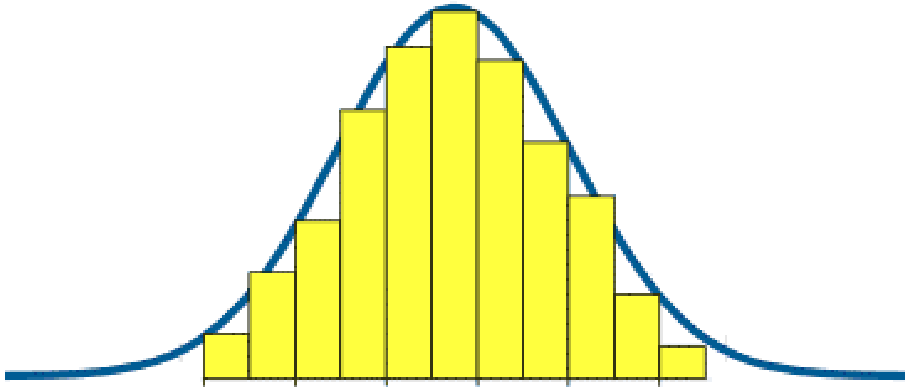
Probability and Statistics MAT 271E

PART 1

Introduction



Assist. Prof. Dr. Ümit KARADOĞAN

Course originally developed by : Prof. Dr. Mehmetçik BAYAZIT, Prof. Dr. Beyhan YEĞEN

INSTRUCTOR:

Assist. Prof. Dr. Ümit KARADOĞAN

karadoganum@itu.edu.tr

Civil Engineering /Geotechnical Department

COURSE OUTLINE:

- Introduction
- Elements of Probability Theory
- Distributions of Random Variables
- Multivariable Distributions
- Parameters of Random Variables, Bernoulli Trials
- Frequency Analysis of Samples, Parameter Estimation
- Probability Distribution Functions (Normal Distribution)
- Probability Distribution Functions (Other Distributions)
- Sampling Distributions
- Statistical Hypotheses
- Hypothesis Tests
- Regression Analysis

TEXTBOOK:

- **Bayazıt, M., Oğuz, B., Probability and Statistics for Engineers, Birsen Yayınevi, 1998**.

OTHER REFERENCES:

- Ross, S., A First Course in Probability, Prentice-Hall International, 1998.
- Walpore, E. W., Myers, R. H., Myers, S. L., Ye, K., Essentials of Probability and Statistics for Engineers and Scientists, Pearson, 2013.
- Murray R. Spiegel, Theory and Problems of Statistics, McGraw-Hill, 1961.
- Bulu, A., İstatistik Problemleri, Teknik Kitaplar Yayınevi, 1986.
- Weiss, N. A. Introductory Statisitics, Pearson, 2008

GRADING:

- **2 Midterm Exam** 40 %
- **2 Homework Assignments** 10 %
- **Final Exam** 50 %

All course material will be uploaded to **Ninova**.

Midterm Exam on

Using and understanding **probability** and **statistics** theories have become required skills in every profession and academic discipline.

" Probability theory is nothing but common sense reduced to calculation. "

P. S. LAPLACE

" Statistics is the grammar of science. "

K. PEARSON

Probability:

The chance that a given event will occur.

A branch of mathematics concerned with developing models to define the **likelihood of an event**.

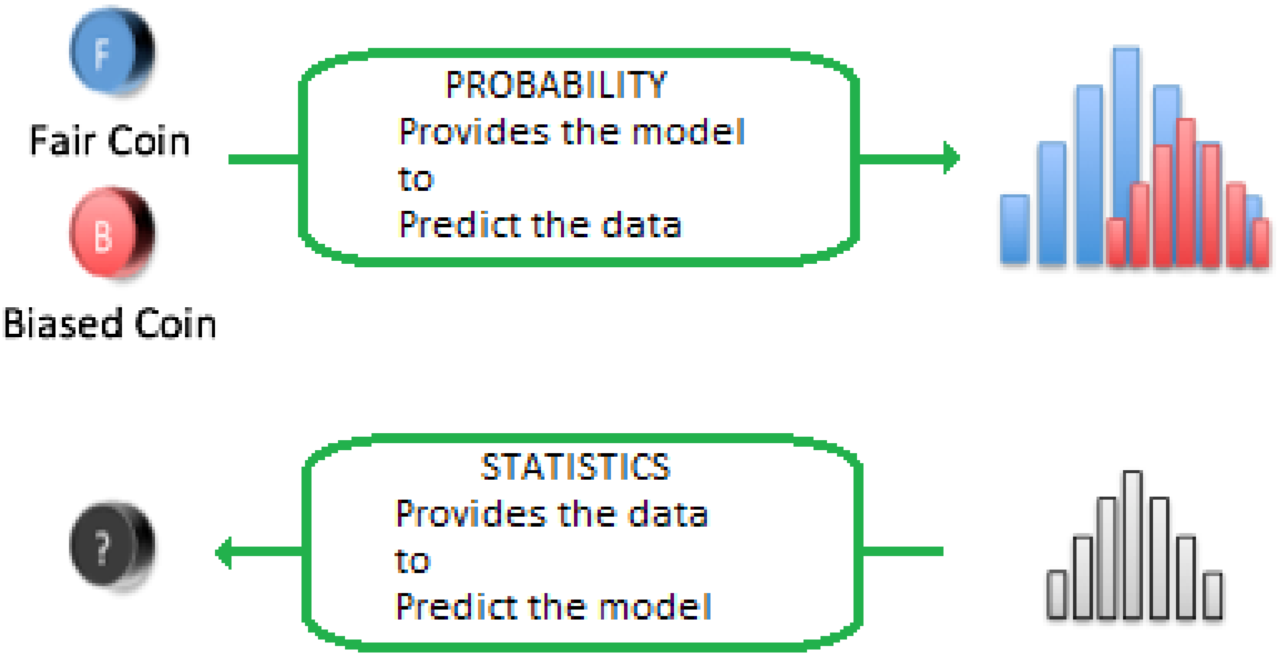
Statistics:

Statistics is based on the **collection, analysis, interpretation,** and **presentation of numerical facts (data)**.

A branch of mathematics dealing with fitting the available data to probability models and thus estimating the properties of the variable.

Probability vs. Statistics

Imagine you flip a coin:



Probability theory is devoted to the study of **uncertainty** and **variability.**
Probability quantifies how **certain/uncertain** we are about future events.

Statistics can be described as the study of how to make decisions in the face of **uncertainty** and **variability.**

Some examples of how **probability** and **statistics** shape your life when you don't even know it.

- Weather Forecasts
- Emergency Preparedness
- Predicting Catastrophes (earthquakes, floods...)
- Medical Studies
- Genetics
- Insurance
- Consumer Goods
- Stock Market
- Quality Control
- etc...

Engineers...

Engineers apply scientific laws and mathematics to design, develop, test, and supervise various products and services.

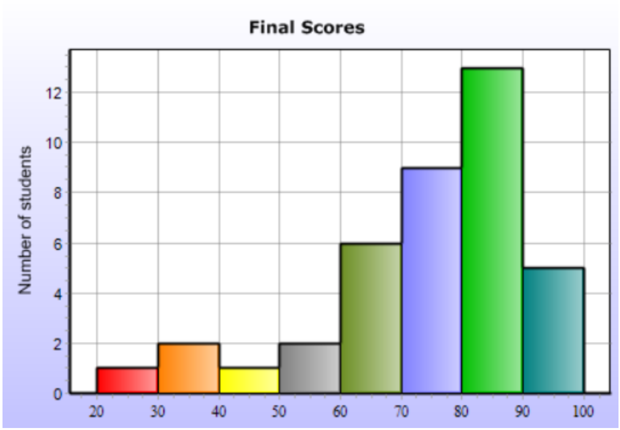
They perform experiments and **collect and analyze data** that can be used to explain relationships better and to reveal information about the quality of the products and services they provide.

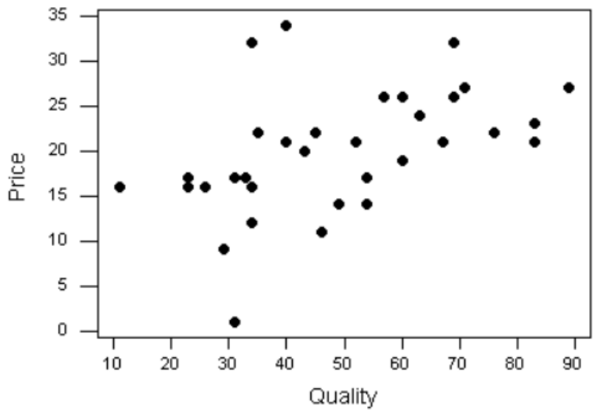
Engineers make use of fundamental **laws of probability** and **statistical results** to draw conclusions about scientific systems.

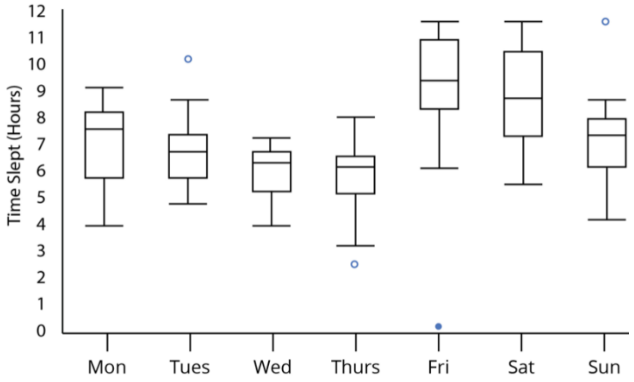
Information is gathered in the form of **sample data** or collections of **observations**.

To be able to better visualize and examine the nature of the available information, several types of tools are often used:

histograms, scatter plots, box plots etc...







How to approach a problem:

Deterministic Approach

- Deterministic approach assumes **certainty in all aspects**.
- A deterministic situation is the one in which the system parameters can be determined **exactly.** This is also called a situation of **certainty.**
- In **engineering systems in reality, such a system rarely exists**. There is usually some uncertainty associated.

Some Examples:

Predicting the amount of money in a bank account.

If you know the initial deposit, the amount of interest and the amount you spent, then: You can determine the amount left in the account.

Finding the acceleration (a) of a body of known mass (m) when a certain force (F) is exerted.

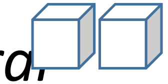
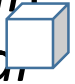
Using the Newton's second law, you can calculate the acceleration (F=ma) and always obtain the same output from the provided input.

Probabilistic Approach

- Probabilistic situation is called a situation of **uncertainty**.
- You know the likelihood that something will happen, but **<u>you don’t know if or when</u>** it is going to happen.

Some Examples:

Predicting **what** number will come up when you roll a die.

(Dice are commonly used to give examples in probability. Dice is plural  *, die is singular*  *)*

Predicting **when** number 6 will come up when rolling dice.

You know that in each roll, each number will come up with the probability of 1/6, but you **cannot exactly predict what will come up and when**.

Some of the commonly used statistical terms...

(which will again be mentioned in detail during the rest of the course)

Mean *(Arithmetic Mean)*

It is computed by **adding** all of the numbers in the data together and **dividing** by the number elements contained in the data set.

$$\bar{x} = \left(\sum_{i=1}^{N} x_i \right) / N$$

Example :

- Data Set = 2, 5, 9, 3, 5, 4, 7
- Mean = (2 + 5 + 9 + 7 + 5 + 4 + 3) / 7 = 5

Median

- Median is the **middle** number in a **sorted** list of numbers.

- How to calculate:

First reorder the data set **from the smallest to the largest.**

Find the **middle value**.

If there are 2 middles, add them up and **divide by 2**.

Example : Odd Number of Elements

- Data Set = 2, 5, 9, 3, 5, 4, 7
- Number of Elements in Data Set = 7
- Reordered = 2, 3, 4, 5, 5, 7, 9

  ^

- Median = 5

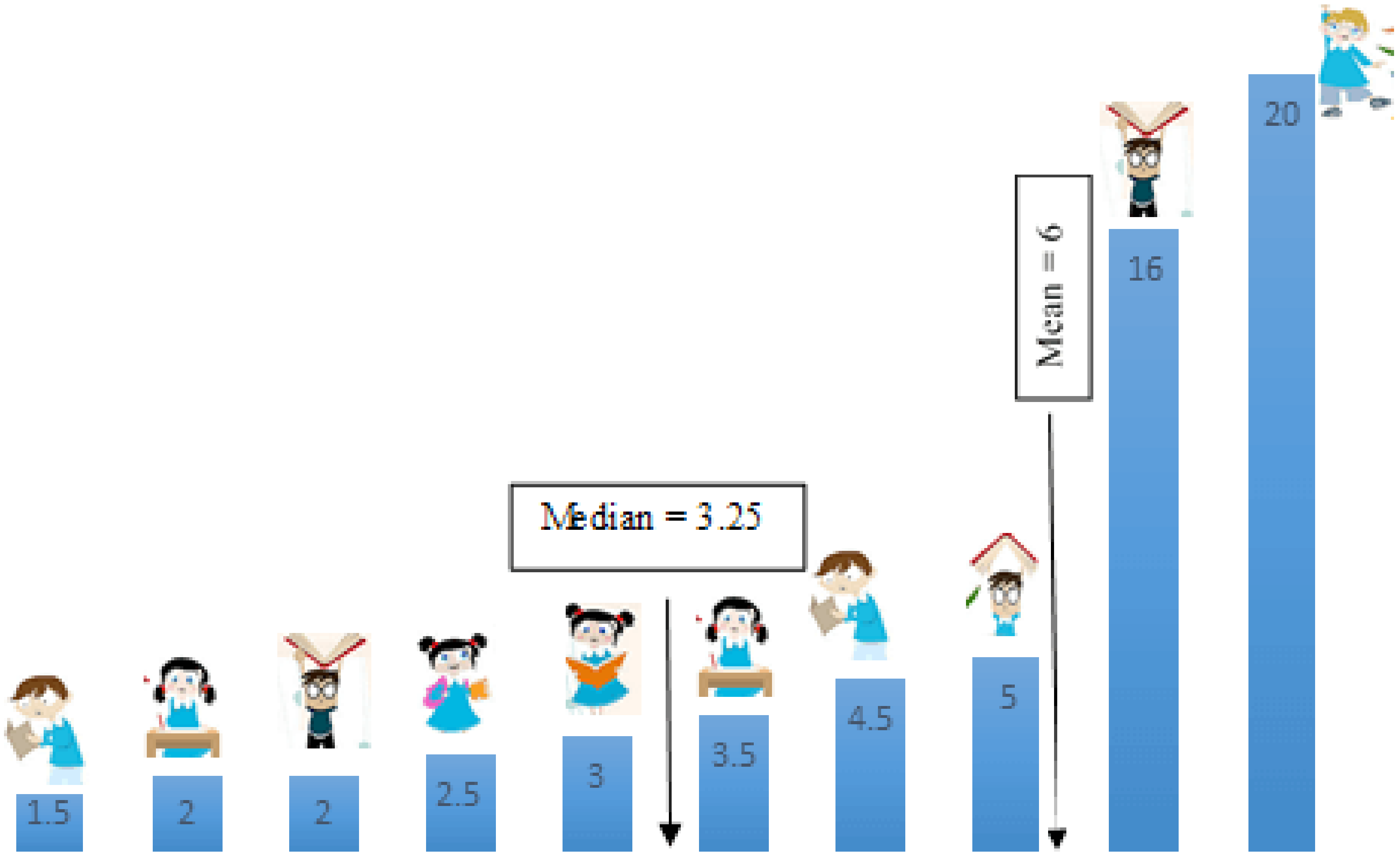
Example: Even Number of Elements

- Data Set = 2, 5, 9, 3, 5, 4
- Number of Elements in Data Set = 6
- Reordered = 2, 3, 4, 5, 5, 9

  ^ ^

- Median = (4 + 5) / 2 = 4.5

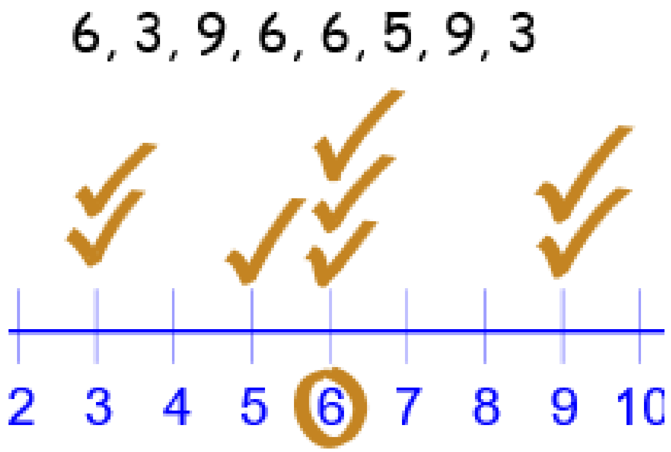
Mean vs. Median



Mode

- The **most frequently occurring** number (or member) found in a set of numbers (members).
- The **mode** is found by collecting and organizing data in order to count the frequency of each result.
- The result with the highest count of occurrences is the **mode** of the set.

Example:



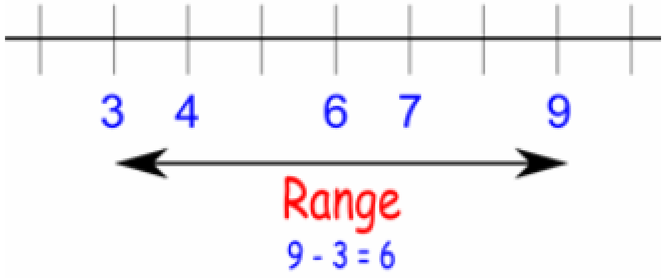
mode: 6

Range :

- The range for a data set is the **difference** between the largest value and smallest value contained in the data set.
- First **reorder** the data set from smallest to largest. Then **subtract** the first element from the last element (or just **subtract** the smallest from the largest).

Example:

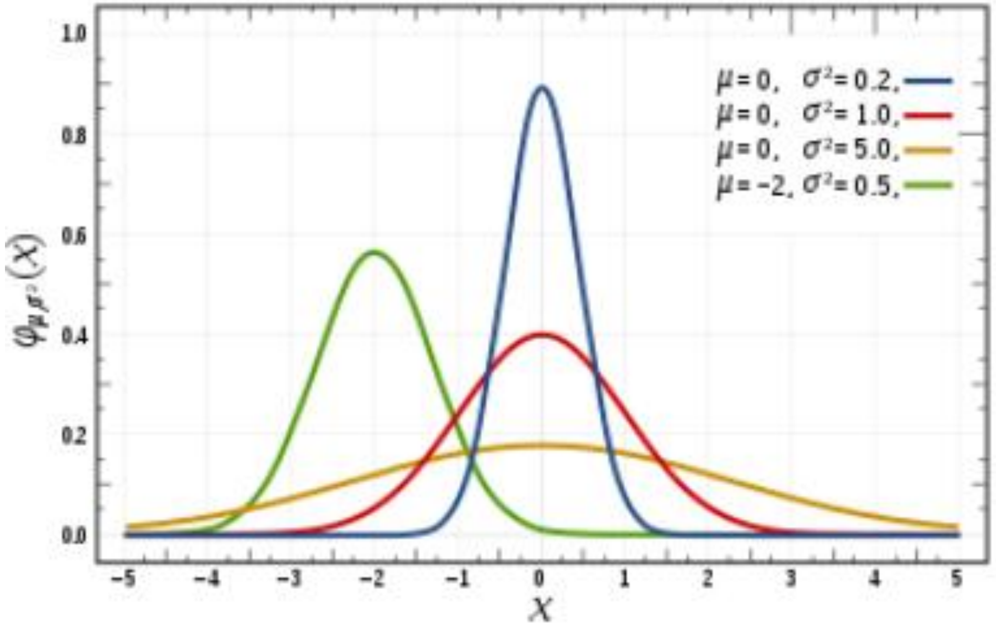
- Data Set = 7,6,4,9,3
- Reordered = 3, 4, 6, 7, 9
- Range = (9 - 3) = 6



Variance :

- The variance measures **how far each number in the set is from the arithmetic mean**.
- Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set.

$$Var\,(X) = \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]/N$$



Standard Deviation :

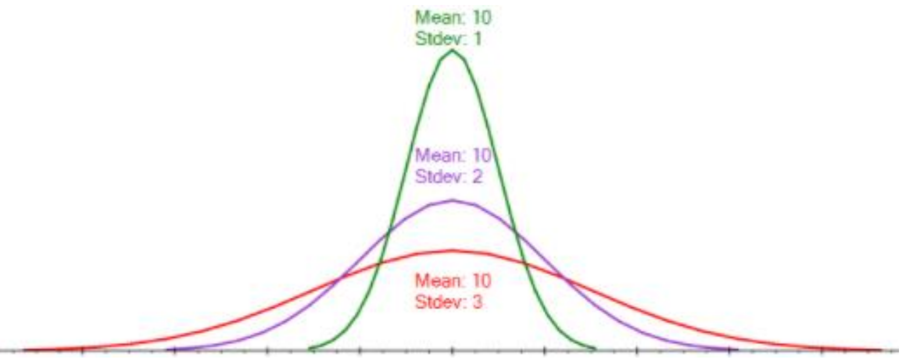
- Standard deviation is a measure of the **dispersion** of a set of data from its arithmetic mean.
- It is calculated as the **square root of variance.**
- If the data points are further away from the mean, there is higher **deviation** within the data set.

$$s_X = [Var(x)]^{1/2}$$

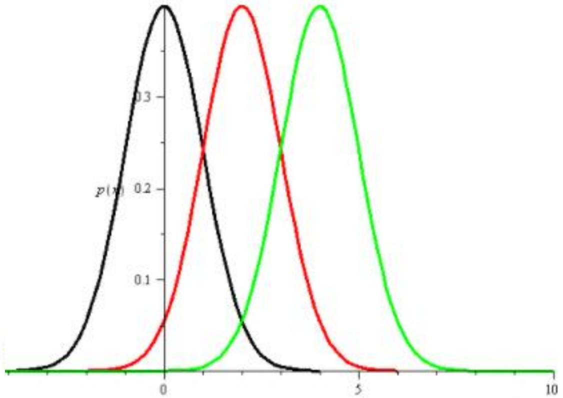
- It is always a **positive quantity**. It has the **same unit (dimension) as the data itself**.
- It would be equal to zero if all the data were equal.

Standard Deviation :

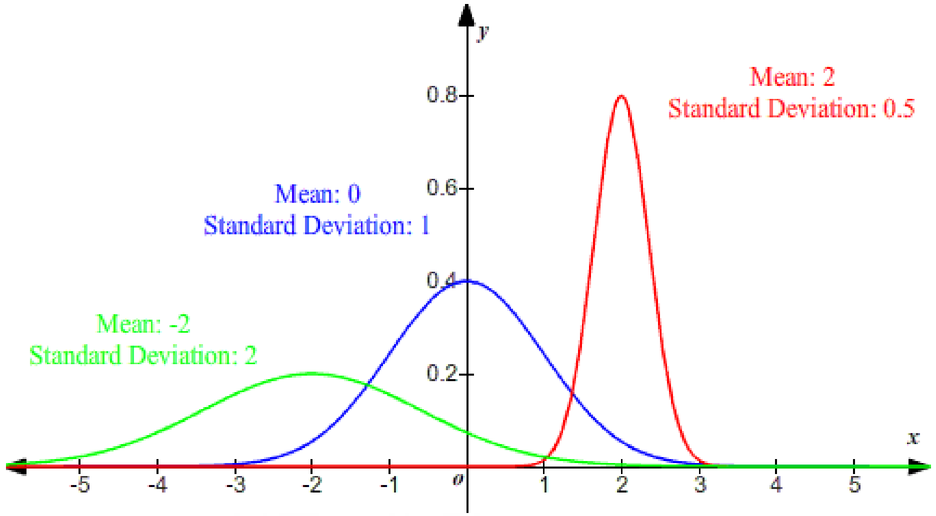
- Standard deviation would not be affected if all the data were increased or decreased by the same amount.



Same Means
Different Standard Deviations



Different Means
Same Standard Deviations



Different Means
Different Standard Deviations

Coefficient of Variation:

- Coefficient of variation is a statistical measure of the **dispersion** of data points in a data series around the mean.
- The coefficient of variation represents the **ratio of the standard deviation to the mean**.
- It is a useful term for comparing the degree of variation from one data series to another, **even if the means are drastically different** from one another.
- Coefficient of variation is **unitless (dimensionless)**.

$$C_{vX} = s_X/\bar{x}$$

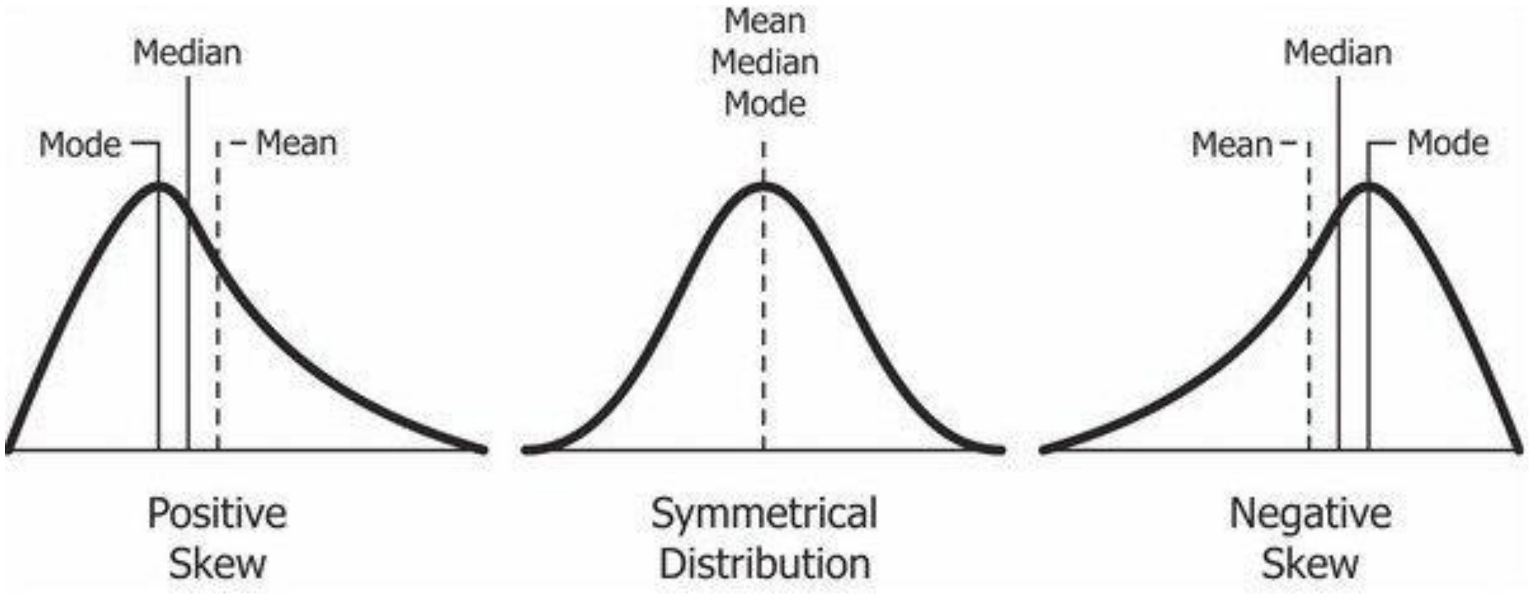
Coefficient of Skewness:

- Skewness can be measured by the **mean of the cubes of the differences** for each term $(x_i - \bar{x})^3$ divided by the **cube of the standard deviation s_x^3**.
- Coefficient of skewness is **unitless (dimensionless)**.
- It is expressed as either a **number smaller than 1 or a percentage**. It can **also** be **negative (negativelyskewed)**.
- If the data is **perfectly symmetrical**, the cube of a positive difference is canceled by the cube of an equal negative difference, and therefore the mean of the cubes (**skew**) is **zero**. Therefore, coefficient of skewness is also **zero**.

$$C_{sX} = \left[\sum_{i=1}^{N} (x_i - \bar{x})^3 / N \right] / s_X^3$$

Coefficient of Skewness:

- Skewness is a term in statistics which is used to describe **asymmetry**.
- Skewness can come in the form of **<u>negative</u> skewness or <u>positive</u> skewness**, depending on whether data points are skewed **to the left or to the right**.



Some introductory examples that show the significance of statistics in engineering problems...

Example:

Annual flow volumes at Keban station (dam) on Fırat river were measured from years 1937 to 1967.

Thirty one recorded values are given below (in 10^9 m^3).

Year	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946	1947
Flow	20.2	24.7	19.3	27.2	27.9	22.7	22.4	24.5	16.7	20.7	15.8

Year	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957
Flow	25.1	15.5	16.8	15.8	22.9	21.6	24.3	13.1	19.7	18.8

Year	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967
Flow	15.0	12.5	19.9	10.1	15.1	30.8	20.8	18.5	26.6	27.6

How can we arrange and analyze the data **to better understand it**?

- We can start by drawing a **Histogram** (step diagram):

First, we can classify the data into **class intervals** (of 3×10^9 m^3).

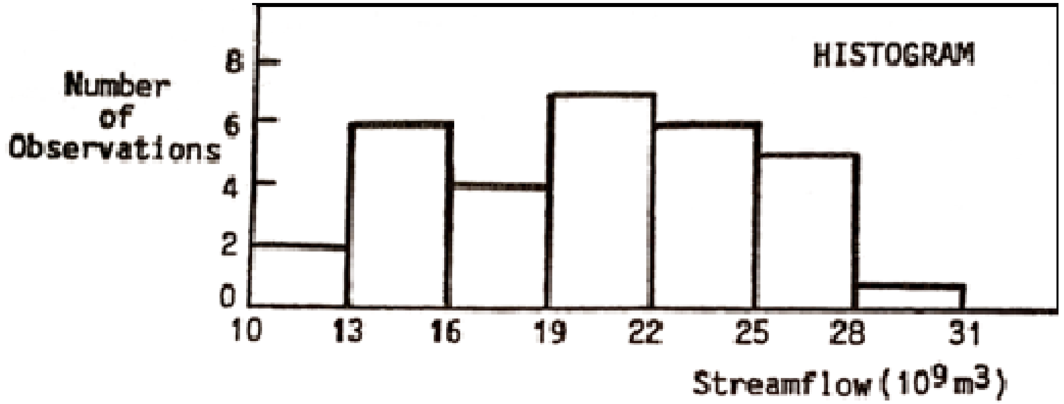
We can then plot the number of observations in each class interval as a horizontal line.

- The **histogram** clearly demonstrates the distribution of the observations (which cannot that clearly be extracted from tabulated values).

Year	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946	1947
Flow	20.2	24.7	19.3	27.2	27.9	22.7	22.4	24.5	16.7	20.7	15.8

Year	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957
Flow	25.1	15.5	16.8	15.8	22.9	21.6	24.3	13.1	19.7	18.8

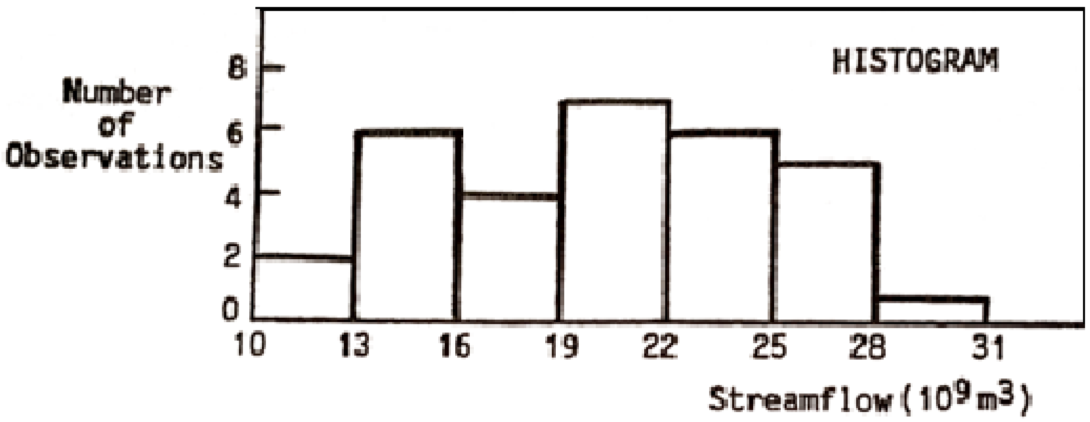
Year	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967
Flow	15.0	12.5	19.9	10.1	15.1	30.8	20.8	18.5	26.6	27.6



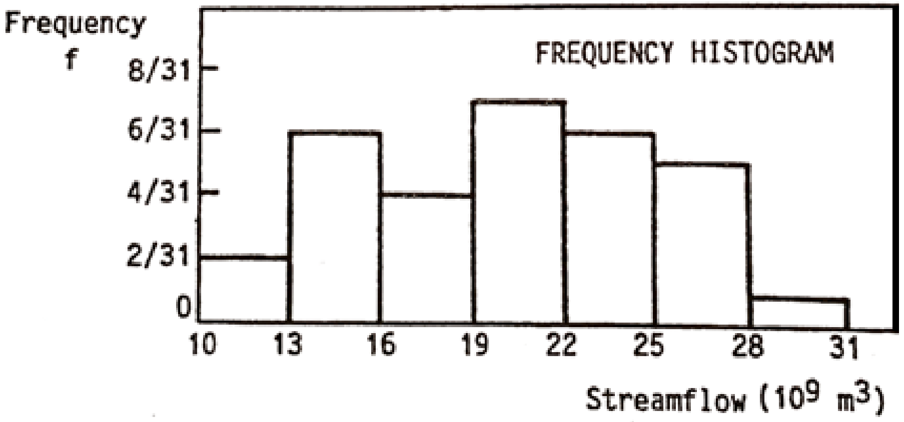
Histogram of Streamflows

For example, we can now easily find out that the streamflow was in the range of 19-22 ($\times 10^9$ m^3) for seven years.

- We can then draw a **Frequency Histogram** by plotting the **frequencies** (defined as the percentage of observations in a class interval) on the vertical axis to have a more meaningful graph.
- The y axis of the **Frequency histogram** is **unitless (dimensionless)**.



Histogram of Streamflows



Frequency Histogram of Streamflows

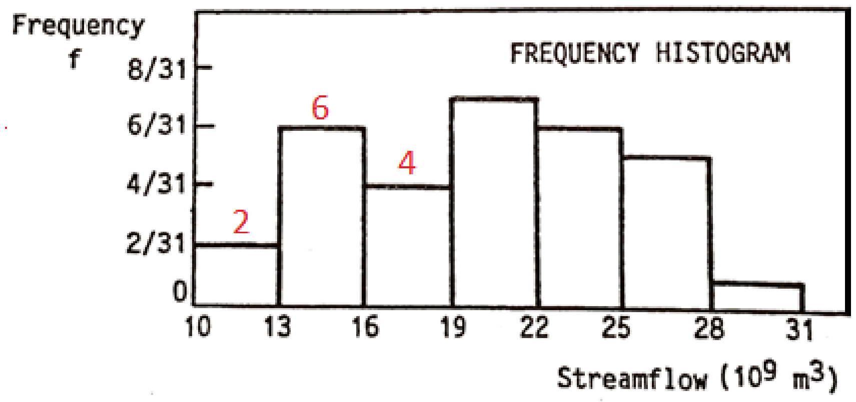
For example, now we can find out that the frequency in the range of 19-22 (x 10^9 m^3) equals 7/31 ≈ 0.23 = 23%.

- We can then also plot the **Cumulative Frequency Distribution** by **adding up** the frequencies in the class intervals **below** that value.
- For certain purposes, it may be required to estimate the frequency below a certain value.

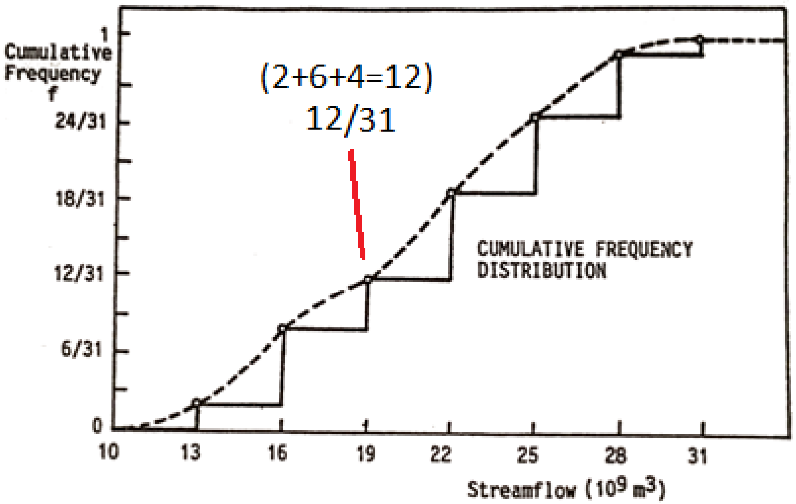
For example, then we can find out the frequency below 19 (x 10^9 m^3).

It is (2+6+4)/31=0.387.

In other words, 38.7 % of the observed flows were smaller than 19 (x 10^9 m^3).



Frequency Histogram of Streamflows



Cumulative Frequency Distributions of Streamflows

We can also summarize the information contained in the tabulated or sketched data by using statistical parameters.

- We can calculate the **Arithmetic Mean** around which the observations are scattered.

$$\bar{x} = \left(\sum_{i=1}^{N} x_i \right) / N \qquad \bar{x} = 20.3 \times 10^9 \text{ m}^3$$

- We can also determine the **median**.
- First, the data should be rearranged in an increasing sequence (in terms of streamflow).

Flow	Year
10,1	1961
12,5	1959
13,1	1955
15,0	1958
15,1	1962
15,5	1949
15,8	1947
15,8	1951
16,7	1945
16,8	1950
18,5	1965
18,8	1957
19,3	1939
19,7	1956
19,9	1960
20,2	1937
20,7	1946
20,8	1964
21,6	1953
22,4	1943
22,7	1942
22,9	1952
24,3	1954
24,5	1944
24,7	1938
25,1	1948
26,6	1966
27,2	1940
27,6	1967
27,9	1941
30,8	1963

16th value among 31

$$M_x = 20.2 \times 10^9 \text{m}^3$$

- It is not sufficient to characterize the set of data by only the mean and/or the median.
- It is necessary to use **at least one more parameter** to define the **uncertainty.**
- In several years, the streamflow is either smaller or higher than the mean. Therefore, **variance** which is a measure of the **scatter (dispersion)** of the data around the mean can be used.

$$Var\ (X) = \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]/N$$

$$Var\ (X) = 26 \times 10^{18}\ m^6$$

- The variance in this example has the unit of m^6 because it is calculated by taking the square of the variable.
- To obtain a parameter that has the **same unit** (dimension) as the data set, we should take the square root of the variance, which is the **standard deviation**.

$$s_X = [Var(x)]^{1/2}$$

$$s_x = 5.5 \times 10^9\ m^3$$

- We can also check the **coefficient of skewness** of the distribution.
- Coefficient of skewness is **zero** for **symmetrical data**.

$$C_{sX} = \left[\sum_{i=1}^{N} (x_i - \bar{x})^3 / N \right] / s_X^3$$

- For the Fırat River flows, coefficient of skewness is equal to 0.075 which means that the data is nearly symmetrical with a small positive skew.

If we want to make a **<u>comparison</u>**:

Example:

Annual flows of the Ceyhan river has:

a mean of $\bar{y}$=7.1x10^9 m^3, and

a standard deviation of s_Y=2.3x10^9 m^3.

Which river (Ceyhan or Fırat) has **<u>more variable</u> (more scattered)** flows?

To **compare** two variables, <u>a (unitless) dimensionless</u> parameter should be used.

Example:

The **Coefficient of Variation** (which is unitless) is the **ratio of the standard deviation** of a variable to its **mean**.

$$\boxed{C_{vX} = s_X/\bar{x}}$$

	Fırat River:	Ceyhan River:
mean	$\bar{x} = 20.3 \times 10^9 \text{ m}^3$	$\bar{y} = 7.1 \times 10^9 \text{ m}^3$
standard deviation	$s_X = 5.5 \times 10^9 \text{ m}^3$	$s_Y = 2.3 \times 10^9 \text{ m}^3$
coefficient of variation	$C_{vX} = \frac{5.5}{20.3} = 0.27$	$C_{vY} = \frac{2.3}{7.1} = 0.32$

$$\bar{x} > \bar{y}$$

$$s_X > s_Y$$

$$C_{vX} < C_{vY}$$

Flows of Ceyhan River show a **higher dispersion**.

Therefore, Ceyhan River has **more variable** flows

(even though its standard deviation is lower).