BDA - Assignment 3

Anonymous

Contents

```
# To install aaltobda, see the General information in the assignment.
library(aaltobda)
library(markmyassignment)
library("ggplot2")
assignment_path <-
  paste("https://github.com/avehtari/BDA_course_Aalto/", "blob/master/assignments/tests/assignment3.yml
set_assignment(assignment_path)
```

```
## Assignment set:
## assignment3: Bayesian Data Analysis: Assignment 3
## The assignment contain the following (6) tasks:
## - mu_point_est
## - mu_interval
## - mu_pred_interval
## - mu_pred_point_est
## - posterior_odds_ratio_point_est
## - posterior_odds_ratio_interval
```

```
data("windshieldy1")
head(windshieldy1)
```

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

1. Inference for normal mean and deviation

a)

We assume that the observed hardness values of windshields, y_i, follow a normal distribution with unknown mean μ and standard deviation σ. Our variable of intrest is μ. We have data with $n = 9$ observations.

Assume that the data is i.i.d.

Model likelihood:

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

Non-informative prior is of form:

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

Posterior density for normal data with non-informative prior is derived in the BDA3 book p. 65-66.

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

So, the posterior is actually

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

density. Where $\bar{y}$ is sample mean, n is sample size and s is sample standard deviation.

Our parameter of intrest is μ. The 95 percent central credible/posterior interval for μ is obtained from t_n marginal posterior distribution of μ -> $\bar{y} \pm t_{\alpha,n-1}\frac{s}{\sqrt{n}}$

Let's implement functions to calculate μ point estimate, and μ posterior interval.

```
# mu point estimate = sample mean y_bar
mu_point_est <- function(data){
  y_bar <- mean(data)
  return(y_bar)
}

mu_interval <- function(data, prob){
  y_bar <- mu_point_est(data)
  n <- length(data)
  # sample variance (BDA3, p. 64)
  s_squared <- ( 1 / (n-1) ) * sum((data-y_bar)^2)
  degrees_of_freedom <- n - 1

  lower_bound <- qt((1-prob)/2, degrees_of_freedom)
  upper_bound <- qt( prob + (1-prob)/2, degrees_of_freedom)

  # Need to scale the interval bounds
  lower_scaled <- y_bar + ( lower_bound*(s_squared^0.5) ) / sqrt(n)
  upper_scaled <- y_bar + ( upper_bound*(s_squared^0.5) ) / sqrt(n)

  return(c(lower_scaled, upper_scaled))
}
```

```
mu_point_est(data = windshieldy1)
```

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

```
mu_interval(data = windshieldy1, prob = 0.95)
```

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

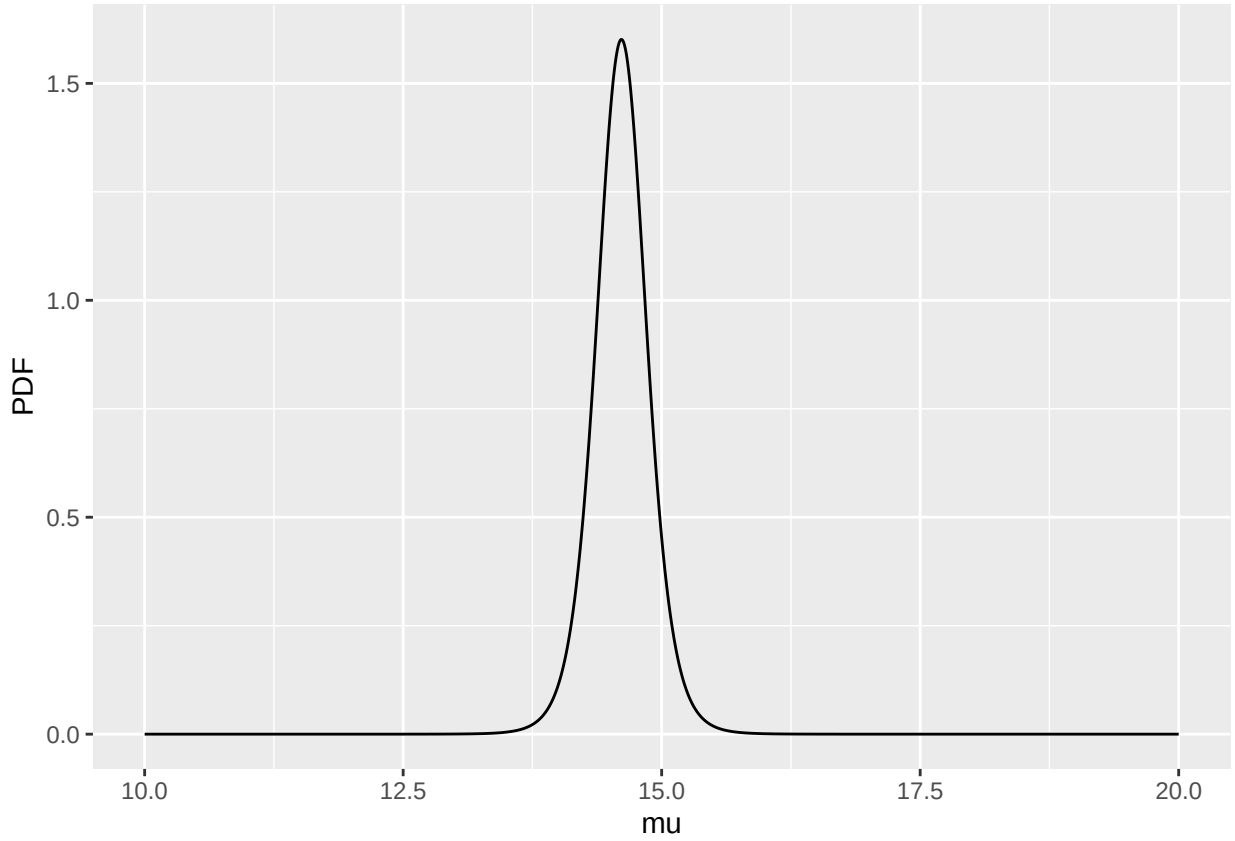
Plotting the density:

```
y_bar <- mu_point_est(windshieldy1)
n = length(windshieldy1)
s_squared <- sum((windshieldy1-y_bar)^2)/(n-1)
mu <- seq(10, 20, 0.01)
degrees_of_freedom <- n - 1

density <- dtnew(mu, degrees_of_freedom, y_bar, s_squared/n)

plot_df <- data.frame(p1 = density, x_var = mu)

ggplot(plot_df, aes(x_var)) + geom_line(aes(y = p1)) + labs(x = "mu") + labs(y = "PDF")
```



b)

Let future observation be notated as $\tilde{y}$.

BDA3, p. 66: "In fact, the posterior predictive distribution of $\tilde{y}$ is a t distribution with location $\hat{y}$, scale $(1+\frac{1}{n})^{\frac{1}{2}}s$, and n-1 degrees of freedom."

```
mu_pred_point_est <- function(data){
  y_bar <- mean(data)
  return(y_bar)
}

mu_pred_interval <- function(data, prob){
  y_bar <- mu_point_est(data)
  n <- length(data)
  s_squared <- ( 1 / (n-1) ) * sum((data-y_bar)^2)
  degrees_of_freedom <- n - 1
  scale <- sqrt((1+1/n))*sqrt(s_squared)

  lower_bound <- qt((1-prob)/2, degrees_of_freedom)
  upper_bound <- qt(prob + (1-prob)/2, degrees_of_freedom)

  lower_scaled <- y_bar + ( lower_bound * scale )
  upper_scaled <- y_bar + ( upper_bound * scale )

  return(c(lower_scaled, upper_scaled))
```

```
}
```

```
mu_pred_point_est(data = windshieldy1)
```

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

```
mu_pred_interval(data = windshieldy1, prob = 0.95)
```

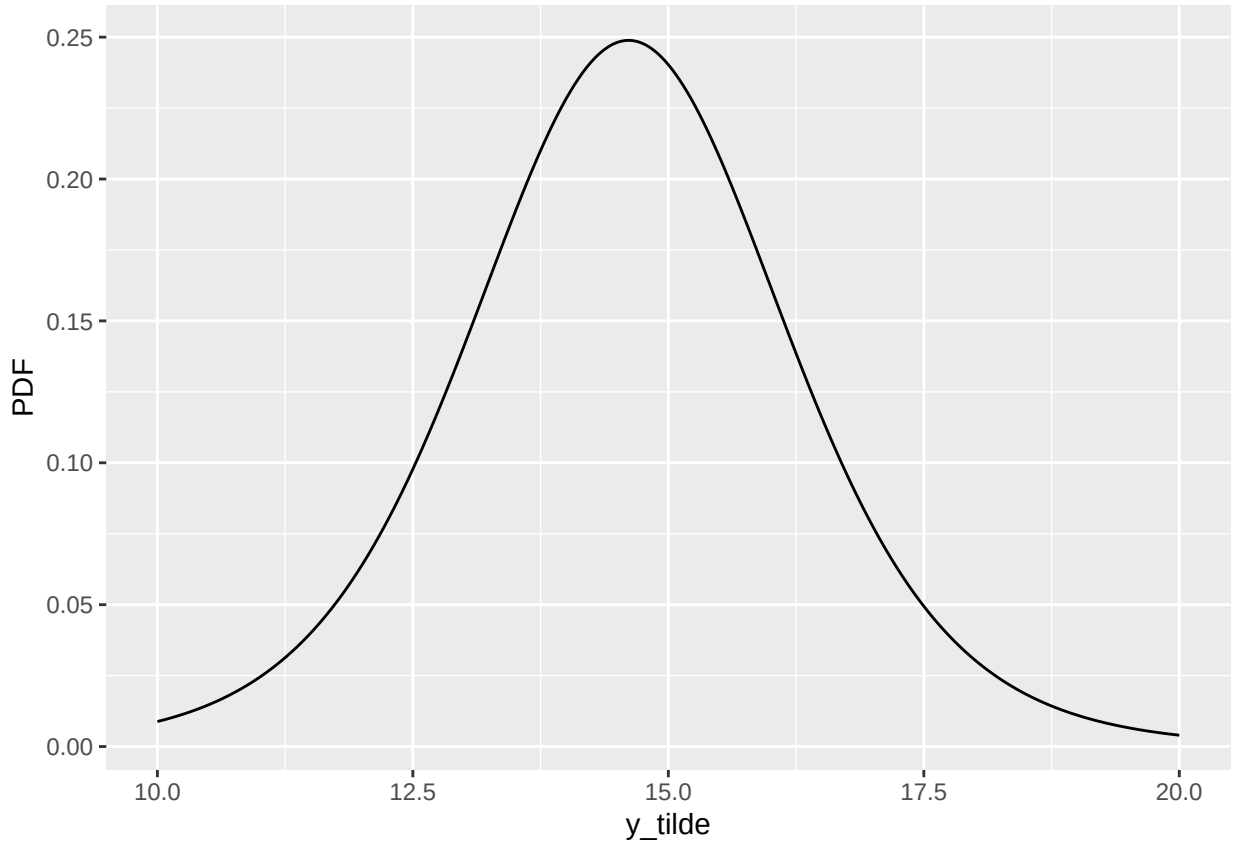
```
## [1] 11.02792 18.19453
```

Plotting the density:

```
s_squared <- sum( (windshieldy1 - y_bar)^2) / (n-1)
x <- seq(10, 20, 0.01)
degrees_of_freedom <- n - 1
s <- sqrt((1+1/n))*sqrt(s_squared)
density <- dtnew(mu, degrees_of_freedom, y_bar, s)

plot_df <- data.frame(p1 = density, x_var = x)

ggplot(plot_df, aes(x_var)) + geom_line(aes(y = p1)) + labs(x = "y_tilde") + labs(y = "PDF")
```



2. Inference for the difference between proportions

a)

Assume that the outcomes in the two groups are independent and binomially distributes. Out parameters of intrest are (p_0, p_1).

Let $i = \{0, 1\}$ be the group indexes.

Let n_i be the amount of patients and y_i the amount of deaths occured in group i.

The data is binomially distributes, thus the likelihood function is of form

$$P(y|p_i) = Bin(y_i|n_i, p_i)$$

.

The prior is non-informative conjugate Beta-prior where both outcomes are equally expected

$$P(p_i) \propto Beta(\alpha = 1, \beta = 1)$$

.

The posterior is also beta distribution of form

$$P(p_i|y_i) \propto Beta(p_i|\alpha + y_i, \beta + n_i - y_i)$$

Resulting...

$$P(p_0|y_0) \propto Beta(p_0|\alpha + y_0, \beta + n_0 - y_0) = Beta(p_0|\alpha' = 40, \beta' = 636)$$

and

$$P(p_1|y_1) \propto Beta(p_1|\alpha + y_1, \beta + n_1 - y_1) = Beta(p_1|\alpha' = 23, \beta' = 659)$$

```
alpha <- 1
beta <- 1

alpha_0 <- 40
beta_0 <- 636
alpha_1 <- 23
beta_1 <- 659

posterior_odds_ratio_point_est <- function(p0,p1){
  ratio <- (p1/(1-p1)) / (p0/(1-p0))
  return(mean(ratio))
}

posterior_odds_ratio_interval <- function(p0, p1, prob){
  ratio <- (p1/(1-p1)) / (p0/(1-p0))
  cred_interval <- as.vector(quantile(ratio, probs = c((1-prob)/2, (1-prob)/2+prob)))
  return(cred_interval)
}

# Sampling from the 2 posterior distributions to get p0 and p1, 10000 simulations
sims <- 10000
p0 <- rbeta(sims, alpha_0, beta_0)
p1 <- rbeta(sims, alpha_1, beta_1)
```

```r
# Point estimate
posterior_odds_ratio_point_est(p0 ,p1)
```

```
## [1] 0.5705463
```

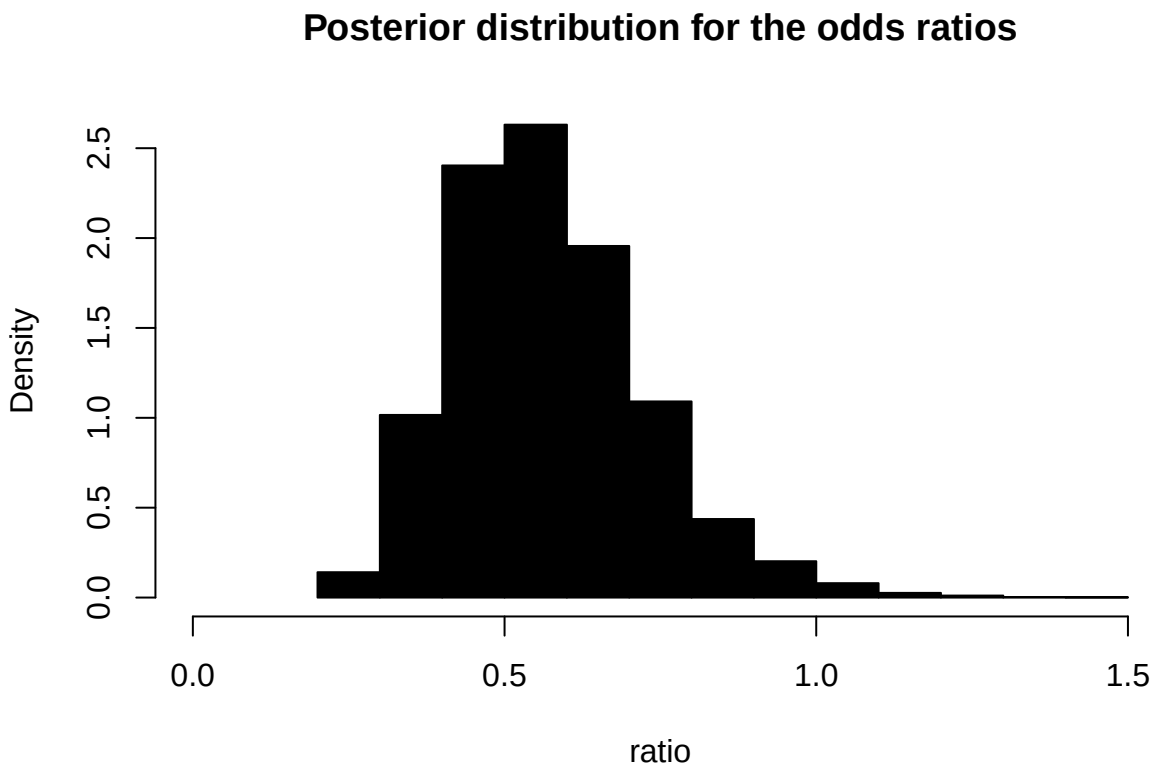
```r
# Credible interval
posterior_odds_ratio_interval(p0, p1, prob = 0.95)
```

```
## [1] 0.3222144 0.9232640
```

```r
# HISTOGRAM

ratio <- (p1/(1-p1))/((p0)/(1-p0))

hist(ratio, main="Posterior distribution for the odds ratios",xlab="ratio",xlim=c(0,1.5),col="black",fr
```



b)

The sensitivity of the inference to the choice of prior density can be analyzed for example by trying out different prior.

```r
alpha <- 3
beta <- 7

alpha_0 <- 3 + 39
beta_0 <- 7 + 674 - 39
alpha_1 <- 3 + 22
```

```r
beta_1 <- 7 + 680 - 22

# Similarly as before, sampling from the 2 posterior distributions to get p0 and p1, 10000 simulations
sims <- 10000
p0 <- rbeta(sims, alpha_0, beta_0)
p1 <- rbeta(sims, alpha_1, beta_1)
```

```r
# Point estimate
posterior_odds_ratio_point_est(p0 ,p1)
```

```
## [1] 0.5888995
```

```r
# Credible interval
posterior_odds_ratio_interval(p0, p1, prob = 0.95)
```

```
## [1] 0.3340802 0.9343667
```

Even though the prior changed from uniform non-informative to weakly informative prior, the point estimate and the 95 % credible inteval didn't change too much.

3. Inference for the difference between normal means

a)

Pretty similar starting point as in exercise 1... This time the parameter of intrest is the difference between the the distibution's mean estimates $\mu_d = \mu_1 - \mu_2$.

Let $i = \{1, 2\}$. We assume that the samples have unknown standard deviations σ_1 and σ_2.

Likelihood function:

$$p(y_i|\mu_i, \sigma_i^2) = N(y_i|\mu_i, \sigma_i^2)$$

Uninformative prior:

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

Similarly, as in the exercise 1, the posterior is of form

$$p(\mu_i|y_i) = t_{n_i-1}\left(\bar{y_i}, \frac{s_i}{n_i^2}\right)$$

```r
# Load data
data("windshieldy1")
data("windshieldy2")

# Start by deriving the needed parameters
y_bar1 = mean(windshieldy1)
n1 = length(windshieldy1)
degrees_of_freedom1 <- n1 -1
var1 = sum( (windshieldy1-y_bar1)^2 ) / (degrees_of_freedom1)
s1 = sqrt(var1/n1)

y_bar2 = mean(windshieldy2)
n2 = length(windshieldy2)
degrees_of_freedom2 <- n2 - 1
var2 = sum( (windshieldy2-y_bar2)^2 ) / (degrees_of_freedom2)
s2 = sqrt(var2/n2)
```

```r
# Sample from the posterior distributions of the mean estimates, 10000 simulations
sims <- 10000
sample1 = rt(sims, degrees_of_freedom1) * s1 + y_bar1
sample2 = rt(sims, degrees_of_freedom2) * s2 + y_bar2
mu_d = sample1 - sample2
```

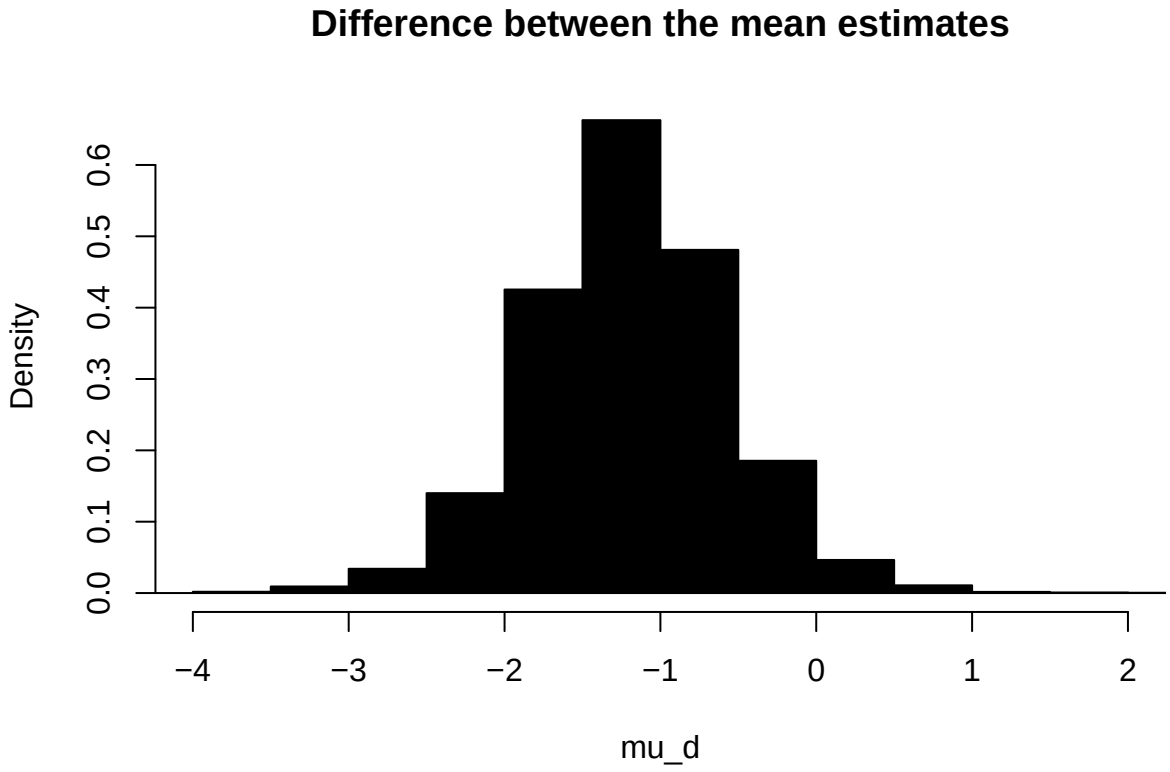
```r
# Bayesian point estimate
mean(mu_d)
```

```
## [1] -1.200806
```

```r
# 95% redible interval estimate
quantile(mu_d, probs = c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -2.47212270  0.06228741
```

```r
# HISTOGRAM
hist(mu_d, main="Difference between the mean estimates", xlab="mu_d", xlim=c(-4,2), col="black", freq=F
```



b)

Analytic solution for the given task of determining the probability that the means are equal, can be studied with a two sample t-test. Our null hypothesis is $\mu_d = 0$ and the alternative hypothesis is $\mu_d \neq 0$.

```r
t.test(sample1, sample2)
```

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -189.31, df = 13966, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.213239 -1.188373
## sample estimates:
## mean of x mean of y
##  14.61553  15.81634
```

Based on the test results, we can conclude that with 95% significance level, the means are not equal.