# Formative Assessment 1

MERCADO, C & SINOCRUZ, A

2025-02-01

## Question 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?

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

$$\text{Skewness} \approx \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

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

### STEPS:

**Load the necessary libraries and data**

```
library(readr)
library(moments)
df <- read_csv("C:\\Users\\CONSUELO B. MERCADO\\OneDrive\\Documents\\r fas\\results.csv")
```

```
## Rows: 93 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): gender
## dbl (4): arch1, prog1, arch2, prog2
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

**We need to specify the file path for this. And then type**

```
library(moments)
```

**This is a required package for the skewness function that we will use for the exact skewness part.Next, we calculated the mean, median, and standard deviation excluding the gender column and the NA values. These are the needed values for Pearson's approximation**

```
sapply(df[2:5], mean, na.rm = TRUE)
```

```
##    arch1    prog1    arch2    prog2
## 68.61111 64.50549 39.40217 41.78409
```

```
sapply(df[2:5], median, na.rm = TRUE)
```

```
## arch1 prog1 arch2 prog2
##  74.5  68.0  41.5  50.5
```

```
sapply(df[2:5], sd, na.rm = TRUE)
```

```
##    arch1    prog1    arch2    prog2
## 23.00742 21.73240 24.10009 28.18268
```

**From the moments package, we use the skewness function to calculate the exact skewness.**

```
numeric_columns <- c("arch1", "prog1", "arch2", "prog2")

df_numeric <- df[, numeric_columns]

exact_skewness <- sapply(df_numeric, skewness, na.rm = TRUE)
exact_skewness
```

```
##      arch1      prog1      arch2      prog2
## -0.7788194 -0.6073172  0.3294140 -0.1683422
```

**Next, let's compute for the Pearson Skewness.**

```
pearson_skewness <- sapply(df_numeric, function(x) {(3 * (mean(x, na.rm = TRUE) -
median(x, na.rm = TRUE))) / sd(x, na.rm = TRUE)})
pearson_skewness
```

```
##      arch1      prog1      arch2      prog2
## -0.7678682 -0.4823911 -0.2611392 -0.9277944
```

**To see the comparison of the two, let's combine them using data frame.**

```
skewness_results <- data.frame(Subject = numeric_columns, Exact_Skewness =
exact_skewness, Pearson_Skewness = pearson_skewness)
skewness_results
```

```
##       Subject Exact_Skewness Pearson_Skewness
## arch1   arch1     -0.7788194       -0.7678682
## prog1   prog1     -0.6073172       -0.4823911
## arch2   arch2      0.3294140       -0.2611392
## prog2   prog2     -0.1683422       -0.9277944
```

**Analysis**

According to the research (Green et al., 2023), skewness is used to describe the lack of symmetries in a data distribution. In this problem, we calculate the *Exact Skewness* and *Pearson Approximation* for the four subjects in results.csv. The first two subjects, arch1 and prog1 showed a little difference in their exact and pearson values. However, in arch2, the exact skewness is positive while the other is negative. Also in the fourth subject, prog2, there is a big difference in their skewness. This shows that Pearson Skewness is reasonable approximation for data distribution but it still needs other statistical method just like the Exact Skewness to check its accuracy.

**Reference:**

Green, J. L., Manski, S. E., Hansen, T. A., & Broatch, J. E. (2023, January 1). *Descriptive statistics* (R. J. Tierney, F. Rizvi, & K. Ercikan, Eds.). ScienceDirect; Elsevier. https://www.sciencedirect.com/science/article/abs/pii/B9780128186305100831