observation is about 14.6, with a probability density of about 0.24.

Inference for the difference between proportions (3 points)

3.(a)

1.Likelihood:

For the control group:

$$L(p_0) = \binom{674}{39} p_0^{39}(1-p_0)^{635}$$

For the treatment group:

$L(p_1) = \binom{680}{22} p_1^{22}(1-p_1)^{658}$ Where $\binom{n}{k}$ represents the binomial coefficient,which means n choose k.

2.The prior: for a noninformative prior using the beta distribution: $p_0 \sim \text{Beta}(1,1)$

$$p_1 \sim \text{Beta}(1,1)$$

The probability density function for a Beta distribution is given by: $f(p; \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$

Where $B(\alpha, \beta)$ is the Beta function, and it acts as a normalization constant. For our noninformative priors:

$$f(p_0) = p_0^0(1 - p_0)^0 = 1$$

$$f(p_1) = p_1^0(1 - p_1)^0 = 1$$

3.The resulting posterior: We can use the Bayes' theorem and the fact that the beta distribution is a conjugate prior for the binomial likelihood, the posterior distributions are also beta distributions. For p_0: $p(p_0|data) \propto L(p_0) \times f(p_0)$

$$\text{Posterior for} p_0 \sim \text{Beta}(1 + 39, 1 + 635) = \text{Beta}(40, 636)$$

For p_1: $p(p_1|data) \propto L(p_1) \times f(p_1)$

$$\text{Posterior for} p_1 \sim \text{Beta}(1 + 22, 1 + 658) = \text{Beta}(23, 659)$$

3.(b)

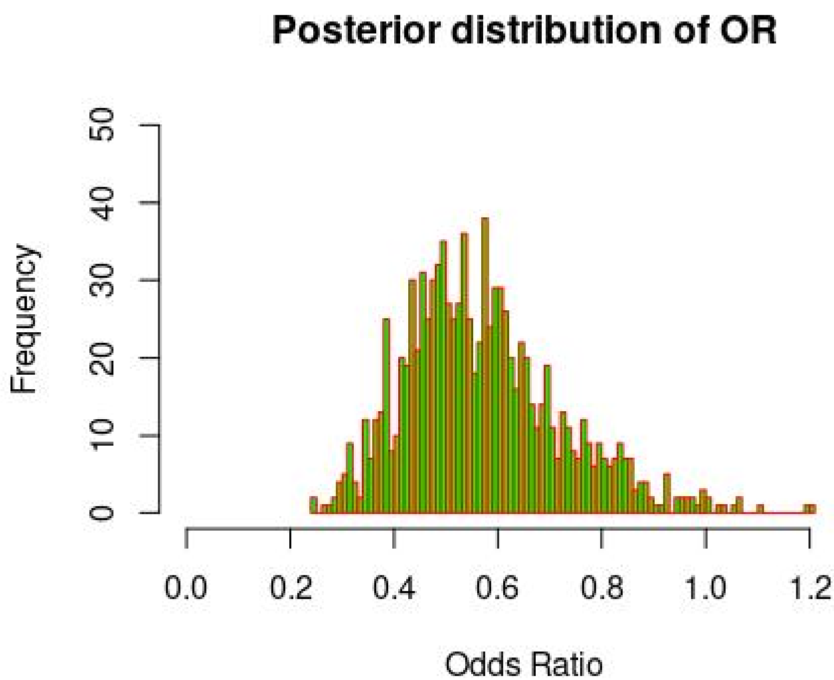
This means that the probability of events occurring in the group we are interested in is 57.1% higher than that in the control group. In other words, the likelihood of events occurring in the group we are concerned about is relatively low. Meanwhile, the 95% confidence interval tells us that there is a 95% chance that the true value of OR will fall between 0.3221829 and 0.9220926.

```
set.seed(4711)
ndraws = 1000
#p0 = rbeta(ndraws, 5, 95)
#p1 = rbeta(ndraws, 10, 90)
p0 = rbeta(ndraws, 40, 636)
p1 = rbeta(ndraws, 23, 659)
```

Keep the below name and format for the functions to work with `markmyassignment`:

```
# Useful function: mean(), quantile()

posterior_odds_ratio_point_est <- function(p0, p1) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    #2.650172
    OR_samples <- (p1 / (1 - p1)) / (p0/ (1 - p0))
    E_OR <- mean(OR_samples)
    return(E_OR)
}
posterior_odds_ratio_point_est(p0,p1)

## [1] 0.5710218
```

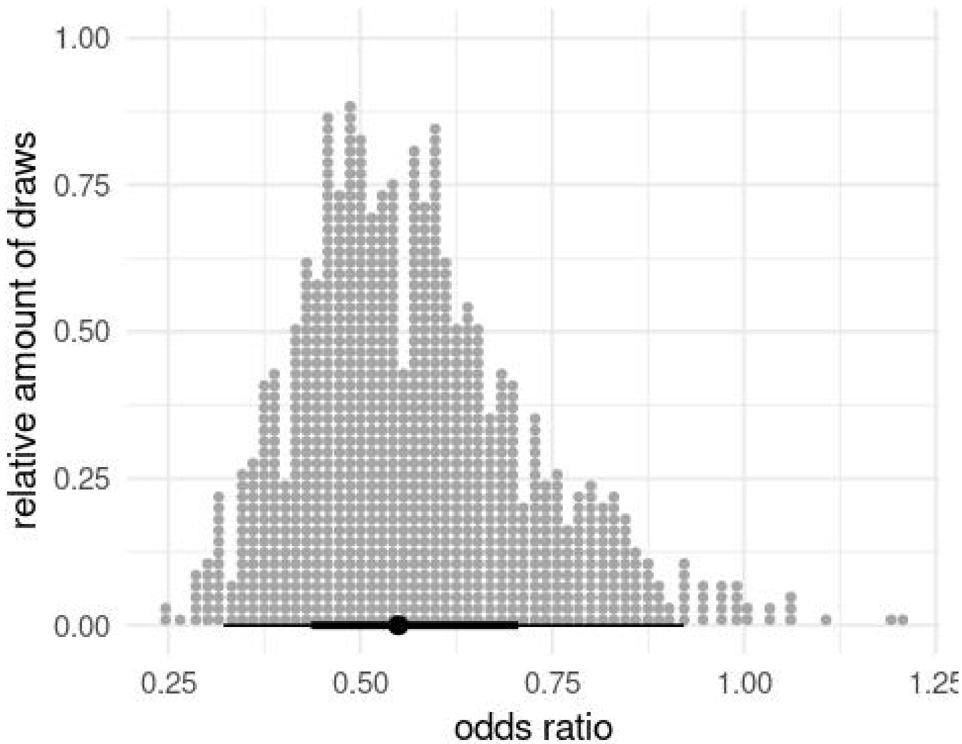
```r
posterior_odds_ratio_interval <- function(p0, p1, prob = 0.95) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    #c(0.6796942,7.3015964)
    OR_samples <- (p1 / (1 - p1)) / (p0/ (1 - p0))
    lower_bound <- quantile(OR_samples, (1-prob)/2)
    upper_bound <- quantile(OR_samples, prob+(1-prob)/2)
    interval <- c(lower_bound,upper_bound)
    hist(OR_samples, main="Posterior distribution of OR", xlab="Odds Ra
tio", border="red", col="green", breaks=100,xlim=c(0,1.3),ylim=c(0,50))
    return(interval)
}
interval<-posterior_odds_ratio_interval(p0,p1,prob = 0.95)
```




```r
cat("Point estimate of OR:", posterior_odds_ratio_point_est(p0,p1), "\n
")

## Point estimate of OR: 0.5710218

cat("95% credible interval for OR:", interval[1], "-", interval[2], "\n
")

## 95% credible interval for OR: 0.3221829 - 0.9220926
```

From the graph, it appears that the most likely value for the odds ratio is around 0.6. This is evident from the highest frequency of this value, as represented by the tallest green bar in the histogram. An odds ratio of 0.6 suggests that the odds of the event occurring in the group of interest are 60% of the odds of the event occurring in the comparison group. In other words, the event is less likely to occur in the group of interest compared to the comparison group.

```
rodds_ratio = (r1/(1-r1))/(r0/(1-r0))
ggplot(data.frame(
    rv=c(rodds_ratio)
)) +
    aes(xdist=rv) +
    labs(x="odds ratio", y="relative amount of draws") +
    stat_dotsinterval()
```



3.(c)

I use two different priors. The first prior assume we have prior knowledge that suggests that the treatment generally has a positive effect. $p0 \sim Beta(20,80), p1 \sim Beta(40,60)$.The next prior assume we want to be skeptical about the treatment's efficacy, we might use:$p0 \sim Beta(40,60), p1 \sim Beta(20,80)$,this prior assumes the treatment is more likely to be ineffective.

```
library(rstan)

## Loading required package: StanHeaders

## rstan (Version 2.21.8, GitRev: 2e1f913d3ca3)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend
 calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend call
ing
## rstan_options(auto_write = TRUE)

##
## Attaching package: 'rstan'

## The following objects are masked from 'package:posterior':
##
##     ess_bulk, ess_tail
```

```
N <- 1000  # Number of samples

sample_OR <- function(p0_samples, p1_samples) {
  OR_samples <- (p1_samples / (1 - p1_samples)) / (p0_samples / (1 - p0
_samples))
  list(
    E_OR = mean(OR_samples),
    lower_bound = quantile(OR_samples, 0.025),
    upper_bound = quantile(OR_samples, 0.975)
  )
}

# 1. Informative Prior
p0_samples_info <- rbeta(N, 20, 80)
p1_samples_info <- rbeta(N, 40, 60)
results_info <- sample_OR(p0_samples_info, p1_samples_info)

# 2. Skeptical Prior
p0_samples_skept <- rbeta(N, 40, 60)
p1_samples_skept <- rbeta(N, 20, 80)
results_skept <- sample_OR(p0_samples_skept, p1_samples_skept)

# Print Results
cat("Informative Prior:\n")
```

```
## Informative Prior:
```

```
cat("Point estimate of OR:", results_info$E_OR, "\n")
```

```
## Point estimate of OR: 2.856836
```

```
cat("95% credible interval for OR:", results_info$lower_bound, "-", res
ults_info$upper_bound, "\n\n")
```

```
## 95% credible interval for OR: 1.421003 - 5.095268
```

```
cat("Skeptical Prior:\n")
```

```
## Skeptical Prior:

cat("Point estimate of OR:", results_skept$E_OR, "\n")

## Point estimate of OR: 0.3840276

cat("95% credible interval for OR:", results_skept$lower_bound, "-", re
sults_skept$upper_bound, "\n")

## 95% credible interval for OR: 0.1889996 - 0.6672246
```

Informative Prior: The point estimate for the odds ratio (OR) is 2.85. This implies that the odds of death in the treatment group are approximately 2.84 times higher than in the control group. The 95% credible interval is [1.42, 5.09]. This relatively narrow interval suggests a higher level of confidence in this estimate. In the context of this informative prior, the treatment may have a certain positive effect. Skeptical Prior: The point estimate for the odds ratio (OR) is 0.384. This means the odds of death in the treatment group are about 61% lower than in the control group (1 - 0.384 = 0.616). The 95% credible interval is [0.19, 0.67]. This interval is narrow and lies entirely below 1, indicating that in the context of this skeptical prior, the treatment effect may indeed be positive. Prior selection has significant sensitivity to our inference. The prior information leads us to believe that treatment is effective, with a relatively positive probability ratio and confidence interval. On the contrary, the skeptical prior leads us to a relatively conservative conclusion that the treatment effect may not be significant. This difference emphasizes the importance of carefully considering and selecting prior knowledge before forming a conclusion.

Inference for the difference between normal means (3 points)

Loading the library and the data.

```
data("windshieldy2")
# The new data are now stored in the variable `windshieldy2`.
# The below displays the first few rows of the new data:
head(windshieldy2)

## [1] 15.980 14.206 16.011 17.250 15.993 15.722
```

4.(a)

Likelihood: The hardness measurements for both samples are assumed to be normally distributed. For the first production line:

$$\mathbf{y}_1|\mu_1,\sigma_1 \sim \mathrm{Normal}(\mu_1,\sigma_1^2)$$

For the second production line:

$$\mathbf{y}_2|\mu_2,\sigma_2 \sim \mathrm{Normal}(\mu_2,\sigma_2^2)$$

2.Prior: For the means of the hardness measurements from both lines:

$$\mu_1, \mu_2 \sim \text{Normal}(mean, variance)$$

For the standard deviations of the hardness measurements:

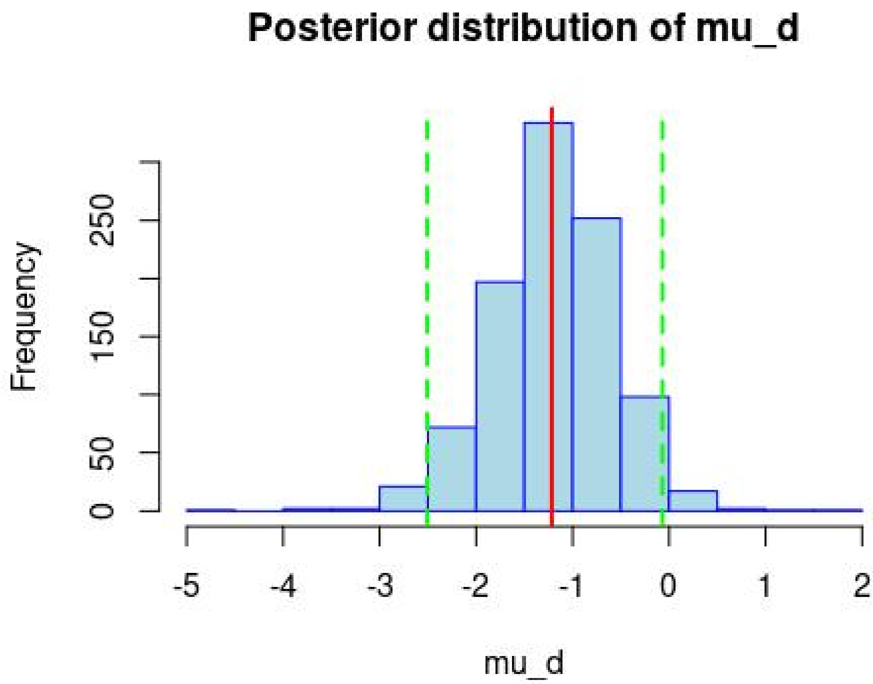
$$\sigma_1, \sigma_2 \sim \text{Uniform}(lower, upper)$$

3.Posterior: Using the Bayes' theorem, the posterior is proportional to the likelihood times the prior:

$$p(\mu_1, \mu_2, \sigma_1, \sigma_2 | \mathbf{y}_1, \mathbf{y}_2) \propto p(\mathbf{y}_1 | \mu_1, \sigma_1) \times p(\mathbf{y}_2 | \mu_2, \sigma_2) \times p(\mu_1) \times p(\mu_2) \times p(\sigma_1) \times p(\sigma_2)$$

4.(b)

```
# Useful functions: mean(), length(), sqrt(), sum(),
# rtnew() (from aaltobda), quantile() and hist().
#windshieldy1
#windshieldy2
E <- function(data1,data2) {
     n1 <- length(data1)
     mean1 <- mean(data1)
     var1 <- var(data1)
     n2 <- length(data2)
     mean2 <- mean(data2)
     var2 <- var(data2)
     # Use rtnew to sample from posterior distributions of mu1 and mu2
     post_samples_mu1 <- rtnew(ndraws, n1 - 1, mean1, sqrt(var1 / n1))
     post_samples_mu2 <- rtnew(ndraws, n2 - 1, mean2, sqrt(var2 / n2))
     #print(post_samples_mu1)
     # Compute the difference
     #print(post_samples_mu2)
     #print(post_samples_mu1)
     post_samples_mud <- post_samples_mu1 - post_samples_mu2

     # Point estimate and credible interval
     point_estimate_mud <- mean(post_samples_mud)
     print(point_estimate_mud)
     cred_interval_mud <- quantile(post_samples_mud, c(0.025, 0.975))

     # Plotting
     hist(post_samples_mud, main="Posterior distribution of mu_d", xlab
="mu_d", border="blue", col="lightblue")
     abline(v = point_estimate_mud, col="red", lwd=2)
     abline(v = cred_interval_mud, col="green", lwd=2, lty=2)
     cat("Point estimate for mu_d:", point_estimate_mud, "\n")
     cat("95% credible interval for mu_d:", cred_interval_mud, "\n")
}

E(windshieldy1,windshieldy2)

## [1] -1.212986
```




```
## Point estimate for mu_d: -1.212986
## 95% credible interval for mu_d: -2.506564 -0.06975627
```

The average hardness of the windshield produced by production line 1 is estimated to be 1.212986 units lower than that of production line 2. 95% confidence indicates that the difference is probably between -2.506564 and -0.0697563. This provides strong evidence that there is a significant difference in the hardness of windshields manufactured by the two production lines, with the hardness of production line 1 being significantly lower.

4.(c)

In the context of continuous distribution, in this case, the normal distribution of mu1 and mu2 used for modeling mean, the probability of any specific value, including the probability of their complete equality, is technically 0.