BDA - Assignment 3

Anonymous

28 9 2019

Contents

Used libraries:

```r
library(tidyverse)
library(ggplot2)
library(aaltobda)
data("windshieldy1")
data("windshieldy2")
windshieldy_test <- c(13.357, 14.928, 14.896, 14.820)
```

Exercise 1

a) What can you say about the unknown μ? Summarize your results using Bayesian point and interval estimates (95%) and plot the density.

uninformative prior:

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

(one can think of this as improper uniform on (μ, $log(\sigma)$))

likelihood (for single observation):

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

posterior (for n observations):

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

(where s^2 is sample variance)

Since the marginal distribution was derived in the materials already, I am just referencing the results here:

$$\mu|y \sim \mathsf{t}_{\mathsf{n-1}}(\mathsf{\bar{y}}, \frac{\mathsf{s}^2}{\mathsf{n}})$$

which is symmetric around $\overline{y}$ (or zero for un-scaled), meaning it has mean value of $\overline{y}$.

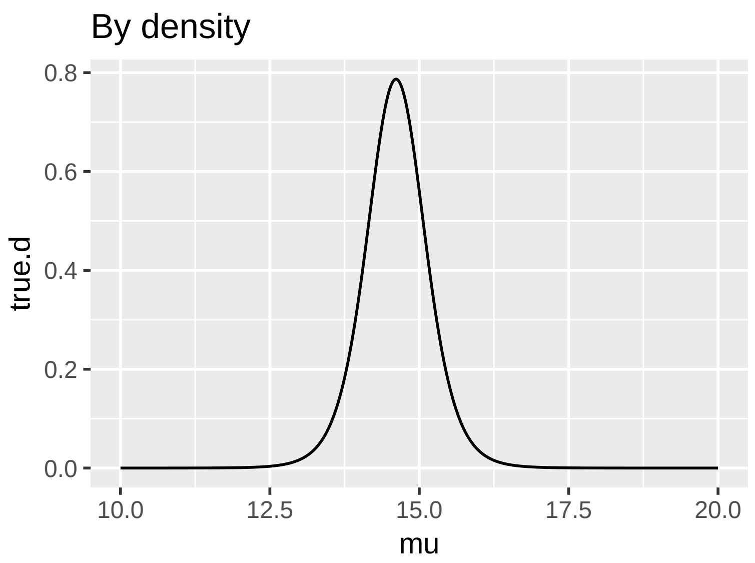
The interval on the other hand can be easily calculated for $\mathsf{t}_{\mathsf{n-1}}(0, 1)$ and shifted by $\overline{y}$ and scaled by $\frac{s}{\sqrt{n}}$.

Here are the functions for the calculations:

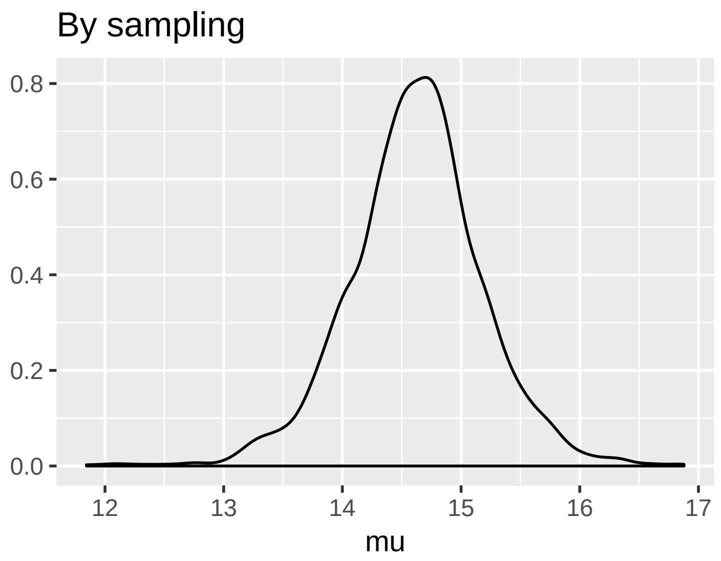
```r
mu_point_est <- function(data= windshieldy_test) {
  mean(data)
}

mu_interval <- function(data = windshieldy_test, prob = 0.95) {

  lo <- (1-prob)/2
  hi <- 1-lo

  #r's var( ...) should work here for sample var
  n <- length(data)
  s2 <- var(data)
  s <- sqrt(s2)

  sc <- s/sqrt(n)
  shift <- mean(data)

  shift + sc * qt(c(lo,hi),df = n-1)
}
```

Here is a function for plotting:

```r
plot_density <- function(data = windshieldy_test, type = 'density', scaling=NULL) {
  n <- length(data)
  s2 <- var(data)
  s <- sqrt(s2)

  sc <- if(is.null(scaling)) s/sqrt(n) else scaling(data)
  stand.t <- rt(2000, df = (n-1))
  true.t <- mean(data) + sc * stand.t
  lab <- if(is.null(scaling)) 'mu' else 'Posterior prediction'

  if(type!='density') {
    qplot(true.t, geom = 'density', xlab = lab, main='By sampling')
  } else {
    true.d <- dtnew(seq(10,20,by = 0.01),mean = mean(data), scale = sc, df = n-1)
    qplot(x=seq(10,20,by = 0.01), true.d, geom = 'line', main='By density', xlab = lab)
  }
}

plot_density(windshieldy1)
```



```
plot_density(windshieldy1, type='')
```



So in short, the **results**:

```
mu_point_est(windshieldy1)
```

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

```
mu_interval(windshieldy1)
```

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

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 and interval estimates (95%) and plot the density.

Assuming a known σ^2, μ can be sampled from $N(\overline{y}, \frac{\sigma^2}{n})$. This of course already implies, that the expected value $E(y|\mu, \sigma)$ has to be $\overline{y}$. Furthermore, since we know the way σ is distributed, i.e., $\sigma^2|y \sim Inv - X^2(n-1, s^2)$, we can sample those and for each sample, construct the corresponding distribution for μ. Using both of these, we can finally get the predictive posterior $p(\tilde{y}|y)$.

However, since we were given the analytic form also for this, I am going to use it (unless I have some time left before the deadline. . .).

$$\tilde{y}|y \sim \mathsf{t}_{\mathsf{n}-1}(\overline{\mathsf{y}}, \mathsf{s}(1 + \frac{1}{\mathsf{n}})^{1/2})$$

Here's a short utility function for the scale argument:

```
t.scale <- function(data) {
  sqrt(var(data))*(1+1/length(data))^.5
}
```

And here are the required functions:

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


mu_pred_interval <- function(data= windshieldy_test, prob = 0.95) {

  n <- length(data)
  lo <- (1-prob)/2
  hi <- 1-lo

  sc <- t.scale(data)

  mean(data) + sc * qt(p = c(lo,hi), df = n-1)

}


#Additional sanity check
mu_pred_interval_2 <- function(data= windshieldy_test, prob = 0.95) {

  n <- length(data)
  lo <- (1-prob)/2
  hi <- 1-lo

  sc <- t.scale(data)

  #I am using random samples here (one could just as well use densities)
  quantile(mean(data) + sc * rt(10000, df = n-1), probs = c(lo,hi))


}
```

Here are the **results**:

```
# Point est
mu_pred_point_est(windshieldy1)
```

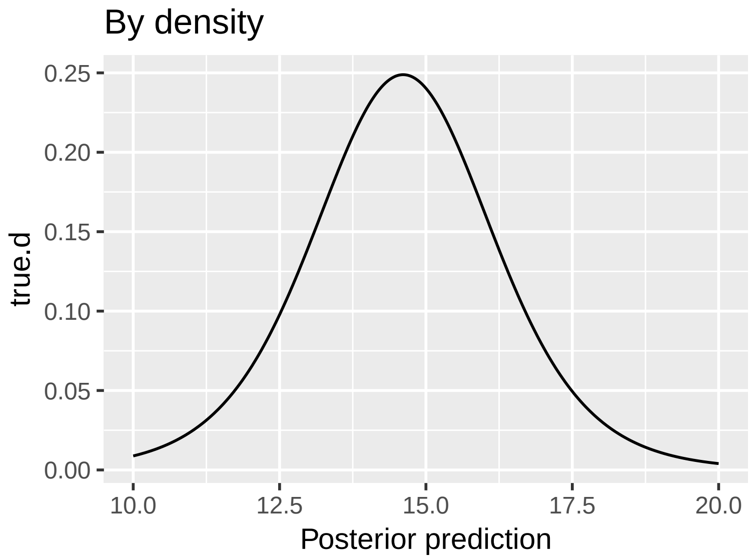
```
## [1] 14.61122
```

```
# Interval:
mu_pred_interval(windshieldy1)
```

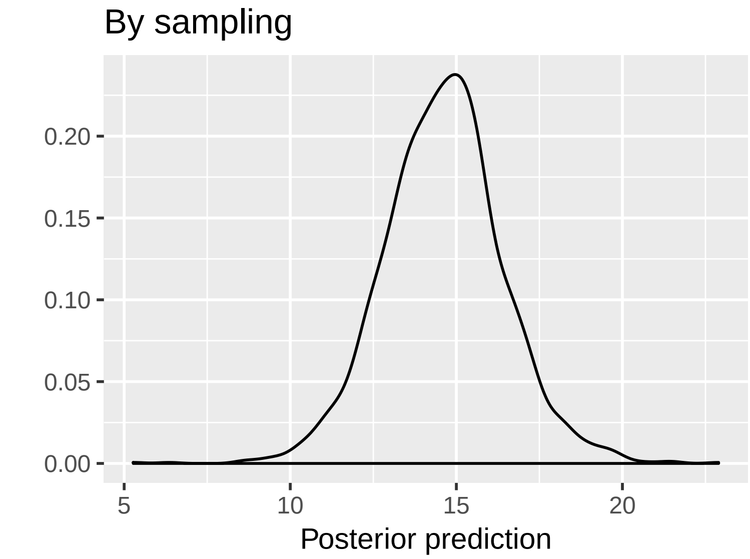
```
## [1] 11.02792 18.19453
```

And here is the plot:

```
plot_density(data=windshieldy1, scaling=t.scale)
```



```
plot_density(data=windshieldy1, type = '', scaling=t.scale)
```



Note: All of the previous t -distributions could have been approximated with simple normal distributions (and with a decent degree of accuracy). However, since the point was to use the "correct" distributions, I did so. Also, I hope I explained in enough detail the reason why mean values could be used directly for some of the results.

Excercise 2

a) Summarize the posterior distribution for the odds ratio, $(p1/(1-p1))/(p0/(1-p0))$ Compute the point and interval estimates (95%) and plot the histogram.

As uninformative **priors**, I am using **Beta(1,1)** (i.e., uniform, same for both groups), and since the patient groups are independent, I can just simply sample both from their posteriors $\mathsf{p_i}|\mathsf{y},\mathsf{n} \sim \mathsf{Beta}(\mathsf{y}+1,\mathsf{n}-\mathsf{y}+1)$, and then use those samples to get the odds ratio (distribution). And as a point estimate, I will use the distribution mean. (**likelihoods** are of course just regular Binomial distributions $y|\mathsf{p_i} \sim \mathsf{Bin}(\mathsf{y}|\mathsf{n},\mathsf{p_i})$- i.e., also Beta's). I am assuming these are clear enough to just use the posteriors directly.

Here is an utility function for sampling (and odds):

```
posterior_sample <- function(y,n,alpha=1, beta=1, sample_size=10000) {

  # default alpha, beta  are the uninformative assumptions wrt. the prior.
  rbeta(sample_size, alpha + y, n-y+beta)

}

odds <- function(p0,p1) { (p1/(1-p1))/(p0/(1-p0))  }
```

Here are the required functions:

```
posterior_odds_ratio_point_est <- function(p0,p1) {
  #p0,p1 are samples from independent posteriors
  mean((p1/(1-p1))/(p0/(1-p0)))
}

posterior_odds_ratio_interval <- function(p0,p1,prob = 0.9) {

  lo <- (1-prob)/2
  hi <- 1-lo
  o_r <- (p1/(1-p1))/(p0/(1-p0))
  quantile(o_r,probs = c(lo,hi))

}
```

Test:

```
set.seed(4711)
p0 <- rbeta(100000, 5, 95)
p1 <- rbeta(100000, 10, 90)
posterior_odds_ratio_point_est(p0,p1)
```

```
## [1] 2.676146
```

```
posterior_odds_ratio_interval(p0,p1)
```

```
##       5%      95%
## 0.875367 6.059110
```

And then with the given values:

```
#control:
nc <- 674
yc <- 39
p0 <- posterior_sample(yc, nc)
#treatment:
nt <- 680
yt <- 22
p1 <- posterior_sample(yt, nt)
```

The **results**; point estimate:

```
posterior_odds_ratio_point_est(p0,p1)
```

```
## [1] 0.5664933
```

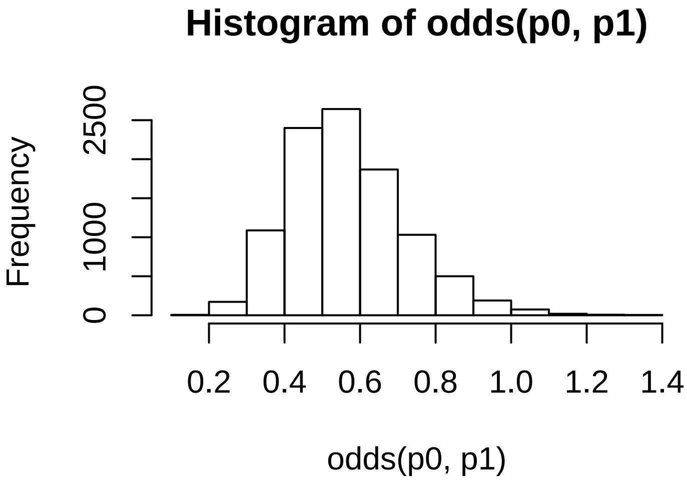
Interval estimate:

```
posterior_odds_ratio_interval(p0,p1,prob=.95)
```

```
##      2.5%     97.5%
## 0.3150374 0.9148377
```

And the histogram:

```
hist(odds(p0,p1))
```



Based on these, it seems reasonable to assume that the treatment has an effect on the deathrate of the patients (on average, treatment halves the deathrate).

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

In simple terms (i.e., couple of sentences), posterior odds are equal to prior odds ratio *times* likelihood ratio. So, given the assumption of set data (~constant), and similar group sizes for treatment and control, the prior odds ratio can shift posterior odds significantly. That being said, when using odds ratio for things like drug efficacy evaluation, it is not obvious why one would allow the priors to differ significantly (rather just assume a single distribution and independent samples from it - in which case prior odds would naturally be 1, i.e., no effect). In my results, the odds ratio is fully explained with the likelihood ratio.

Excercise 3

Consider a case where the same factory has two production lines for manufacturing car windshields. Independent samples from the two production lines were tested for hardness. The hardness measurements for the two samples y1 and y2 are given in the files windshieldy1.txt and windshieldy2.txt.

We assume that the samples have unknown standard deviations σ_1 and σ_2. Use uninformative or weakly informative priors and answer the following questions:

a) What can you say about $\mu_{\mathsf{d}} = \mu_1 - \mu_2$? Summarize your results using Bayesian point and interval estimates (95%) and plot the histogram.

For both $\mu's$, the following holds independently.

Prior(s):

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

likelihood(s) (for single observation):

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

(naturally we take a product of multiple of this form)

posterior (for n observations):

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

(where s^2 is sample variance)

From the posterior, one can get the marginal distribution for σ^2 by integrating over μ. We get a closed form:

$$\sigma^2|y \sim Inv - X^2(n-1, s^2)$$

and using samples from it, we can get the conditional(s):

$$\mu|\sigma^2, y \propto N(\overline{y}, \sigma^2/n)$$

This is the distribution, from which one should sample both μ_1, μ_2, and construct $\mu_{\mathsf{d}} = \mu_1 - \mu_2$. And from the resulting distribution, it is simple to get the summary statistics.

However (again…), since we were given the analytic form also for these, I am going to use them. Both $\mu's$ can be directly sampled from marginal posteriors:

$$\mu|y \sim \mathsf{t}_{\mathsf{n}-1}(\overline{\mathsf{y}}, \mathsf{s}^2/\mathsf{n})$$

Here is a function for the individual samples for $\mu's$:

```
  #I am using random samples rather than densities dt( ...)
mu_sample <- function(data, sample_size = 10000) {
  n <- length(data)
  s2 <- var(data)

  sc <- sqrt(s2/n)
  shift <- mean(data)

  shift + sc * rt(sample_size,df = n-1)
}
```

And then a function that constructs the combined distribution:

```
dif_sample <- function(data1, data2, sample_size = 10000) {

  mu_sample(data1, sample_size) - mu_sample(data2, sample_size)

}
```

And here are the results:

```
#     point estimate:
mean(dif_sample(windshieldy1, windshieldy2))
```

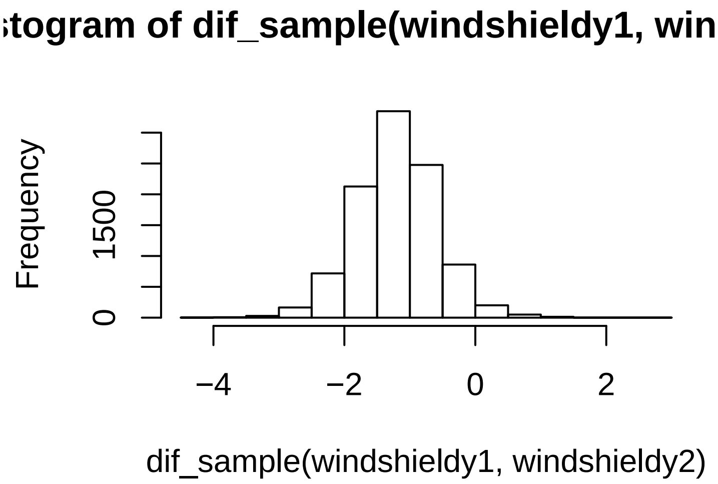
```
## [1] -1.217705
```

NOTE: This is a best estimate for the difference in μ_1(smaller) and μ_2. But this does not mean, that the analysis would necessarily conclude the means different (that is, significantly different). See the b) -part below.

```
#     interval:
quantile(dif_sample(windshieldy1, windshieldy2, 100000), probs = c(0.025,0.975))
```

```
##        2.5%       97.5%
## -2.44744278  0.02608835
```

```
#     histogram:
hist(dif_sample(windshieldy1, windshieldy2))
```



b) Are the means the same? Explain your reasoning with a couple of sentences.

This is also continuation to the a -part ("what can you say about . . . ").

If one should assert a hypothesis $H : \mu_{\mathrm{d}} = 0, (\text{i.e.}, \mu_1 = \mu_2)$, the results would *not* give enough evidence to discard it: Clearly, $\mu_{\mathrm{d}} = 0$ is still within the 95% confidence.

Just for comparison, with 90% confidence interval:

```
quantile(dif_sample(windshieldy1, windshieldy2), probs = c(0.1,0.9))
```

```
##         10%         90%
## -1.9705809 -0.4354238
```

and now one would have to discard the hypothesis (i.e., conclude that the means might actually be different).