Biostatistical Methods HW1

Pei Tian, pt2632

9/14/2023

Problem 1

Please classify each of the following variables as qualitative (specify if binary, nominal, or ordinal) or quantitative (specify if discrete or continuous):

a) homework feedback, labeled as "poor", "fair", "good", "very good"
qualitative, ordinal variable

b) homework feedback, labeled as "fail", "pass"
qualitative, binary variable

c) country of birth
qualitative, nominal variable

d) the quantity of grapes (in lbs) to make 3 liters of wine
quantitative, continuous variable

e) number of TAs in the P8130 course
quantitative, discrete variable

Problem 2

In a study of 133 individuals with a recent bike crash history, depression scores were measured using a standardized test. The depression scores for 14 of these individuals are as follows:

45, 39, 25, 47, 49, 5, 70, 99, 74, 37, 99, 35, 8, 59

a) Compute the following descriptive summaries of these data: mean, median, range, SD.

b) Describe the box plot and the underlying distribution of the data. Use some of the following terms: left-skewed, right-skewed, symmetric, bimodal, unimodal distribution.

Additionally, 140 individuals with a recent car crash history also participated in the study. The depression scores for 13 of these individuals are given below:

67, 50, 85, 43, 64, 35, 47, 97, 58, 58, 10, 56, 50

a) Using R, make a side-by-side box plot of the depression scores stratified by type of accident. Make sure you label your figure appropriately.

b) Describe each of the box plots and the underlying distribution of the data. Use some of the following terms: left-skewed, right-skewed, symmetric, bimodal, unimodal distribution.

c) Comparing the 2 box plots, which group appears to have a lower typical depression score?

Q1 - a)

$X = \{x_i | i = 1, 2, 3, ..., 14\}, n = 14$

$mean = \sum_{i=1}^{n} x_i / n = 49.35714$

$median = \arg\min_{a} \sum_{i=1}^{n} |x_i - a| = (45 + 47)/2 = 46$

$range = max(X) - min(X) = 94$

$SD = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 / (n-1)} = 28.84603$

```
# mean
mean(data1)
```

```
## [1] 49.35714
```

```
# median
median(data1)
```

```
## [1] 46
```

```
# range
max(data1) - min(data1)
```

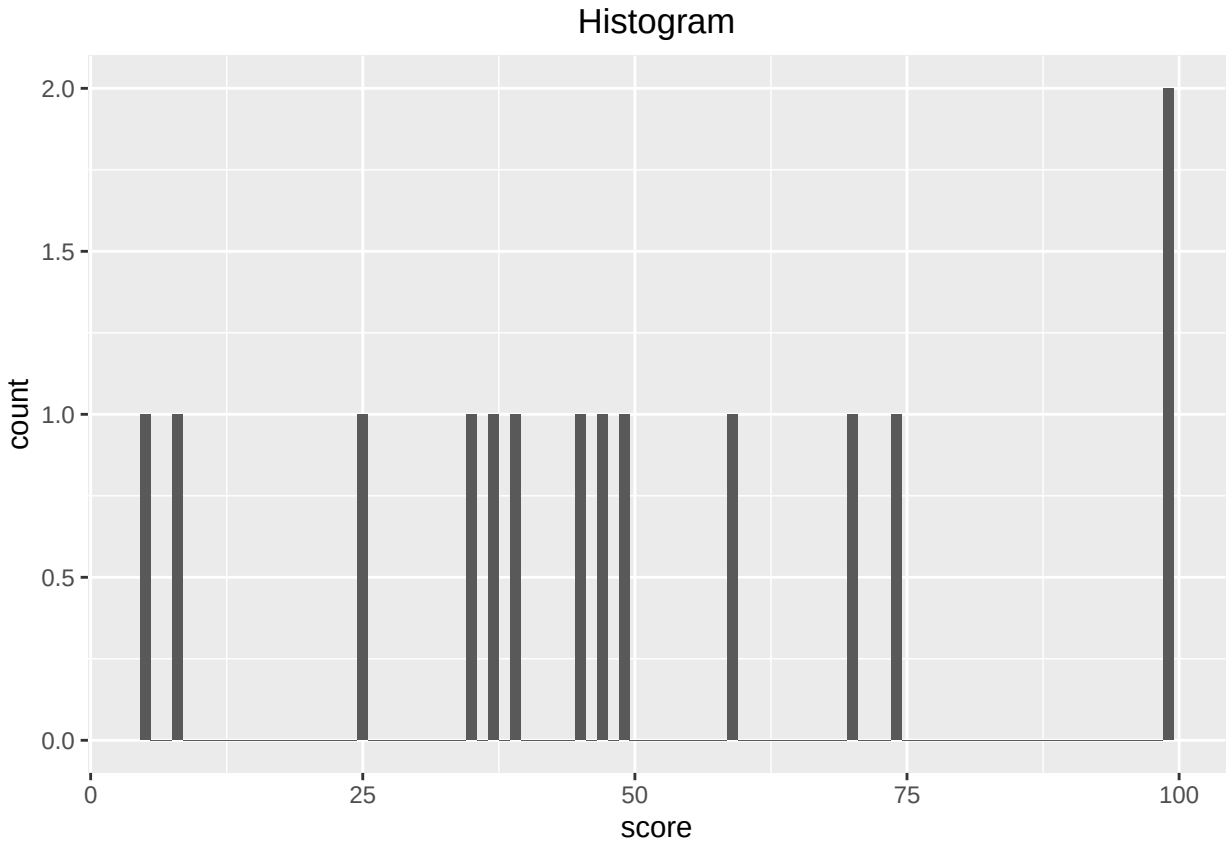
```
## [1] 94
```

```
# SD
sd(data1) # sqrt(sum((data1 - mean(data1)) ^ 2 / (length(data1) - 1)))
```

```
## [1] 28.84603
```

Q1 - b)

```
# Histogram for Peak Detection
df <- tibble(score = data1)
ggplot(df, aes(x = score)) + geom_histogram(binwidth = 1) +
  ggtitle("Histogram") + theme(plot.title = element_text(hjust = 0.5))
```

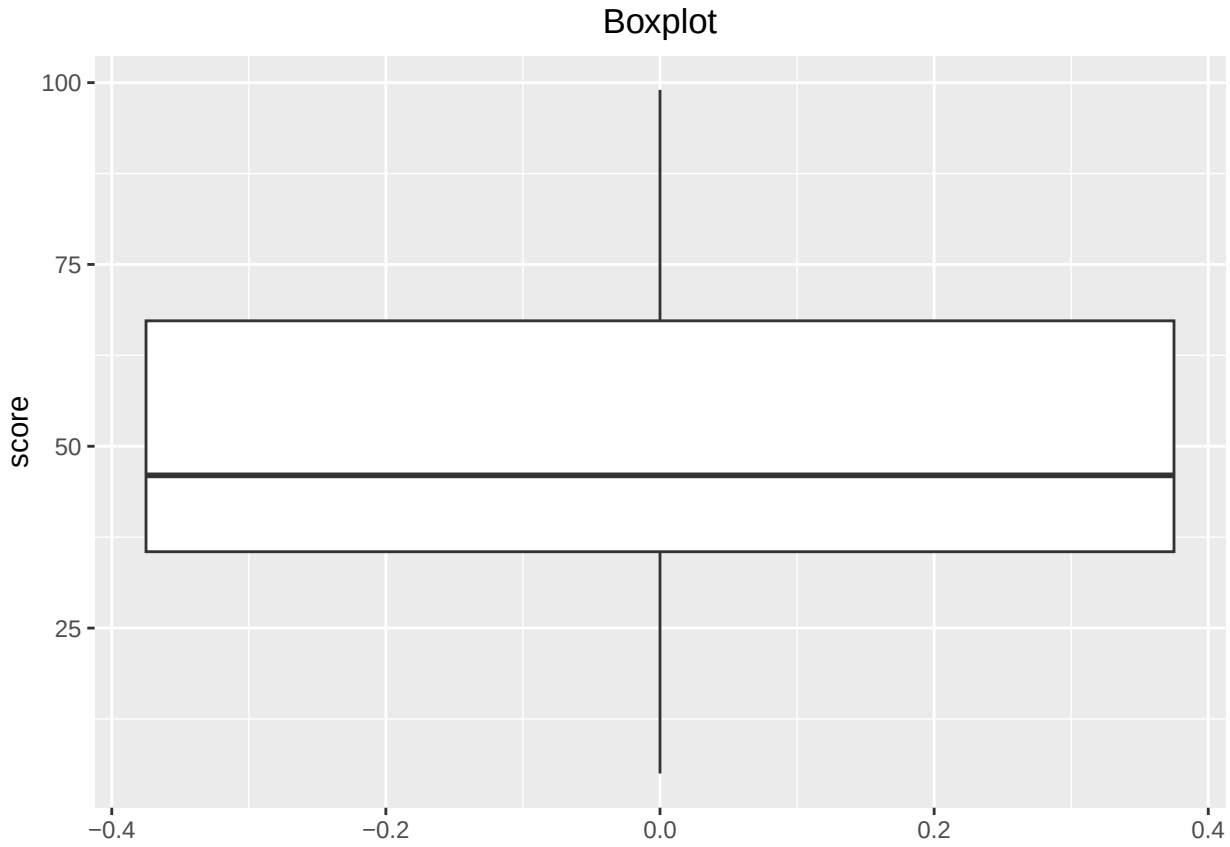


Boxplot Description: Q1 is 35.5, Q3 is 67.25, median is 46, max is 99, min is 5, no outlier

Observation & Conclusion:

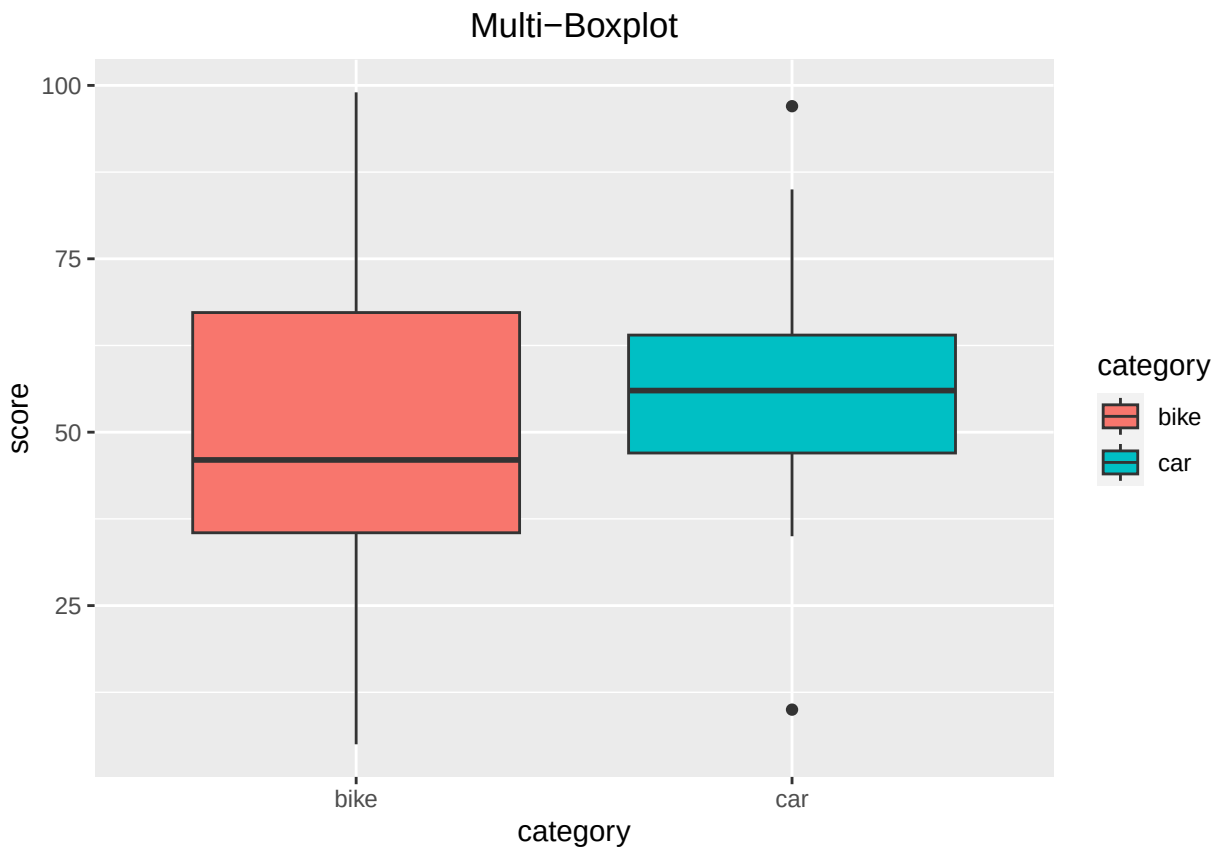
- mean > median, right-skewed distribution
- only one peak in histogram, unimodal distribution

```
# Histogram for Peak Detection
df <- tibble(score = data1)
ggplot(df, aes(y = score)) + geom_boxplot() +
  ggtitle("Boxplot") + theme(plot.title = element_text(hjust = 0.5))
```



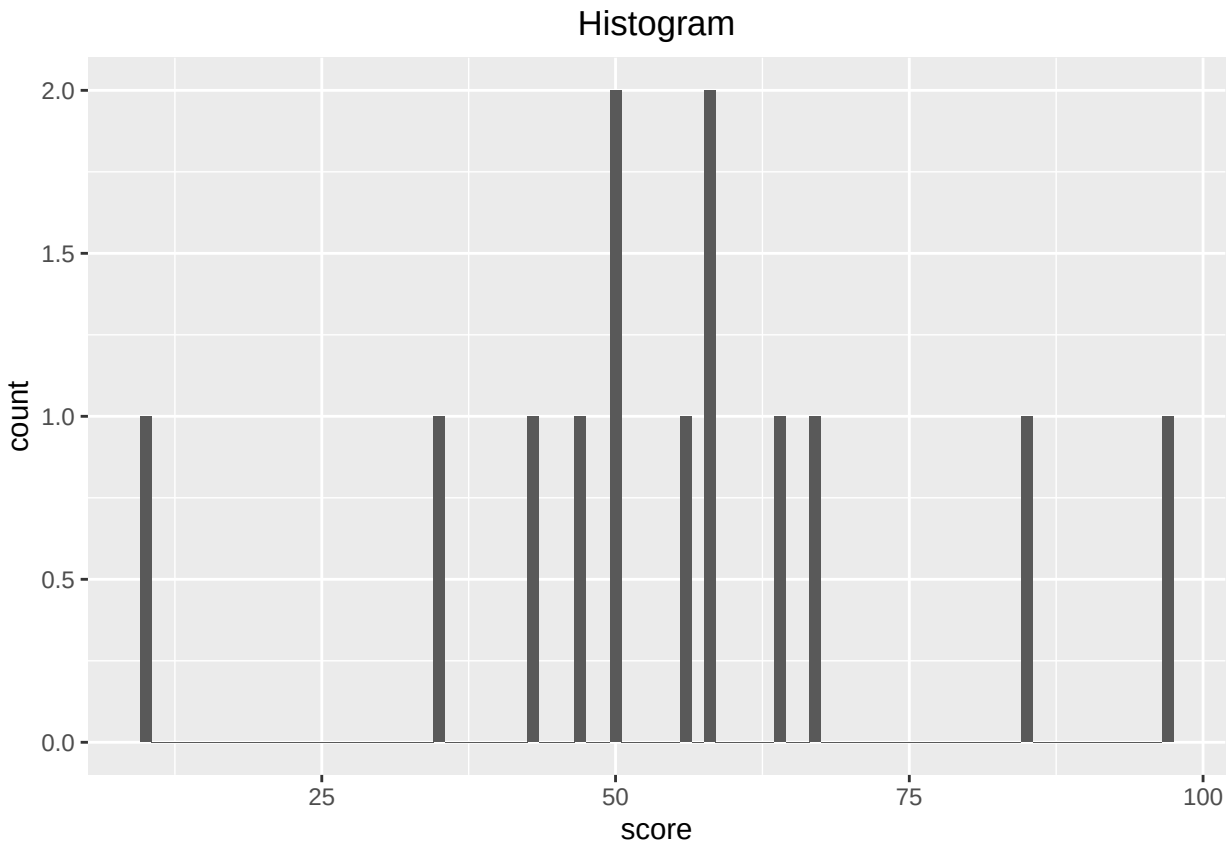
Q2 - a)

```
# DataFrame Construction
data2 <- c(67, 50, 85, 43, 64, 35, 47, 97, 58, 58, 10, 56, 50)
df <- tibble(
  score = c(data1, data2),
  category = c(rep("bike", length(data1)), rep("car", length(data2)))
)
# Multi-Boxplot
ggplot(df, aes(y = score, x = category, fill = category)) + geom_boxplot() +
  ggtitle("Multi-Boxplot") + theme(plot.title = element_text(hjust = 0.5))
```



Q2 - b)

```
# Histogram for Peak Detection
df <- tibble(score = data2)
ggplot(df, aes(x = score)) + geom_histogram(binwidth = 1) +
  ggtitle("Histogram") + theme(plot.title = element_text(hjust = 0.5))
```



Boxplot Description:

Bike: Q1 is 35.5, Q3 is 67.25, median is 46, max is 99, min is 5, no outlier

Car: Q1 is 47, Q3 is 64, median is 56, max is 97, min is 10, 2 outliers

Observation & Conclusion:

Bike:

- mean > median, right-skewed distribution
- only 1 peak in histogram, unimodal distribution

Car:

- mean < median, left-skewed distribution
- 2 peaks in histogram, bimodal distribution

```
mean(data2)
```

```
## [1] 55.38462
```

```
median(data2)
```

```
## [1] 56
```

Q1 - c)

Individuals with bike crash history have a lower typical depression score. Because depression score median of individuals with bike crash history is less than that of individuals with car crash history.

Problem 3

Suppose we toss one fair 12-sided die:

a) Let's define the event A as "an even number appears". What is the probability of the event A?
$P(A) = \frac{1}{2}$

b) Let's define the event B as "number 10 appears". What is the probability of the event B?
$P(B) = \frac{1}{12}$

c) Compute P(B U A).
$B \subset A \Rightarrow P(B \cup A) = P(A) = \frac{1}{2}$

d) Are events A and B independent? Why? Prove your answer.
A and B are **NOT** independent.
Proof:
$\because B \subset A$
$\therefore P(AB) = P(A \cap B) = P(B) = \frac{1}{12}$
$\therefore P(AB) \neq P(A) * P(B)$
$\therefore A, B \; are \; NOT \; independent$

Problem 4

5% of women above age of 75 have dementia. Among women (75+ years old) with dementia, 80% have positive findings on their CT scan. Among women (75+ years old) who don't have dementia, 10% will have a positive CT scan findings. A randomly-selected woman (75+ years old) had a positive CT scan findings.

What is the probability that she actually has dementia? Compute by hand and show the key steps. The answer can be hand written.

Solution.

Define:

A: a woman(75+ years old) actually have dementia

B: a woman(75+ years old) have postitive CT scan finding

According to the problem description,

$P(A) = 0.05$, $P(B|A) = \frac{P(AB)}{P(A)} = 0.8$, $P(B|\overline{A}) = \frac{P(\overline{A}B)}{P(\overline{A})} = 0.1$

$\therefore P(AB) = P(B|A) * P(A) = 0.04$, $P(\overline{A}B) = P(B|\overline{A}) * (1 - P(A)) = 0.095$

$\therefore P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(AB)}{P(AB)+P(\overline{A}B)} = \frac{0.04}{0.04+0.095} = \frac{8}{27}$

The probability that a randomly-selected woman (75+ years old) with a positive CT scan findings actually has dementia is $\frac{8}{27}$.