BDA - Assignment 3

Contents

Exercise 1.

a)

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

My model goes as follows. We have our prior distribution $P(\mu, \sigma^2) \propto \sigma^{-2}$, the likelihood $P(y|\mu, \sigma^2) = (\frac{1}{\sigma\sqrt{2\pi}})^n exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2)$, and the resulting posterior $P(\mu, \sigma^2|y) \propto \sigma^{-n-2}exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\hat{y} - \mu)^2])$.
Now if we margin out σ we can get our marginal posterior $P(\mu|y) = \int_0^\infty \sigma^{-n-2}exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\hat{y} - \mu)^2])d\sigma^2 \propto [1 + \frac{n(\mu-\overline{y})^2}{(n-1)s^2}]^{-n/2}$, which means that $P(\mu|y) = t_{n-1}(\mu|\overline{y}, s^2/n)$

```r
mu_point_est<-function(data){
n_total<-length(data)
data_sum<-sum(data)
loc<-data_sum/n_total
scale<-var(data)/n_total
sample<-rt(100000,n_total-1)
sample<-sample*sqrt(scale)+loc
return(mean(sample))
}

mu_interval<-function(data,prob){
n_total<-length(data)
data_sum<-sum(data)
loc<-data_sum/n_total
scale<-var(data)/n_total
sample<-rt(100000,n_total-1)
sample<-sample*sqrt(scale)+loc
interval<-quantile(sample,probs=c((1-prob)/2,1-(1-prob)/2))
return (interval)
}

mu_point_est(data = windshieldy1)
```

```
## [1] 14.61019
```

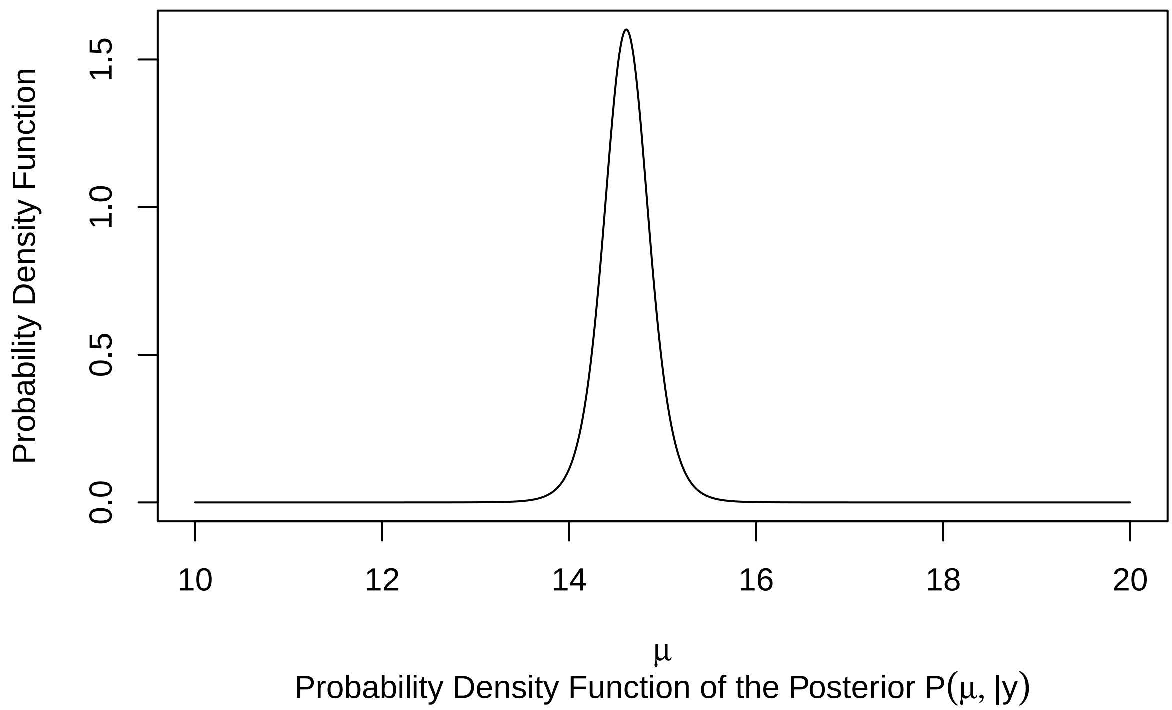
```r
mu_interval(data = windshieldy1, prob = 0.95)
```

```
##      2.5%    97.5%
```

```
## 13.48313 15.73906
```

So for the value of the unknown μ according to the observations and the prior knowledge, the point estimate is 14.6 and the 95% interval estimate is [13.5,15.7].
And the plot of the density of the posterior is shown below.

```
n_total<-length(windshieldy1)
data_sum<-sum(windshieldy1)
loc<-data_sum/n_total
scale<-var(windshieldy1)/n_total
x<-seq(10,20,len=10000)
y<-dtnew(x,n_total-1,loc,scale)
xlab<-expression(mu)
ylab<-"Probability Density Function"
sub<-expression(paste("Probability Density Function of the Posterior ",paste(P(mu,paste("|",y)))))
plot(x,y,'l',xlab=xlab,ylab=ylab,sub=sub)
```



b)

The posterior predictive distribution of the hardness of the next windshield coming from the production line is $P(\tilde{y}|y) = t_{n-1}(\tilde{y}|\overline{y}, (1+\frac{1}{n})s^2)$

```
mu_pred_point_est<-function(data){
n_total<-length(data)
data_sum<-sum(data)
loc<-data_sum/n_total
scale<-var(data)*(1+1/n_total)
sample<-rt(100000,n_total-1)
```

```
sample<-sample*sqrt(scale)+loc
return(mean(sample))
}

mu_pred_interval<-function(data,prob){
n_total<-length(data)
data_sum<-sum(data)
loc<-data_sum/n_total
scale<-var(data)*(1+1/n_total)
sample<-rt(100000,n_total-1)
sample<-sample*sqrt(scale)+loc
interval<-quantile(sample,probs=c((1-prob)/2,1-(1-prob)/2))
return (interval)
}

mu_pred_point_est(data = windshieldy1)
```

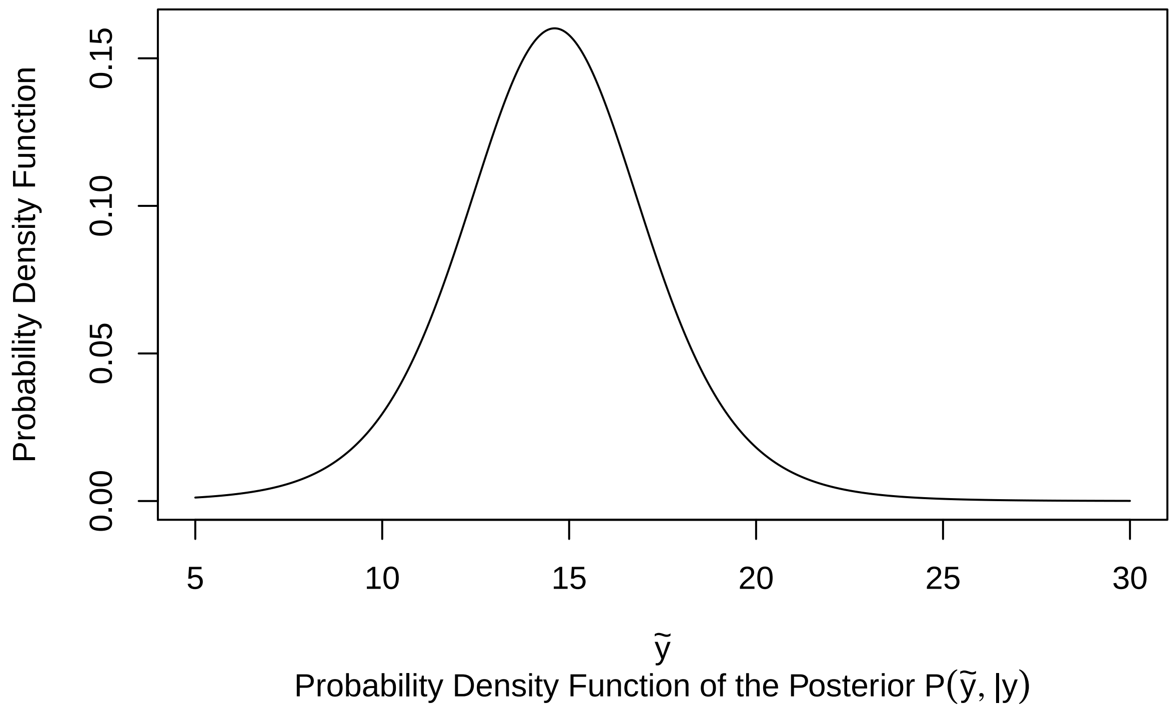
```
## [1] 14.61571
```

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

```
##     2.5%    97.5%
## 11.02843 18.19279
```

So for the value of the hardness of the next windshield coming from the production line, the point estimate is 14.6 and the 95% interval estimate is [11.0,18.2]. And the plot of the density of the posterior is shown below.

```
n_total<-length(windshieldy1)
data_sum<-sum(windshieldy1)
loc<-data_sum/n_total
scale<-var(windshieldy1)*(1+1/n_total)
x<-seq(5,30,len=10000)
y<-dtnew(x,n_total-1,loc,scale)
xlab<-expression(tilde(y))
ylab<-"Probability Density Function"
sub<-expression(paste("Probability Density Function of the Posterior ",paste(P(tilde(y),paste("|",y))))
plot(x,y,'l',xlab=xlab,ylab=ylab,sub=sub)
```



Exercise 2.

a)

My model goes as follows. We have our prior distributions $p_0, p_1 \sim Beta(1,1)$, which are non-informative(uniform prior). the likelihood $P(y_0|p_0) \sim binomial(n_0, p_0) = Binomial(674, 39)$ and $P(y_1|p_1) \sim binomial(n_1, p_1) = Binomial(680, 22)$, and the resulting posteriors $P(p_0|y_0) \sim Beta(1 + y_0; 1 + n_0 - y_0) = Beta(40, 636)$ $P(p_1|y_1) \sim Beta(1 + y_1; 1 + n_1 - y_1) = Beta(23, 659)$.

```
p0_sample<-function(prior_alpha, prior_beta){
n_total<-674
n_death<-39
sample<-rbeta(10000,prior_alpha+n_death, prior_beta+n_total-n_death)
return(sample)
}

p1_sample<-function(prior_alpha, prior_beta){
n_total<-680
n_death<-22
sample<-rbeta(10000,prior_alpha+n_death, prior_beta+n_total-n_death)
return(sample)
}

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

```
}
posterior_odds_ratio_interval<-function(p0,p1,prob=0.9){
est<-(p1/(1-p1))/(p0/(1-p0))
interval<-quantile(est,probs=c((1-prob)/2,1-(1-prob)/2))
return (interval)
}

p0<-p0_sample(1,1)
p1<-p1_sample(1,1)

posterior_odds_ratio_point_est(p0,p1)
```

```
## [1] 0.5687514
```

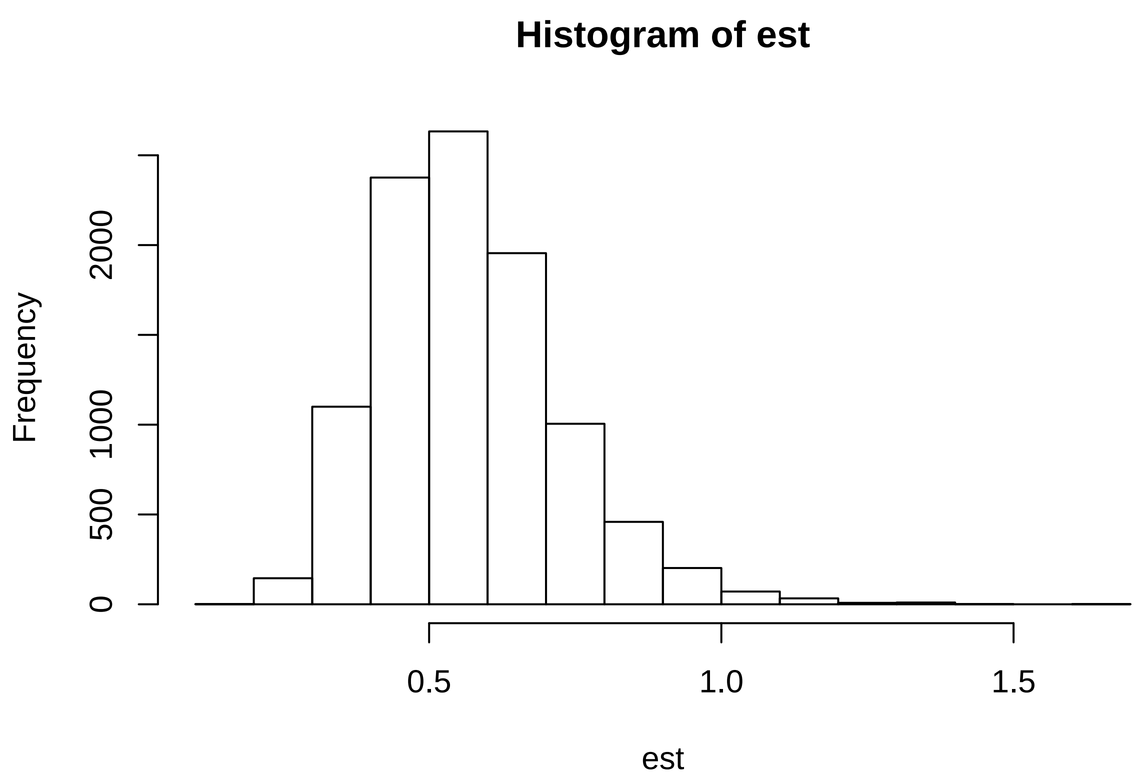
```
posterior_odds_ratio_interval(p0,p1, prob = 0.95)
```

```
##      2.5%     97.5%
## 0.3185748 0.9309333
```

So for the value of the odds ratio $\frac{p_1/(1-p_1)}{p_0/(1-p_0)}$ according to the observations and the prior knowledge, the point estimate is 0.57 and the 95% interval estimate is [0.32,0.93].

Now we plot the histogram of the distribution of the odds ratio.

```
est<-(p1/(1-p1))/(p0/(1-p0))
hist(est)
```

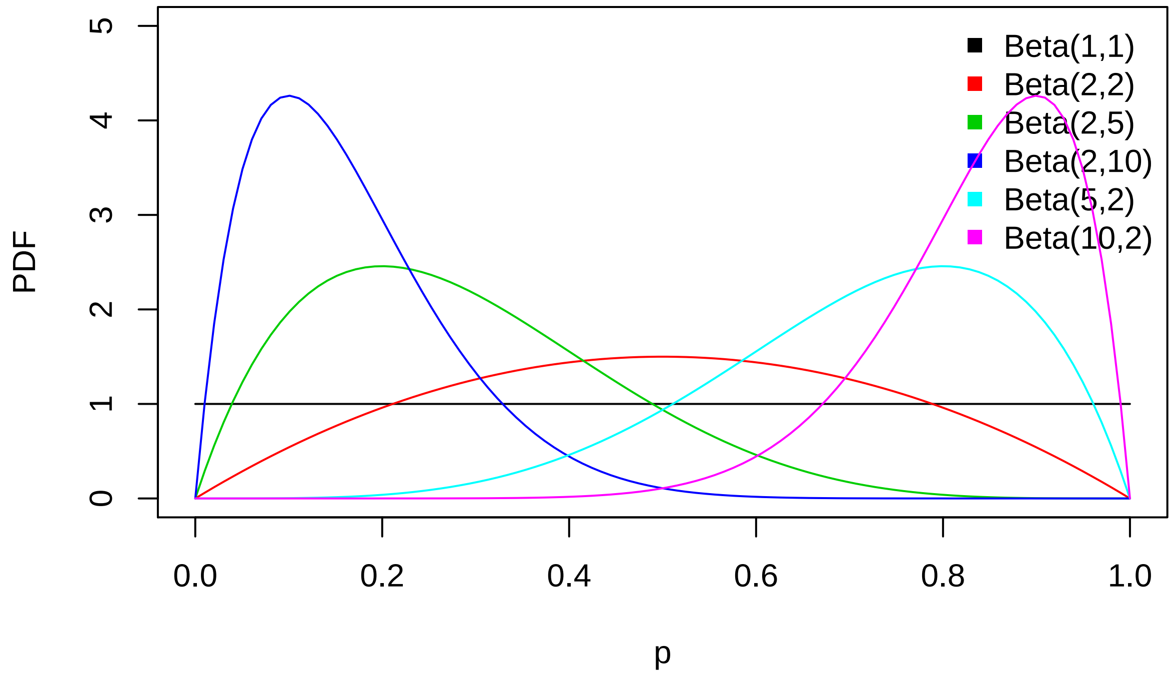


b)

For uniform prior it has been discussed above. Now let's use more priors and take a look at how can they influcence the posterior of the odds ratio.

Let's first visualize the priors we use.

```
# visualize the priors we use
x<-seq(0,1,length=100)
prior1<-dbeta(x,1,1)
prior2<-dbeta(x,2,2)
prior3<-dbeta(x,2,5)
prior4<-dbeta(x,2,10)
prior5<-dbeta(x,5,2)
prior6<-dbeta(x,10,2)
x_l<-"p"
y_l<-"PDF"
st<-"Visualization of priors we use"
plot(x,prior1,col=1,xlab=x_l,sub=st,ylab=y_l,ylim=c(0,5),"l")
legend("topright",pch=c(15,15),legend=c("Beta(1,1)","Beta(2,2)","Beta(2,5)","Beta(2,10)","Beta(5,2)","Be
lines(x,prior2,col=2)
lines(x,prior3,col=3)
lines(x,prior4,col=4)
lines(x,prior5,col=5)
lines(x,prior6,col=6)
```



Visualization of priors we use

Using beta(2,2) as prior

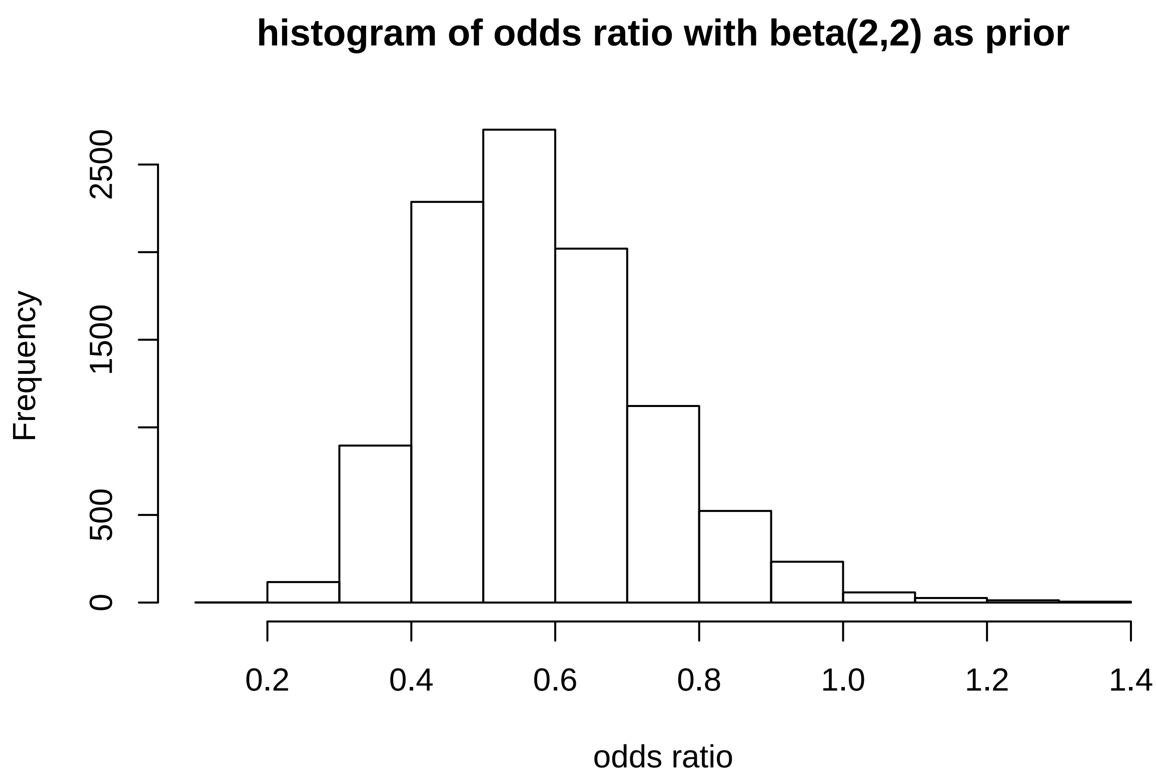
```r
# with beta(2,2) as prior
p0<-p0_sample(2,2)
p1<-p1_sample(2,2)
paste("point estimation of the odds ratio is: ",posterior_odds_ratio_point_est(p0,p1))
```

```
## [1] "point estimation of the odds ratio is:  0.579551717300972"
```

```r
paste("interval estimation of the odds ratio is: ",posterior_odds_ratio_interval(p0,p1, prob = 0.9))
```

```
## [1] "interval estimation of the odds ratio is:  0.363065596869528"
## [2] "interval estimation of the odds ratio is:  0.863142388849291"
```

```r
est<-(p1/(1-p1))/(p0/(1-p0))
x_l<-"odds ratio"
st<-"histogram of odds ratio with beta(2,2) as prior"
hist(est,xlab=x_l,main=st)
```



Using beta(2,5) as prior

```r
# with beta(2,5) as prior
p0<-p0_sample(2,5)
p1<-p1_sample(2,5)
paste("point estimation of the odds ratio is: ",posterior_odds_ratio_point_est(p0,p1))
```

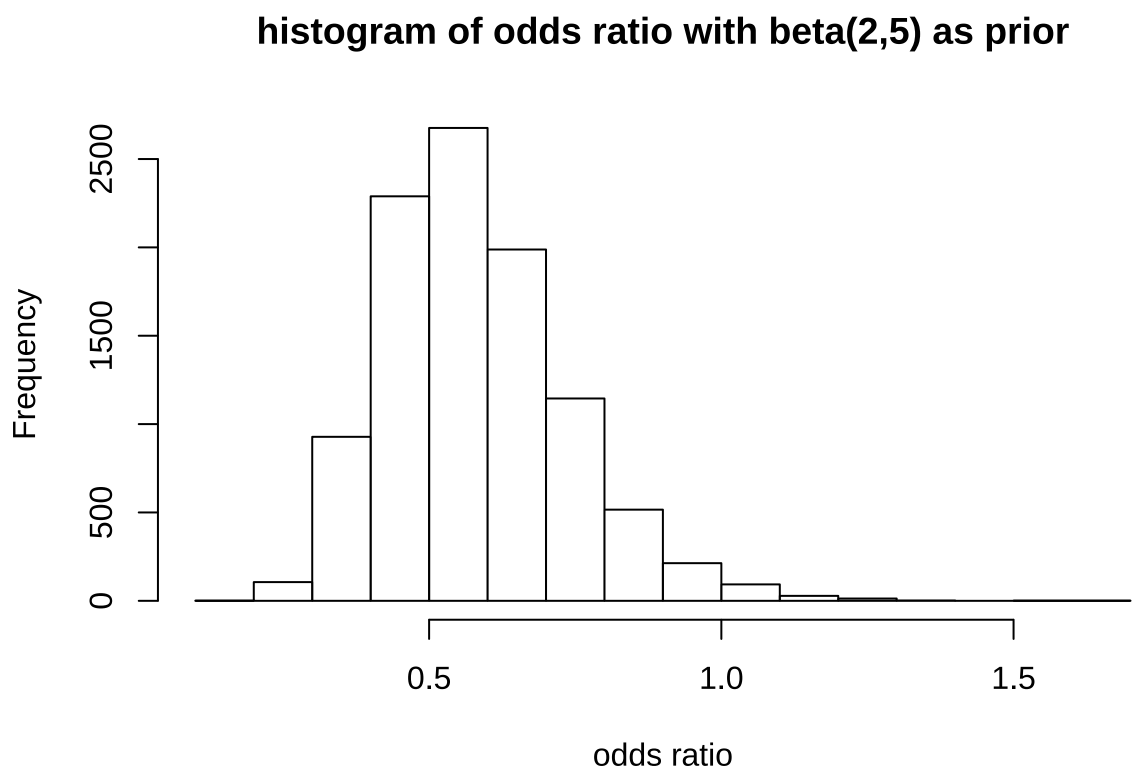
```
## [1] "point estimation of the odds ratio is:  0.579914635389897"
```

```r
paste("interval estimation of the odds ratio is: ",posterior_odds_ratio_interval(p0,p1, prob = 0.9))
```

```
## [1] "interval estimation of the odds ratio is:  0.362602293193583"
```

```
## [2] "interval estimation of the odds ratio is:  0.859484954993169"
```

```r
est<-(p1/(1-p1))/(p0/(1-p0))
x_l<-"odds ratio"
st<-"histogram of odds ratio with beta(2,5) as prior"
hist(est,xlab=x_l,main=st)
```



Using beta(2,10) as prior

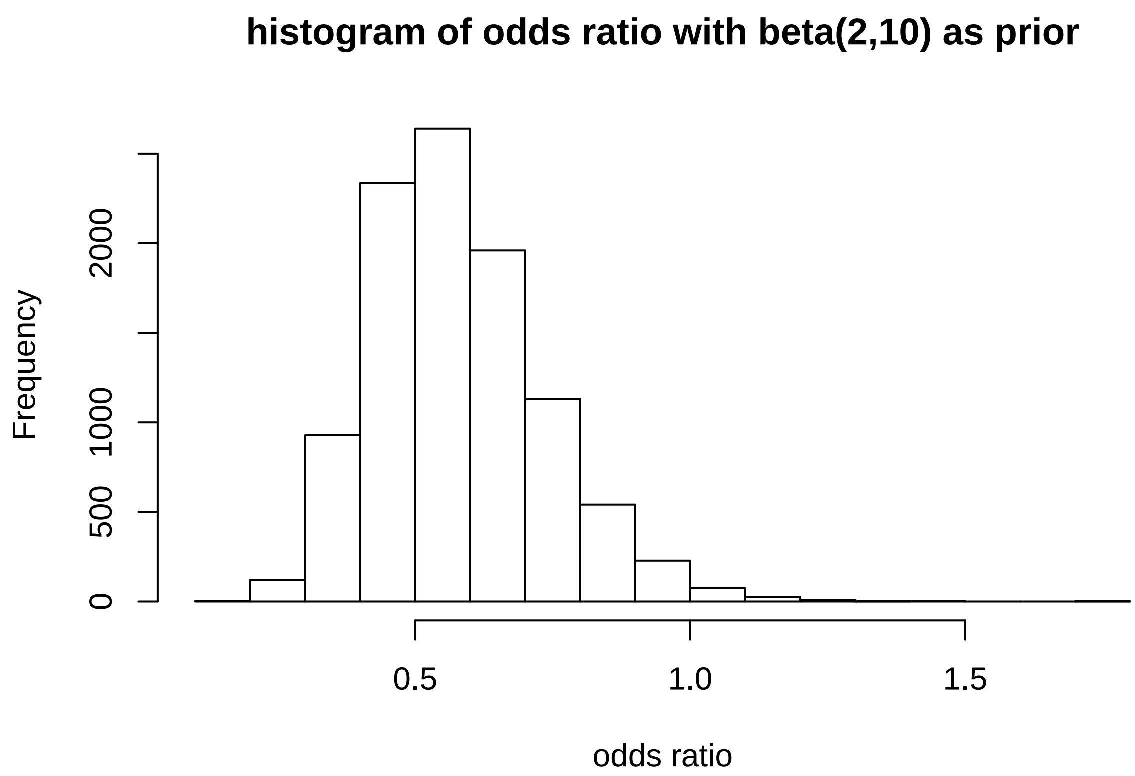
```r
# with beta(2,10) as prior
p0<-p0_sample(2,10)
p1<-p1_sample(2,10)
paste("point estimation of the odds ratio is: ",posterior_odds_ratio_point_est(p0,p1))
```

```
## [1] "point estimation of the odds ratio is:  0.578656024459409"
```

```r
paste("interval estimation of the odds ratio is: ",posterior_odds_ratio_interval(p0,p1, prob = 0.9))
```

```
## [1] "interval estimation of the odds ratio is:  0.35969077747665"
## [2] "interval estimation of the odds ratio is:  0.861638598664006"
```

```r
est<-(p1/(1-p1))/(p0/(1-p0))
x_l<-"odds ratio"
st<-"histogram of odds ratio with beta(2,10) as prior"
hist(est,xlab=x_l,main=st)
```



Using beta(5,2) as prior

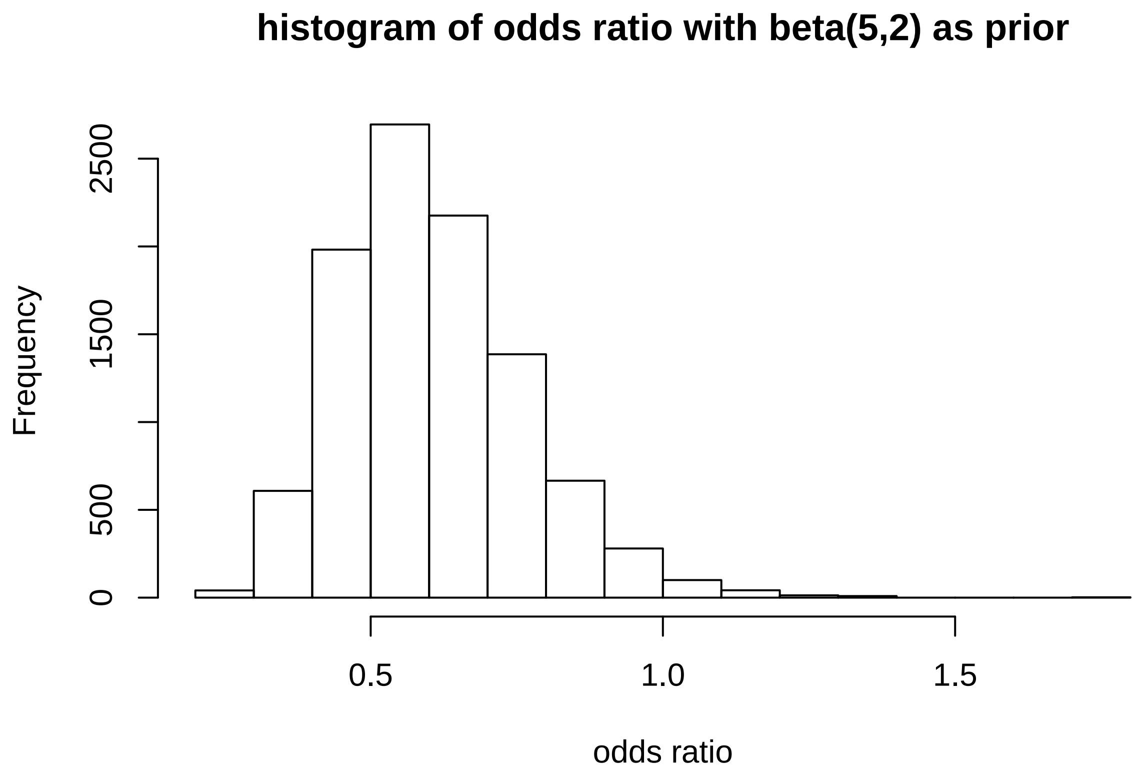
```
# with beta(5,2) as prior
p0<-p0_sample(5,2)
p1<-p1_sample(5,2)
paste("point estimation of the odds ratio is: ",posterior_odds_ratio_point_est(p0,p1))
```

```
## [1] "point estimation of the odds ratio is:  0.607042007614516"
```

```
paste("interval estimation of the odds ratio is: ",posterior_odds_ratio_interval(p0,p1, prob = 0.9))
```

```
## [1] "interval estimation of the odds ratio is:  0.386088097073164"
## [2] "interval estimation of the odds ratio is:  0.887597605565896"
```

```
est<-(p1/(1-p1))/(p0/(1-p0))
x_l<-"odds ratio"
st<-"histogram of odds ratio with beta(5,2) as prior"
hist(est,xlab=x_l,main=st)
```



Using beta(10,2) as prior

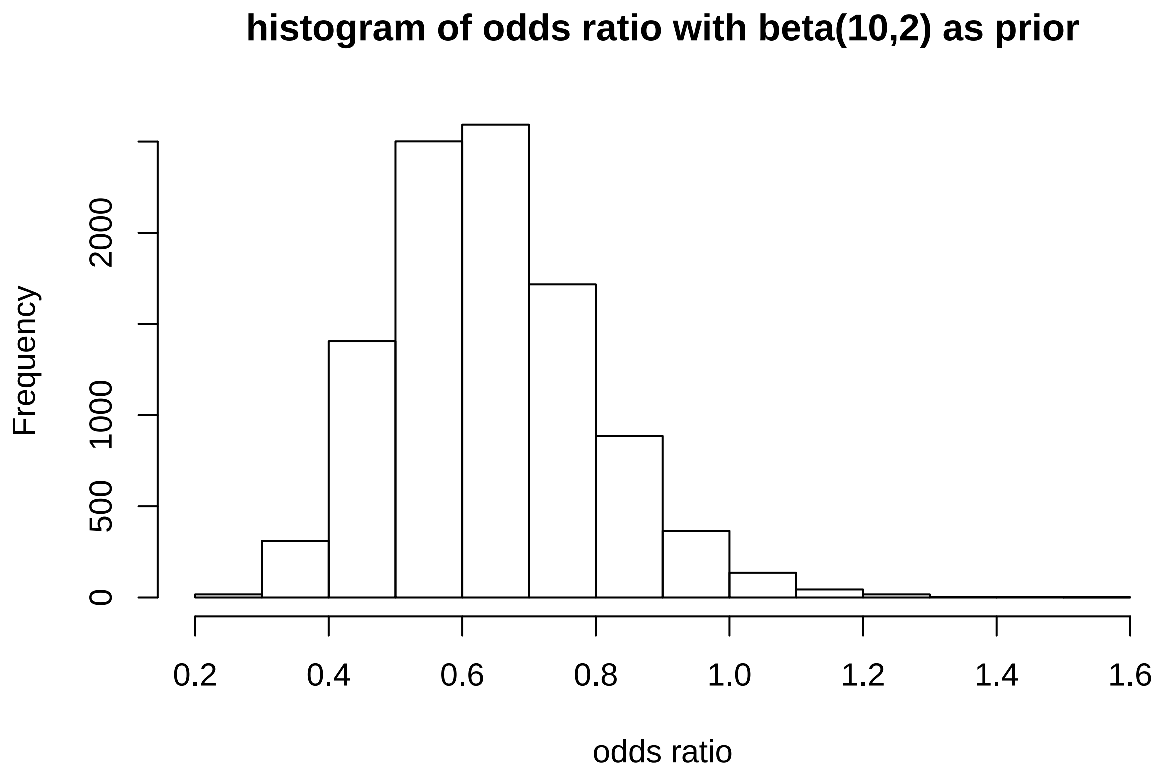
```
# with beta(10,2) as prior
p0<-p0_sample(10,2)
p1<-p1_sample(10,2)
paste("point estimation of the odds ratio is: ",posterior_odds_ratio_point_est(p0,p1))
```

```
## [1] "point estimation of the odds ratio is:  0.642327093422211"
```

```
paste("interval estimation of the odds ratio is: ",posterior_odds_ratio_interval(p0,p1, prob = 0.9))
```

```
## [1] "interval estimation of the odds ratio is:  0.422016557361165"
## [2] "interval estimation of the odds ratio is:  0.911563739981766"
```

```
est<-(p1/(1-p1))/(p0/(1-p0))
x_l<-"odds ratio"
st<-"histogram of odds ratio with beta(10,2) as prior"
hist(est,xlab=x_l,main=st)
```



We can see that even if we use different priors, we still got similar point estimations and interval estimations for the odds ratio without oscillating too much. Thus we can say that the sensitivity of the inference to different choices of prior density is low.

Exercise 3.

a)

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

The model for this question is the same as exercise 1.We have our prior distribution $P(\mu_1, \sigma_1^2) \propto \sigma_1^{-2}$ and $P(\mu_2, \sigma_2^2) \propto \sigma_2^{-2}$, the likelihood $P(\mathbf{y_0}|\mu_0, \sigma_0^2) = (\frac{1}{\sigma_0\sqrt{2\pi}})^n exp(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i^0 - \mu_0)^2)$ and $P(\mathbf{y_1}|\mu_1, \sigma_1^2) = (\frac{1}{\sigma_1\sqrt{2\pi}})^n exp(-\frac{1}{2\sigma_1^2}\sum_{i=1}^n (y_i^1 - \mu_1)^2)$, and the resulting posterior $P(\mu_0, \sigma_0^2|\mathbf{y_0}) \propto {\sigma_0}^{-n-2} exp(-\frac{1}{2\sigma_0^2}[(n-1)s_0^2 + n(\hat{\mathbf{y}}_\mathbf{0} - \mu_0)^2])$ and $P(\mu_1, \sigma_1^2|\mathbf{y_1}) \propto \sigma_1^{-n-2} exp(-\frac{1}{2\sigma_1^2}[(n-1)s_1^2 + n(\hat{\mathbf{y}}_\mathbf{1} - \mu_1)^2])$.
Now if we margin out σ_0 and σ_1 we can get our marginal posterior $P(\mu_0|\mathbf{y_0}) = t_{n-1}(\mu_0|\overline{\mathbf{y}}_\mathbf{0}, s_0^2/n)$ and $P(\mu_1|\mathbf{y_1}) = t_{n-1}(\mu_1|\overline{\mathbf{y}}_\mathbf{1}, s_1^2/n)$

```
mu_sample<-function(data){
n_total<-length(data)
data_sum<-sum(data)
loc<-data_sum/n_total
scale<-var(data)/n_total
sample<-rt(100000,n_total-1)
sample<-sample*sqrt(scale)+loc
return(sample)
```

```
}

u_1<-mu_sample(windshieldy1)
u_2<-mu_sample(windshieldy2)
probs=0.95
paste("the point estimation of mu_d is: ", mean(u_1-u_2))
```

```
## [1] "the point estimation of mu_d is:  -1.206297868708"
```

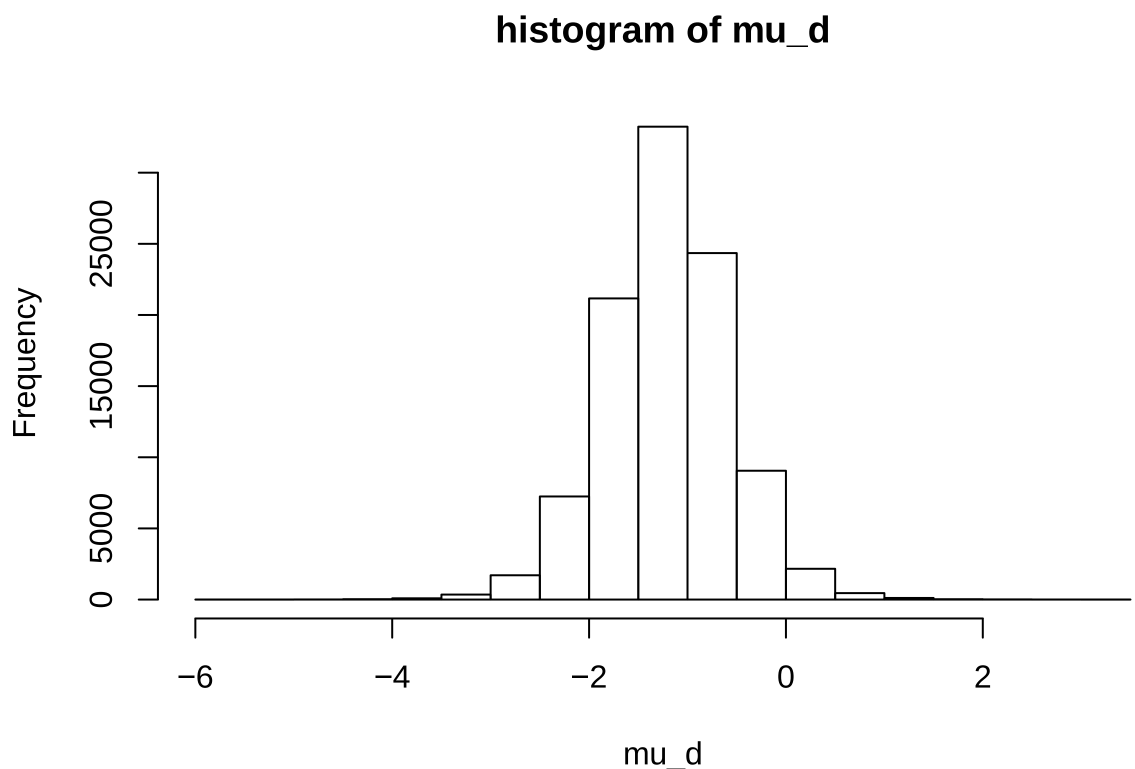
```
print("the interval estimation of mu_d is: ")
```

```
## [1] "the interval estimation of mu_d is: "
```

```
print(quantile(u_1-u_2,probs=c((1-probs)/2,1-(1-probs)/2)))
```

```
##        2.5%       97.5%
## -2.45718124  0.03234614
```

```
hist(u_1-u_2,xlab="mu_d",main="histogram of mu_d")
```



So the point estimation of μ_d is -1.21 and the interval estimation of μ_d is [-2.46,0.03]

b)

No They are not. The point estimation indicate that the mean of μ_d is -1.2 and also the interval esimation shows that at least 95% of μ_d is negative. Consequently we can say that μ_1 is smaller than μ_2 for high probability.
Also we can show this claim by t test. With confidence level $= 0.9$ we can argue that μ_1 is smaller than μ_2.

```r
t.test(windshieldy1, windshieldy2, alternative="two.sided", mu=0,var.equal=FALSE,conf.level = 0.9)
```

```
##
##  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
## 90 percent confidence interval:
##  -2.1868925 -0.2328169
## sample estimates:
## mean of x mean of y
##  14.61122  15.82108
```