Biostatistics Method I - Homework 1

Yongyan Liu (yl6107)

Sep 10, 2025

Problem 1 (5 points)

Classify each of the following variables as binary, nominal, ordinal, discrete, or continuous:

a. Patient survival status recorded as "alive" or "deceased"
b. Stage of cancer at diagnosis (I, II, III, IV)
c. Type of vaccine received (e.g., Pfizer, Moderna, Johnson & Johnson)
d. Body temperature measured in degrees Celsius
e. Number of emergency room visits in the past year
f. Self-reported pain level ("none", "mild", "moderate", "severe")
g. Systolic blood pressure (in mmHg)
h. Diabetes status ("no diabetes", "pre-diabetes", "diabetes")

- binary: a
- nominal: c
- ordinal: b f h
- discrete: e
- continuous: d g

Problem 2 (10 points)

In a study on recovery times following two types of minor surgeries, recovery duration (in days) was recorded for 15 patients who underwent laparoscopic surgery and for 14 patients who underwent open surgery. The recovery times are as follows:

Laparoscopic surgery:

12, 15, 10, 14, 13, 16, 20, 18, 9, 11, 17, 19, 14, 13, 15

Open surgery:

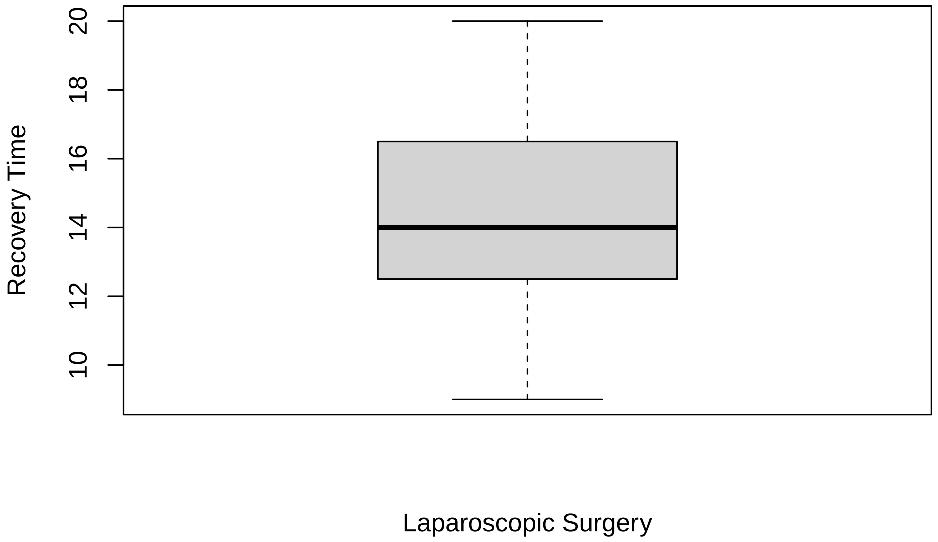
20, 25, 22, 30, 18, 27, 24, 21, 29, 26, 22, 23, 28, 24

a. Calculate the mean, median, range, and standard deviation of recovery times for the laparoscopic surgery group.

- mean: 14.4
- median: 14
- range: 9, 20
- standard deviation: 3.2249031

b. Construct and describe a box plot for the laparoscopic surgery group. Use terms such as skewness and modality to describe the distribution.

```r
boxplot(
  scopic,
  xlab = "Laparoscopic Surgery",
  ylab = "Recovery Time"
  )
```



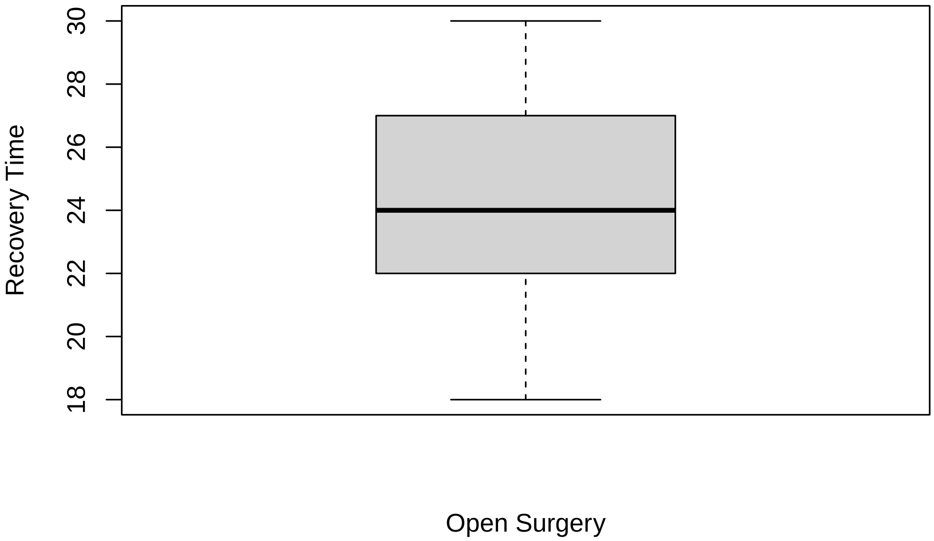
The distribution appears roughly symmetric, without strong skewness or multiple modes.

c. Calculate the mean, median, range, and standard deviation of recovery times for the open surgery group.

- mean: 24.2142857
- median: 24
- range: 18, 30
- standard deviation: 3.5121453

d. Construct and describe a box plot for the open surgery group using similar descriptive terms.

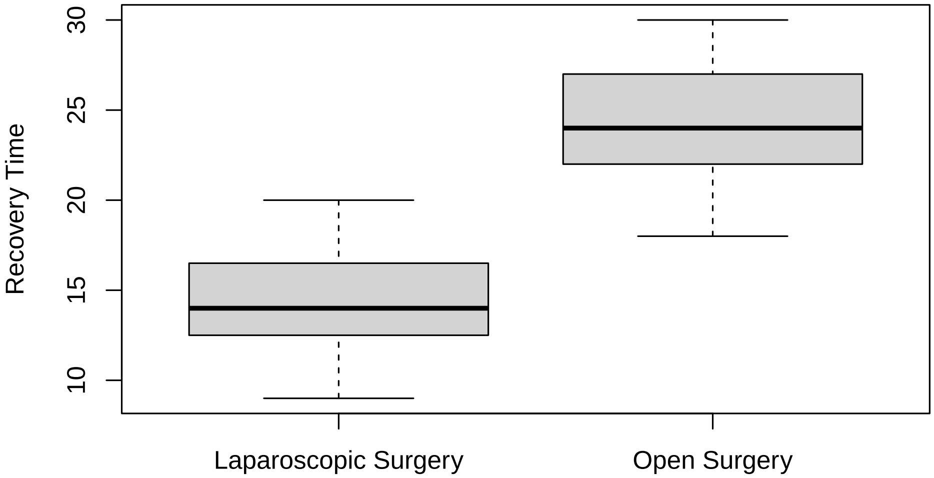
```
boxplot(
  open_surgery,
  xlab = "Open Surgery",
  ylab = "Recovery Time"
  )
```



The distribution appears roughly symmetric, without strong skewness or multiple modes. The range is slightly larger than laparoscopic surgery, but still appears normal.

e. Create side-by-side box plots comparing recovery times between the laparoscopic and open surgery groups. Label the plots clearly.

```r
boxplot(
  scopic, open_surgery,
  names = c("Laparoscopic Surgery", "Open Surgery"),
  ylab = "Recovery Time"
  )
```



f. Compare the two groups' recovery times based on the box plots and summary statistics. Which group tends to recover faster?

The recovery time for Laparoscopic surgery is usually much shorter than open surgery, by about 10 days in average.

g. Describe any potential outliers or unusual values visible in the box plots and suggest how they might affect interpretation.

From the data, the values are grouped fairly tight around the median for both laparoscopic surgery and open surgery. There are no obvious outliers presented. The differences within the group (about 10 days) are most likely due to other factors, for example, the differences in patients' baseline health status.

That strengthens the conclusion that laparoscopic surgery recovers much faster than traditional open surgery.

h. Discuss limitations of judging recovery differences based on these samples and suggest additional analyses that could be performed.

- The sample set is relatively small. Only 14-15 patients is not enough to detect long recovery outliers.

- Missing critical characteristics of patients, e.g. age, weight, health status, etc. These factors could impact the recovery time singificantly.
- Additional analyses suggested:
    - more samples
    - age analysis
    - break down on different reasons for the surgery

Problem 3 (5 points)

We say that events A and B are independent if P(A|B) = P(A). Show that if P(A|B) = P(A) then also P(B|A) = P(B).

Assume P(A) > 0 and P(B) > 0 (otherwise P(A|B) or P(B|A) is meaningless)

Proof.

from P(A|B) = $\dfrac{P(A \cap B)}{P(B)}$ and P(A|B) = P(A)

Then $P(A \cap B)$ = P(A)P(B)

So P(B) = $\dfrac{P(A \cap B)}{P(A)}$ = P(B|A)

Problem 4 (10 points)

In a community health study of 200 adults, data were collected on two behaviors: regular exercise (at least 3 times a week) and daily consumption of sugary drinks.

Define the following events:

- Event A: An adult exercises regularly.
- Event B: An adult consumes sugary drinks daily.

Survey results showed that:

- 120 adults exercise regularly.
- 70 adults consume sugary drinks daily.
- 40 adults both exercise regularly and consume sugary drinks daily.

a. What is P(A) and P(B)?

$$\text{P(A)} = \frac{120}{200} = 60\%$$

$$\text{P(B)} = \frac{70}{200} = 35\%$$

b. Calculate P($A \cup B$), the probability that a randomly selected adult either exercises regularly or consumes sugary drinks daily (or both).

P($A \cup B$) = P(A) + P(B) - P($A \cap B$) = 60% + 35% - 20% = 75%

c. Calculate P(B|A).

$$\text{P(B|A)} = \frac{P(A \cap B)}{P(A)} = \frac{20\%}{60\%} = \frac{1}{3}$$

d. Calculate $P(A|B^c)$

$$P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{P(A) - P(A \cap B)}{1 - P(B)} = \frac{60\% - 20\%}{1 - 35\%} = \frac{8}{13}$$

e. Are events A and B independent? Justify your answer.

They are not indenpendent.

From the calcualtion above, P(B) = 35% but P(B|A)=$\frac{1}{3}$

thus $P(B) \neq P(B|A)$

f. If two adults are selected at random independently, what is the probability that both exercise regularly but neither consumes sugary drinks daily?

The probability of a adult who exercise regularly and don't consumes sugary drinks daily is

$P(A \cap B^c) = \frac{8}{13} \approx 61.5\%$

For 2 adaults, the probabilty is $P(A \cap B^c)^2 = 37.8698225\%$

Problem 5 (10 points)

Among women aged 75 and older in a certain population, 20% have dementia. Among women in this group who have dementia, 70% show positive findings on a certain brain CT scan. Among women without dementia, about 15% also show positive findings (false positives).

a. If a randomly selected woman in this population has a positive CT scan finding, what is the probability that she actually has dementia?

Let 'D' be the event of dementia and 'T' be the event of positive findings on a CT scan.

From the problem statement, we know P(D) = 20%, $P(T|D) = 70\%$ and $P(T|D^c) = 15\%$.

So $P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} = \frac{70\% * 20\%}{70\% * 20\% + 15\% * 80\%} = \frac{14}{26} \approx 53.8\%$

b. If a randomly selected woman in this population has a NEGATIVE CT scan finding, what is the probability that she DOES NOT have dementia?

Where $P(T^c|D^c) = 1 - P(T|D^c) = 85\%$ and $P(T^c|D)$ = 1 - $P(T|D) = 30\%$

$$P(D^c|T^c) = \frac{P(T^c|D^c)P(D^c)}{P(T^c|D^c)P(D^c) + P(T^c|D)P(D)} = \frac{85\% * 80\%}{85\% * 80\% + 30\% * 20\%} \approx 91.9\%$$

c. Discuss the implications of these probabilities for interpreting test results in this population. Show all calculations. This answer may be hand-written.

The results show that the test is not equally reliable for positive and negative findings. A positive CT scan corresponds to only about a 54% chance of actually having dementia, which means the test by itself is not very strong for confirming the disease. On the other hand, a negative CT scan corresponds to roughly a 92% chance of not having dementia, suggesting that the test is much better at ruling out dementia.

In practice, this implies that a positive result should be interpreted with caution and ideally supported by additional diagnostic evaluations, while a negative result can provide more reassurance that dementia is unlikely.