BDA - Assignment 3

1. Inference for normal mean and deviation

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

Formulate

- **the likelihood p(y|μ, σ^2)**
- **the prior p(μ)**
- **the resulting posterior p(μ, σ^2|y)**

Solution.

The likelihood, the prior and the resulting prior are stated respectively as,

$$\begin{aligned} p(y|\mu,\sigma^2) &\propto \mathcal{N}(\mu,\sigma^2) \\ p(\mu,\sigma^2) &\propto (\sigma^2)^{-1} \\ p(\mu,\sigma^2|y) &\propto \sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2+n(\overline{y}-\mu)^2]\right) \end{aligned}$$

such that,

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

a) What can you say about the unknown μ? Summarize your results using Bayesian point estimate, a posterior interval, and plot the density.

Solution.

After marginalization of $p(\mu,\sigma^2|y)$ with respect to σ^2 (as we don't care about the value of σ^2), we get that the posterior distribution μ has the form,

$$\frac{\mu-\bar{y}}{s/\sqrt{n}} \mid y \sim t_{n-1}$$

Let us now compute the Bayesian point estimate and a posterior interval,

```
mu_point_est <- function(data) {
  mean(data)
}

mu_interval <- function(data,prob) {
  low_int = (1-prob)/2
  high_int = 1 - low_int
```

```
  y_bar = mean(data)
  n = length(data)
  s = sqrt((1/(n-1)) *sum((data-y_bar)^2))

  qtnew(c(low_int,high_int), n-1, mean = y_bar,scale = s/sqrt(n))
}

mu_point_est(data = windshieldy1)
```

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

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

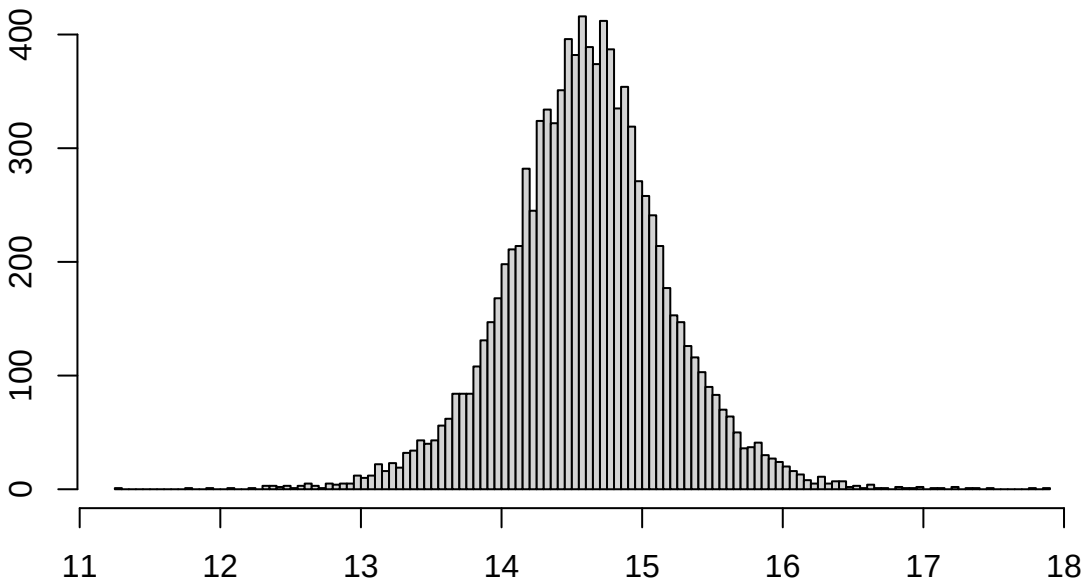
```
## [1] 13.47808 15.74436
```

So our point estimate is 14.6 and the 95% posterior interval is [13.5,15.7]. Finally for the density plot we sample data using the rtnew() command from the aaltobda library,

```
data <- windshieldy1

n = length(data)
y_bar = mean(data)
s = sqrt((1/(n-1)) *sum((data-y_bar)^2))
sample_data = rtnew(10000,n-1, mean = y_bar,scale = s/sqrt(n))

hist(sample_data,breaks = 100, main ="", ylab = "",xlab = "")
```



b) What can you say about the hardness of the next windshield coming from the production line before actually measuring the hardness? Summarize your results using Bayesian point estimate, a predictive interval, and plot the density.

Solution.

Now that we have estimated the μ we are able to simulate the hardness of the windshield from a closed form distribution. Referring to chapter 3.2 on the book, we can find the posterior predictive distribution of $\tilde{y}$. In fact $\tilde{y}$ is a t distribution with location $\bar{y}$, scale $(1+\frac{1}{n})^{0.5} \cdot s$ and $n-1$ degrees of freedom,

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

mu_pred_interval <- function(data,prob) {
  low_int = (1-prob)/2
  high_int = 1 - low_int
  y_bar = mean(data)
  n = length(data)
  s =sqrt((1/(n-1)) *sum((data-y_bar)^2))
  qtnew(c(low_int,high_int), n-1, mean = y_bar,scale = (1+1/n)^0.5*s)
}

mu_pred_point_est(data = windshieldy1)
```

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

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

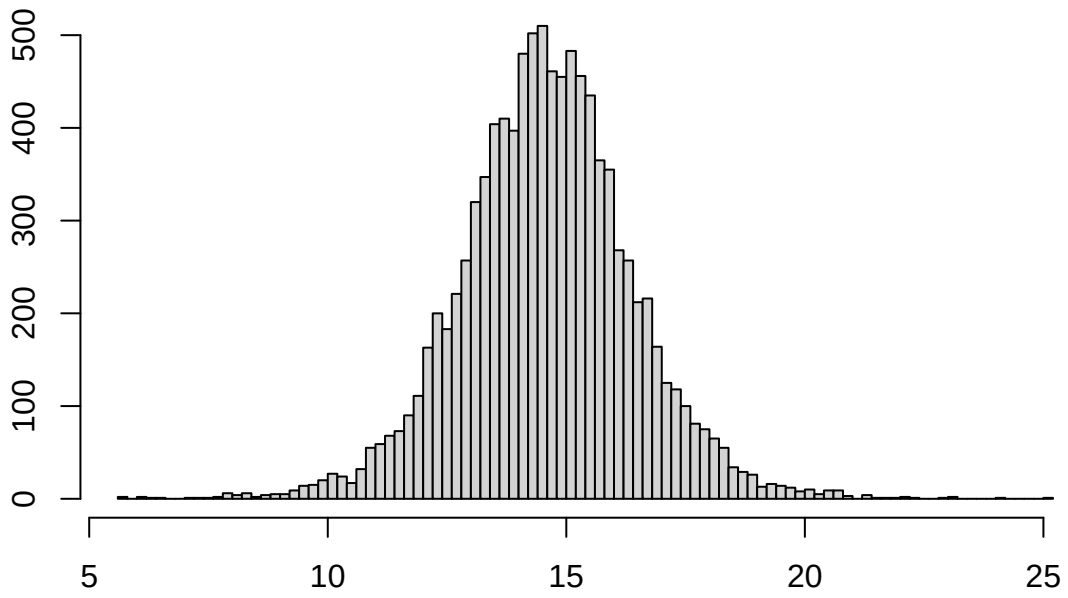
```
## [1] 11.02792 18.19453
```

So our point estimate is 14.6 and the 95% predictive interval is [11.0,18.2]. Now plotting the density of the predictive distribution,

```r
  data <- windshieldy1

  n = length(data)
  y_bar = mean(data)
  s = sqrt((1/(n-1)) *sum((data-y_bar)^2))
  sample_data = rtnew(10000,n-1, mean = y_bar,scale = (1+1/n)^0.5*s)

  hist(sample_data,breaks = 100, main ="", ylab = "",xlab = "")
```



2. Inference for the difference between proportions

a) Summarize the posterior distribution for the odds ratio. Compute the point estimate, a posterior interval, and plot the histogram.

Solution.

As we assume that the outcomes are independent and binomially distributed, we get the likelihoods for p_0 and p_1

$$p(p_0|\pi_0) \sim Bin(n_0, \pi_0) = Bin\left(674, \frac{39}{674}\right)$$
$$p(p_1|\pi_1) \sim Bin(n_1, \pi_1) = Bin\left(680, \frac{22}{680}\right)$$

Then we set up an uninformative prior, $\pi_i \sim Beta(1, 1)$. Thus the resulting posterior density is similar to the result from last week's assignment,

$$p(\pi_0|p_0) \sim Beta(1 + 39, 1 + 674 - 39) = Beta(40, 636)$$
$$p(\pi_1|p_1) \sim Beta(1 + 22, 1 + 680 - 22) = Beta(23, 659)$$

For the posterior for the odds ratio we have,

$$\begin{aligned} p(\pi_0, \pi_1|p_0, p_1) &\sim \frac{Beta(23, 659)}{Beta(40, 636)} \\ &= \frac{p_1^{22}(1-p_1)^{658}B(40, 636)}{p_0^{39}(1-p_0)^{635}B(23, 659)} \end{aligned}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

Now let us compute the point estimate and posterior interval,

```
sampled_p0 = rbeta(100000, 1 + 39, 1+ 674-39)
sampled_p1 = rbeta(100000, 1 + 22, 1+ 680-22)

posterior_odds_ratio_point_est <- function(p0, p1){
  odd_ratio = (p1/(1-p1))/(p0/(1-p0))
  mean(odd_ratio)
}

posterior_odds_ratio_interval <- function(p0, p1, prob){
  odd_ratio = (p1/(1-p1))/(p0/(1-p0))
  low_int = (1-prob)/2
  high_int = 1- low_int
  quantile(odd_ratio, probs = c(low_int,high_int))
}

posterior_odds_ratio_point_est(sampled_p0, sampled_p1)
```

```
## [1] 0.5709003
```

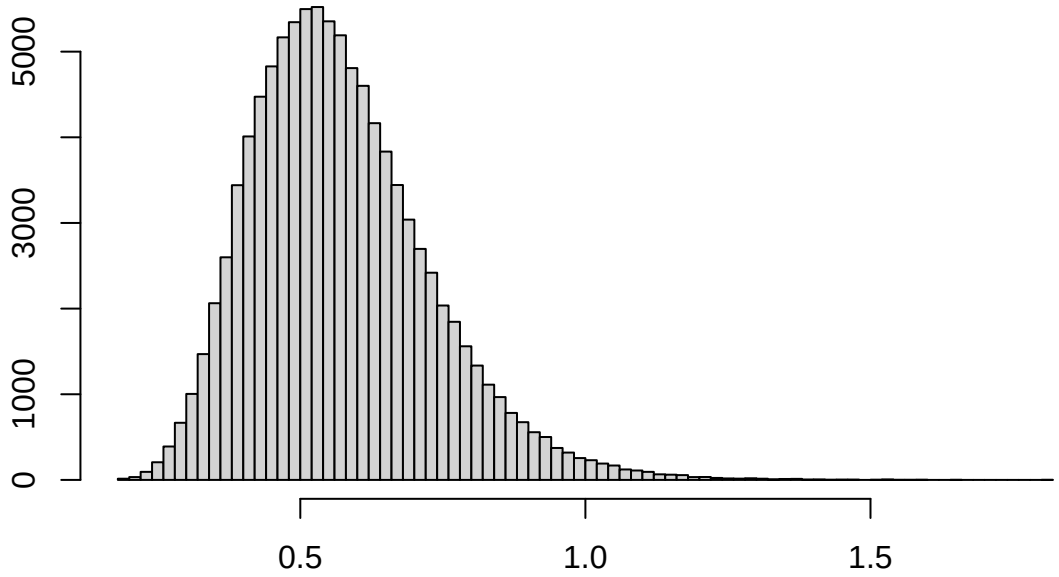
```
posterior_odds_ratio_interval(sampled_p0, sampled_p1, prob = 0.95)
```

```
##      2.5%     97.5%
## 0.3215127 0.9296650
```

So the point estimate is 0.57 and the interval [0.32,0.93]. The fact all of these values are less than 1 imply that event of dying from being assigned control group is higher than when getting treatment. Now for the histogram,

```
p0 = sampled_p0
p1 = sampled_p1

odd_ratio = (p1/(1-p1))/(p0/(1-p0))
hist(odd_ratio,breaks = 100, main ="", ylab = "",xlab = "")
```



b) Discuss the sensitivity of your inference to your choice of prior density with a couple of sentences.

Solution.

Looking at the posterior distribution we see that changing the parameters (α, β) in the prior distribution will not have an huge effect on the posterior distribution of the odd's ratio due to cancellation of the parameters. So the model is robust in terms of being insensitive to uninformative or weakly informative priors.

3. Inference for the difference between normal means

Formulate

- **the likelihood**
- **the prior**
- **the resulting posterior**

Solution.

The likelihood, the prior and the resulting prior are stated respectively as,

$$\begin{aligned} p(y_i|\mu_i,\sigma_i^2) &\propto \mathcal{N}(\mu_i,\sigma_i^2) \\ p(\mu_i,\sigma_i^2) &\propto (\sigma_i^2)^{-1} \\ p(\mu_i,\sigma_i^2|y_) &\propto \sigma_i^{-n-2}\exp\left(-\frac{1}{2\sigma_i^2}[(n-1)s_i^2+n(\overline{y_i}-\mu_i)^2]\right) \end{aligned}$$

such that,

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}((y_i)_j-\overline{y_i})^2$$

a) What can you say about μ_d? Summarize your results using a Bayesian point estimate, a posterior interval, and plot the histogram.

Solution.

Due to the linearity of distribution the difference μ_d is also distributed according the t-distribution with the proper parameters. Now let us compute the point estimate and interval,

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

mu_point_est_2 <- function(data1, data2) {
      mean(data1-data2)
}

mu_interval_2 <- function(data1,data2,prob) {
  low_int = (1-prob)/2
  high_int = 1 - low_int
  y_bar1 = mean(data1)
  n1 = length(data1)
  s1 = sqrt((1/(n1-1)) *sum((data1-y_bar1)^2))

  y_bar2 = mean(data2)
  n2 = length(data2)
  s2 = sqrt((1/(n2-1)) *sum((data2-y_bar2)^2))

  mu1 = rtnew(10000, n1-1, mean = y_bar1,scale = s1/sqrt(n1))

  mu2 = rtnew(10000, n2-1, mean = y_bar2,scale = s2/sqrt(n2))
  quantile(mu1-mu2,c(low_int,high_int) )
```

```
}

mu_point_est_2(data1 = windshieldy1, data2 = windshieldy2)
```

```
## Warning in data1 - data2: longer object length is not a multiple of shorter
## object length
```

```
## [1] -1.207308
```

```
mu_interval_2(data1 = windshieldy1,data2 = windshieldy2, prob = 0.95)
```

```
##         2.5%        97.5%
## -2.456565687  0.002727103
```

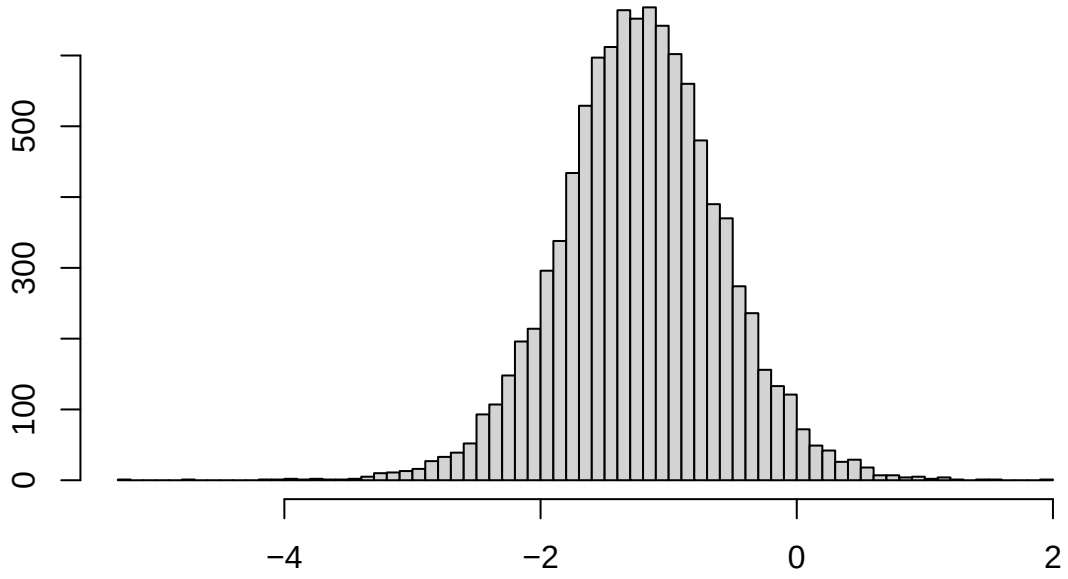
So the point estimate is -1.2 and the interval [-2.4,-0.03]. As all of these values are negative this implies that in in general we have hardness of the windshields in more for the second production company. This is further emphasized in the following histogram,

```
data1 = windshieldy1
data2 = windshieldy2
y_bar1 = mean(data1)
n1 = length(data1)
s1 = sqrt((1/(n1-1)) *sum((data1-y_bar1)^2))

y_bar2 = mean(data2)
n2 = length(data2)
s2 = sqrt((1/(n2-1)) *sum((data2-y_bar2)^2))

mu1 = rtnew(10000, n1-1, mean = y_bar1,scale = s1/sqrt(n1))
mu2 = rtnew(10000, n2-1, mean = y_bar2,scale = s2/sqrt(n2))

hist(mu1-mu2,breaks = 100, main ="", ylab = "",xlab = "")
```



b) Given this specific model, what is the probability that the means are exactly the same ($\mu_1 = \mu_2$)?

Solution.

We know that the posterior distributions of μ_1 and μ_2 are continuous which means that the posterior distribution of μ_d is also continuous. Now for we know that for any continuous distribution, the probability that it equals to exactly some value is always **zero**. This holds for any continuous distribution, we only can get non-zero values when dealing with the probability of getting values in an interval.