# ASSIGNMENT- 3

## 1. Introduction

Statistics plays an important role in analyzing and interpreting numerical data. Two important measures used in statistics are:

Mean (μ) – Measure of central tendency

Standard Deviation (σ) – Measure of dispersion

While the mean gives the average value of a dataset, it does not explain how the data is distributed. Two datasets may have the same mean but very different spreads. Therefore, standard deviation is used to measure the variability of data around the mean.
In this study:

Mean (μ) = 55

$\sigma_a = 4$

$\sigma_\beta = 10$

$\sigma = 15$

2.1 Mean (μ)

The mean is calculated as:

Where:

X = individual observations

N = total number of observations

The mean represents the central value of the dataset.

2.2 Standard Deviation (σ)

Standard deviation is defined as the square root of the variance. It is calculated as:

It indicates how much individual data values differ from the mean.

Small σ  Data tightly packed around mean

Large σ  Data widely spread

## 3. Importance of Standard Deviation in Real Life

Quality control in industries

Stock market risk analysis

Academic performance comparison

Medical research studies

Business forecasting

For example: If two classes have the same average marks (55), but one class has $\sigma = 4$ and another has $\sigma = 15$, the first class performed more consistently.

## 4. Comparative Analysis of Given Data

Dataset A ($\sigma = 4$)

Very small variation

Data values are close to 55

High consistency

More reliable

Dataset B ($\sigma = 10$)

Moderate variation

Some deviation from mean

Moderate consistency

Dataset C ($\sigma = 15$)

High variation

Values widely spread

Less predictable

## 5. Graphical

## Explanation

Smaller σ Narrow and tall curve

Larger σ Wide and flat curve

If drawn graphically:

Dataset A Most concentrated around mean

Dataset B Moderately spread

Dataset C Widely dispersed

This visually proves Dataset A is more stable.

## 6. Statistical Comparison Table

| Source of Variation | DF | Sum of Squares | Mean Square | F Value |
|---|---|---|---|---|
| | | | | |
| Runs | 41 | 15.42 | 0.38 | 34.44** |
| Query | 49 | 46.25 | 0.94 | 86.46** |
| Error | 2009 | 21.93 | 0.01 | |
| Total | 2099 | 83.60 | | |

**Probability of F < .0001.

Table 1 presents the results of an Analysis of Variance (ANOVA) conducted to evaluate the effect of different sources of variation on the dataset. ANOVA is a statistical method used to compare group means and determine whether the differences observed are statistically significant. It separates total variation into different components and examines their individual contributions.

The table includes the following columns: Source of Variation, Degrees of Freedom (DF), Sum of Squares (SS), Mean Square (MS), and F-value.

The total variation in the dataset in divided into four components: Runs, Query, Error, and Total. Each source represents a different factor contributing to variability in the results.

The Degrees of Freedom (DF) represent the number of independent observations available for estimating variation. In this table:Runs has 41 degrees of freedom.

Query has 49 degrees of freedom.

Error has 2009 degrees of freedom.

Total degrees of freedom is 2099.

The total degrees of freedom equals the sum of the individual components, indicating proper partitioning of variation.

The Sum of Squares (SS) measures the total variability contributed by each source. It shows how much each factor contributes to the overall variation in the dataset.

Runs contributed 15.42 to the total variation.

Query contributed 46.25, which is the highest among all sources.

Error contributed 21.93.

The total sum of squares is 83.60.

This indicates that the Query factor explains the largest portion of variation in the data.

The Mean Square (MS) is calculated by dividing the Sum of Squares by the corresponding Degrees of Freedom:

From the table:

Mean Square for Runs = 0.38

Mean Square for Query = 0.94

Mean Square for Error = 0.01

The F-value is calculated by dividing the Mean Square of the factor by the Mean Square of the error:

The F-values obtained are:

Runs: 34.44

Query: 86.46

Both F-values are marked with significance indicators, and the table notes that the probability of F is less than 0.0001. This means the likelihood of obtaining such high F-values by chance is extremely low.

Since the p-value is less than 0.0001, the results are highly statistically significant. Therefore, the null hypothesis (which assumes no difference between groups) is rejected.

Comparing the F-values, Query (86.46) has a much higher F-value than Runs (34.44). This indicates that the Query factor has a stronger and more significant impact on the dataset than Runs.

Additionally, the Error mean square is very small (0.01), suggesting that unexplained variation is minimal. This indicates that the model fits the data well.

## 7. Interpretation

Since:

$4 < 10 < 15$

Dataset A has the least dispersion around the mean.

This indicates:

More stability

Less fluctuation

Higher reliability

Dataset C shows the maximum variability and is therefore less consistent.

## 8. Conclusion

After comparing the three datasets with equal mean ($\mu = 55$), it is concluded that Dataset A ($\sigma = 4$) is the best among the three.

Because:

It has the smallest standard deviation

It shows minimum variation

It ensures higher consistency and predictability

Dataset B shows moderate variability, while Dataset C shows the highest

dispersion.