EDA A4 Q3

Imagine that a virulent virus has led to a world wide pandemic and that the case fatality rate (proportion of those infected who die) is huge (say 6%) in some age group.

Suppose that through a concerted and collaborative effort of health scientists worldwide, three different treatments have been developed for this group. All three treatments have been used at one time or another on numerous patients in this group from 100 different cities worldwide.

We imagine that recovery rates (as a percent) for the patients treated by each of the three treatments were recorded for each of the hundred cities and are available for analysis as the R data frame `pandemic`.

```
# Either
#load(path_concat(dataDirectory, "pandemic.rda"))
# or
pandemic <- read.csv("/Users/rudranibhadra/Downloads/pandemic.csv")
#pandemic
```

Note again that this data is **not real** and city names are attached just to make it look more **realistic**.

a. One way to determine whether a given treatment is better than doing nothing is to test the hypothesis that its mean recovery rate is the same as the that doing nothing.

In R, there is a handy function called `t.test()` that can be used to test such hypotheses.

i. *(7 marks)* Using this function, test this hypothesis for each of the three treatments (`A`, `B`, and `C`) **individually**.

- Explain your choice of arguments to `t.test()`
- Report any conclusions you draw from each test (with supporting evidence).
- What do you conclude about the three treatments based on these tests?

```
a<-pandemic[,c(2)]
b<-pandemic[,c(3)]
c<-pandemic[,c(4)]
n<-nrow(pandemic)
#z<-rep(94,n)
t.test(a,mu=94)
```

```
##
##  One Sample t-test
##
## data:  a
## t = 7.7033, df = 99, p-value = 1.031e-11
## alternative hypothesis: true mean is not equal to 94
## 95 percent confidence interval:
##  94.51227 94.86773
## sample estimates:
## mean of x
##     94.69
```

```
t.test(b,mu=94)
```

```
##
##  One Sample t-test
##
## data:  b
```

```
## t = 2.306, df = 99, p-value = 0.0232
## alternative hypothesis: true mean is not equal to 94
## 95 percent confidence interval:
##  94.10646 95.41954
## sample estimates:
## mean of x
##    94.763
```

```
t.test(c,mu=94)
```

```
##
##  One Sample t-test
##
## data:  c
## t = 4.8342, df = 99, p-value = 4.897e-06
## alternative hypothesis: true mean is not equal to 94
## 95 percent confidence interval:
##  94.90024 96.15376
## sample estimates:
## mean of x
##    95.527
```

Each of the columns corresponding to the three treatments along with mu=94 is passed in the `t.test` function. mu=94 since the mean recovery rate of doing nothing is 100-fatality rate=100-6=94.

From each test we see that the true mean of any of the treatments is not equal to 94 as they are all greater than 94 and the p values for all three tests are smaller than 0.05 and so evidence against the null hypothesis is strong.

Therefore, any of the treatments A, B or C is better than doing nothing.

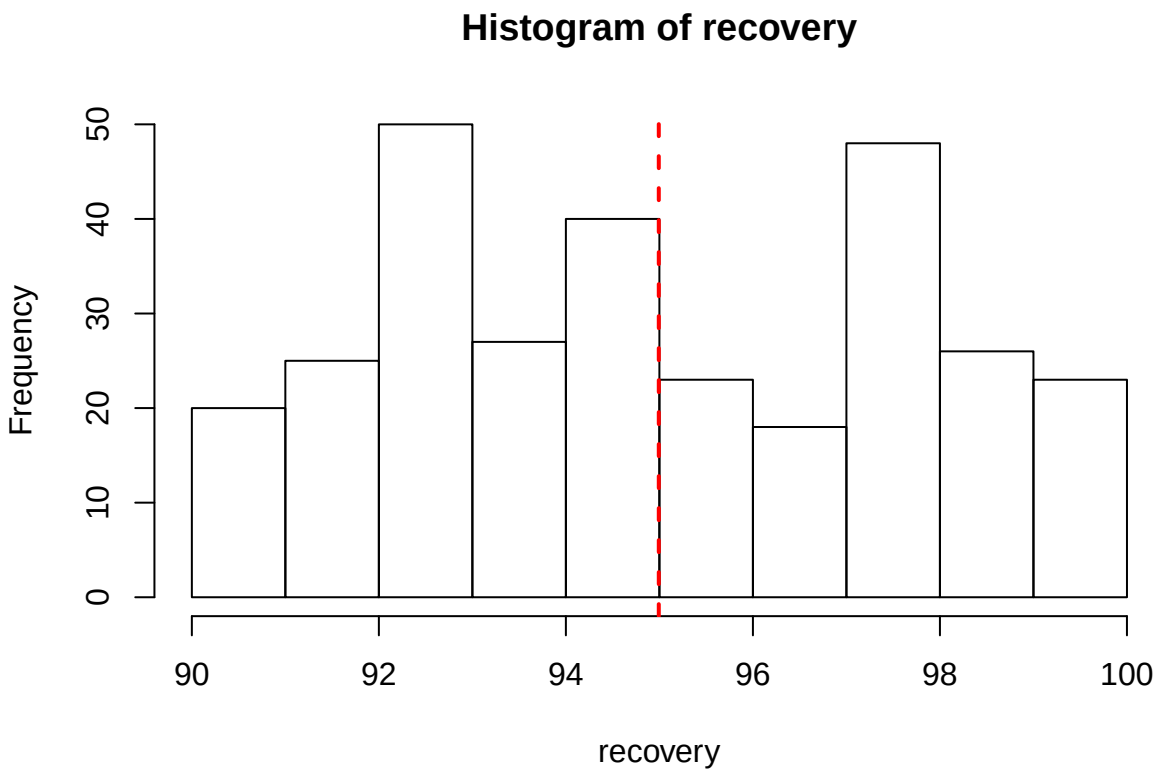
ii. *(3 marks)* Among those tests you think were statistically significant above, which, if any, would you consider to be **practically** significant? Why or why not?

I think the one sample test for treatment can be practically significant since its mean recovery rate is the highest out of all three and higher than 94 and that its p value is small enough to give evidence against the null hypothesis.

iii. *(5 marks)* Construct a vector called `recovery` containing all `300` recovery rates (i.e. all three treatments together).

- Show your code.
- Draw a **nicely labelled** histogram of the recovery rates. Add a red dashed vertical line at the average recovery rate.
- Using `t.test()`, examine whether using any of these treatments would be better than doing nothing.

```
recovery<-as.vector(as.matrix(pandemic[,c("A", "B", "C")]))
#recovery
hist(recovery)
abline(v=mean(recovery),lty=2,lwd=2,col="red")
```




```
n<-length(recovery)
#z<-rep(94,n)
t.test(recovery,mu=94)
```

```
##
##  One Sample t-test
##
## data:  recovery
## t = 6.3514, df = 299, p-value = 7.926e-10
## alternative hypothesis: true mean is not equal to 94
## 95 percent confidence interval:
##  94.68555 95.30111
## sample estimates:
## mean of x
##  94.99333
```

From `t.test()`, using any of these treatments is better than doing nothing since the total mean is greater than 94 since the p values is extremely small enough to reject the null hypothesis.

iv. *(2 marks)* What more would you like to know about the recovery rate without treatment in order to draw your conclusions?

I would like to know if the recovery rate without treatment is the same or different for all age groups and in different cities.

b. Given three treatments to choose from (viz. `A`, `B`, and `C`), it is of interest to know which of these is best overall. One population attribute of interest might be the number of cities in which treatment `A` has a higher recovery rate than treatment `B`, et cetera.

c. *(2 marks)* Determine the fraction of cities which have higher recovery rates for `A` than for `B`.

- Show your code and the resulting proportion.
- Which of the two treatments does this proportion suggest should be preferred? Why?

```
x1<-length(pandemic[pandemic$A>pandemic$B,1])/nrow(pandemic)
x1
```

```
## [1] 0.66
```

Based on the result above, treatment A should be preferred as 66% cities (more than 50%) out of all have higher recovery rates for `A` than for `B`.

ii. *(2 marks)* Repeat the above question but not to determine the proportion of cities which have higher recovery rates for `B` than for `C`.

```
x1<-length(pandemic[pandemic$B>pandemic$C,1])/nrow(pandemic)
x1
```

```
## [1] 0.59
```

Based on the result above, treatment B should be preferred as 59% (more than 50%) cities out of all have higher recovery rates for `B` than for `C`.

iii. *(2 marks)* Repeat the above question but not to determine the proportion of cities which have higher recovery rates `C` than for `A`.

```
x1<-length(pandemic[pandemic$C>pandemic$A,1])/nrow(pandemic)
x1
```

```
## [1] 0.7
```

Based on the result above, treatment C should be preferred as 70% (more than 50%) cities out of all have higher recovery rates for `C` than for `A`.

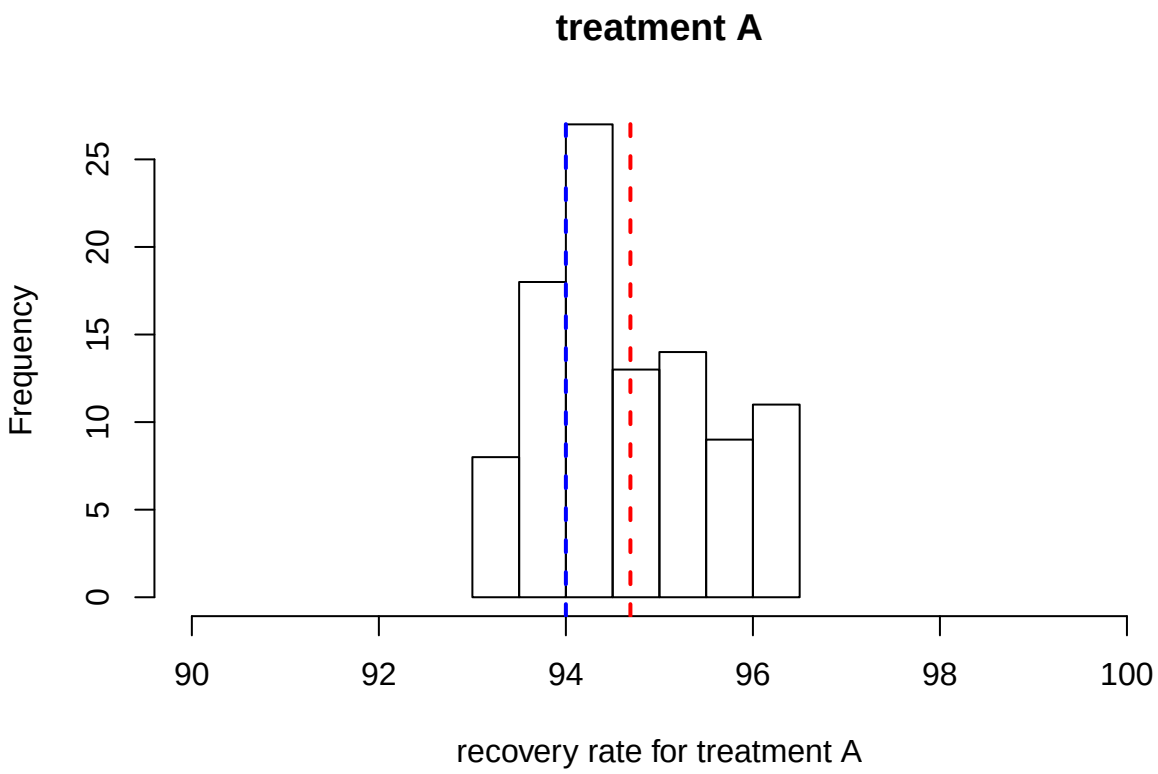
iv. *(2 marks)* Given your answers to each of the previous three questions, what do you conclude about which of the three treatments should be preferred? Why?

According to the results above, I think treatment C should be preferred because the proportion of cities having sucees with C over A is huge (70%). Even though the proportion of cities having success with B is higher than C, proportion of cities having success with A higher than B is also large (66%). And since C is greater than A, C should be preferred.

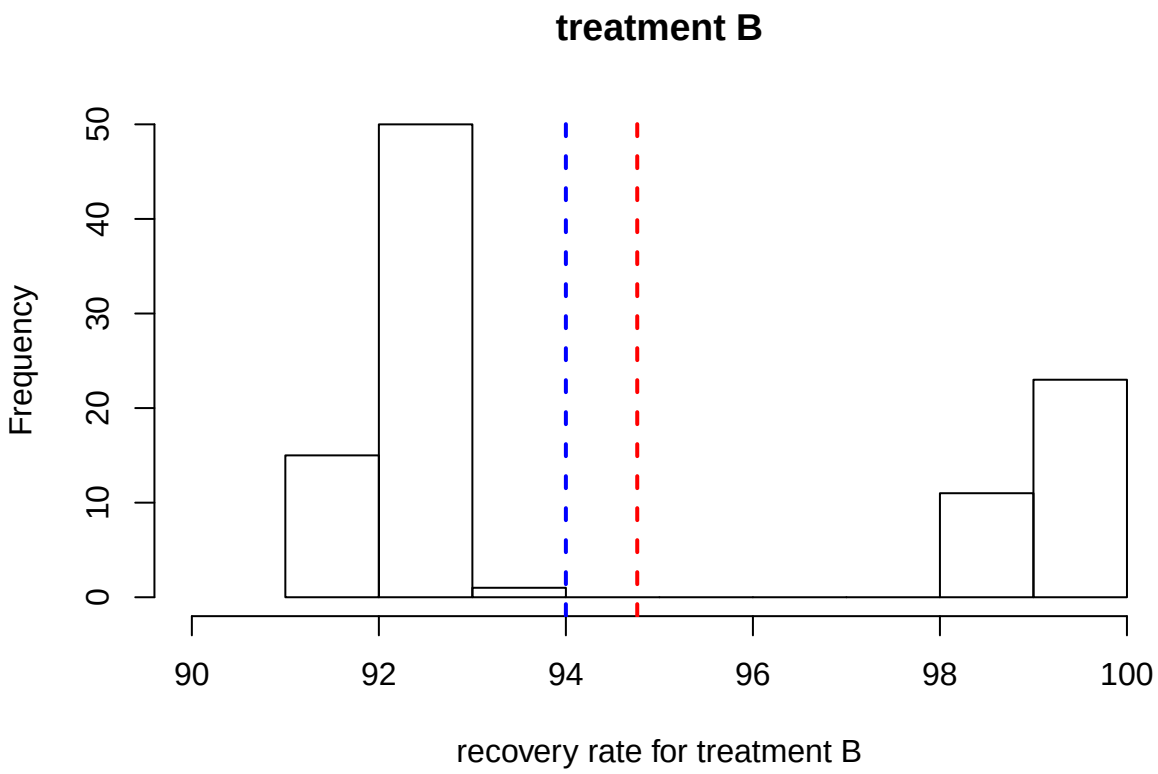
c. *(6 marks)* For each treatment draw a histogram of its recovery rates.

- Show your code.
- Use a common `xlim` in all three histograms.
- Label each histogram meaningfully.
- Add a red vertical dashed line at the average for recovery rate for that treatment **and** a blue one for the recovery rate if nothing is done.

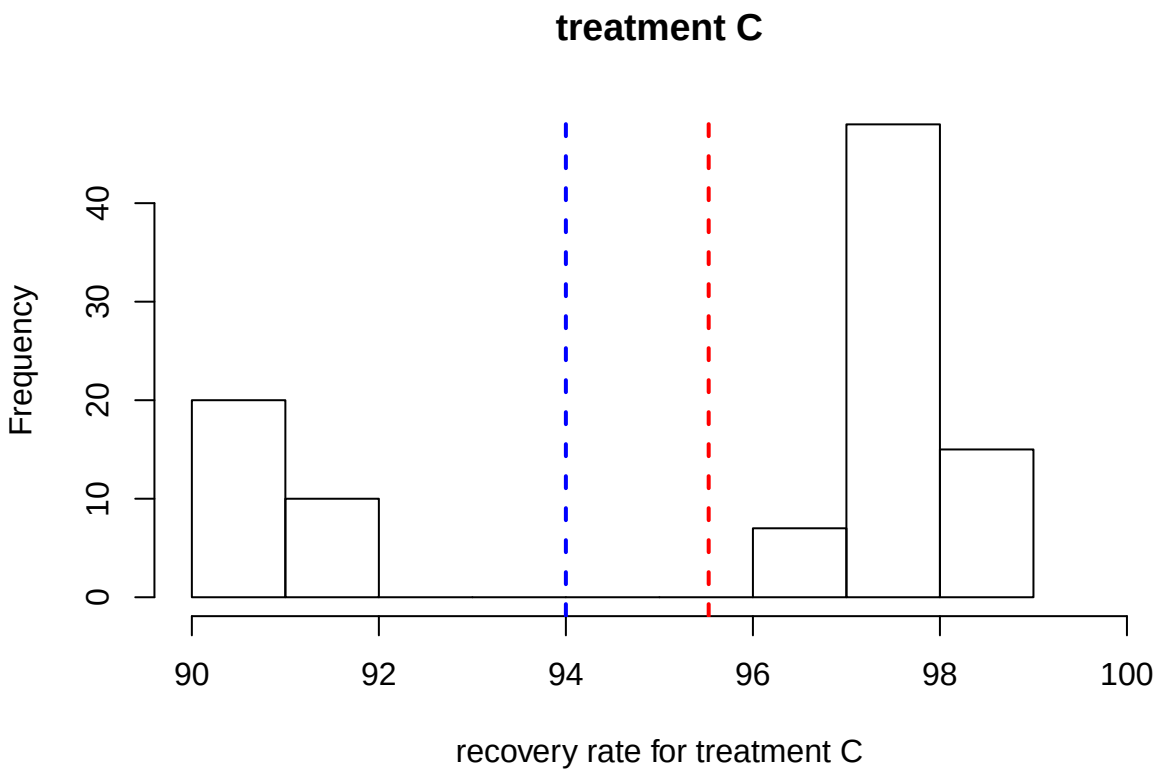
```
h1<-hist(pandemic$A, main="treatment A",xlim=c(90,100),xlab='recovery rate for treatment A')
abline(v=mean(pandemic$A),lty=2,lwd=2,col="red")
abline(v=94,lty=2,lwd=2,col="blue")
```




```
h1<-hist(pandemic$B, main="treatment B",xlim=c(90,100),xlab='recovery rate for treatment B')
abline(v=mean(pandemic$B),lty=2,lwd=2,col="red")
abline(v=94,lty=2,lwd=2,col="blue")
```




```
h1<-hist(pandemic$C, main="treatment C",xlim=c(90,100),xlab='recovery rate for treatment C')
abline(v=mean(pandemic$C),lty=2,lwd=2,col="red")
abline(v=94,lty=2,lwd=2,col="blue")
```



d. *(5 marks)* Based on all of your findings above, what recommendations would you report to health scientists worldwide? What would you urge them to do? Explain your reasoning. Write in language that a knowledgeable person unfamiliar with statistics might understand.

Based on the findings above, I would recommend that it is definitely better to give either of these treatments rather then doing nothing since the death rate of the disease is pretty high. Even if treatment C has the highest mean out of all three treatments, for some cities other treatments have worked better hence the treatment should be given accordingly to the recovery rate in each of those cities. More data should be collected according to gender and age groups as well in each city to see which treatment is most effective for each group as the data given is too general.