Formative Assessment 1

Lindsy Rossel C. Masicat

2024-02-01

R Markdown

FIRST PROBLEM

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?

- Pearson has given an approximate formula for the skewness that is easier to calculate than the exact formula given in Equation 2.1.
- Write a program to calculate this and apply it to the data in results.txt (results.csv). Is it a reasonable approximation?

```
results = read.csv ("C:/Users/Lindsy/Desktop/MASICAT_R-LANGUAGE/results2.csv" , header = T)
results$gender <- as.factor(results$gender)
summary(results)
```

```
##  gender      arch1            prog1           arch2           prog2      
##  f: 19   Min.   :  3.00   Min.   :12.00   Min.   : 6.00   Min.   : 5.00  
##  m:100   1st Qu.: 46.75   1st Qu.:40.00   1st Qu.:40.00   1st Qu.:30.00  
##          Median : 68.50   Median :64.00   Median :48.00   Median :57.00  
##          Mean   : 63.57   Mean   :59.02   Mean   :51.97   Mean   :53.78  
##          3rd Qu.: 83.25   3rd Qu.:78.00   3rd Qu.:61.00   3rd Qu.:76.50  
##          Max.   :100.00   Max.   :98.00   Max.   :98.00   Max.   :97.00  
##          NA's   :3        NA's   :2       NA's   :4       NA's   :8      
```

First method using the code from the book of Jane Horgan:

```
skw <- function(j) {
  jbar <- mean(j, na.rm = T )
  add2 <- sum((j-jbar)^2, na.rm = T)
  add3 <- sum((j-jbar)^ 3, na.rm = T)
  skw <- (sqrt(length(j))* add3)/(add2^(1.5))
  skw}
cat("Skewness for arch1:", skw(results$arch1), "\n")
```

```
## Skewness for arch1: -0.5195368
```

```
cat("Skewness for prog1:", skw(results$prog1), "\n")
```

```
## Skewness for prog1: -0.3362643
```

```
cat("Skewness for arch2:", skw(results$arch2), "\n")
```

```
## Skewness for arch2: 0.4558875
```

```
cat("Skewness for prog1:", skw(results$prog1), "\n")
```

```
## Skewness for prog1: -0.3362643
```

Second method using Pearson’s coefficient of skewness:

```
skewness2 <- function(i) {
  mn <- mean(i, na.rm = T)
  mdn <- median(i, na.rm = T)
  standarddev <- sd(i, na.rm = T)
  skwness <- (3*(mn-mdn))/standarddev
  skwness}
skewness2(results$arch1)
```

```
## [1] -0.6069042
```

```
skewness2(results$prog1)
```

```
## [1] -0.643229
```

```
skewness2(results$arch2)
```

```
## [1] 0.5421286
```

```
skewness2(results$prog2)
```

```
## [1] -0.3562908
```

Analysis

Skewness Calculation Methods:

Method 1 (Jane Horgan’s Formula):

The first technique uses Jane Horgan’s formula to calculate skewness using the third standardized moment. It requires complex computations based on higher-order statistical moments.

Method 2 (Pearson’s Coefficient of Skewness):

The second technique, which employs Pearson’s coefficient of skewness, is an estimated formula that makes the calculation simpler. It operates on mean, median, and standard deviation.

Comparing the two methods:

Both methods may produce comparable results, but the method that employs Pearson’s coefficient of skewness is an approximation that trades some accuracy for simplicity.

If precision is critical, particularly in instances with non-normal distributions, the approach of the method from Jane Horgan’s book may be preferred.

The choice of approach is determined by the data’s qualities and the trade-off between accuracy and processing efficiency.It is critical to interpret skewness numbers in context with the data. Positive skewness suggests a right-skewed distribution, while negative skewness shows a left-skewed distribution. Understanding skewness is useful in determining the shape and symmetry of a data distribution.

SECOND PROBLEM

2. 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;
- construct a box-plot for each gender and discuss the findings.

```
females=c(57,59,78,79,60,65,68,71,75,48,51,55,56,41,43,44,75,78,80,81,83,83,85)
males=c(48,49,49,30,30,31,32,35,37,41,86,42,51,53,56,42,44,50,51,65,67,51,56,58,64,64,75)
```

STEM-AND-LEAF FOR FEMALES:

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

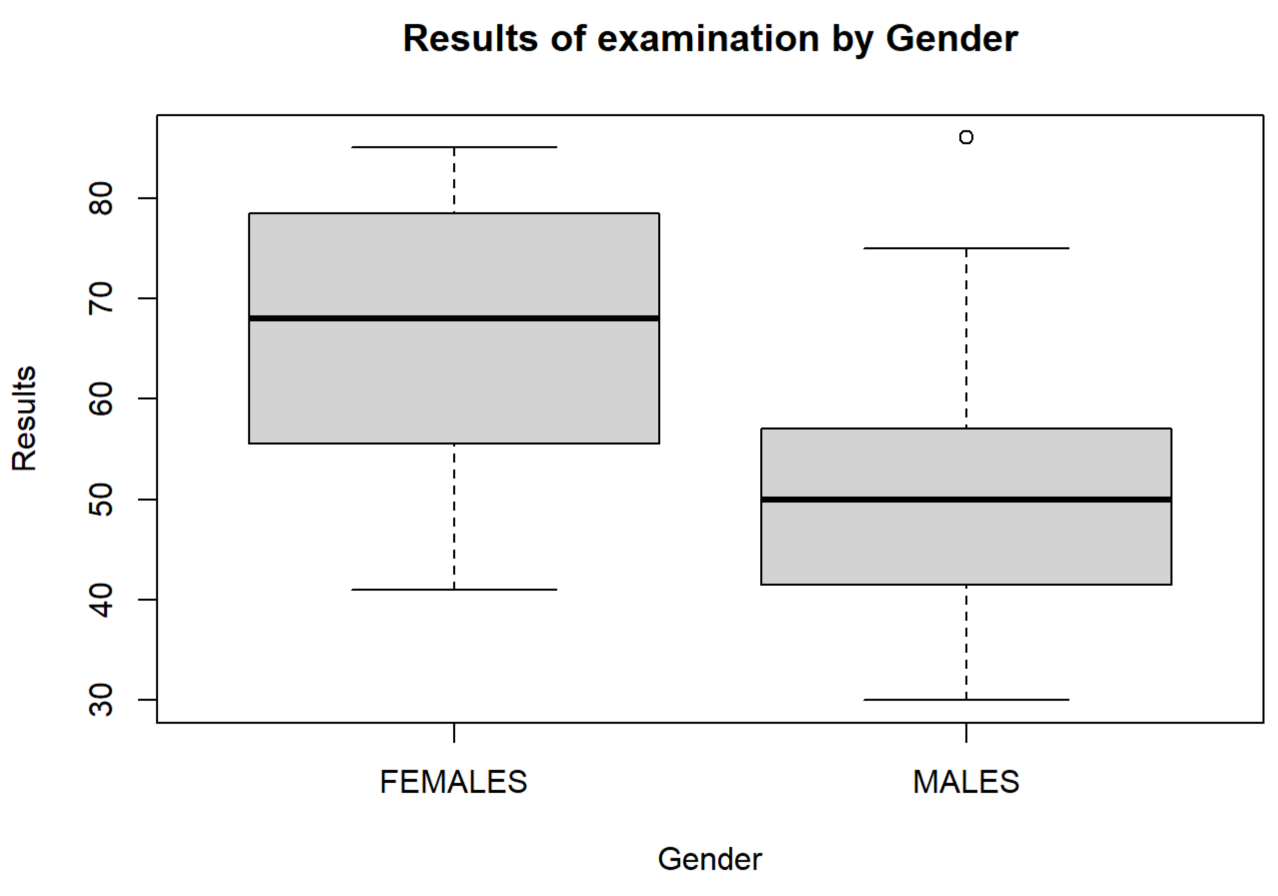
STEM-AND-LEAF FOR MALES:

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

Stem-and-leaf displays various advantages over histograms: A) They give increased granularity, allowing for a more thorough depiction of the distribution by separately displaying each data point. B) Stem-and-leaf plots preserve the original data values, making them easier to interpret because they keep exact values. C) Stem-and-leaf displays are compact, making them especially useful for smaller datasets, as they provide a brief yet relevant summary of data distribution without the possible loss of detail found in bigger, more sophisticated visualizations such as histograms.

BOXPLOT FOR BOTH GENDERS:

```
boxplot(females, males, names=c("FEMALES", "MALES"),
        xlab="Gender", ylab="Results",
        main="Results of examination by Gender")
```



The box plot visually captures the central tendency, which is principally represented by the median, and provides an immediate comprehension of the dataset’s midpoint and its dispersion as defined by the interquartile range. Furthermore, the box plot is an effective tool for identifying outliers, allowing for the simple discovery of individual points that fall outside the plot’s “whiskers”. This capacity is critical for recognizing any outliers that could have a substantial impact on the general distribution.