Pandemic

An artificial data set

---

38 marks

See the pandemic background file for context.

Assuming you have downloaded the data files into a directory called `dataDirectory`, you can create the `data.frame` called `trtPan` in either of the following two ways:

```
# Either
#load(file.path(dataDirectory, "trtPan.rda"))
# or
trtPan <- read.csv(file.path(dataDirectory, "trtPan.csv"))
```

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

a. **(10 marks)** To begin, ignore any difference between treatments and interpret all recovery rates in `trtPan` as if they were from a single treatment.

   i. *(3 marks)* Based on the recovery rates in `trtPan`, use `t.test()` to test the hypothesis that the mean treatment recovery rate is the same as that for untreated patients (versus that it is greater).

   On the basis of these results, would you recommend treatment? Why?

   Show your code.

```
mean_treatment_rate <- mean(trtPan$Recovery)
t.test(trtPan$Recovery, mu=96, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  trtPan$Recovery
## t = 18.136, df = 299, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 96
## 95 percent confidence interval:
##  97.31444      Inf
## sample estimates:
## mean of x
##    97.446
```

- Yes, I would recommend the treatment because after testing with t.test(), we got a evidence against the hypothesis that the mean treatment recovery rate is the same as that for untreated patients. This indicates that by taking a treatment, your recovery rate is likely to be greater than untreated patients.

   i. *(5 marks)* Create an interactive histogram of the recovery rates.

   The histogram should:

   – be assigned to the variable `h_recovery`
   – have `linkingGroup = "pandemic"`
   – show scales on the axes

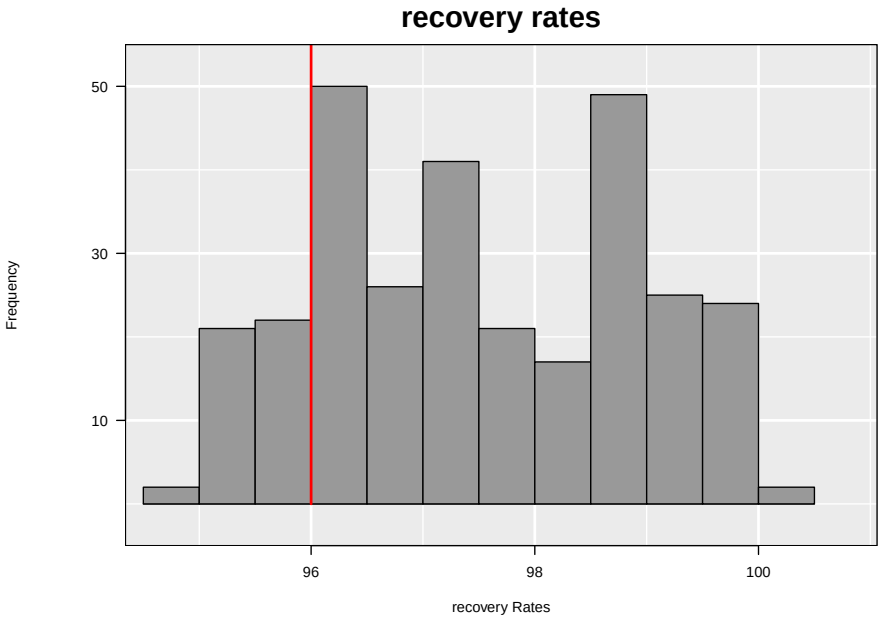
- have `binwidth` of 0.5
- have `origin` 0
- have a vertical line at the recovery rate for **untreated** patients
  - the line should be red
  - the line should have width 2
  - use `l_layer_line()`

Show your code **and** display the plot.

```
h_recovery <- l_hist(trtPan$Recovery, binwidth = 0.5, origin = 0, showScales = TRUE ,linkingGroup = "pa
l_layer_line(widget = h_recovery, x=c(96,96),y = c(0,500), linewidth = 2, color = "red")
```

```
## loon layer "line" of type line of plot .l0.hist
## [1] "layer0"
```

```
plot(h_recovery)
```



i. *(2 marks)* Determine the proportion of recovery rates greater than the untreated recovery rate.

Based only on this value, would you recommend treatment? Why?

Show your code.

```
num <-trtPan$Recovery > 96
table_greater <- table(num)
table_greater[[2]]/length(trtPan)
```

```
## [1] 83.33333
```

- By calculating the proportion of the cases that have recovery rate greater than 96. There is 83.3 percent chance of getting a better recovery rate if one has taken treatment. I would recommend the treatment because it significantly increases the chance of getting healed. Also, the worst case in the trtPan dataset shows that taking a treatment is not likely to have negative effect on the recovery rate.

b. **(8 marks)** Now consider the recovery rates for each treatment.

i. *(2 marks)* Determine the proportion of recovery rates greater than the untreated recovery rate for:

- treatment "A" only

- treatment “B” only
- treatment “C” only

Show your code.

```
treatA <- trtPan[trtPan$Treatment == "A", ]
treatA_96 <- treatA[treatA$Recovery > 96, ]
nrow(treatA_96)/nrow(treatA)
```

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

```
treatB <- trtPan[trtPan$Treatment == "B", ]
treatB_96 <- treatB[treatB$Recovery > 96, ]
nrow(treatB_96)/nrow(treatB)
```

```
## [1] 0.8
```

```
treatC <- trtPan[trtPan$Treatment == "C", ]
treatC_96 <- treatC[treatC$Recovery > 96, ]
nrow(treatC_96)/nrow(treatC)
```

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

i. *(1 mark)* Based only on the above values, order the treatments from best to worst.

- A -> B -> C

i. *(2 marks)* Determine the average recovery rate for each treatment.

Show your code.

```
#A
mean(treatA$Recovery)
```

```
## [1] 97.312
```

```
#B
mean(treatB$Recovery)
```

```
## [1] 97.321
```

```
#C
mean(treatC$Recovery)
```

```
## [1] 97.705
```

i. *(1 mark)* Based only on the above values, order the treatments from best to worst.

- C -> B -> A

i. *(2 marks)* Given the calculations only in part (b) above, which of the three treatments would you recommend? Why?

- Given the calculations in part (b), I would recommend treatment A because of two reasons. After calculation, we know that it is 100 percent to get better recovery rate than untreated recovery rate. This could indicate that the treatment itself is stable and well-developed. Even though treatment C has the highest average recovery rate, but it the average recovery rate of C is only 0.4 more than treatment A and there is 30 percent that the treatment does not increase the recovery rate.

c. **(14 marks)** Recall that recovery rates are available for all treatments in **every** city.

i. *(1 mark)* Why might it be a good idea to compare treatment recovery rates **within** each city?

- This may be a good idea because different city is different in many ways, and could have impact on the treatment recovery rates.

i. *(4 marks)* To compare treatments **within** city, it will be convenient to construct a new data frame, called `recovery`, that has the following characteristics:

   - it is assigned to `recovery`
   - it contains 100 rows, one for each unique city
   - has 4 variables `City`, `A`, `B`, and `C` in that order with values:
     * `City`, containing the name of the city for that row
     * `A`, containing the recovery rate for treatment `A` in that row's city
     * `B`, containing the recovery rate for treatment `B` in that row's city
     * `C`, containing the recovery rate for treatment `C` in that row's city
   - the rows should be sorted by `City` in alphabetical order.

   Construct the above data frame.

   Show your code.

   Show the results of

   - `head(recovery, 2)` and
   - `recovery[recovery$City == "Toronto", ]`.

```
joinAB <- merge(treatA, treatB, by = 'City', all=TRUE)
recovery <- merge(joinAB, treatC, by = 'City', all=TRUE)
recovery$Treatment.x <- NULL
recovery$Treatment.y <- NULL
recovery$Treatment <- NULL
colnames(recovery) <- c("City", "A", "B","C")

head(recovery, 2)
```

```
##        City    A    B    C
## 1   Abidjan 97.2 96.4 98.5
## 2 Ahmadabad 97.6 96.1 98.5
```

```
recovery[recovery$City == "Toronto", ]
```

```
##       City    A    B    C
## 94 Toronto 97.3 99.9 98.7
```

i. *(2 marks)* Determine the fraction of cities whose recovery rate for `A` is greater than that for `B`.

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

```
A_greater_B <- recovery$A > recovery$B
tableA_B <- table(A_greater_B)
tableA_B[[2]]/nrow(recovery)
```

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

- After calculating the fraction of cities whose recovery rate for A is greater than B, it shows that there are 66 percent chance that the recovery of treatment A is greater than treatment B. As a result, this proportion suggests that treatment A is preferred over treatment B because it has a higher chance of getting a higher recovery rate.

i. *(2 marks)* Determine the fraction of cities whose recovery rate for `B` is greater than that for `C`.

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

```
B_greater_C <- recovery$B > recovery$C
tableB_C <- table(B_greater_C)
tableB_C[[2]]/nrow(recovery)
```

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

- After calculating the fraction of cities whose recovery rate for B is greater than C, it shows that there are 59 percent chance that the recovery of treatment B is greater than treatment C. As a result, we conclude that this proportion suggest that treatment B is preferred over treatment B because it has a higher chance of getting a higher recovery rate.

    i. *(1 mark)* Based only on the two values just calculated, order the treatments from best to worst.

- A -> B -> C

    i. *(2 marks)* Determine the fraction of cities whose recovery rate for `C` is greater than that for `A`.

        – Show your code and the resulting proportion.

        – Which of the two treatments does this proportion suggest should be preferred? Why?

```
C_greater_A <- recovery$C > recovery$A
tableC_A <- table(C_greater_A)
tableC_A[[2]]/nrow(recovery)
```

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

- After calculating the fraction of cities whose recovery rate for C is greater than A, it shows that there are 70 percent chance that the recovery of treatment C is greater than treatment A. As a result, we conclude that this proportion suggests that treatment C is preferred over treatment A because it has a higher chance of getting a higher recovery rate.

    i. *(2 marks)* How does this new result affect the ordering of `A`, `B`, and `C`? What treatment would you recommend based on these last three proportions?

- This affect the ordering by a lot, it seems like A better than B, B is better than C, but also C is better than A. Based on these proportions, I would recommend C because it has the highest percentage(70 percent) over treatment A.

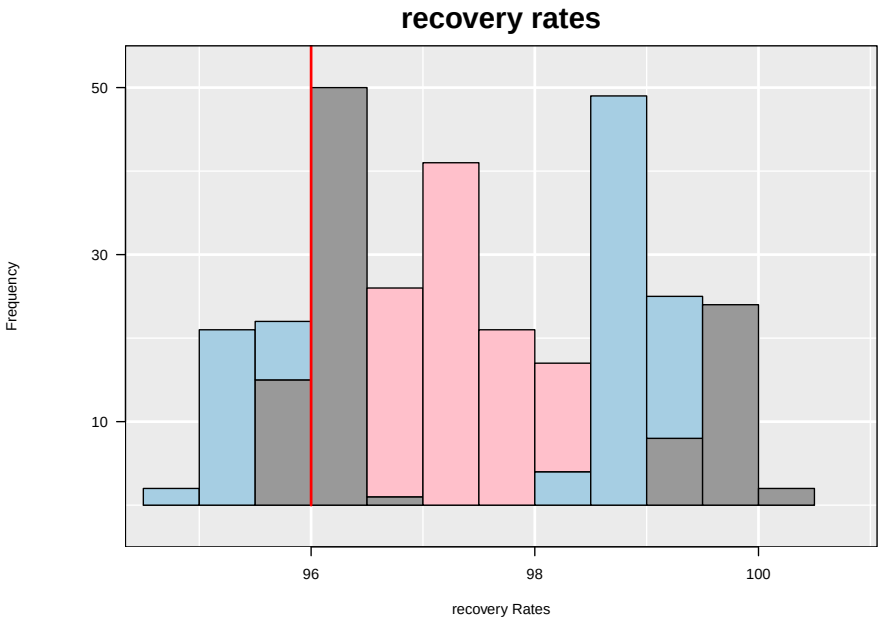
d. **(6 marks)** Recall the histogram `h_recovery` from part (a).

    i. *(1 mark)* Colour the histogram (programmatically) according to the treatment.

        - Show your code.

        - Show the resulting histogram

```
h_recovery['color'] <- trtPan$Treatment
plot(h_recovery)
```

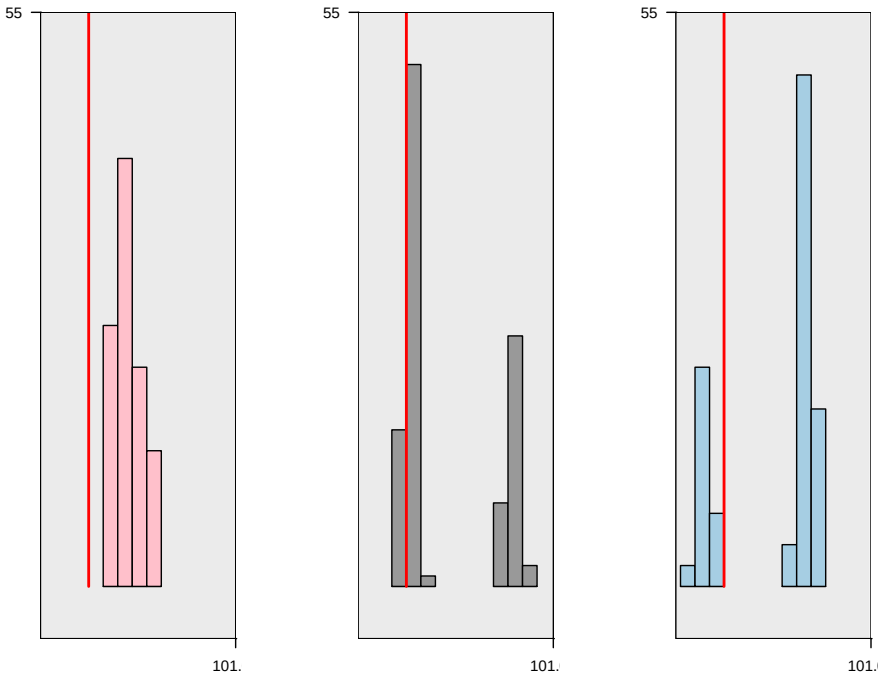


i. *(1 mark)* Facet the histogram `h_recovery` above by treatment (i.e., not by `"color"`).

- Show your code.
- Show the resulting histogram.

(Treatment `A` will be pink, `B` grey, `C` light blue.)

```
facet_treatment <- l_facet(h_recovery, by = trtPan$Treatment)
plot(facet_treatment)
```



i. *(4 marks)* Base on the above faceted histogram explain each of the following:

(colours should be as in parentheses):

- Why `A` (pink) might be preferred to `B` (grey)?

- Because A (pink) does not have recovery rate below the red line (96), while there is a protion of B(grey) that is below the red line

  - Why `B` (grey) might be preferred to `A` (pink)?

- B might be preferred to A is because that for some of the B, the recovery rates are really close to 100 percent recovery rate.

  - Why `B` (grey) might be preferred to `C` (light blue)?

- For both B and C, they are separated into two portions. When we look at the higher recovery rates for B and C, it looks like B has higher recovery rate than C. When we look at the lower recovery rates for B and C, B still has a higher recovery rate than C.

  - Why `C` (light blue) might be preferred to `B` (grey)?

- C might be preferred to B because when we look at the histogram, C has higher frequency at the higher end than B.