BDA - Assignment 3

Anonymous

Contents

Load packages 1

Exercise 1 1

Exercise 2 4

Exercise 3 6

Load packages

```
library(aaltobda)
data("windshieldy1")
data("windshieldy2")
head(windshieldy1)
```

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

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

Exercise 1

The observations follow a normal distribution with an unknown standard deviation σ. We wish to obtain information about the unknown average hardness μ.

the model likelihood follows a normal distribution:

$$p(y|\mu, \sigma^2) \propto N(\mu, \sigma^2)$$

The prior distribution is as following:

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

```
n <- length(windshieldy1)
samples_mean <- mean(windshieldy1)
samples_var <- var(windshieldy1)
samples_mean
```

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

```
samples_var
```

```
## [1] 2.173153
```

In the given case, the prior distribution is:

$$p(\mu|y) = t_{n-1}(\bar{y}, \frac{s^2}{n}) = t_8(14.61, 0.52)$$

Finally, the posterior, as derived in the BDA3 book:

$$p(\mu, \sigma^2|y) \propto \sigma^{-n-2} exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

Where:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

So, as computed before, in the given case:

$$s^2 = 2.1731$$

$$\bar{y} = 14.6112$$

a)

```
#The following function computes the point estimate
t_point_estimate <- function(data){
  samples_mean <- mean(data)
  return(samples_mean)
}

point_estimate <- t_point_estimate(windshieldy1)
point_estimate
```

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

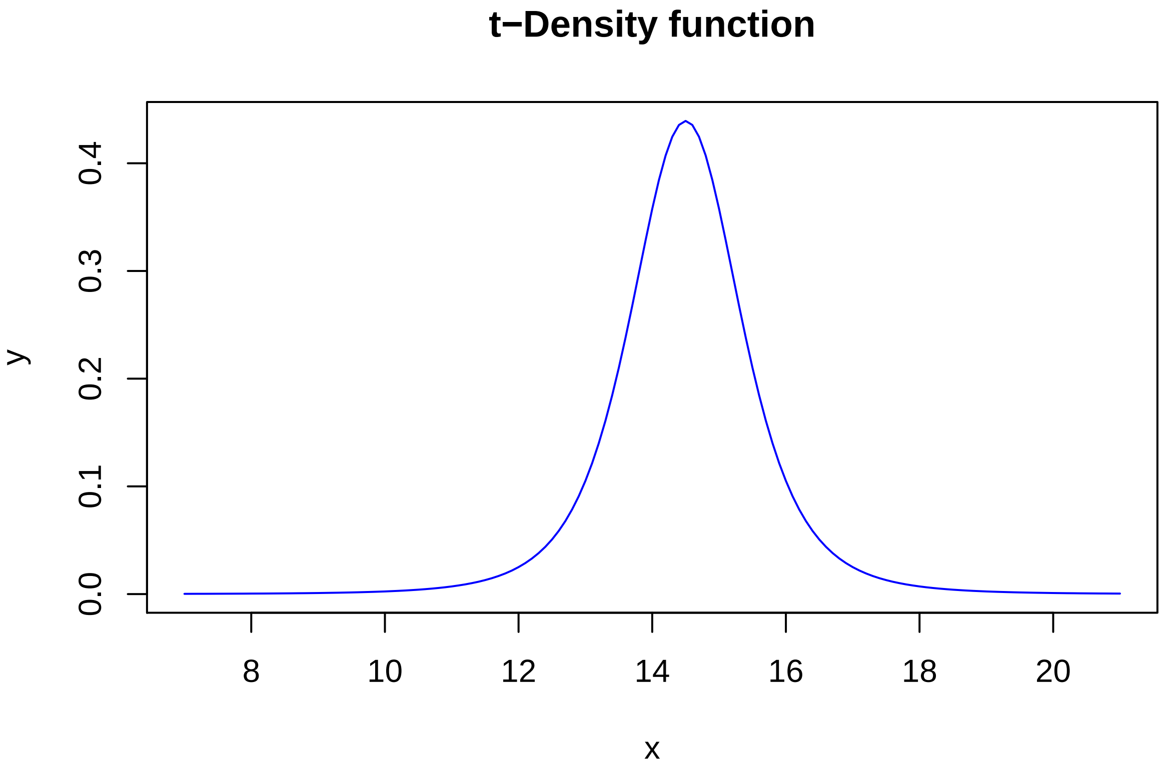
The point estimate is 14.61, meaning that this is the expected value for the mean.

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

  limit_1 <- (1-prob)/2
  limit_2 <- limit_1 + prob

  samples_mean <- mean(data)
  samples_var <- var(data)
  n <- length(data)

  estimate_1 <- qtnew(limit_1, n-1, samples_mean, scale = sqrt(samples_var)/sqrt(n))
  estimate_2 <- qtnew(limit_2, n-1, samples_mean, scale = sqrt(samples_var)/sqrt(n))
  results <- c(estimate_1, estimate_2)
  return(results)
}

estimate_95 <- mu_interval(windshieldy_test, 0.95)
estimate_95
```

```
## [1] 13.28533 15.71517
```

The 95% posterior interval is [13.29, 15.72]. This is the interval in which the mean should be contained.

```r
plot_density <- function(data){
  x <-seq(from = 7, to = 21, by = 0.1)
  samples_mean <- mean(data)
  samples_var <- var(data)
  n <- length(data)
  y <- dtnew(x, n, samples_mean, scale = sqrt(1+1/n)*sqrt(samples_var))
  plot(x, y, type="l", col="blue", main = "t-Density function")
}
plot_density(windshieldy_test)
```



b)

```r
#Predictive interval
mu_pred_interval <- function(data, prob){
  samples_mean <- mean(data)
  samples_var <- var(data)
  n <- length(data)
  limit_1 <- (1-prob)/2
  limit_2 <- limit_1 + prob
  pred_1 <- qtnew(limit_1, n-1, samples_mean, scale = sqrt(1+1/n)*sqrt(samples_var))
  pred_2 <- qtnew(limit_2, n-1, samples_mean, scale = sqrt(1+1/n)*sqrt(samples_var))
  result <- c(pred_1, pred_2)

  return(result)
}

pred_interval <- mu_pred_interval(windshieldy_test, 0.95)
pred_interval
```

```
## [1] 11.78361 17.21689
```

The hardness of the next windshield should be contained in the 95% predictive interval given by [11.78, 17.22]. The density was plotted in the previous exercise 1a.

Exercise 2

Considering as noninformative prior:

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$$

Since the observations follow a Binomial model, the likelihood would be as following:

$$p(y|\theta) \propto \binom{n}{y}\theta^y(1-\theta)^{n-y} = Binomial(n, y)$$

Finally, from the previous expressions, the one for the posterior distribution can be derived as following:

$$p(\theta|y) \propto Beta(\theta|\alpha + y, \beta + n - y)$$

For the control group: 39 died out of 674. The likelihood is as following:

$$p(y|\theta) \propto Binomial(674, 39)$$

Given the prior:

$$Beta(1, 1)$$

The posterior would be:

$$p(\theta|y) \propto Beta(1 + 39, 1 + 674 - 39) = Beta(40, 636)$$

Following the same procedure for the treatment group, the likelihood is:

$$p(y|\theta) \propto Binomial(680, 22)$$

Finally, the posterior:

$$p(\theta|y) \propto Beta(1 + 22, 1 + 680 - 22) = Beta(23, 659)$$

Given the two posterior distributions, the samples can be extracted:

```
#Test set
#set.seed(4711)
#p0 <- rbeta(10000, 5, 95)
#p1 <- rbeta(10000, 10, 90)

p0 <- rbeta(10000, 40, 636)
p1 <- rbeta(10000, 23, 659)
```

a)

```
posterior_odds_ratio_point_est <- function(p0, p1){
  p2 <- (p1/(1-p1))/(p0/(1-p0))
  est <- mean(p2)
  return(est)
}

point_estimate <- posterior_odds_ratio_point_est(p0, p1)
point_estimate
```

```
## [1] 0.5721283
```

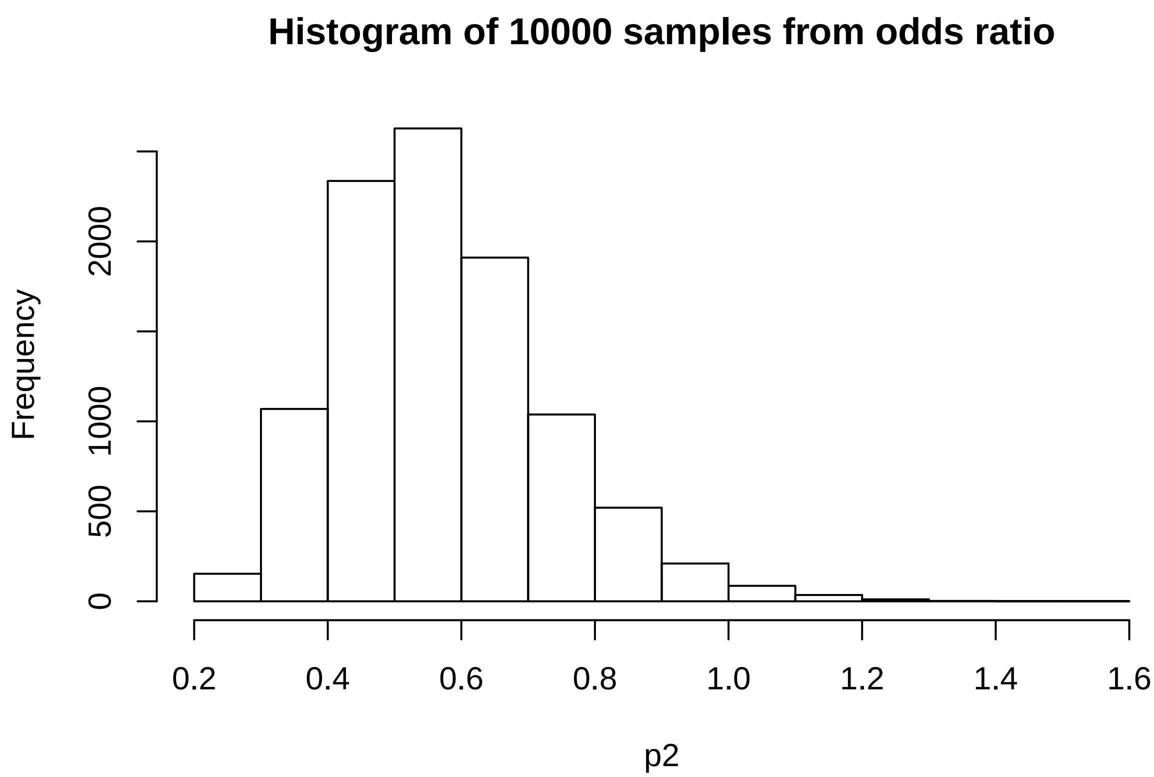
The point estimate is 0.57.

```r
posterior_odds_ratio_interval<-function(p0, p1, prob){
  p2 <- (p1/(1-p1))/(p0/(1-p0))
  limit_1 <- (1-prob)/2
  limit_2 <- limit_1 + prob
  a <- quantile(p2, probs = limit_1)
  b <- quantile(p2, probs = limit_2)
  result <- c(a, b)
  return(result)
}

posterior_interval <- posterior_odds_ratio_interval(p0,p1,0.95)
posterior_interval
```

```
##      2.5%     97.5%
## 0.3188575 0.9370284
```

The 95% posterior interval is [0.32, 0.93].

```r
p2 <- (p1/(1-p1))/(p0/(1-p0))
hist(p2, main = "Histogram of 10000 samples from odds ratio")
```



b)

The chosen prior Beta(1,1) is a noninformative prior, meaning that the role it plays in the posterior distribution must be the minimum possible. Actually, in the given case, considering how the posterior is derived:

$$p(\theta|y) \propto Beta(\theta|\alpha + y, \beta + n - y)$$

If the number of points (n) and the number of positives (y) are large enough, the influence of the chosen prior is very small. From a numerical point, in this example, α and β are equal to 1, while n is 674 and y is equal to 39, so the prior is not relevant enough to change the tendency of the posterior distribution.

Exercise 3

Assuming that the samples have unknown standard deviations σ_1 and σ_2. Considering that the samples are drawn from a normal distribution, the noninformative prior distribution is:

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

The likelihood function is as following:

$$p(y|\mu, \sigma^2) \propto \sigma^{-n} exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

Where:

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Finally, the joint posterior distribution is:

$$p(\mu, \sigma^2|y) \propto \sigma^{-n} \sigma^{-2} exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

For the given case, the interesting expression is the marginal posterior distribution for μ:

$$p(\mu|y) \propto [1 + \frac{n(\mu - \bar{y})^2}{(n-1)s^2}]^{-n/2}$$

Which corresponds with a t distribution:

$$p(\mu|y) \propto t_{n-1}(\bar{y}, s^2|n)$$

Now, for each of the measurements sets:

```
n1 <- length(windshieldy1)
n2 <- length(windshieldy2)

sample_mean_1 <- mean(windshieldy1)
sample_mean_2 <- mean(windshieldy2)

sample_var_1 <- var(windshieldy1)
sample_var_2 <- var(windshieldy2)
print("n")
```

```
## [1] "n"
```

```
n1
```

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

```
n2
```

```
## [1] 13
```

```
print("Sample Means")
```

```
## [1] "Sample Means"
```

```
sample_mean_1
```

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

```
sample_mean_2
```

```
## [1] 15.82108
```

```
print("Sample Variances")
```

```
## [1] "Sample Variances"
```

```
sample_var_1
```

```
## [1] 2.173153
```

```
sample_var_2
```

```
## [1] 0.7614481
```

```
print("s^2/n")
```

```
## [1] "s^2/n"
```

```
s_n_1 <- sample_var_1^2/n1
s_n_2 <- sample_var_2^2/n2
s_n_1
```

```
## [1] 0.5247326
```

```
s_n_2
```

```
## [1] 0.04460024
```

For the two data sets:

$$p(\mu_1|y_1) = t_8(14.611, 0.525)$$

$$p(\mu_2|y_2) = t_{12}(15.821, 0.045)$$

```
p1 <- rtnew(10000, n1, sqrt(1+1/n1)*sqrt(sample_var_1))
p2 <- rtnew(10000, n2, sqrt(1+1/n2)*sqrt(sample_var_2))
p_d <- p1-p2
```

```
posterior_d_interval<-function(p, prob){
  limit_1 <- (1-prob)/2
  limit_2 <- limit_1 + prob
  a <- quantile(p, probs = limit_1)
  b <- quantile(p, probs = limit_2)
  result <- c(a, b)
  return(result)
}

posterior_interval_d <- posterior_d_interval(p_d, 0.95)
posterior_interval_d
```

```
##      2.5%     97.5%
## -2.426276  3.718779
```

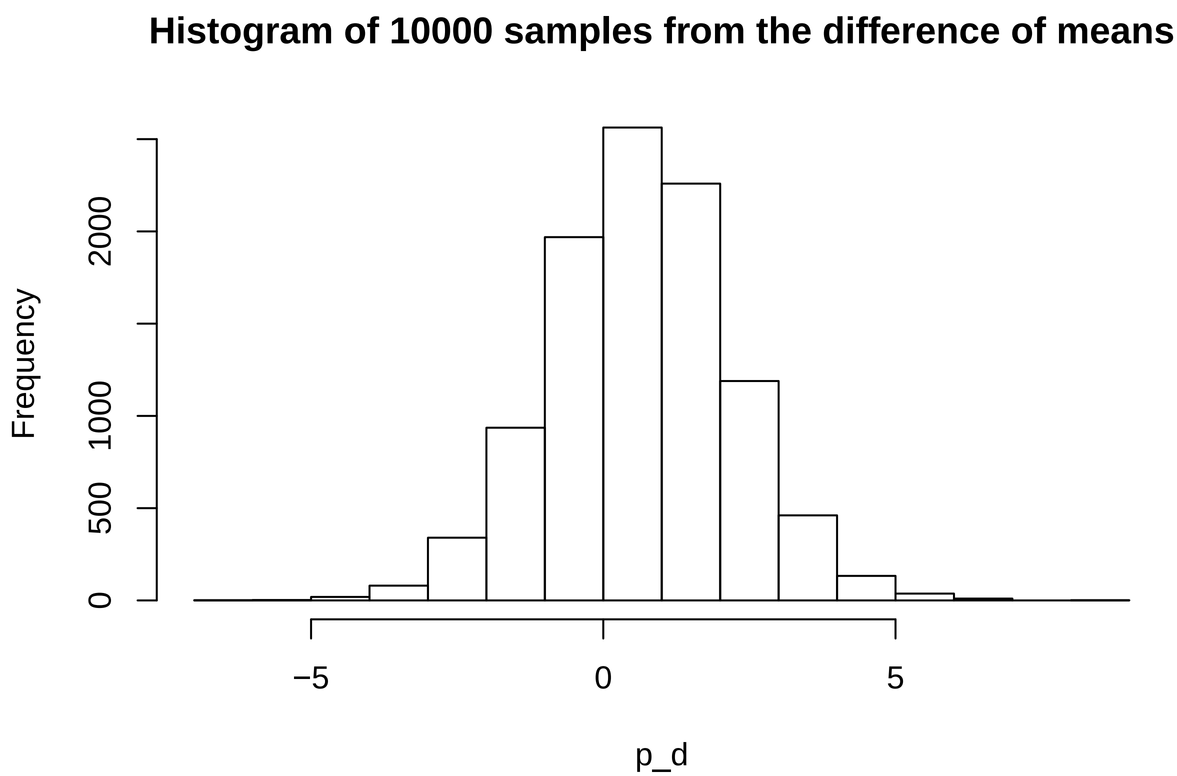
The previous two numbers are the limits for the 95% posterior interval.

```r
posterior_d_point_est <- function(p){
  est <- mean(p)
  return(est)
}

point_d_estimate <- posterior_d_point_est(p_d)
point_d_estimate
```

```
## [1] 0.6480908
```

The point estimate is the obtained number.

```r
hist(p_d, main = "Histogram of 10000 samples from the difference of means")
```



b)

The probability that the means are the same is zero. This is a matter of how the problem is defined. It is a hypothesis testing problem, so the null hypothesis consists in assuming that the computations from the data sets are true, while the alternative hypothesis can be that the subtraction of the means is either greater than zero or smaller than zero or that the subtraction of the means is different than zero. For this reason, the concept of the means taking the same value is not considered in the distribution.