BDA - Assignment 3

Anonymous

Contents

Exercise 1) Inference for normal mean and deviation

a)

The data to be used in this assignment is `windshieldy1`. This data includes the values for the observed hardness values y_1 of a sample of windshields.

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

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

Here, we have some test samples to check the correctness of our results against some given tests.

```
windshieldy_test <- c(13.357, 14.928, 14.896, 14.820)
```

The sufficient statistics for this data are

```
n <- length(windshieldy1)
y <- sum(windshieldy1)
s2 <- var(windshieldy1)
my <- mean(windshieldy1)
```

We assume the data to be normally distributed $\mathrm{N}(\mu, \sigma^2)$ with an unknown standard deviation σ and an unknown average hardness μ. Assuming prior independence of location and scale parameters, we know that the prior distribution is

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

The likelihood can be infered to be

$$p(y \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \quad .$$

Under the aforementioned prior, the joint posterior distribution is proportional to the likelihood function, resulting in

$$p(\mu, \sigma^2 \mid y) = \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) \quad .$$

Since we are interested in estimating σ, we can derive the analytical form for the marginal posterior distribution of μ by integrating the joint posterior density over σ^2

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

The result of this is an unnormalized gamma integral, which gives that the posterior follows the standard t density $p(\mu \mid y) \sim t_{n-1}(\bar{y}, s^2/n)$ with $n-1$ degrees of freedom.

Based on the formulation of the prior, the model likelihood and the resulting posterior, we can obtain our estimates for μ

```
mu_point_est <- function(data) {
  n <- length(data)
  df <- n - 1

  mean <- mean(data)
  s2 <- var(data)
  scale <- s2 / n

  point_est <- mean(rtnew(100000, df, mean = mean, scale = scale))
  return(point_est)
}
mu_point_est(data = windshieldy1)
```

```
## [1] 14.61135
```

Thus, $E(\mu \mid y) = 14.61$ (result may change slightly due to the randomness in the simulation).

The 95% posterior interval can be found as follows.

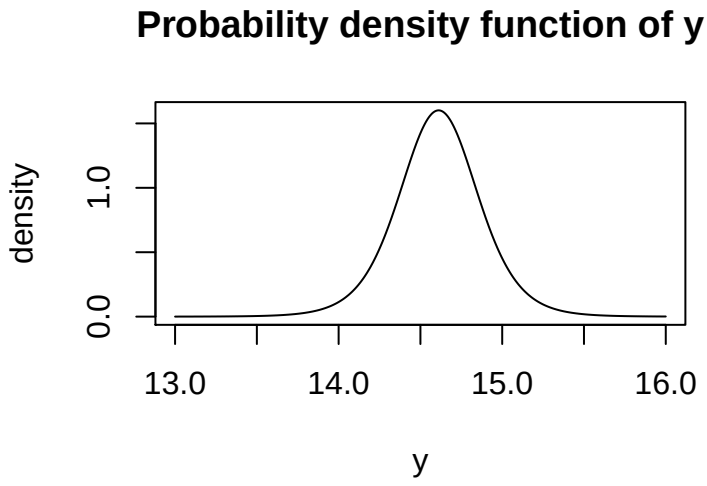
```
mu_interval <- function(data, prob) {
  n <- length(data)
  df <- n - 1

  mean <- mean(data)
  s2 <- var(data)
  scale <- s2 / n

  # lower and upper intervals
  lower <- (1-prob)/2.0
  upper <- 1-lower

  sample<-rt(100000, df)*sqrt(scale)+mean

  return (quantile(sample, c(lower, upper)))
  }
  mu_interval(data = windshieldy1, prob = 0.95)
```

```
##     2.5%    97.5%
## 13.48979 15.72877
```

The 95% posterior interval is $[14.0531, 15.17]$.

The density function is seen in the next plot.

```
plot_density <- function(data){
  n <- length(data)
  df <- n - 1

  mean <- mean(data)
  s2 <- var(data)
  scale <- s2 / n

  y <- seq(from = 13, to = 16, length.out = 1000)
  density <- dtnew(y, df, mean = mean, scale = scale)
  plot(y, density, type = "l")
  title("Probability density function of y")
}
plot_density(windshieldy1)
```



b)

Next, we focus on the estimate for the hardness of the next windshield, that is, the posterior predictive distribution for a future observation.

Based on derivations from BDA3, we know that the posterior predictive distribution of $\bar{y}$ is a t distribution with location $\bar{y}$, scale $\left(1 + \frac{1}{n}\right)^{1/2} s$, and $n - 1$ degrees of freedom.

Based on this information, we can make the same computations as in Section 1.a with the new parameters.

```
mu_pred_point_est <- function(data) {
  n <- length(data)
  df <- n - 1

  mean <- mean(data)
  s2 <- var(data)
  scale <- sqrt(1+(1/n))*sqrt(s2)
```

```
  point_est <- mean(rtnew(100000, df, mean = mean, scale = scale))
  return(point_est)
}
mu_pred_point_est(data = windshieldy1)
```

```
## [1] 14.60928
```

Thus, $E(\mu \mid y) = 14.61$.

The 95% posterior interval can be found as follows.

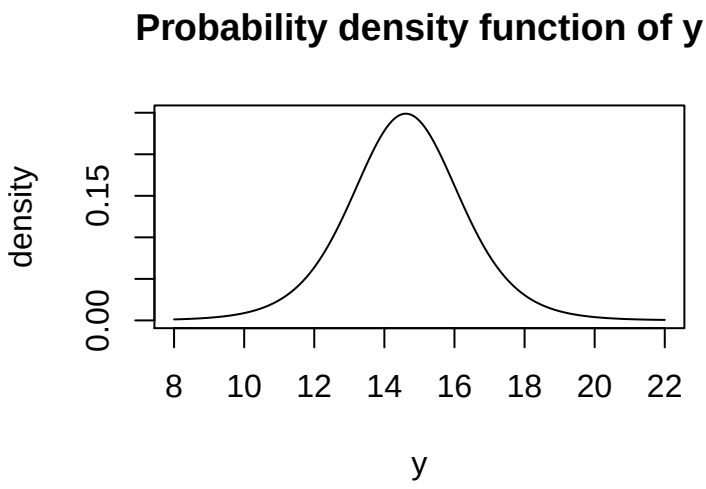
```
mu_pred_interval <- function(data, prob) {
  n <- length(data)
  df <- n - 1

  mean <- mean(data)
  s2 <- var(data)
  scale <- sqrt(1+(1/n))*sqrt(s2)

  # lower and upper intervals
  lower <- (1-prob)/2.0
  upper <- 1-lower

  sample <- rtnew(100000, df, mean = mean, scale = scale)

  return (quantile(sample, c(lower, upper)))
  }
  mu_pred_interval(data = windshieldy1, prob = 0.95)
```

```
##     2.5%    97.5%
## 10.99781 18.19373
```

The 95% predictive interval is [11.0118.23]. This interval is considerably wider than the one in Section 1.a.

The density function is seen in the next plot.

```
plot_pred_density <- function(data){
  n <- length(data)
  df <- n - 1

  mean <- mean(data)
  s2 <- var(data)
  scale <- sqrt(1+(1/n))*sqrt(s2)

  y <- seq(from = 8, to = 22, length.out = 1000)
  density <- dtnew(y, df, mean = mean, scale = scale)
  plot(y, density, type = "l")
  title("Probability density function of y")
}
plot_pred_density(windshieldy1)
```



As was suspected from the predictive interval, this density function is much wider than the one seen before. Our confidence in this prediction is lower than in the previous.

Exercise 2) Inference for the difference between proportions

a)

We have an experiment with two groups: control and treatment. The probability of death with control is p_0 and the probability of death with the treatment is p_1.

We can set up a uniform prior distribution for both $p_0 \sim \text{Beta}(1,1)$ and $p_1 \sim \text{Beta}(1,1)$.

The likelihood for both probabilities is $p(y_i|p_i)$ with $i = 0, 1$ is given by

$$p(y_i|p_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

Thus, the likelihoods are $p(y_0|p_0) = \text{Beta}(39, 635)$ and $p(y_1|p_1) = \text{Beta}(22, 658)$.

The posterior density for p_i with $i = 0, 1$ is

$$p(p_i|y) = \text{Beta}(p_i|\alpha + y, \beta + n - y).$$

based on the derivation of the posterior in p.35 of the BDA3. Therefore, the resulting posteriors are $p(p_0|y) = \text{Beta}(p_0|40, 636)$ and $p(p_1|y) = \text{Beta}(p_1|23, 659)$.

```
p0 <- rbeta(100000, 40, 636)
p1 <- rbeta(100000, 23, 659)

posterior_odds_ratio_point_est <- function (p0, p1) {
  num <- p1 / (1 - p1)
  den <- p0 / (1 - p0)
  odds_ratio <- num / den

  return (mean (odds_ratio))
}

posterior_odds_ratio_point_est(p0 = p0, p1 = p1)
```

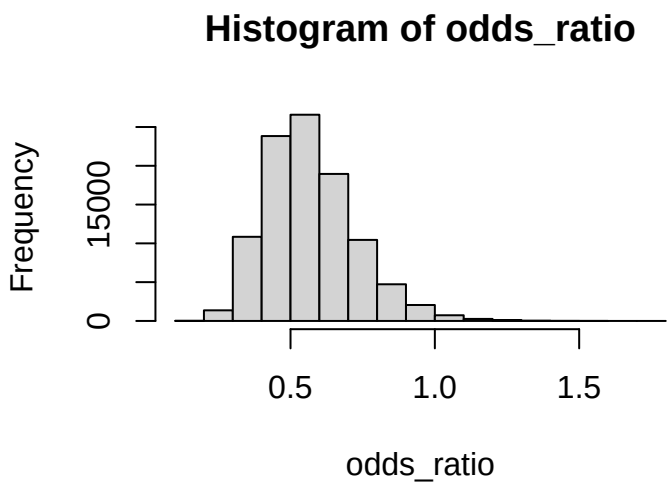
```
## [1] 0.569016
```

```
posterior_odds_ratio_interval <- function (p0, p1, prob) {
  num <- p1 / (1 - p1)
  den <- p0 / (1 - p0)
  odds_ratio <- num / den

  # lower and upper intervals
  lower <- (1-prob)/2.0
  upper <- 1-lower

  return (quantile(odds_ratio, c(lower, upper)))
}
posterior_odds_ratio_interval(p0 = p0, p1 = p1, prob = 0.95)
```

```
##      2.5%     97.5%
## 0.3205474 0.9237889
```

The histogram is seen in the next plot.

```
num <- p1 / (1 - p1)
den <- p0 / (1 - p0)
odds_ratio <- num / den
hist(odds_ratio)
```

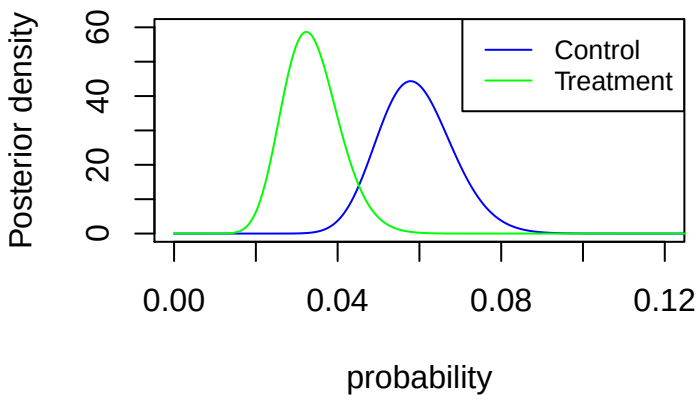


The two prior distributions were chosen so that it is equally likely that the treatment is effective as it is not. The point estimate for posterior distribution of the odds ratio is 0.57 and the 0.95 posterior interval is [0.32, 0.93]. Assuming priors p_0 and p_1, there is a 95% probability that the true treatment effect is in the interval [0.32, 0.93]. Furthermore, there is a 57% probability that the treatment results are different from those in the control group.

The posterior density functions for p_0 and p_1 can be seen below.

```
n_0 <- 674
y_0 <- 39

n_1 <- 680
y_1 <- 22
```

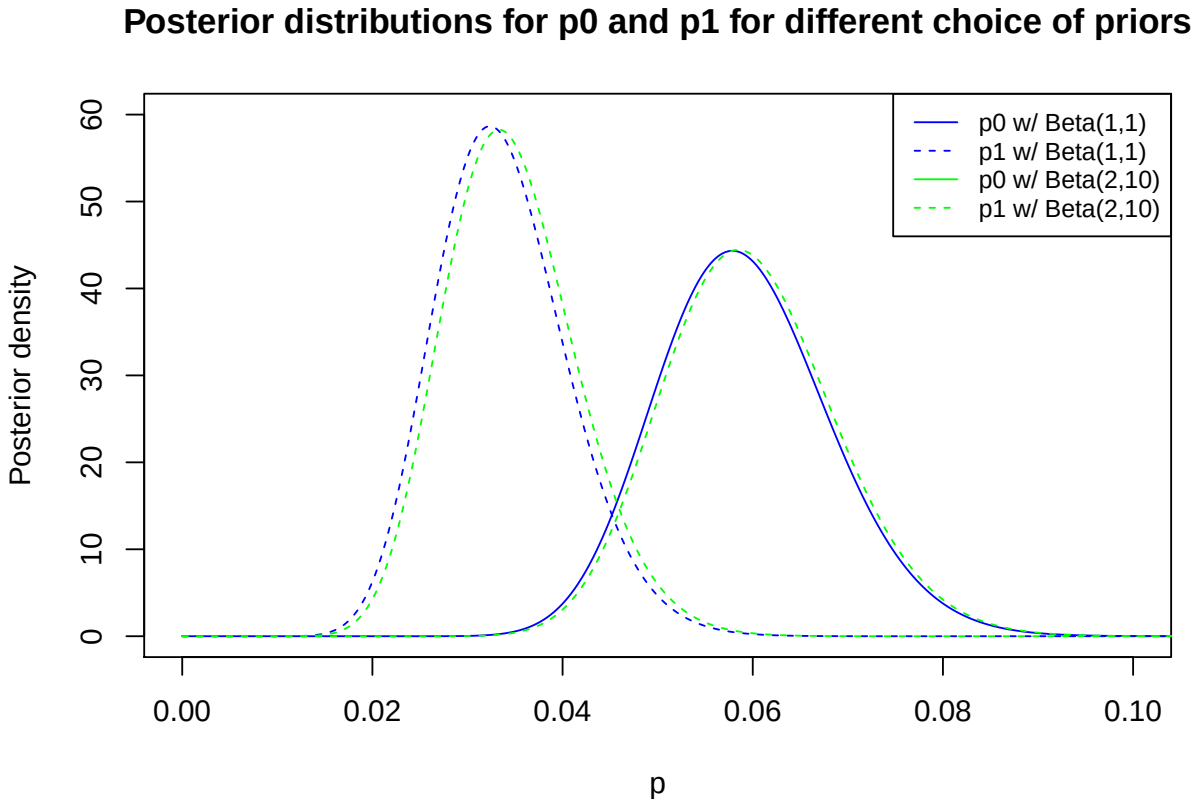
```r
prior_alpha = 1
prior_beta = 1
x <- seq(0, 0.3, length = 10000)
density_0 <- dbeta(x, prior_alpha + y_0, prior_beta + n_0 - y_0)
density_1 <- dbeta(x, prior_alpha + y_1, prior_beta + n_1 - y_1)
plot(x, density_0, type = "l", col = "blue", xlab = "probability", ylab = "Posterior density",
     ylim = c(0, 60), xlim = c(0, 0.12))


lines(x, density_1, type = "l", col = "green")
legend("topright", legend = c("Control", "Treatment"),
       col = c("blue", "green"), lty = 1, cex = 0.8)
```



b)

```r
prior_alpha = 1
prior_beta = 1
x <- seq(0, 0.3, length = 10000)
density_0 <- dbeta(x, prior_alpha + y_0, prior_beta + n_0 - y_0)
density_1 <- dbeta(x, prior_alpha + y_1, prior_beta + n_1 - y_1)
plot(x, density_0, type = "l", col = "blue", xlab = "p", ylab = "Posterior density",
     ylim = c(0, 60), xlim = c(0, 0.1))
lines(x, density_1, type = "l", lty = 2, col = "blue")


prior_alpha = 2
prior_beta = 10
density_0 <- dbeta(x, prior_alpha + y_0, prior_beta + n_0 - y_0)
density_1 <- dbeta(x, prior_alpha + y_1, prior_beta + n_1 - y_1)
lines(x, density_0, type = "l", lty = 2, col = "green")
lines(x, density_1, type = "l", lty = 2, col = "green")



legend("topright", legend = c("p0 w/ Beta(1,1)", "p1 w/ Beta(1,1)", "p0 w/ Beta(2,10)",
                              "p1 w/ Beta(2,10)"),
       col = c("blue", "blue","green", "green"), lty = c(1,2, 1, 2), cex = 0.8)
```

```
title("Posterior distributions for p0 and p1 for different choice of priors")
```



We can see the results with different priors are plotted in different colors. In this case, the posterior distributions for p_0 and p_1 do not seem to be very sensitive to the choice of prior. Thus, the uniform prior seems to be a sensible choice for this case given the amount of data we have.

Exercise 3) Inference for the difference between normal means

a)

In this case, there are two production lines with hardness measurements y_1 and y_2 as oppose to the one in exercise one. The prior, likelihood and posterior follow the same distributions as in Exercise 1.

The prior distribution is

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

with $i = 1, 2$. The likelihood can be infered to be

$$p(y_i \mid \mu_i, \sigma_i^2) = \left(\frac{1}{2\pi\sigma_i^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma_i^2}\sum_{j=1}^{n}(y_i^j - \mu_i)^2\right) \quad .$$

The posterior follows the standard t density $p(\mu_i \mid y_i) \sim t_{n_i-1}(\bar{y_i}, s_i^2/n_i)$ with $n_i - 1$ degrees of freedom.

The point estimate for $\mu_d = \mu_1 - \mu_2$ is

```
mu <- function (data) {
  n <- length(data)
  df <- n - 1

  mean <- mean(data)
  s2 <- var(data)
  scale <- s2 / n

  return (rtnew(100000, df, mean = mean, scale = sqrt(scale)))
}
mu_1 <- mu(windshieldy1)
mu_2 <- mu(windshieldy2)

mu_d_est <- mean(mu_1-mu_2)
mu_d_est
```

```
## [1] -1.206452
```

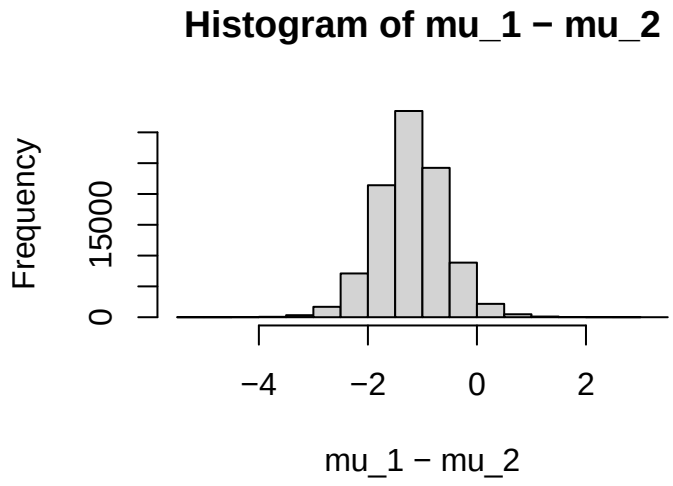
The 95% posterior interval is

```
prob <- 0.95

# lower and upper intervals
lower <- (1-prob)/2.0
upper <- 1-lower

(quantile(mu_1-mu_2, c(lower, upper)))
```

```
##        2.5%       97.5%
## -2.44641003  0.03740226
```

The histogram is plotted here

```
hist(mu_1-mu_2)
```



Given the priors that were chosen, the point estimate for μ_d is -1.21 with 95% posterior intervals $-2.44, 0.04]$.

b)

Assuming the priors p_1 and p_2, there is more than a 95% probability that the means are different as the posterior intervals show. Given that the upper value of the interval is inferior to 0, it is very likely that the values of the means differ.