

Homework One(P8130)

Problem 1

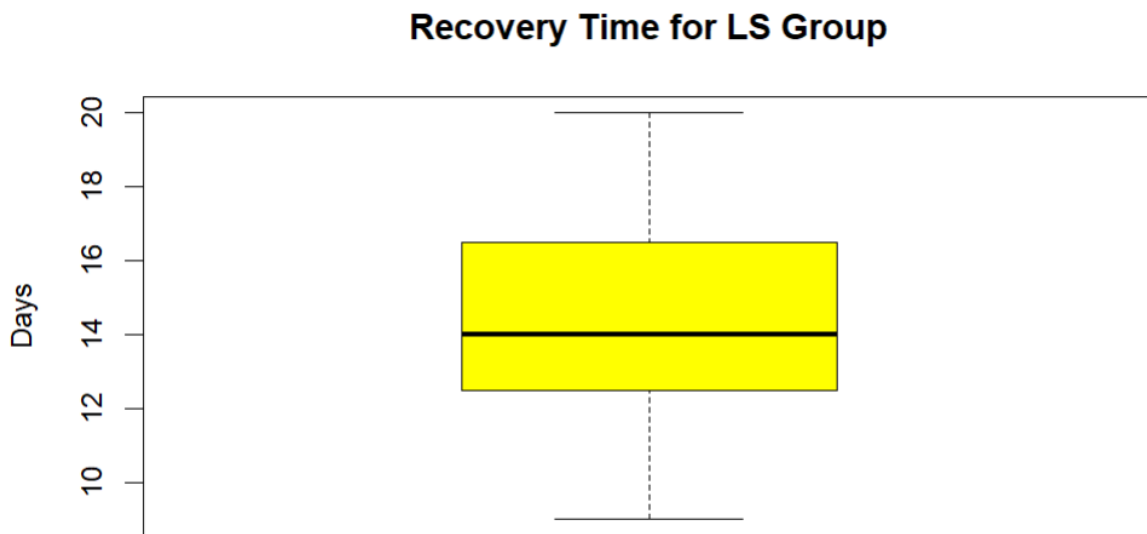
- a. The survival status is a **binary** variable because there are only two options.
- b. The stage of cancer is an **ordinal** variable because there is an order.
- c. The type of vaccine is a **nominal** variable because there isn't an order.
- d. The body temperature is a **continuous** variable.
- e. The number of emergency room visits is a **discrete** variable.
- f. The pain level is an **ordinal** variable.
- g. The systolic blood pressure is a **continuous** variable.
- h. The diabetes status is an **ordinal** variable.

Problem 2

- a. After calculation in R, the mean is 14.4, the median is 14, the range is $20 - 9 = 11$ and the standard deviation is 3.224903.

```
23 ## Question a
24 {r}
25 ls_mean = mean(vec_ls)
26 ls_median = median(vec_ls)
27 ls_range = range(vec_ls)
28 ls_sd = sd(vec_ls)
29
30 ls_mean
31 ls_median
32 ls_range
33 ls_sd
34
35
36 The mean of recovery time for the laparoscopic group is ls_mean, median is
37 ls_median, the minimum is min(vec_ls), the maximum is max(vec_ls)
and standard deviation is ls_sd.
```

b. The boxplot for LS group



The distribution is more like symmetric, because the upper part of this plot is nearly the same size of the lower one and there are no outliers. As for modality, we cannot find the result from this boxplot, but what we can find out from the data set itself is that most data concentrates between 12 and 16 and there is only one peak.

Here is the R code below.

```
## Question b
# {r drawing a boxplot}
boxplot(vec_ls,
  main = "Recovery Time for LS Group",
  ylab = "Days",
  col = "yellow",
  border = "black"
)
```

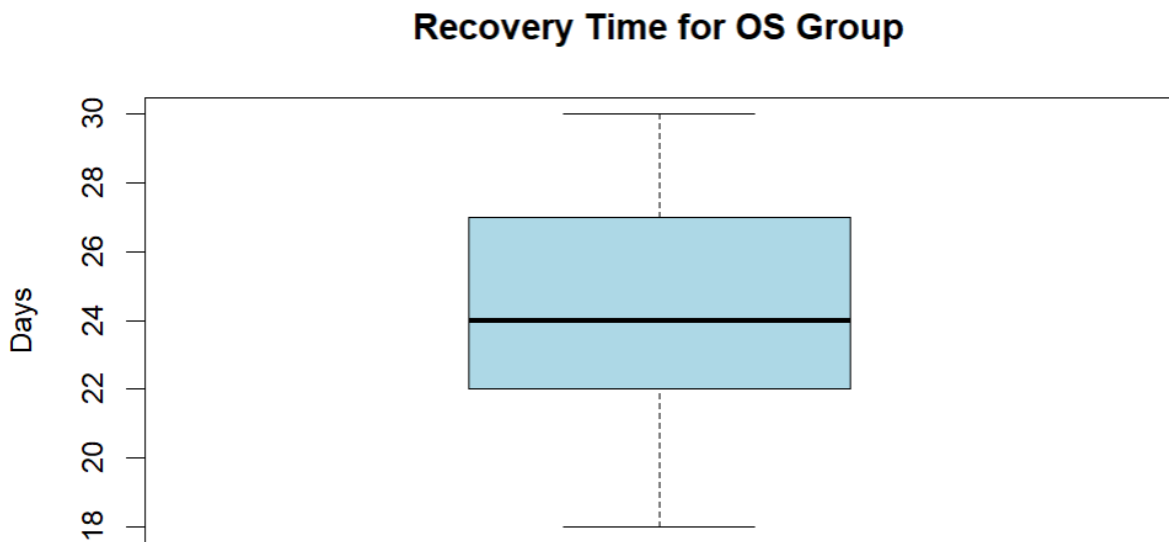
c. The mean for OS group is 24.21429, the median is 24, the range is $30 - 18 = 12$

and the standard variance is 3.512145.

```
49 ## Question c
50 {r}
51 os_mean = mean(vec_os)
52 os_median = median(vec_os)
53 os_range = range(vec_os)
54 os_sd = sd(vec_os)
55
56 os_mean
57 os_median
58 os_range
59 os_sd
60
[1] 24.21429
[1] 24
[1] 18 30
[1] 3.512145

61 The mean of recovery time for the laparoscopic group is os_mean, median is
62 os_median, the minimum is min(vec_os), the maximum is max(vec_os) and standard deviation is os_sd.
```

d. The boxplot for OS group

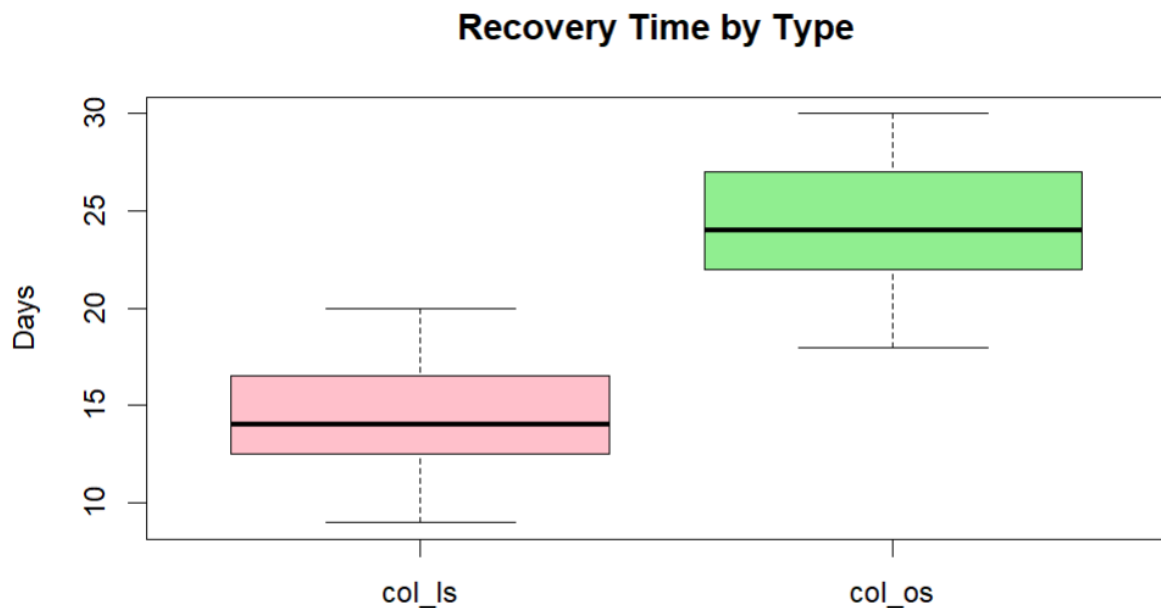


I think this one is also symmetric which means there are no outliers and the width of each part is similar. As for modality, I suppose this distribution is unimodal and has only one peak.

Here is the R code below.

```
64 ## Question d
65 {r}
66 boxplot(vec_os,
67         main = "Recovery Time for OS Group",
68         ylab = "Days",
69         col = "lightblue",
70         border = "black"
71       )
72 }
```

e. The boxplots for both groups



f. We can see clearly from the plots that patients in the LS group tend to recover faster than those in the other group because of the means and medians.

g. We usually describe outliers as those which are bigger than $Q3 + 1.5 * (Q3 - Q1)$ or smaller than $Q1 - 1.5(Q3 - Q1)$. The outliers usually don't affect medians but do have an obvious impact on means, which can make the average location of a dataset away from the normal one.

```
> summary(vec_ls)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9.0   12.5   14.0   14.4   16.5   20.0

> summary(vec_os)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00  22.00  24.00  24.21  26.75  30.00
```

Based on the calculation results above, there are no outliers in both groups.

h. There are some limitations. For example, the sample is small, which means that the outcomes are sensitive to outliers and the estimations are not stable. Besides, there are some other factors that we have not considered like age and medical history which do have huge impacts on the recovery time.

Perhaps we can add more samples and paint histograms.

Problem 3

Proof: We know that $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, if $P(A \mid B) = P(A)$, then we can get: $P(A) = \frac{P(A \cap B)}{P(B)}$.

Thus, $P(A \cap B) = P(A)P(B)$. Therefore, $P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$.

Problem 4

a. There are 120 people exercising regularly, so $P(A) = 120/200 = 0.6$. There are 70 people consume sugary drinks daily, so $P(B) = 70/200 = 0.35$.

b. First, we need to calculate $P(A \cap B)$.

There are 40 adults both exercise regularly and consume sugary drinks daily, so $P(A \cap B) = 0.2$.

Then, based on the equality $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, we can get $P(A \cup B) = 0.6 + 0.35 - 0.2 = 0.75$

c. $P(B \mid A) = \frac{P(A \cap B)}{P(A)} = 0.2/0.6 = \frac{1}{3}$

d.

We know that $P(A)=0.6$, $P(B)=0.35$, $P(B|A)=\frac{1}{3}$

$$P(A|B^c) = \frac{P(B^c|A) \cdot P(A)}{1 - P(B)} = \frac{(1 - P(B|A)) \cdot P(A)}{1 - P(B)} = \frac{(1 - \frac{1}{3}) \times 0.6}{1 - 0.35} \approx 0.6154$$

e.

A and B are not independent.

Proof: $P(A \cap B) = 0.2$, $P(A) = 0.6$, $P(B) = 0.35$

It's clear that $0.2 \neq 0.6 \times 0.35$, then $P(A \cap B) \neq P(A)P(B)$

Therefore, A and B are dependent.

f.

$$P(A \cap B^c) = P(A) - P(A \cap B) = 0.6 - 0.2 = 0.4$$

We've learned that two adults're selected

independently, then the probability is $0.4 \times 0.4 = 0.16$

Problem 5

a.

Let women having dementia be event A and showing positive finding be event B.

Then we can get $P(A) = 0.2$, $P(B|A) = 0.7$, $P(B|A^c) = 0.15$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)} = \frac{0.7 \times 0.2}{0.7 \times 0.2 + 0.15 \times 0.8} \approx 0.5385$$

b.

$$P(A^c|B^c) = \frac{P(B^c|A^c) \cdot P(A^c)}{1 - P(B)} = \frac{(1 - 0.15) \times (1 - 0.2)}{1 - (0.7 \times 0.2 + 0.15 \times 0.8)} \approx 0.9189$$

c. If someone has a negative finding, then it is very likely that she doesn't have dementia ($P = 0.9189$). However, if the finding is positive, then the possibility of having dementia and not having dementia is nearly fifty-fifty ($0.5385 - 0.4615$). Thus, the negative findings are more reliable than the positive ones when it comes to diagnose.

The R markdown file can be found here:

https://github.com/ChrisW12372/2025-09-15_p8130-homework01_r-codes.git