p8130_homework1_lq2250

Lanlan_Qing

2024-09-14

This is homework 1 of Biostatistics P8130

The solutions are as follows:

Problem 1

Answers

a) qualitative ordinal

b) qualitative binary

c) qualitative nominal

d) quantitative continuous

e) quantitative discrete

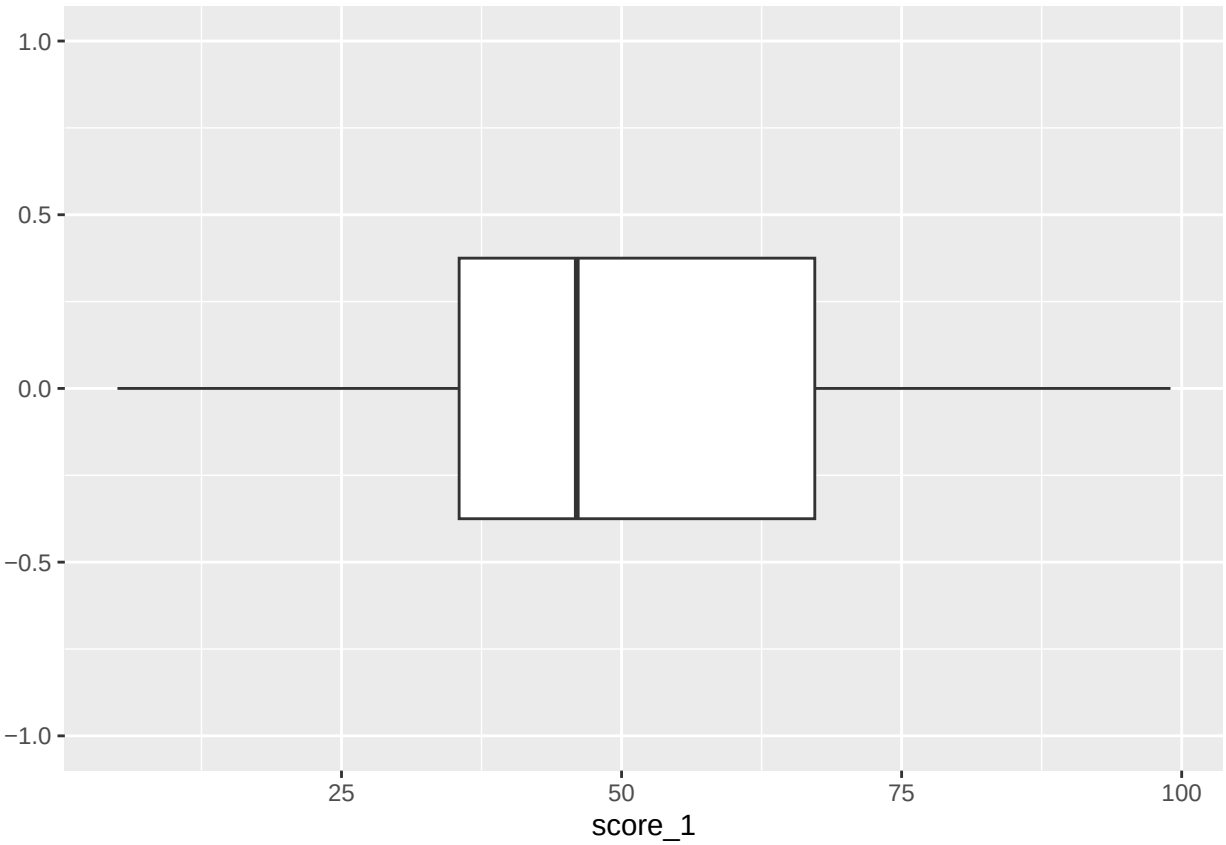
Problem 2

Codings for the first question

```
# Set score variable of bike crash
score_1 = c(45, 39, 25, 47, 49, 5, 70, 99, 74, 37, 99, 35, 8, 59)
score_sample = data.frame(score_1)
# Compute descriptive data of the first sample (bike crash)
mean_score1 = mean(score_1)
median_score1 = median(score_1)
range_score1 = max(score_1)-min(score_1)
sd_score1 = sd(score_1)
summary(score_1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   35.50   46.00   49.36   67.25   99.00
```

```
#Create box plot
library(ggplot2)
ggplot(score_sample, aes(x = score_1))+
  geom_boxplot()+
  ylim(-1, 1)
```




```
  labs(title = 'Box Plot of Depression Scores',
       x = 'Score')
```

```
## $x
## [1] "Score"
##
## $title
## [1] "Box Plot of Depression Scores"
##
## attr(,"class")
## [1] "labels"
```

Answers

(a)

Mean = 49.36; **Median** = 46; **Range** = 94; **SD** = 28.85

(b)

1.According to the box plot and summary:

- Minimum(the lower bound of whisker): 5
- Q1(the lower bound of the box): 35.5
- Median: 46
- Q3(the upper bound of the box): 67.25
- Maximum (the upper bound of whisker): 99

2.Distribution Description:

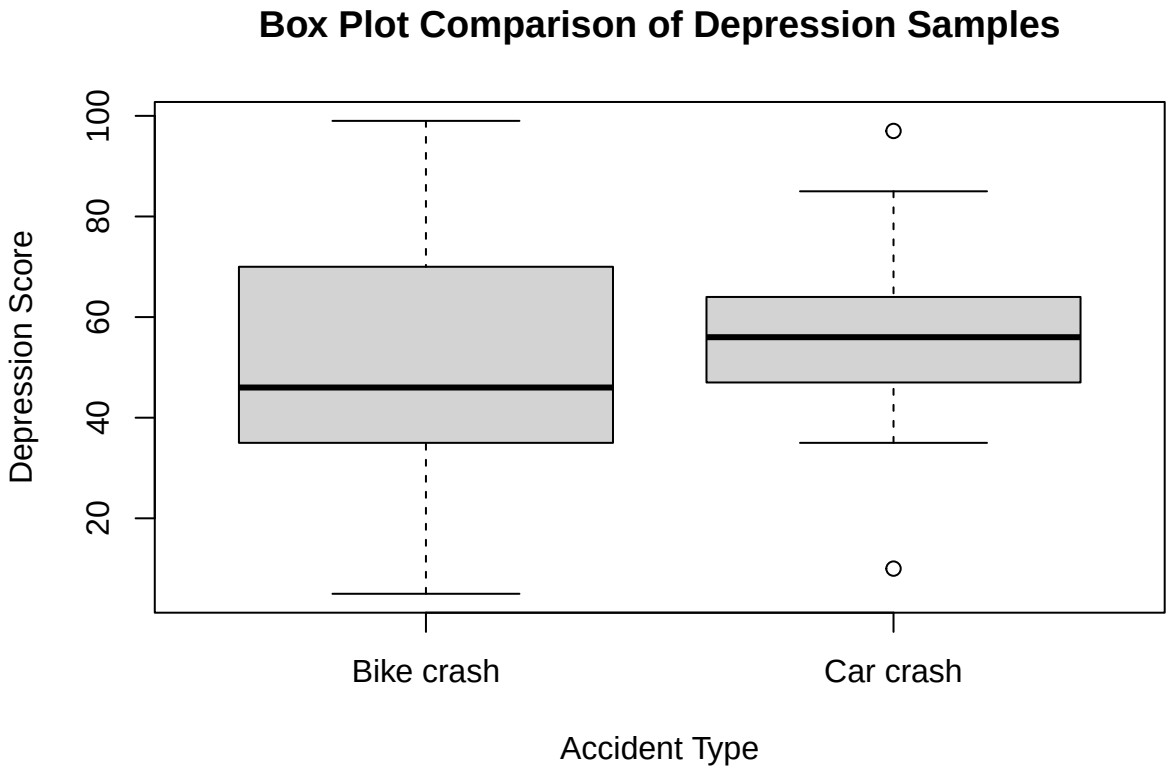
- Score mean = 49 is higher than median = 46, indicating mean on the right side of the median. High values pull the distribution to the right side
- Most data are concentrated on the lower half (on the left side of the median)
- Suggest a right-skewed, non-symmetric and unimodal distribution

Codings for the second question

```
# Set Set score variable of car crash
score_2 = c(67, 50, 85, 43, 64, 35, 47, 97, 58, 58, 10, 56, 50)

# Create a list of data used for creating box plot
depression_data = list('Bike crash' = score_1,
            'Car crash' = score_2)

# Create a side-by-side box plot of samples
box_plot_comparison = boxplot(depression_data, main = 'Box Plot Comparison of Depression Samples',
        xlab = 'Accident Type',
        ylab = 'Depression Score')
```




```
# Make a table of two sets of results
table_box_plot = cbind(summary(score_1), summary(score_2))
table_box_plot
```

```
##             [,1]     [,2]
## Min.     5.00000 10.00000
## 1st Qu. 35.50000 47.00000
## Median  46.00000 56.00000
```

```
## Mean    49.35714 55.38462
## 3rd Qu. 67.25000 64.00000
## Max.    99.00000 97.00000
```

Answers

(b)

Description of box plots:

1.For bike crash accident:

- Minimum(the lower bound of whisker): 5
- Q1(the lower bound of the box): 35.5
- Median: 46
- Q3 (the upper bound of the box): 67.25
- Maximum (the upper bound of whisker): 99

2.For car crash accident:

- Minimum(the lower bound of whisker): 10
- Q1(the lower bound of the box): 35.47
- Median: 56
- Q3(the upper bound of the box):64
- Maximum (the upper bound of whisker): 97

Description of the underlying distribution:

1.For bike crash accident:

- It follows a right-skewed, non-symmetric and unimodal distribution without any outliers.

2.For car crash accident:

- It follows a relatively left-skewed distribution, as the mean is lower than the median, and the upper 25th percentile is larger than the lower 25th percentile, though the upper whisker is longer than the lower one, which might be influenced by the large extreme value.
- It follows a non-symmetric and unimodal distribution with 1 upper outlier and 1 lower outlier.

(c) The bike group appears to have a lower typical depression score.

Problem 3

Answers

(a) A = {2,4,6,8,10,12}

P(A) = 6/12 = 1/2

(b) B = {10}

P(B) = 1/12

(c) Since B is the subset of A

P(BUA) = P(A) = 1/2

(d) P(A intersect B) = P(A) + P(B) - P(AUB) = 1/2 + 1/12 - 1/2 = 1/12

P(A)P(B) = 1/2*1/12 is not equal to 1/12, thus A and B are not independent.

Problem 4

Answers

According to the text: P(de) = 0.05; P(+|de) = 0.8; P(+|no de) = 0.1

Codings

```
p_de = 0.8*0.05/(0.8*0.05+0.1*0.95)
```

P(de|+) = [P(+|de)*P(de)]/[P(+|de)*P(de)+P(+|no de)*P(no de)] = 0.2962963