P8130 HW1

Chhiring Lama

2024-09-22

Problem 1 (5 points)

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"
   - Homework feedback labeled as above is a qualitative (ordinal) variable.

b) homework feedback, labeled as "fail", "pass"
   - Homework feedback here is a qualitative (binary) variable.

c) country of birth
   - Country of birth is a qualitative variable. It is a nominal variable as the values here cannot be ordered.

d) the quantity of grapes (in lbs) to make 3 liters of wine
   - Quantity of grapes (in lbs) to make 3 L of wine is a quantitative (continuous) variable.

e) number of TAs in the P8130 course
   - Number of TAs in this course if quantitative (discrete) variable.

Problem 2 (15 points)

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.

- Here, number of items (n) = 14, sum of items (S) = $\sum_{i=1}^{n} x_i = 45 + 39 + 25 + 47 + 49 + 5 + 70 + 99 + 74 + 37 + 99 + 35 + 8 + 59 = 691$.

  Mean = S/n = $\sum_{x=1}^{n} \frac{x_i}{n} = \frac{691}{14} = 49.36$.

  For median, we need to reorder the items in the ascending order so. reordered list = {5, 8, 25, 35, 37, 39, 45, 47, 49, 59, 70, 74, 99, 99}. Median = $\frac{(\frac{n}{2})^{th} + (\frac{n+1}{2})^{th}}{2} = \frac{(\frac{14}{2})^{th} + (\frac{14+1}{2})^{th}}{2} = \frac{45+47}{2}. = 46$.

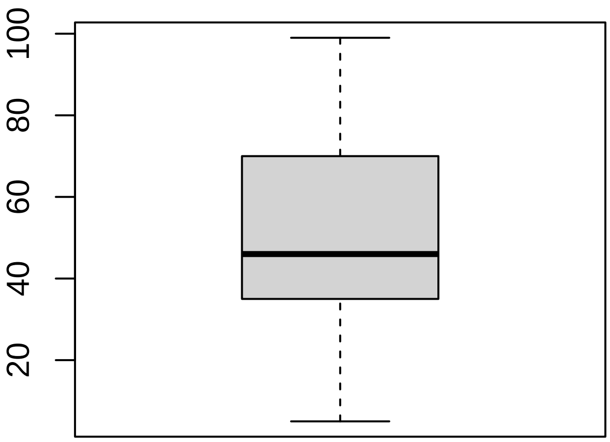
Range $= max - min = 94$.

Standard Deviation $(\sigma) = \sqrt{\sigma^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ where x is an element in the list and i is the index.
$\sigma = \sqrt{\frac{1}{14-1}\sum_{i=1}^{14}(x_i - \frac{691}{14})^2} = 28.85$.

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.

```r
bike_crash_list <- list(`Bike Crash` = bike_crash_lst)

boxplot(bike_crash_list)
```



- As shown in the box plot above, the underlying distribution of the data is right-skewed. The distribution is unimodal since the mode is only 1 element (99). However by the nature of the distribution being right-skewed, the median is less the average (mean) and hence the distribution is not symmetric.

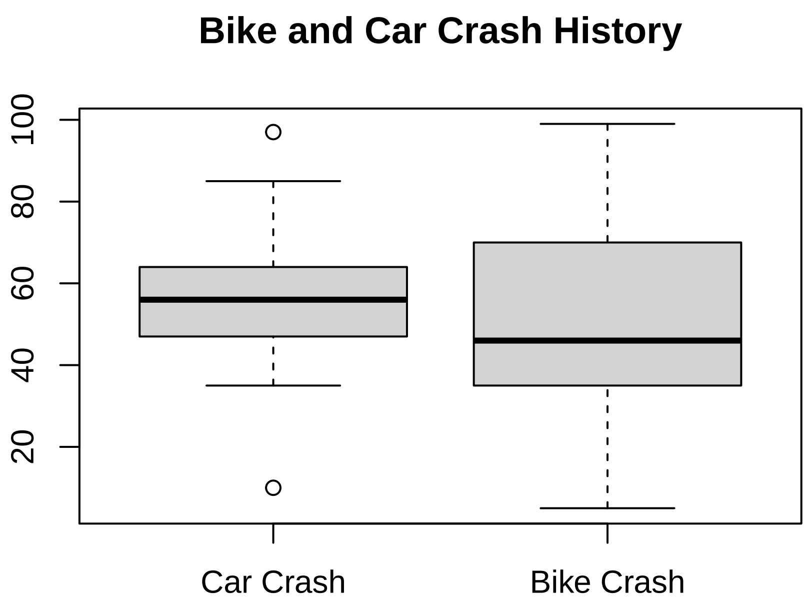
Additionally, 140 individuals with a recent car crash history also participated in the study. The depression scores for 13 of these individuals are given below:
67, 50, 85, 43, 64, 35, 47, 97, 58, 58, 10, 56, 50

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.

```r
car_crash_lst = c(67, 50, 85, 43, 64, 35, 47, 97, 58, 58, 10, 56, 50)

crash_list <- list(`Car Crash` = car_crash_lst,
                   `Bike Crash` = bike_crash_lst)

boxplot(crash_list, main = "Bike and Car Crash History")
```



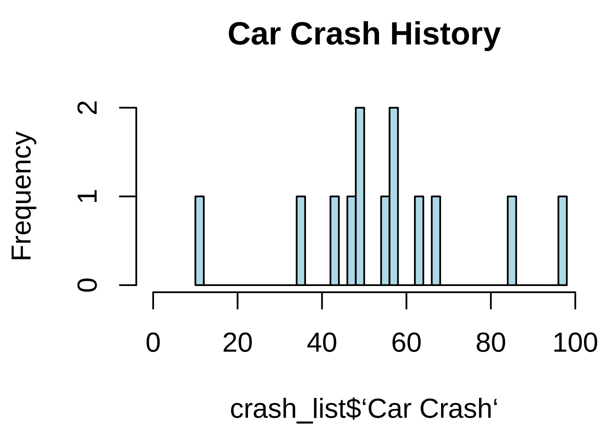
b) Describe each of the box plots and the underlying distribution of the data. Use some of the following terms: left-skewed, right-skewed, symmetric, bimodal, unimodal distribution.

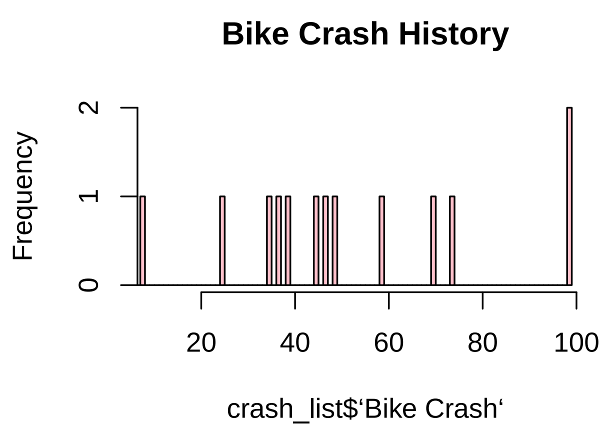
- The distribution of bike crash history is right skewed while that of the car crash is symmetric. To check their modes we can use histogram.

```
par(mfrow = c(1, 2))

hist(crash_list$`Car Crash`, main = "Car Crash History",
     breaks = (2*IQR(crash_list$`Car Crash`)) / 3*sqrt(14),
     col = "lightblue",
     xlim = c(0, 100))

hist(crash_list$`Bike Crash`, main = "Bike Crash History",
     breaks = (2*IQR(crash_list$`Bike Crash`)) / 3*sqrt(13),
     col = "pink", xlim = c(10, 100))
```





- The distribution for car crash is a bimodal distribution since there are two modes - 50 and 58 (also shown in the histogram). As mentioned before, distribution for bike crash history is unimodal.

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

- Comparing the two groups, group with bike crash history has a lower typical depression score. Both Median (as shown in the box-plot) and mean (55.38 for car crash and 49.36 for bike crash) are lower for bike crash history.

Problem 3 (10 points)

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?

- Here, sample size (Ω) = {1,2, 3, ..., 12}
  An even number appears (A) = {0, 2, 4, 6, 8, 10, 12}
  In a roll of a dice, the possibilities are all discrete random variable whereby there are total 12 possibility and 6 possibilities for event A. Therefore, $P(A) = \frac{6}{12} = 0.5$. Hence, probability of the event A is 0.5.

b) Let's define the event B as "number 10 appears". What is the probability of the event B?

- $B = \{10\}$ (only one element) . So, $P(B) = 1/12 = 0.08$. Hence, the probability of the number 10 to appear is 0.08.

c) Compute P(B U A).

- After listing both A and B, we know that $B \subset A$. This means $A \cap B = P(B)$. So, P(B U A) = $P(A)$ + $P(B)$ - $P(A \cap B)$ = $P(A)$ + $P(B)$ - $P(B)$ = $P(A) = 0.5$.

d) Are events A and B independent? Why? Prove your answer.

- No, A and B are not independent because if event A and B come from the sample sample space Ω where $P(A)$ affects $P(B)$ and vice versa.
  Mathematically, when two events are independent, $P(A \cap B) = P(A).P(B)$
  In our case, $P(A).P(B) = (\frac{1}{2}).(\frac{1}{12}) = \frac{1}{24}$. However, from c), we know that $P(B \cap A) = P(B) = \frac{1}{12}$. Since $\frac{1}{24} \neq \frac{1}{12}$, A and B are not independent.

Problem 4 (10 points)

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.

- Here, Probability of dementia in 75+ women ($P(D)$) = 0.05 and probability of healthy cases ($P(H)$) = 1- 0.05 = 0.95
  Probability of positive findings in women with dementia = Probability of Positive test given true dementia cases ($P(T^+|D)$) = 0.80
  Probability of positive findings in women with dementia = Probability of Positive test given non-dementia cases ($P(T^+|H)$) = 0.80

  A 75+ years women is already tested positive (T^+), and we want to find out the probability that she has dementia given she tested positive ($P(D|T^+)$).

  Through the requirement of conditional probability, we know that: ($P(D|T^+) = \frac{P(D \cap T^+)}{P(T^+)}$)
  Using multiplication rule, $P(D \cap T^+) = P(D).P(T^+|D)$, and using total law of probability testing positive is possible either when you have dementia and test positive ($D \cap T^+$) or when you don't have dementia and test positive ($H \cap T^+$).

  Replacing this is the main equation:

  $P(D|T^+) = \frac{P(D).P(T^+|D)}{P(D \cap T^+)+P(H \cap T^+)} = \frac{P(D).P(T^+|D)}{P(D).P(T^+|D)+P(H).P(T^+|H)} = \frac{0.05*0.80}{0.05*0.80+0.95*0.10} = 0.296.$

  Hence, the probability that she actually has dementia is 0.296.