BAYBAYON, DARLYN B.-FA1

2024-02-01

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?

```
# read data
results <- read.table ("D:/SCHOOL FILES/3RD YR 2ND SEM/APM1110/results.txt", header = T)
# load data
head(results)
```

```
##   gender arch1 prog1 arch2 prog2
## 1      m    99    98    83    94
## 2      m    NA    NA    86    77
## 3      m    97    97    92    93
## 4      m    99    97    95    96
## 5      m    89    92    86    94
## 6      m    91    97    91    97
```

```
#load library
library(moments)

attach(results)
```

Scores in Architecture-1

```
mean_a1 <- mean(arch1, na.rm = TRUE)
median_a1 <- median(arch1, na.rm = TRUE)
sd_a1 <- sd(arch1, na.rm = TRUE)
skew_a1 <- (3*(mean_a1 - median_a1))/sd_a1
skew_a1_exact <- skewness(arch1, na.rm = TRUE)
cat("Approx Skewness: ", skew_a1, "\nExact Skewness: ", skew_a1_exact)
```

```
## Approx Skewness:  -0.6069042
## Exact Skewness:  -0.5129462
```

Scores in Architecture-2

```r
mean_a2 <- mean(arch2, na.rm = TRUE)
median_a2 <- median(arch2, na.rm = TRUE)
sd_a2 <- sd(arch2, na.rm = TRUE)
skew_a2 <- (3*(mean_a2 - median_a2))/sd_a2
skew_a2_exact <- skewness(arch2, na.rm = TRUE)
cat("Approx Skewness: ", skew_a2, "\nExact Skewness: ", skew_a2_exact)
```

```
## Approx Skewness:  0.5421286
## Exact Skewness:  0.44816
```

Scores in Programming-1

```r
mean_p1 <- mean(prog1, na.rm = TRUE)
median_p1 <- median(prog1, na.rm = TRUE)
sd_p1 <- sd(prog1, na.rm = TRUE)
skew_p1 <- (3*(mean_p1 - median_p1))/sd_p1
skew_p1_exact <- skewness(prog1, na.rm = TRUE)
cat("Approx Skewness: ", skew_p1, "\nExact Skewness: ", skew_p1_exact)
```

```
## Approx Skewness:  -0.643229
## Exact Skewness:  -0.3334265
```

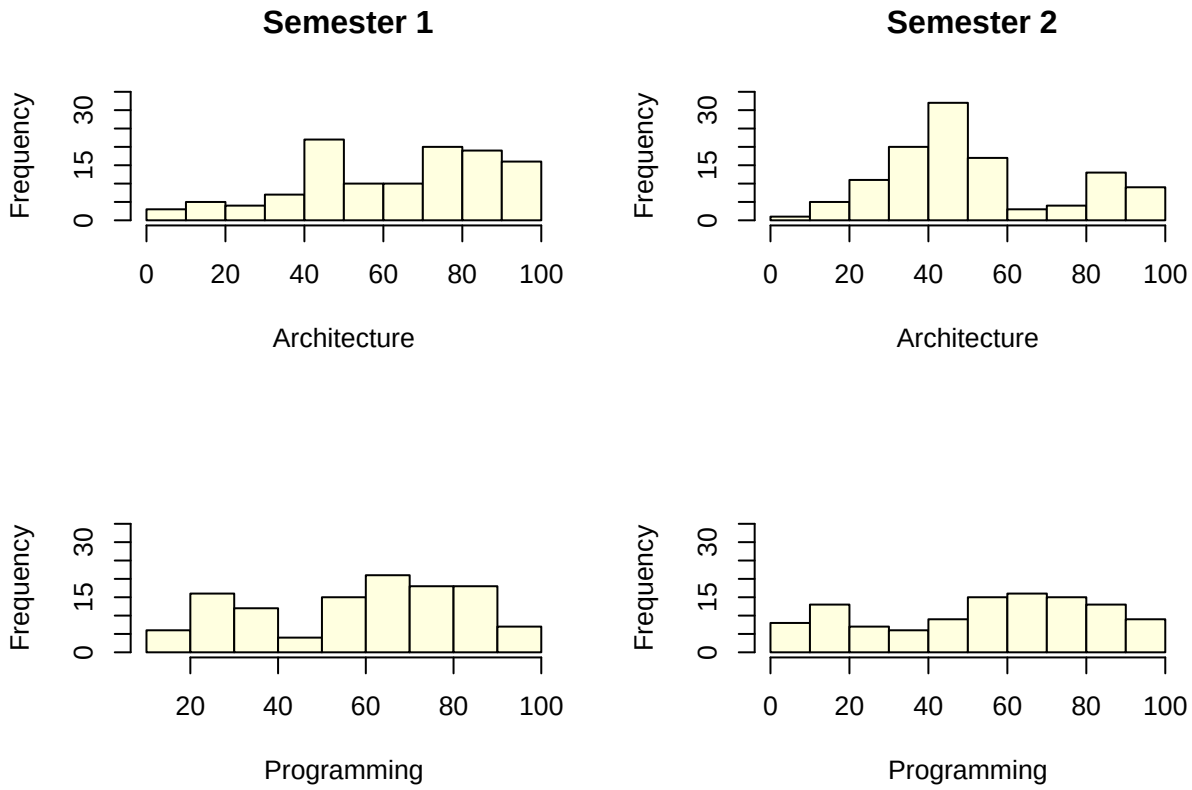
Scores in Programming-2

```r
mean_p2 <- mean(prog2, na.rm = TRUE)
median_p2 <- median(prog2, na.rm = TRUE)
sd_p2 <- sd(prog2, na.rm = TRUE)
skew_p2 <- (3*(mean_p2 - median_p2))/sd_p2
skew_p2_exact <- skewness(prog2, na.rm = TRUE)

cat("Approx Skewness: ", skew_p2, "\nExact Skewness: ", skew_p2_exact)
```

```
## Approx Skewness:  -0.3562908
## Exact Skewness:  -0.3018269
```

```r
par (mfrow = c(2,2))
hist(arch1, xlab = "Architecture",
main = " Semester 1", ylim = c(0, 35), col="lightyellow")
hist(arch2, xlab = "Architecture",
main = " Semester 2", ylim = c(0, 35), col="lightyellow")
hist(prog1, xlab = "Programming",
main = " ", ylim = c(0, 35), col="lightyellow")
hist(prog2, xlab = "Programming",
main = " ", ylim = c(0, 35), col="lightyellow")
```



Compared to the skewness coefficients obtained using R packages, the formula for approximation provided by Pearson yielded different but close values. The approximations were off by around 0.05 to 0.31. Though not exact, this can be reasonable enough to describe the asymmetry of a distribution by showing which direction it is skewed.

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;

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

```
# read scores

f_scores <- c(57, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43,
              44, 75, 78, 80, 81, 83, 83, 85)
m_scores <- 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 Plot (Female Students)

```
stem(f_scores)
```

```
##
##   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 Plot (Male Students)

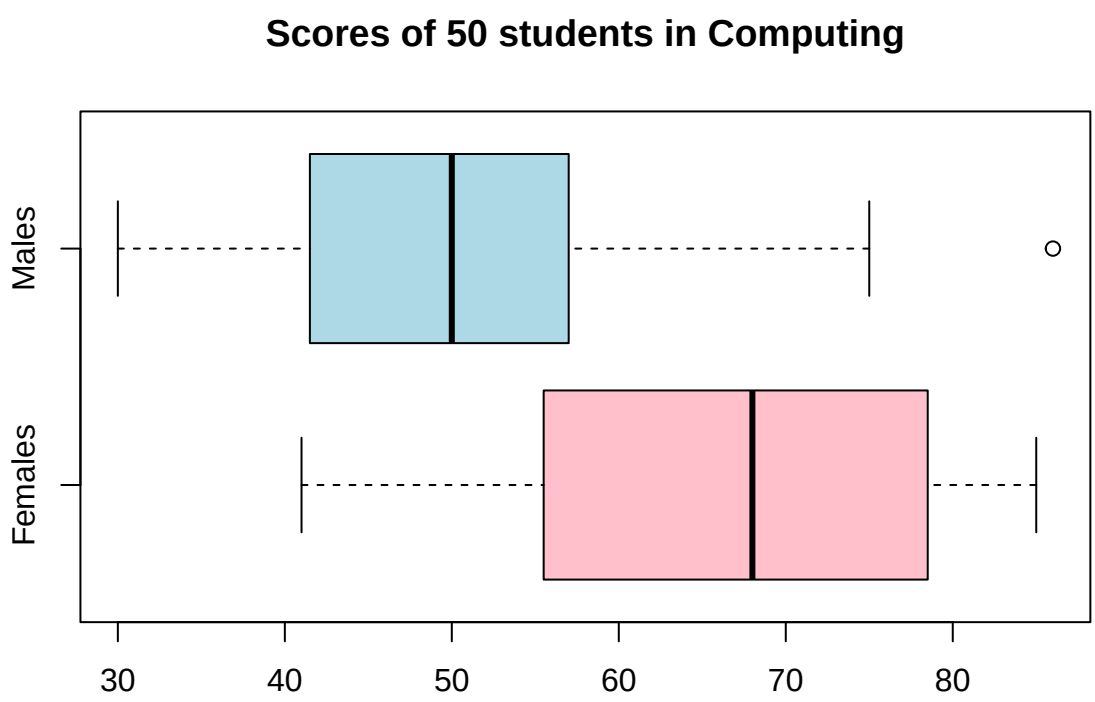
```r
stem(m_scores)
```

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

Using the stem and leaf plot allows us to see immediately the overview of the data and identify the shape of the distribution as well as the min and max values. Unlike the histogram, data points are preserved in the plot, allowing us to precisely identify exact values quickly. The stem and leaf plot is also more appropriate for smaller data sets such as in this example, while using histograms will be more suitable for visualizing and analyzing larger data sets.

Box Plot

```r
boxplot(f_scores, m_scores, horizontal = TRUE,
        main="Scores of 50 students in Computing",
        names = c("Females", "Males"), col = c("pink", "lightblue"))
```



As seen in the diagram, the box plot of scores for female students lie on a higher range compared to that of the male students. This indicates that most male students scored comparably lower than female students. The distribution of the scores of both male and female students are slightly negatively skewed. The female students also obtained more varied scores compared to the males. It is also notable that there is a single outlier in the male students who scored the highest among all the students.