BDA - Assignment 2

Anonymous

Contents

Loaded packages **1**

Loaded data **1**

Exercise 1) Inference for normal mean and deviation **1**

Exercise 2) Inference for the difference between proportions **4**

Exercise 3) Inference for the difference between normal means **7**

Loaded packages

```r
library(aaltobda)
library(stats)
library(tinytex)
library(ggplot2)
```

Loaded data

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

Exercise 1) Inference for normal mean and deviation

Formulate model likelihood

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

Formulate the prior

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

Formulate the resulting posterior

The joint posterior density is the product of conditional and marginal posterior densities, shown as follows:

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

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

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

a)

```
#point estimate
mu_point_est <-function(data){
  mdata <-sum(data)/length(data)
  s <-sqrt(sum((data-mdata)^2)/(length(data)-1))
  t <-qt(seq(0.01, 0.99, length = 100), df=(length(data)-1))
  temp <-mdata-sum(t)/100*s/sqrt(length(data))
  return(temp)}

mu_point_est(data=windshieldy1)
```

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

The Bayesian point estimate is 14.61122

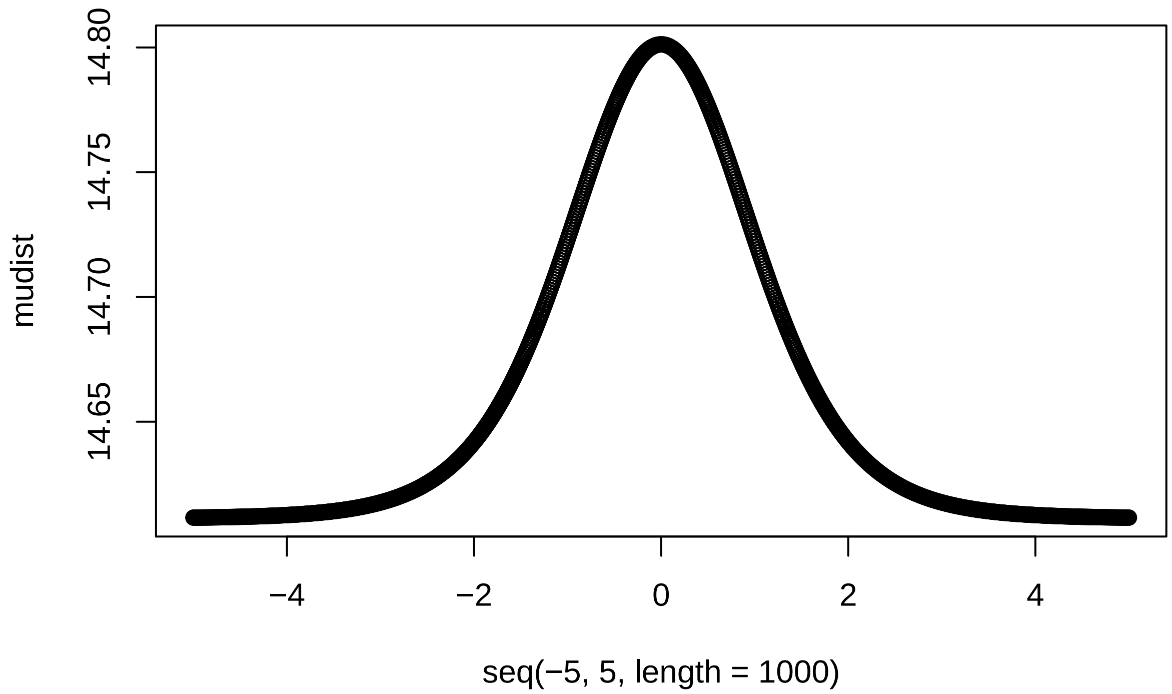
```
#posterior interval 95%
mu_interval <-function(data, prob){
  mdata <-sum(data)/length(data)
  s <-sqrt(sum((data-mdata)^2)/(length(data)-1))
  t <-qt(c((1-prob)/2, prob+(1-prob)/2), df=(length(data)-1))
  temp <-c(mdata+t[1]*s/sqrt(length(data)),
           mdata+t[2]*s/sqrt(length(data)))
  return(temp)}

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

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

The 95% posterior interval is [13.47808, 15.74436]

```
#plot the density
mdata <-sum(windshieldy1)/length(windshieldy1)
s <-sqrt(sum((windshieldy1-mdata)^2)/(length(windshieldy1)-1))
tdist<-dt(seq(-5,5,length=1000),length(windshieldy1)-1)
mudist<-mdata+tdist*s/sqrt(length(windshieldy1))
plot(seq(-5,5,length=1000),mudist)
```



b)

```r
#point estimate
mu_pred_point_est <-function(data){
  mdata <-sum(data)/length(data)
  s <-sqrt(sum((data-mdata)^2)/(length(data)-1))
  t <-qt(seq(0.01, 0.99, length = 100), df=(length(data)-1))
  temp <-mdata-sum(t)/100*s*(1+1/length(data))
  return(temp)}

mu_pred_point_est(data=windshieldy1)
```

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

The Bayesian point estimate is 14.61122

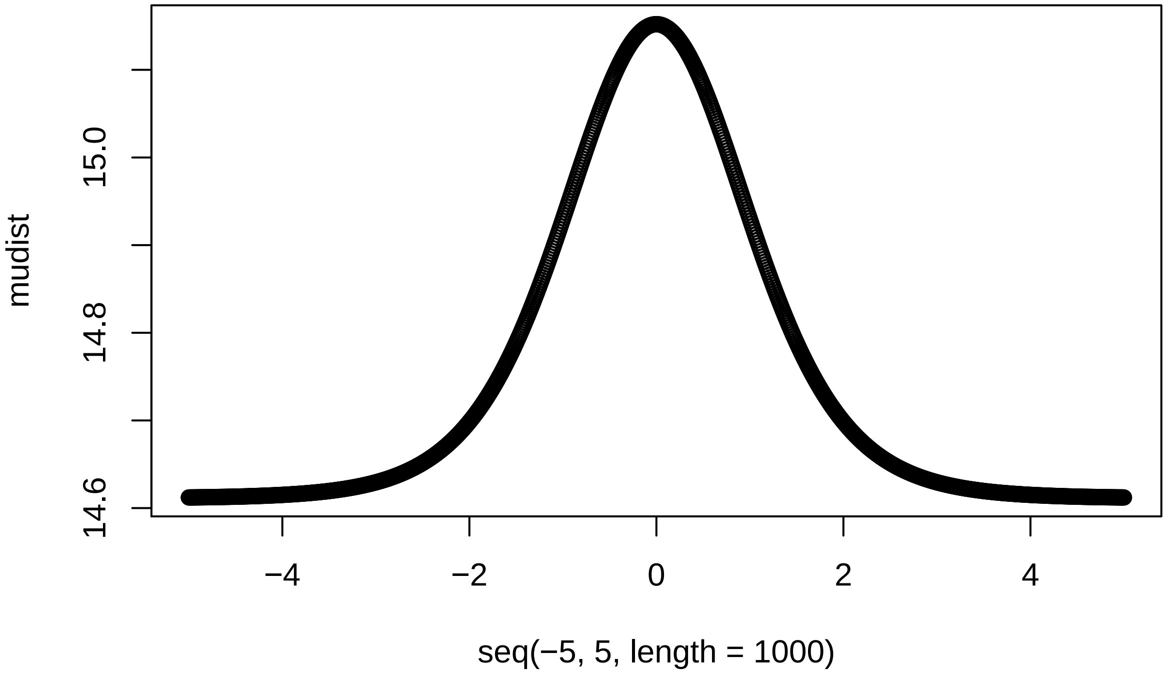
```r
#predictive interval 95%
mu_pred_interval <-function(data, prob){
  mdata <-sum(data)/length(data)
  s <-sqrt(sum((data-mdata)^2)/(length(data)-1))
  t <-qt(c((1-prob)/2, prob+(1-prob)/2), df=(length(data)-1))
  temp <-c(mdata+t[1]*s*sqrt(1+1/length(data)),
           mdata+t[2]*s*sqrt(1+1/length(data)))
  return(temp)}

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

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

The 95% predictive interval is [11.02792, 18.19453]

```
#plot the density
mdata <-sum(windshieldy1)/length(windshieldy1)
s <-sqrt(sum((windshieldy1-mdata)^2)/(length(windshieldy1)-1))
tdist<-dt(seq(-5,5,length =1000),length(windshieldy1)-1)
mudist<-mdata+tdist*s/sqrt(1+1/length(windshieldy1))
plot(seq(-5,5,length =1000),mudist)
```



Exercise 2) Inference for the difference between proportions

Formulate model likelihood

Formulate the prior

Formulate the resulting posterior

The posterior distribution for p_0 is Beta(40,636)

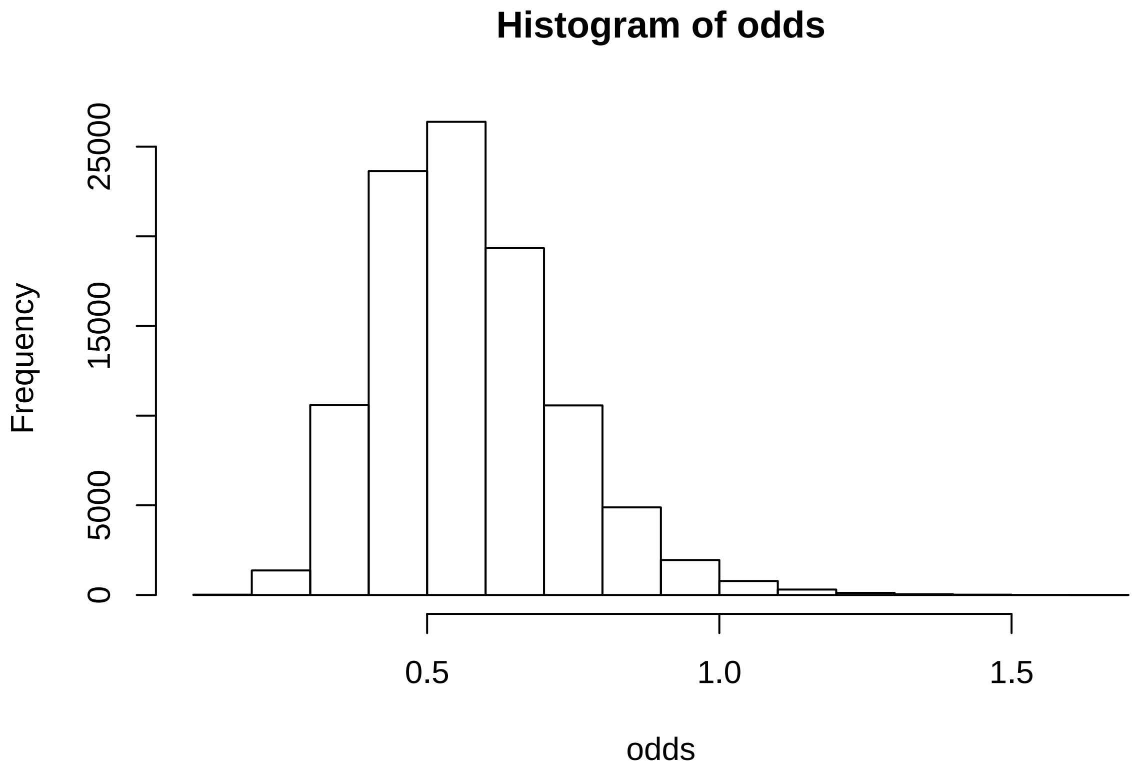
The posterior distribution for p_1 is Beta(23,659)

a)

```
set.seed(4711)
p0 <-rbeta(100000,1+39,1+674-39)
p1 <-rbeta(100000,1+22,1+680-22)
```

The histogram of 100000 draws from the posterior density of the odds ratio is attached.

```
#plot the histogram
odds <- (p1/(1-p1))/(p0/(1-p0))
hist(odds)
```



```
#point estimate
posterior_odds_ratio_point_est <-function(p0,p1){
  temp <-mean((p1/(1-p1))/(p0/(1-p0)))
  return(temp)
}

posterior_odds_ratio_point_est(p0, p1)
```

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

```r
#95% posterior interval
posterior_odds_ratio_interval <-function(p0,p1,prob){
  temp <- (p1/(1-p1))/(p0/(1-p0))
  temp <-quantile(temp,probs=c((1-prob)/2,1-(1-prob)/2))
  return(temp)
}

posterior_odds_ratio_interval(p0, p1, prob = 0.95)
```

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

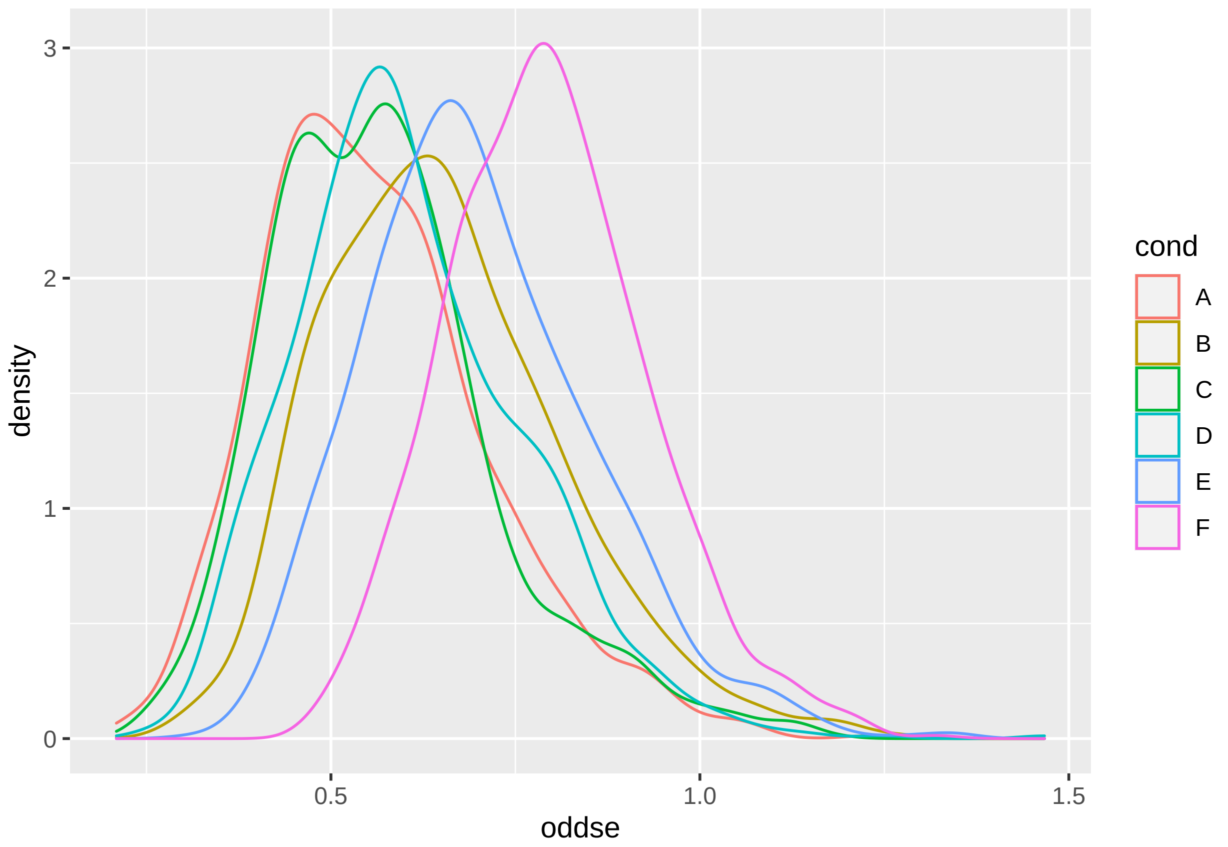
Based on 100000 simulated values of (μ, σ^2), we estimate the posterior mean of the odd to be 0.570978 and a 95% central posterior interval for the odds ratio to be [0.321063, 0.924998], close to the analytically calculated interval.

b)

```r
posterior_odds_ratio_gen <-function(alpha,beta){
  p0 <-rbeta(1000,alpha+39,beta+674-39)
  p1 <-rbeta(1000,alpha+22,beta+680-22)
  temp <- (p1/(1-p1))/(p0/(1-p0))
  return(temp)
}
```

```r
#different parármeters of beta distribution
beteseq0 <-posterior_odds_ratio_gen(1,1)   #alpha = beta = 1
beteseq1 <-posterior_odds_ratio_gen(10,1)  #alpha = 10, beta = 1
beteseq2 <-posterior_odds_ratio_gen(1,10)  #alpha = 1, beta = 10
beteseq3 <-posterior_odds_ratio_gen(5,5)   #alpha = beta = 5
beteseq4 <-posterior_odds_ratio_gen(20,20) #alpha = beta = 20
beteseq5 <-posterior_odds_ratio_gen(50,20) #alpha = 50, beta = 20

dataa <-data.frame(cond = factor(rep(c("A","B","C","D","E","F"),each=1000)),
                   oddse =c(beteseq0,beteseq1,beteseq2,beteseq3,beteseq4,beteseq5))
ggplot(dataa,aes(x=oddse, colour=cond))+ geom_density()
```



Here I made some changes in the parameters of Beta distribution. Based on the graph, we can say that the choice of prior density will affect the results sightly based on the distributions.

Exercise 3) Inference for the difference between normal means

Formulate model likelihood

- $p(y_1|\mu_1, \sigma_1^2) \sim N(\mu_1, \sigma_1^2)$
- $p(y_2|\mu_2, \sigma_2^2) \sim N(\mu_2, \sigma_2^2)$

Formulate the prior

- $p(\mu_1, \sigma_1^2) \propto (\sigma_1^2)^{-1} \propto \sigma_1^{-2}$
- $p(\mu_2, \sigma_2^2) \propto (\sigma_2^2)^{-1} \propto \sigma_2^{-2}$

Formulate the resulting posterior

The joint posterior density is the product of conditional and marginal posterior densities, shown as follows:

- $p(\mu_1, \sigma_1^2|y_1) = p(\mu_1|\sigma_1^2, y_1)p(\sigma_1^2|y_1) = \sigma_1^{-n_1-2}\exp(-\frac{1}{2\sigma_1^2}[(n_1-1)s_1^2 + n(\bar{y_1} - \mu_1)^2])$

where $\bar{y_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i}$

and $s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y_1})^2$

- $p(\mu_2, \sigma_2^2|y_2) = p(\mu_2|\sigma_2^2, y_2)p(\sigma_2^2|y_2) = {\sigma_2}^{-n_2-2}\exp(-\frac{1}{2\sigma_2^2}[(n_2-1)s_2^2 + n(\bar{y_2} - \mu_2)^2])$

where $\bar{y_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i}$

and $s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y_2})^2$

a)

Assuming the noninformative prior distribution $p(\mu, \sigma^2) \propto \sigma^{-2}$

```
#point estimate
mu_point_est_3 <-function(data){
  mdata <-sum(data)/length(data)
  s <-sqrt(sum((data-mdata)^2)/(length(data)-1))
  t <-rt(100000,length(data)-1)
  temp <- mdata-t*s/sqrt(length(data))
  return(temp)}

mu1<-mu_point_est_3(data=windshieldy1)
mu2<-mu_point_est_3(data=windshieldy2)

mean(mu1-mu2)
```

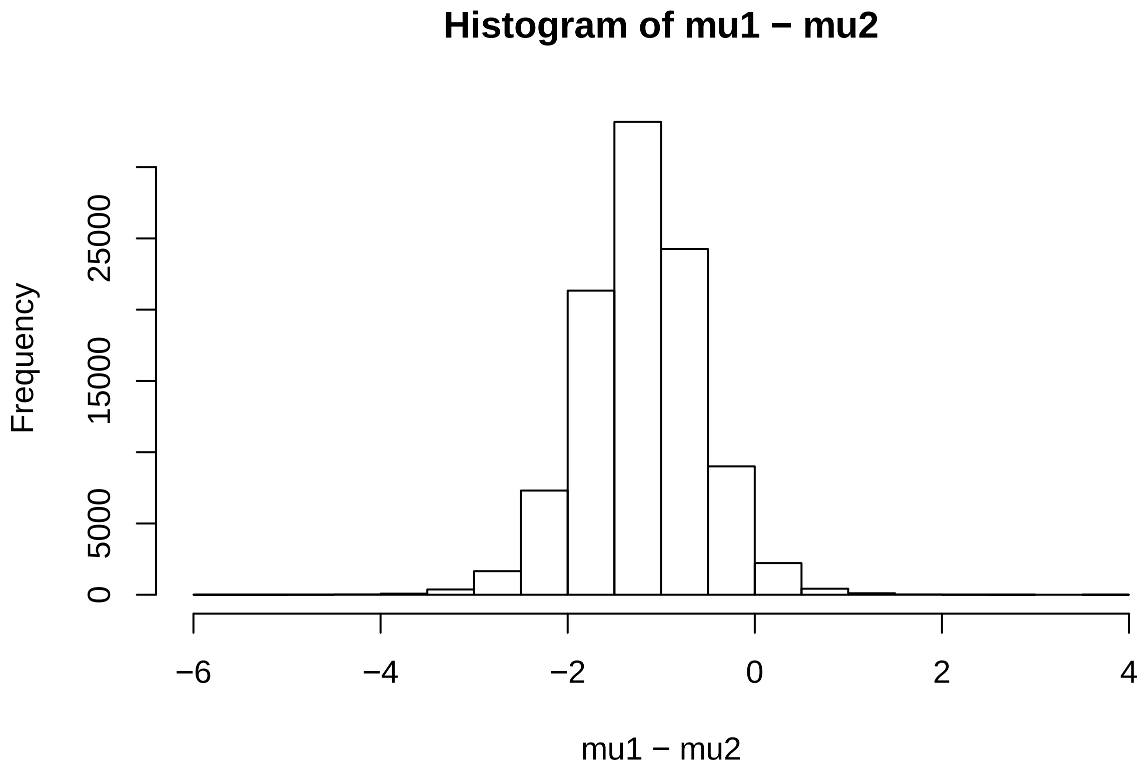
```
## [1] -1.208278
```

```
#95% posterior interval
probs=0.95
quantile(mu1-mu2,probs=c((1-probs)/2,1-(1-probs)/2))
```

```
##        2.5%       97.5%
## -2.45021188  0.03564232
```

Based on 100000 simulated values of difference between μ_1 and μ_2, we estimate the posterior mean to be -1.208278 and a 95% central posterior interval to be [-2.45021188, 0.03564232].

```
#plot the histogram
hist(mu1-mu2)
```



The posterior distribution of μ, given σ^2, we simply use the result for the mean of a normal distribution with known variance and a uniform prior distribution: $\mu|\sigma^2, y \sim N(y, \frac{\sigma^2}{n})$

The joint posterior density is the product of conditional and marginal posterior densities:

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

$= \frac{1}{\sqrt{(2\mu)}\sigma} exp(-\frac{1}{2\sigma^2}(y-\mu)^2)$

$= \frac{1}{\sqrt{(2\mu_1)}\sigma_1} exp(-\frac{1}{2\sigma_1^2}(y_1-\mu_1)^2)$

$= \frac{1}{\sqrt{(2\mu_2)}\sigma_2} exp(-\frac{1}{2\sigma_2^2}(y_2-\mu_2)^2)$

b)

```r
t.test(windshieldy1,windshieldy2,var.equal=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  windshieldy1 and windshieldy2
## t = -2.2088, df = 11.886, p-value = 0.04759
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.40458201 -0.01512739
## sample estimates:
## mean of x mean of y
##  14.61122  15.82108
```

Based on the test, we reject the null hypothesis that the difference in means is equal to 0. In othe words, we can say that the means could not be the same.

In complicated problems,for example, analyzing the results of many survey questions simultaneously, the number of multinomial categories, and thus parameters, becomes so large that it is hard to usefully analyze a dataset of moderate size without additional structure in the model. Formally, additional information can enter the analysis through the prior distribution or the sampling model.

```r
library(markmyassignment)
```

```
## Warning: package 'markmyassignment' was built under R version 3.5.3
```

```r
assignment_path <- paste("https://github.com/avehtari/BDA_course_Aalto/",
                         "blob/master/assignments/tests/assignment3.yml", sep="")
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
```

```r
# To check your code/functions, just run
mark_my_assignment()
```

```
## v |  OK F W S | Context
```

```
## Warning: package 'testthat' was built under R version 3.5.3
```

```
##
/ |   0       | mu_point_est()
v |   4       | mu_point_est()
##
/ |   0       | mu_interval()
v |   5       | mu_interval()
##
/ |   0       | mu_pred_interval()
v |   5       | mu_pred_interval()
##
/ |   0       | mu_pred_point_est()
v |   4       | mu_pred_point_est()
##
/ |   0       | posterior_odds_ratio_point_est()
```

```
v |   6       | posterior_odds_ratio_point_est()
##
/ |   0       | posterior_odds_ratio_interval()
\ |   6       | posterior_odds_ratio_interval()
v |   8       | posterior_odds_ratio_interval() [0.1 s]
##
## == Results ====================================================================================
## Duration: 0.4 s
##
## OK:       32
## Failed:   0
## Warnings: 0
## Skipped:  0
## You're a coding rockstar!
```