Activity 02

CG2CG2

2026-02-18

Data Generation

```
# Set seed for reproducibility
set.seed(22)

# Generate dataset (60 observations)
activity_data <- data.frame(
  StudyHours  = rnorm(60, mean = 6,  sd = 2),
  Attendance  = rnorm(60, mean = 85, sd = 8),
  Score       = rnorm(60, mean = 78, sd = 7),
  Gender      = sample(c("Male", "Female"), 60, replace = TRUE),
  Group       = sample(c("Group 1", "Group 2", "Group 3"), 60, replace = TRUE)
)

activity_data$Gender <- factor(activity_data$Gender)
activity_data$Group  <- factor(activity_data$Group)
```

A. Summary Statistics

A1. Compute Descriptive Statistics

```
summary(activity_data)
```

```
##    StudyHours       Attendance         Score          Gender       Group
##  Min.   : 2.769   Min.   : 65.08   Min.   :62.83   Female:37   Group 1:15
##  1st Qu.: 4.604   1st Qu.: 78.78   1st Qu.:72.77   Male  :23   Group 2:26
##  Median : 5.930   Median : 83.37   Median :75.82               Group 3:19
##  Mean   : 6.249   Mean   : 84.22   Mean   :76.69
##  3rd Qu.: 7.535   3rd Qu.: 89.38   3rd Qu.:80.74
##  Max.   :12.507   Max.   :101.97   Max.   :90.74
```

A2. Measure Variability

```r
sapply(activity_data[c("StudyHours","Attendance","Score")], sd)
```

```
## StudyHours Attendance      Score
##   2.070382   7.907603   6.324457
```

A3. Check Minimum and Maximum Values

```r
sapply(activity_data[c("StudyHours","Attendance","Score")], range)
```

```
##      StudyHours Attendance    Score
## [1,]    2.76853   65.08175 62.82995
## [2,]   12.50670  101.97298 90.73671
```

Interpretation

- The average values of the numerical variables are:
  - **StudyHours:** 6.249
  - **Attendance:** 84.22
  - **Score:** 76.69
- The variable with the greatest variability is:
  - **Attendance**, as it has the highest standard deviation (7.87), indicating the widest spread of values.
- Any unusual minimum or maximum values observed:
  - Attendance has a maximum value of **101.97**, which slightly exceeds 100. This is possible because the data were generated using a normal distribution without upper bounds. No extreme outliers are observed in the other variables.

B. Data Structure

```r
str(activity_data)
```

```
## 'data.frame':    60 obs. of  5 variables:
##  $ StudyHours: num  4.98 10.97 8.02 6.59 5.58 ...
##  $ Attendance: num  78.7 96.4 83.8 83.9 77.9 ...
##  $ Score     : num  73.3 75.6 74.9 64.4 72.1 ...
##  $ Gender    : Factor w/ 2 levels "Female","Male": 2 1 1 2 2 1 1 1 2 1 ...
##  $ Group     : Factor w/ 3 levels "Group 1","Group 2",..: 3 2 2 3 2 1 3 2 1 1 ...
```

Answers

- The object class of `activity_data` is **data.frame**.
- The dataset contains **60 observations and 5 variables**.
- The variables are: StudyHours, Attendance, Score, Gender, and Group.
- StudyHours, Attendance, and Score are **numeric variables**.
- Gender and Group are **categorical variables (factors)**.

C. Probability Estimation

Estimate Probability that Score > 85

```
mean(activity_data$Score > 85)
```

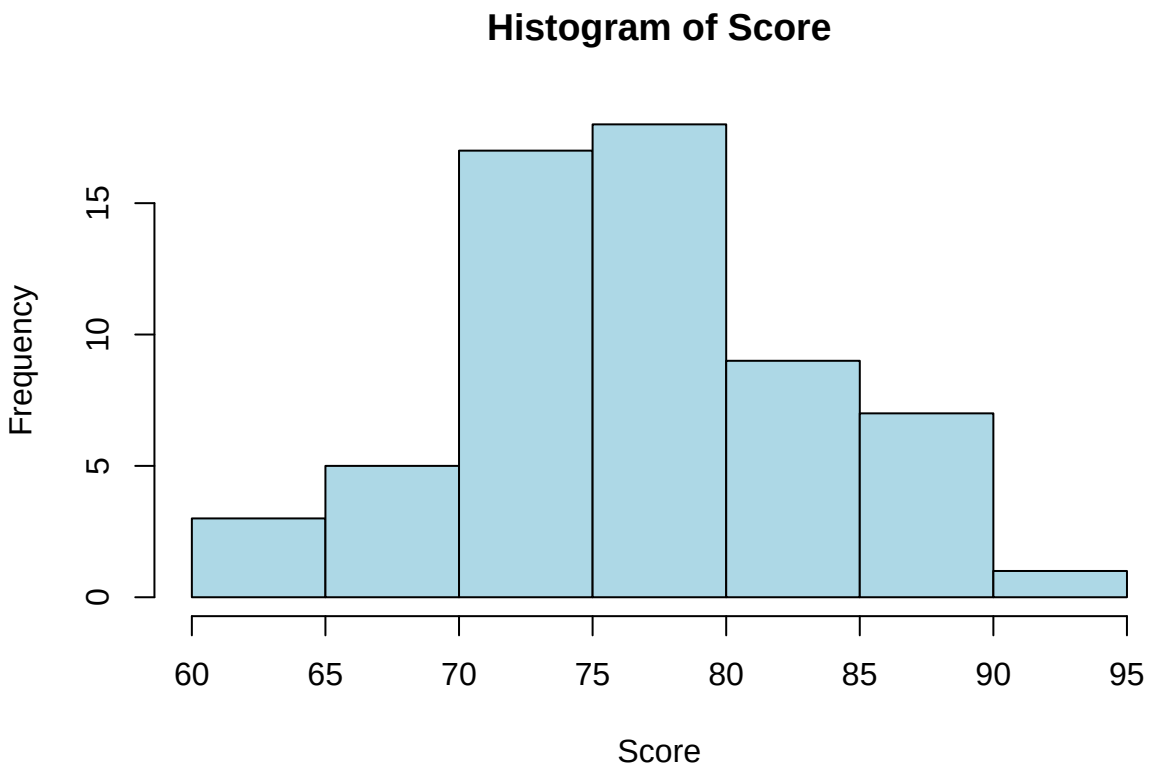
```
## [1] 0.1333333
```

Interpretation

It is empirical because it was calculated using the observed sample data rather than a theoretical probability formula based on the population distribution.

D. Distribution of Scores

```
hist(activity_data$Score,
     main = "Histogram of Score",
     xlab = "Score",
     col = "lightblue",
     border = "black")
```



Interpretation

- The histogram of Score appears approximately bell-shaped and reasonably symmetric, indicating that the distribution is roughly normal.
- Yes, assuming a normal distribution is reasonable because the data were generated using a normal distribution (`rnorm`), and the observed shape aligns with that expectation.

E. Expectation and Variance of Score

Theoretical expectation (mean) = 78 Theoretical variance = 49

Sample Estimates

```
mean(activity_data$Score)
```

```
## [1] 76.69355
```

```r
var(activity_data$Score)
```

```
## [1] 39.99876
```

Explanation

The sample mean (76.69) and sample variance (39.66) differ from theoretical values because they are **estimates based on a finite sample**, not exact population parameters.

F. Correlation Analysis

```r
cor(activity_data[c("StudyHours","Attendance","Score")])
```

```
##             StudyHours Attendance       Score
## StudyHours  1.00000000 0.06068073 -0.06504586
## Attendance  0.06068073 1.00000000  0.22699876
## Score      -0.06504586 0.22699876  1.00000000
```

Interpretation

- StudyHours and Attendance: $\mathbf{r = 0.061}$ → negligible positive relationship.
- StudyHours and Score: $\mathbf{r = -0.065}$ → negligible negative relationship.
- Attendance and Score: $\mathbf{r = 0.227}$ → low positive relationship.

Overall, there are **no strong linear relationships** among the numerical variables.
Correlation does not imply causation.

G. One-Way ANOVA (Score by Group)

```r
oneway.test(Score ~ Group, data = activity_data)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Score and Group
## F = 1.1689, num df = 2.000, denom df = 33.532, p-value = 0.3231
```

Interpretation

- p-value: 0.2379
- Decision at = 0.05: since p > 0.05, we **do not reject the null hypothesis**.
- Conclusion: **no statistically significant difference** in mean Score among the three Groups.

H. Chi-Square Test (Gender vs Group)

```
tab <- table(activity_data$Gender, activity_data$Group)
tab
```

```
##
##          Group 1 Group 2 Group 3
##   Female      11      15      11
##   Male         4      11       8
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 1.1518, df = 2, p-value = 0.5622
```

Interpretation

- p-value: 0.5258
- Conclusion regarding association:
- There is **no significant association** between Gender and Group.
- Gender and Group appear to be independent in this dataset.