FA-1 R Language

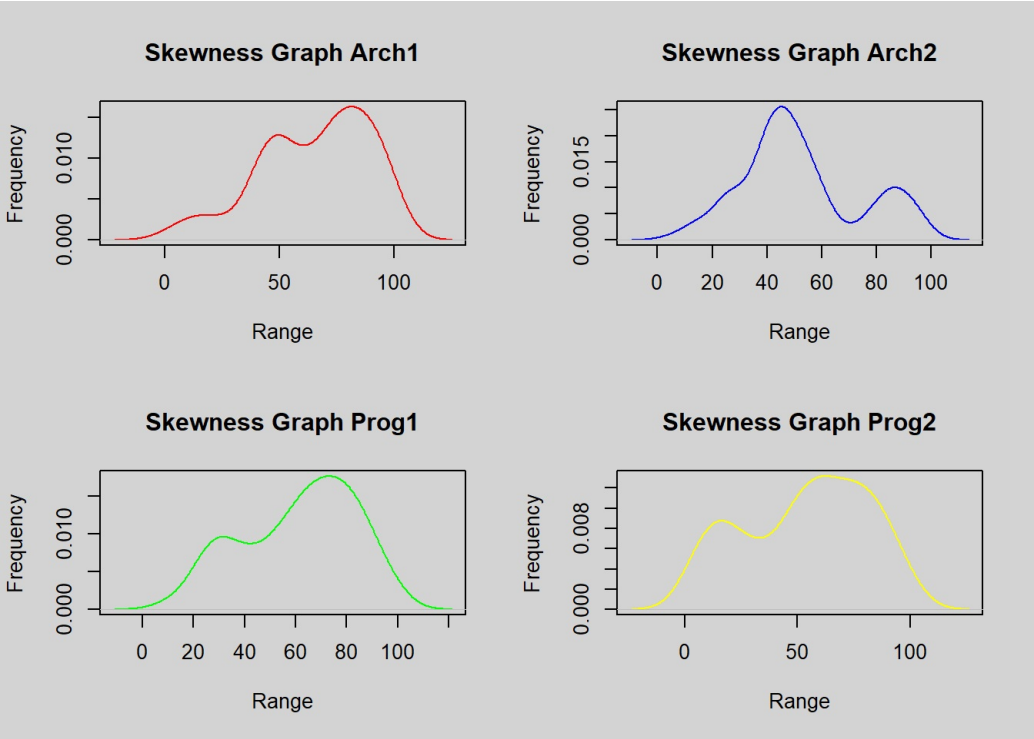
Espiritu, Joseph Raphael M.

2024-01-31

FA1 Questions

1.1 Write the skewness program, and use it to calculate the skewness coefficient of the four examination subjects in results.txt (results.csv). What can you say about these data?

There were four subjects consisting of 119 scores Arch1, Arch2, Prog1, and Prog2 and the NA rows of the students are Omitted




```
## Negative is Right Values and Positive is Left Values
```

```
## The Arch1 Examination Skewness =  -0.602  Most Skewed Data to the Right
```

```
## The Arch2 Examination Skewness =  0.46
```

```
## The Prog1 Examination Skewness =  -0.389
```

```
## The Prog2 Examination Skewness =  -0.274  Nearest Symmetrical Data
```

Analysis and Hypothesis

- Peaks on the data via the high scores on the four examination are evident on the graphs and affects the results of skewness.
- All the graphs still represent a large amount of right side peaks and values or above the mean of score meaning the examiners have been given adequate education or have study habits to pass the examination.
- On the contrary the examination questions and essays might have been to easy for the examiners and due to this many are able to pass over the points and attain high scores.
- The most Symmetrical Data which is Arch2 Examination is interesting to know what made it different from the other three which are teachers/professors or even the examiners themselves or the examination might have been made by a different person hence the difference.

1.2 Write the skewness program, and use it to calculate the skewness coefficient of the four examination subjects in results.txt (results.csv). What can you say about these data?

Pearson has given an approximate formula for the skewness that is easier to calculate than the exact formula given in Equation 2.1.

$$skew = \frac{3(mean - median)}{standard\ deviation}$$

Write a program to calculate this and apply it to the data in results.txt (results.csv). Is it a reasonable approximation?

```
##  arch1  prog1  arch2  prog2
## -0.715 -0.632  0.536 -0.389
```

Explained

Pearson Skewness Use Case:

- The Pearson skewness is a standardized measure of skewness and is often used when comparing the skewness of different distributions.
- It is less sensitive to extreme values compared to the sample skewness and if not often accurate and used for easily computed approximations of the skewness.
- It is suitable for normally distributed data.

Regular Skewness Use Case:

- The sample skewness is a measure of the asymmetry of a sample distribution.
- It is based on moments and is sensitive to extreme values in the data.
- It is often used when working with real-world data samples.

2.1 For the class of 50 students of computing detailed in Exercise 1.1, use R to form the stem-and-leaf display for each gender, and discuss the advantages of this representation compared to the traditional histogram

```
## Stem-and-leaf display for Females:
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   4 | 1348
##   5 | 15679
##   6 | 058
##   7 | 155889
##   8 | 01335
```

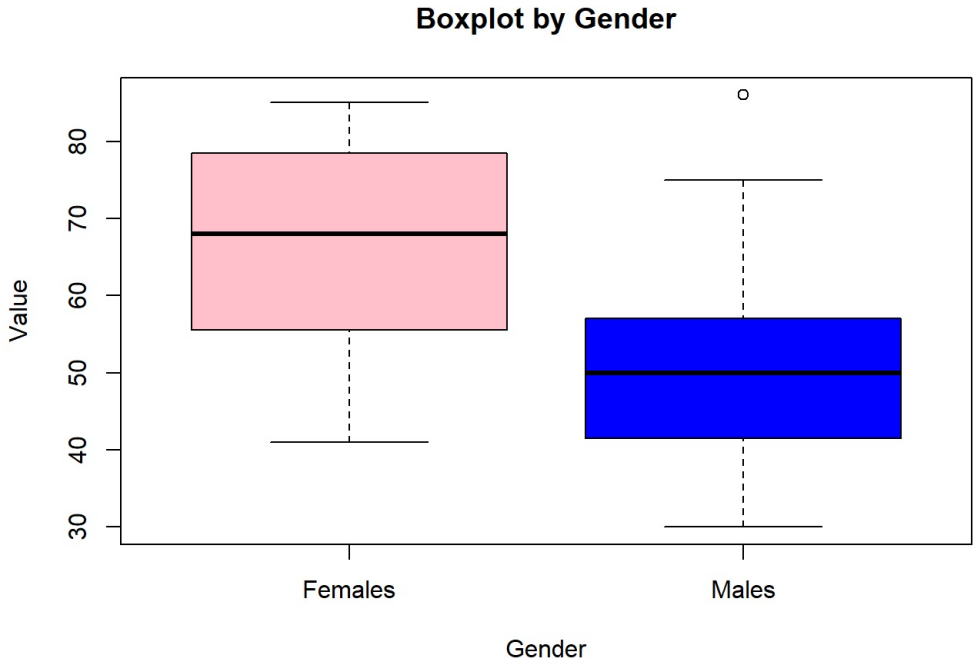
```
## Stem-and-leaf display for Males:
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   3 | 001257
##   4 | 1224899
##   5 | 01113668
##   6 | 4457
##   7 | 5
##   8 | 6
```

Advantages

- Preservation of Data: preserve the individual data points, allowing you to see the exact values in the data set.
- Useful for Small Data sets: particularly useful for small data sets where the granularity of individual data points is important.
- Flexibility: can be easily created by hand or with software tools, providing flexibility in their use.
- Sorting Information: naturally sort the data, making it easier to identify patterns and trends and then serve as both a display of individual data points and a summary of the distribution.

2.2 For the class of 50 students of computing detailed in Exercise 1.1, use R to construct a box-plot for each gender and discuss the findings.



Analysis

- General Distribution: The box plots suggest that the distribution of scores for females tends to be higher, as indicated by quartile 1 nearing the quartile 3 of the males' scores. The overall range of scores for females appears to be higher than that of males.
- Length and Height Differences: The noticeable length and height differences in the boxplots indicate that females, on average, have higher scores than males. The interquartile range (IQR) for females is wider, suggesting a greater spread of scores.
- Means and Quartiles: The means of the box plots reveal a substantial difference between genders. Females have a higher mean (around 68 to 70) compared to males (around 51 to 55), indicating a higher average score for females. The quartile positions also highlight the differences in the distribution of scores between genders.
- Minimum Values: The minimum value for females corresponds to the quartile 1 of the males' boxplots. This implies that even the lower range of female scores is comparable to the lower quartile of male scores.
- Outliers: The males' box plot indicates the presence of an outlier value above its whiskers. This outlier suggests the existence of a male student with an exceptionally high score compared to the rest of the male group.