SEC1-GROUP 16-CESISTA, KJ-FA1

2025-02-02

Github Link: https://github.com/jelaenas/CESISTA-KJ-APM1110/blob/main/SEC1-GROUP%2016-CESISTA%2C%20KJ-FA1.Rmd

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?

```
results <- read.csv(params$filename, header = TRUE)

summary(results)
```

```
##     gender              arch1            prog1            arch2      
##  Length:119         Min.   :  3.00   Min.   :12.00   Min.   : 6.00  
##  Class :character   1st Qu.: 46.75   1st Qu.:40.00   1st Qu.:40.00  
##  Mode  :character   Median : 68.50   Median :64.00   Median :48.00  
##                     Mean   : 63.57   Mean   :59.02   Mean   :51.97  
##                     3rd Qu.: 83.25   3rd Qu.:78.00   3rd Qu.:61.00  
##                     Max.   :100.00   Max.   :98.00   Max.   :98.00  
##                     NA’s   :3        NA’s   :2       NA’s   :4      
##      prog2      
##  Min.   : 5.00  
##  1st Qu.:30.00  
##  Median :57.00  
##  Mean   :53.78  
##  3rd Qu.:76.50  
##  Max.   :97.00  
##  NA’s   :8      
```

```
skew <- function(x) {
  xbar <- mean(x, na.rm = TRUE)
  sum2 <- sum((x - xbar)^2, na.rm = TRUE)
  sum3 <- sum((x - xbar)^3, na.rm = TRUE)
  skew <- (sqrt(length(x)) * sum3) / (sum2^(3/2))
  skew
}

skew_arch1 <- skew(results$arch1)
skew_prog1 <- skew(results$prog1)
skew_arch2 <- skew(results$arch2)
skew_prog2 <- skew(results$prog2)

skew_arch1
```

```
## [1] -0.5195368
```

```
skew_prog1
```

```
## [1] -0.3362643
```

```
skew_arch2
```

```
## [1] 0.4558875
```

```
skew_prog2
```

```
## [1] -0.3125144
```

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?

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

```
pearsons <- function(x) {
  mean_value <- mean(x, na.rm = TRUE)
  median_value <- median(x, na.rm = TRUE)
  st_dv <- sd(x, na.rm = TRUE)

  skew_value <- (3 * (mean_value - median_value)) / st_dv
  skew_value
}

skew_pearsons_arch1 <- pearsons(results$arch1)
skew_pearsons_prog1 <- pearsons(results$prog1)
skew_pearsons_arch2 <- pearsons(results$arch2)
skew_pearsons_prog2 <- pearsons(results$prog2)

# Display the Pearson's skewness values
skew_pearsons_arch1
```

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

```
skew_pearsons_prog1
```

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

```
skew_pearsons_arch2
```

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

```
skew_pearsons_prog2
```

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

The data includes the gender of the students along with their marks for Arch1, Prog1, Arch2, and Prog2. NA values may indicate that those students were not able to take the examination for that course. For both functions Arch1, Arch2, and Prog2 all resulted in a negative value or a left-tail skew. This indicates that more students had scores on the higher-end. Meanwhile, for Prog1 it has a positive value or right-tail skew which indicates that more students had scores on the lower-end. Pearson's formula appears to be a reasonable approximation since when the values are compared they are not far off. This is because the data is not large and when taking the summary of the data the mean and median were close.

2. For the class of 50 students of computing detailed in Exercise 1.1, use R to

(a) form the stem-and-leaf display for each gender, and discuss the advantages of this representation compared to the traditional histogram;

```
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 Display of Female Students:

```
stem(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 of Female Students:

```
stem(males)
```

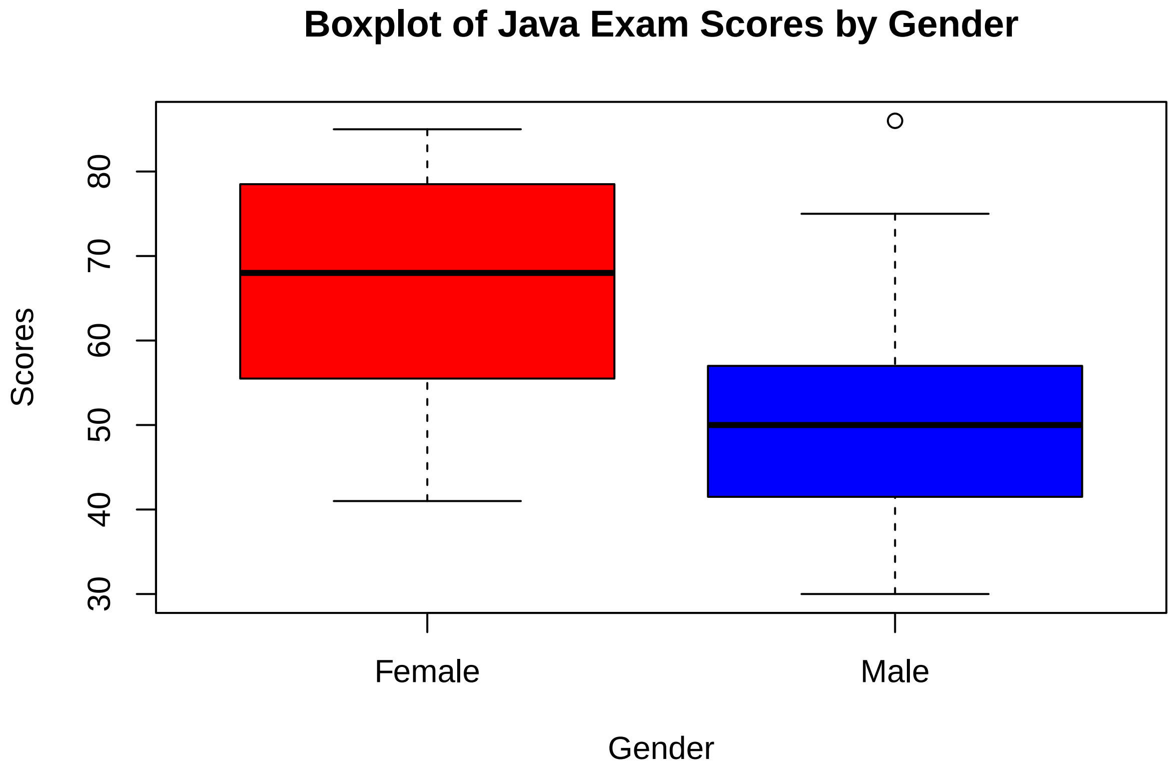
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   3 | 001257
##   4 | 1224899
##   5 | 01113668
##   6 | 4457
##   7 | 5
##   8 | 6
```

For this example, stem-and-leaf display is more appropriate as the data is discrete. Another advantage is that stem-and-leaf is faster compared to historagm, especially here as it would have to be plotted. Additionally stem-and-lead also shows individual data points.

(b) construct a box-plot for each gender and discuss the findings.

```
scores <- c(females, males)
gender <- rep(c("Female", "Male"), times = c(length(females), length(males)))
students <- data.frame(scores = scores, gender = gender)

boxplot(scores ~ gender, data = students, main = "Boxplot of Java Exam Scores by Gender", ylab = "Score
```



From the box-plot it looks as though the female students performed better than the male students. It shows that their median, range, maximum score, and minimum score are all higher compared to the male students. However it appears that the highest score came from a male student, however in the box plot it is considered an out-lier as it is only represented by a point.

Github Link: https://github.com/jelaenas/CESISTA-KJ-APM1110/blob/main/SEC1-GROUP%2016-CESISTA%2C%20KJ-FA1.Rmd