Chemometrics Essentials

Data Presentation and Statistics

Recommended books

1. James N. Miller, Jane C. Miller and Robert D. Miller: Statistics and Chemometrics for Analytical Chemistry, 7th ed., 2018
2. Zeev B. Alfassi, Zvi Boger and Yigal Ronen: Statistical Treatment of Analytical Data, eBook, 2009.
3. Douglas A. Skoog, Donald M. West, F. James Holler and Stanley R. Crouch: Fundamentals of Analytical Chemistry, 10th ed., 2021.

Data Presentation and Statistics

Presenting data - tables and plots, mean, median, mean absolute deviation, standard deviation, concepts of probability distribution functions (PDF), random PDF - normal (Gaussian); lognormal; rectangular; triangular; exponential; gamma, Discrete PDF - binomial and Poisson, derived PDF-student t; F; χ^2, shape of distribution - skewness and kurtosis, sampling distributions-central limit theorem.

Data

Definition

- **Data** are pieces of information collected through observation, measurement, or experimentation.
- They can be quantitative (numerical) or qualitative (descriptive).
- In scientific research, data are crucial for testing hypotheses, drawing conclusions, and validating theories.

Data

Quantitative Data

- **Discrete Data**
  Countable, e.g., No of student in class, no of car in a park
- **Continuous Data**
  Any value within a range, e.g., height of students, temp of a room

Qualitative Data

- **Nominal Data**
  categorical data without any order, e.g., hair color (blonde, brown, black) or types of fruit (apple, banana, orange)
- **Ordinal Data**
  categorical data with a meaningful order, e.g., customer satisfaction ratings (satisfied, neutral, dissatisfied) or education levels (high school, bachelor's, master's

Data Presentation

Definition

- **Data presentation** refers to the process of organizing and displaying data in a visual or textual format to make it easier to understand and interpret.
- This can involve using various tools and methods to highlight trends, patterns, and insights from the data.
- Effective data presentation helps communicate complex information clearly and concisely, facilitating informed decision-making and deeper analysis.

Some Common Methods of Data Presentation

Tables

Graphs and Charts

Infographics

Tables

Definition

- Tables are organized form for displaying data collected during experiments or research.
- They typically consist of rows and columns, where each column represents a variable, and each row represents a different observation or data point.

Key features of scientific data tables include:

1. **Title**: Each table should have a clear, descriptive title that explains the content and purpose of the table.
2. **Numbering**: If you have multiple tables, number them sequentially (e.g., Table 1, Table 2).
3. **Column Headings**: Use clear and informative headings for each column, including units of measurement where applicable.
4. **Consistency**: Ensure consistent formatting throughout the table, including the use of the same units and decimal places.
5. **Data Organization**: Arrange data logically, typically with the independent variable in the leftmost column and dependent variables in subsequent columns.
6. **Footnotes and Annotations**: Include any necessary footnotes or annotations to explain abbreviations, symbols, or specific data points.

Tables

Definition

- Tables are organized form for displaying data collected during experiments or research.
- They typically consist of rows and columns, where each column represents a variable, and each row represents a different observation .

Key features of scientific data tables include:

1. **Title**: Each table should have a clear, descriptive title that explains the content and purpose of the table.
2. **Numbering**: If you have multiple tables, number them sequentially (e.g., Table 1, Table 2).
3. **Column Headings**: Use clear and informative headings for each column, including units (quotient of a quantity and a unit) of measurement where applicable.
4. **Consistency**: Ensure consistent formatting throughout the table, including the use of the same units and decimal places.
5. **Data Organization**: Arrange data logically, typically with the independent variable in the leftmost column and dependent variables in subsequent columns.
6. **Footnotes and Annotations**: Include any necessary footnotes or annotations to explain abbreviations, symbols, or specific data points.

Tables

Example:

Table 1 Measured length, width and height of a given solid[1]

No. of observation	Length, l/cm	Width, w/cm	Height, h/cm
1	3.86	2.51	1.22
2	3.92	2.6	13
3	3.88	2.55	1.27

[1]Dimensions are measured using meter scale.

Table 2 Measured volume of a given solid[2]

No. of observation	Volume without solid, V/m^3	Volume with solid, V/m^3	Volume of solid, V/m^3
1	10.1	12.2	2.1
2	10	12	2
3	10.2	12.4	2.2

[2]Volumes are measured using cylinder by displacement method.

Chart/Graph/Plot

Chart

- A **chart** is a visual representation of data designed to make the information easier to understand.
- Charts can take various forms, such as bar charts, pie charts, and line charts.
- They are often used in business, education, and media to present data in a clear and concise manner.

Plot

- A **plot** is a type of chart that specifically represents data points on a coordinate system.
- Plots are used to show the relationship between two or more variables. Common types of plots include scatter plots, line plots, and histograms.
- They are frequently used in scientific and statistical analysis to identify trends, patterns, and correlations.

Graph

- A **graph** is a broader term that encompasses both charts and plots.
- It refers to any visual representation of data that uses points, lines, bars, or other symbols to convey information.

Chart/Graph/Plot

Chart/Graph/Plot

1	**Title**	Clearly state what the graph represents, such as “The Relationship Between Temperature and Reaction Rate.”
2	**Axes**	• **X-Axis (Horizontal)**: Represents the independent variable. Label it with the variable name and units. • **Y-Axis (Vertical)**: Represents the dependent variable. Label it with the variable name and units (quotient of a quantity and a unit) . • **Scale**: Choose an appropriate scale for both axes to ensure the data is spread out and the graph is as large as possible
3	**Data Points**	• **Plot Accurately**: Mark each data point precisely according to its coordinates. • **Line of Best Fit**: If applicable, draw a line of best fit to show trends. Do not connect the dots directly unless it’s a time series
4	**Labels and Units**	• **Axis Labels**: Include the name of the variable and the unit of measurement (e.g., “Temperature (°C)”). • **Legend**: If multiple data sets are present, use a legend to differentiate them
5	**Clarity and Neatness**	• **Use Pencil and Ruler**: Draw axes, bars, and lines neatly. • **Avoid Clutter**: Ensure the graph is easy to read by avoiding overlapping labels and data points
6	**Review**	• **Check for Errors**: Verify that all data points are plotted correctly and that the graph accurately represents the data. • **Adjust if Necessary**: Make any necessary adjustments to improve clarity and accuracy

Chart/Graph/Plot

1	**Title**	Clearly state what the graph represents, such as “The Relationship Between Temperature and Reaction Rate.”
2	**Axes**	• **X-Axis (Horizontal)**: Represents the independent variable. Label it with the variable name and units. • **Y-Axis (Vertical)**: Represents the dependent variable. Label it with the variable name and units (quotient of a quantity and a unit) . • **Scale**: Choose an appropriate scale for both axes to ensure the data is spread out and the graph is as large as possible
3	**Data Points**	• **Plot Accurately**: Mark each data point precisely according to its coordinates. • **Line of Best Fit**: If applicable, draw a line of best fit to show trends. Do not connect the dots directly unless it’s a time series
4	**Labels and Units**	• **Axis Labels**: Include the name of the variable and the unit of measurement (e.g., “Temperature (°C)”). • **Legend**: If multiple data sets are present, use a legend to differentiate them
5	**Clarity and Neatness**	• **Use Pencil and Ruler**: Draw axes, bars, and lines neatly. • **Avoid Clutter**: Ensure the graph is easy to read by avoiding overlapping labels and data points
6	**Review**	• **Check for Errors**: Verify that all data points are plotted correctly and that the graph accurately represents the data. • **Adjust if Necessary**: Make any necessary adjustments to improve clarity and accuracy

Plotting Scientific Graph

- Plotting
  a) Calculate N_{SDX} and N_{SDY} using the following relation

$$N_{SDX} = \frac{X_i - X_{NL}}{X_{SD}}$$

$$Y_{SDY} = \frac{Y_i - Y_{NL}}{Y_{SD}}$$

where,

X_i and Y_i are the values of X and Y coordinates of i-th experimental data point.

X_{NL} and Y_{NL} are the nearest lower labelled values of X and Y coordinates from i-th experimental data point.

a) Place i-th point at N_{SDX} and N_{SDY} smallest square unit right and up from X_{NL} and Y_{NL} respectively.

Let us consider the following experimental data point

X/mole	0.1212	0.1398	0.1584	0.1771	0.1956	0.2142	0.2328	0.2514	0.2702	0.2886	0.3072
Y/cm^3	52.03	52.62	53.05	53.74	54.07	54.71	55.31	55.71	56.43	56.74	57.38

$$X_{SD} = \frac{0.3072 - 0.1212}{70} = 0.002657 = 2.657 \times 10^{-3} \approx 3.0 \times 10^{-3}$$

$$Y_{SD} = \frac{57.38 - 52.03}{50} = 0.107 = 1.07 \times 10^{-1} \approx 1.5 \times 10^{-1}$$

Plotting Scientific Graph

Intersection point of X and Y axes

$$(0.1212 - 2 \times 3.0 \times 10^{-3}, 52.03 - 2 \times 6.0 \times 10^{-2}) \equiv (0.1152, 51.91)$$

Consider a point $(X_i, Y_i) \equiv (0.1584, 53.05)$

$$N_{SDX} = \frac{0.1584 - 0.1452}{0.003} = 4.4\ SD \qquad N_{SDY} = \frac{53.03 - 51.73}{0.15} = 8.67\ SD$$

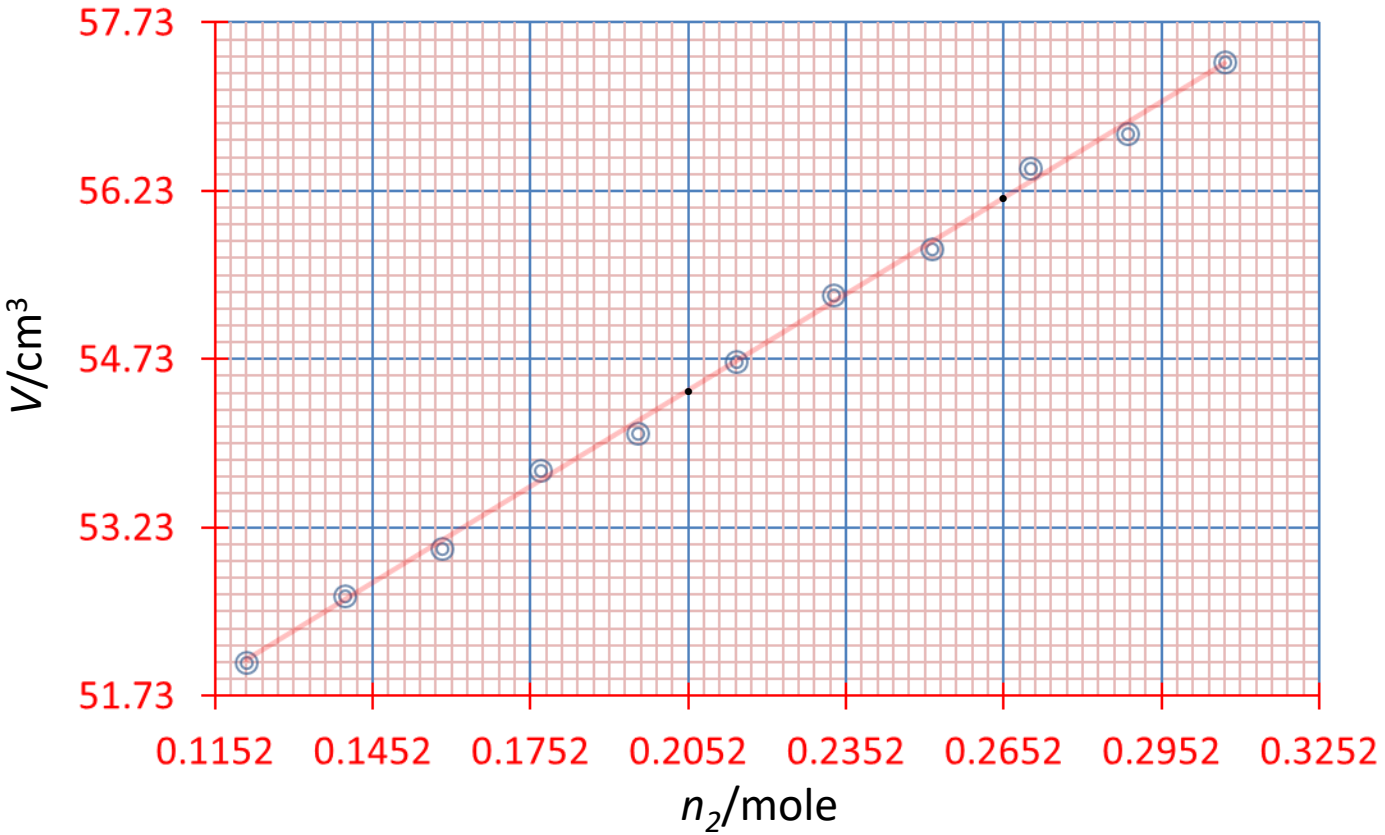


Figure X.XX Volume of ethanol-water solution against mole number of ethanol

1. Dot Plot

•**Description**: A dot plot represents data points along a number line, with each dot representing a value or frequency.

•**Merits**:

- Simple to construct and interpret, especially for small data sets.
- Clearly displays clusters, gaps, and outliers.
- Allows for easy comparison between multiple data sets.

•**Demerits**:

- Inefficient for large data sets, as overcrowding can make individual values hard to distinguish.
- Does not work well for continuous data or large ranges of values.

2. Stem-and-Leaf Plot

•**Description**: A stem-and-leaf plot splits data values into a "stem" (typically the leading digit(s)) and a "leaf" (the trailing digit). It’s useful for displaying the shape of the distribution.

•**Merits**:

- Retains actual data values, so exact values are visible.
- Useful for small to medium data sets, providing insight into data distribution.
- Simple to construct by hand.

•**Demerits**:

- Difficult to use for very large data sets or when data has many decimal places.
- Can be confusing for data with a wide range of values.

3. Box Plot (Box-and-Whisker Plot)

•**Description**: A box plot displays data based on five summary statistics: minimum, first quartile, median, third quartile, and maximum. Outliers may be shown as individual points.

•**Merits**:

- Highlights central tendency and spread (quartiles) in the data, as well as potential outliers.
- Allows for easy comparison of distributions across different data sets.
- Effective in summarizing large data sets without showing individual values.

•**Demerits**:

- Limited detail, as specific data points are not shown.
- May not be helpful for small data sets.
- Interpretation requires understanding of quartiles.

4. Histogram

•**Description**: A histogram groups data into bins (intervals) and uses bars to display the frequency or relative frequency of data within each bin.

•**Merits**:

- Useful for understanding the distribution and spread of large data sets.
- Shows the shape of the data distribution (e.g., normal, skewed).
- Effective for continuous data.

•**Demerits**:

- Choice of bin width can affect interpretation; different bin widths may reveal or obscure patterns.
- Does not retain individual data points.
- Not ideal for small data sets, as results may be misleading.

Measure of Location

Measure of location-

- can help to summarize large data by providing a single value that represents the center or typical value of the data.
- help to reduce the complexity and variability of the data, and
- facilitate comparison and interpretation.

- The main measures of location are the
  - Mean (average)
  - Median
  - Mode
  - Quartiles
  - Percentiles

Mean (Average)

The means are of the

- Arithmetic
- Weighted
- Geometric
- Harmonic

Arithmetic mean: The arithmetic mean of a set of observations is the total sum of the observation divided by the number of observations. This is the most commonly used measure of location and it is often simply referred to as "the mean".

It is mathematically represented by

$$\bar{x} = \frac{\sum x_i}{n}$$

In this formula, remember that:

- $\bar{x}$ represents the arithmetic mean of the sample of observations;
- x_i represent the sample of observations;
- $\sum x_i$ represents the sum of the sample of observation;
- n represents the number of observations.

Mean (Average)

Example 1: Find the mean of following data that are obtained from the titration of nitric acid by standard NaOH solution.

9.69, 9.72, 9.41, 9.01, 9.62, 9.56, 9.74, 9.20, 9.43, 9.29, 9.61, 9.19, 9.23, 9.32 9.23, 9.54, 9.55, 9.64, 8.92, 9.05

Solution:

$$\sum x_i = 169.4, n = 19$$

$$\bar{x} = \frac{169.4}{19} = 9.41$$

Mean (Average)

Example 2: Table 1 is the frequency distribution of the number of days on which 100 1st year students were late in the class in a given month. Using this data, find the mean number of days on which a student is late in a month.

Let

x = the possible values for the number of days late

f = the frequencies associated with each possible values of x

Then,

Total number of days late $= \sum fx = 247$

Total number students $= \sum f = 100$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{247}{100} = 2.47$$

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Table 1

No. of days late (x)	No. of students (f)	No. of Days
1	32	32
2	25	50
3	18	54
4	14	56
5	11	55

Mean (Average)

Example 3: Table 2 is the frequency distribution of the volume of standard 1.0 M NaOH solution required to titrate 15.00 mL nitric acid solution. Using this data, find the mean volume of NaOH solution.

Table 2: Volume of NaOH

Class Boundaries (mL)	Class midpoints (CM) (x)	Frequency (f)	(fx)
13.76 - <13.88	13.82	1	13.82
13.88 - <14.00	13.94	1	13.94
14.00 - <14.12	14.06	8	112.48
14.12 - <14.24	14.18	10	141.80
14.24 - <14.36	14.30	13	185.90
14.36 - <14.48	14.42	14	201.88
14.48 - <14.60	14.54	3	43.62

Solution: The measured variable, x is grouped into a number of classes. Frequency only tells us the number of observations but not value of x. To evaluate mean, x is chosen the class midpoints (CM).

Let x = the class midpoints

f = frequencies associated with each class.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{713.44}{50} = 14.27 \text{ mL}$$

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Mean (Average)

The arithmetic mean is a common measure of central tendency that represents the average value of a data set. Here are some advantages and disadvantages of arithmetic mean with examples:

- **Advantages:**
- It is easy to understand and calculate. For example, the arithmetic mean of 2, 4, and 6 is (2 + 4 + 6) / 3 = 4.
- It is easy to work with and use in further analysis. For example, the arithmetic mean can be used to calculate other statistics like variance and standard deviation.
- It works with different types of data, such as interval, ratio, and some nominal data. For example, the arithmetic mean can be used to find the average temperature, income, or rating.
- **Disadvantages:**
- It is sensitive to extreme values or outliers, which can skew the mean and give a misleading impression of the data. For example, the arithmetic mean of 2, 4, 6, and 100 is (2 + 4 + 6 + 100) / 4 = 28, which is much higher than the median of 5.
- It is not suitable for time series data, which are data collected over time. For example, the arithmetic mean of the daily stock prices of a company may not reflect the trend or volatility of the market.
- It assumes that all values are equally important, which may not be true in some cases. For example, the arithmetic mean of the test scores of a class may not account for the difficulty or weight of each test.

Weighted Mean (Average)

- The weighted mean is used when the values in a data set are not all of equal importance.
- The weight can be thought of as a measure of the importance of each value in the data set.

For instances:

- ✓ Atomic masses are computed by weighting the natural abundance to the isotopic masses.
- ✓ Grade point average (GPA) is computed by weighting the credit point of courses to grade point earned.

Weighted Mean (Average)

Example 4: The isotopic masses and relative abundance of sulfur are given in Table 3. Using this data, find the atomic mass of sulfur.

Table 3: Isotopic masses and relative abundance of sulfur

Isotope of sulfur	Isotopic mass (x)	Relative abundance (%) (w)	(wx)
^{32}S	31.972071	95.020	3037.986186
^{33}S	32.971459	0.750	24.728594
^{34}S	33.967867	4.210	143.004720
^{36}S	35.967081	0.020	0.719342

Solution:

Let x = the atomic mass of isotopes

w = Relative abundance of isotopes.

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{3206.438842}{100.000} = \mathbf{32.064388}$$

Weighted Mean (Average)

Example 5: Toma got semester grade points (SGP) in 1st year odd semester examination as shown in Table 4. Using this data, find Toma's SGPA in the examination.

Table 4: SGP and credit points (CP) of different courses

Course	SGP (x)	CP (w)	(wx)
0531-1111	3.75	2	7.50
0531-1121	2.50	2	5.00
0531-1122	3.00	2	6.00
0531-1131	3.75	2	7.50
0533-1132	2.25	2	4.50
0541-1101	3.50	3	10.50
0542-1102	4.00	3	12.00
0231-1103	2.00	0	0.00
0531-1100L	3.75	3	11.25

Solution:

Let x = the SGP earned

w = CP of associated courses.

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{64.25}{19} = 3.381$$

Geometric Mean (Average)

The geometric mean is used to measure the central tendency of a set of data that have exponential growth or decay.

For example, the geometric mean is suitable for calculating

- the average growth rate of an investment,
- the average speed of a vehicle, or
- the average bacteria population in a culture.

- It is calculated by taking the n^{th} root of the product of n numbers, where n is the total number of values.

$$\bar{x}_{\mathrm{G}} = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots x_n} = \sqrt[n]{\prod_i x_i} \quad \text{or}, \bar{x}_{\mathrm{G}} = \operatorname{Antilog}\left(\frac{\sum \log_{10} x_i}{n}\right)$$

- The weighted geometric mean (WGM) is a generalization of the geometric mean that uses weights (w_i) to reflect the importance of each data value (x_i).

$$\bar{x}_{\mathrm{WG}} = \sqrt[\sum w_i]{\prod_i x_i^{w_i}}$$

Geometric Mean (Average)

Example 6: The following values of hydrogen ion activity for different trials are obtained from electromotive force measurement (emf) using glass electrodes: 1.122×10^{-4} mol/L, 1.115×10^{-4} mol/L, 1.125×10^{-4} mol/L, 1.130×10^{-4} mol/L, and 1.120×10^{-4} mol/L. Find the average pH of solution.

Solution: The hydrogen ion activity changes ten times for every unit of pH variation. The arithmetic mean yields biassed results. The geometric mean provides a correct answer in this instance.

1st method:

$$[H^+]_{GM} = \sqrt[5]{1.122 \times 10^{-4} \cdot 1.115 \times 10^{-4} \cdot 1.125 \times 10^{-4} \cdot 1.130 \times 10^{-4} \cdot 1.120 \times 10^{-4}}$$

$$= \sqrt[5]{1.781 \times 10^{-20}} = 1.122 \times 10^{-4} \text{ molL}^{-1}$$

$$pH = -\log_{10}[1.122 \times 10^{-4}] = 3.950$$

2nd method:

$$[H^+] = 10^{\frac{-4+\log_{10} 1.122-4+\log_{10} 1.115-4+\log_{10} 1.125-4+\log_{10} 1.130-4+\log_{10} 1.120}{5}}$$

$$= 10^{-3.9499} = 1.122 \times 10^{-4} \text{ molL}^{-1}$$

$$pH = -\log_{10}[1.122 \times 10^{-4}] = 3.950$$

Geometric Mean (Average)

Example 7: Find mean ionic molality of the following solutions: (i) 0.2 mol/kg NaCl solution; (ii) 0.25 mol/kg $CaCl_2$ solution and (iii) 0.15 mol/kg $Ca_3(PO_4)_2$ solution.

Solution: Mean ionic molality is geometric mean of ionic molality by weighting the absolute ionic charges, i.e., $m_{\pm} = \sqrt[|\nu_+|+|\nu_-|]{m_+^{|\nu_+|} \cdot m_-^{|\nu_-|}}$

(i) m_{Na} = 0.1 mol/kg; m_{Cl} = 0.1 mol/kg (Assuming 100% dissociation)

$$m_{\pm} = \sqrt[|+1|+|-1|]{(0.1)^{|+1|} \times (0.1)^{|-1|}} = 0.1 \text{ mol/kg}$$

(ii) m_{Ca} = 0.25 mol/kg; m_{Cl} = 0.50 mol/kg (Assuming 100% dissociation)

$$m_{\pm} = \sqrt[|+2|+|-1|]{(0.25)^{|+2|} \times (0.50)^{|-1|}} = 0.315 \text{ mol/kg}$$

(iii) m_{Ca} = 0.45 mol/kg; m_{PO4} = 0.30 mol/kg (Assuming 100% dissociation)

$$m_{\pm} = \sqrt[|+2|+|-3|]{(0.45)^{|+2|} \times (0.30)^{|-3|}} = 0.353 \text{ mol/kg}$$

Harmonic Mean (Average)

- The Harmonic Mean (HM) is defined as the reciprocal of the average of the reciprocals of the data values.

$$\bar{x}_{\mathrm{H}} = \frac{n}{\sum(1/x_i)}$$

- Whereas, the weighted harmonic mean (WHM) is a variation of the harmonic mean that takes into account different weights (w_i) assigned to each value.

$$\bar{x}_{\mathrm{WH}} = \frac{\sum w_i}{\sum(w_i/x_i)}$$

- These means give less weightage to the larger values and larger weightage to the smaller values to balance the values correctly.
- The harmonic mean is more appropriate for data that are inversely proportional to some quantity, such as speed, frequency, or resistance.
- These are used in chemistry when the average of rates or ratios is needed, such as the rate of reaction, the rate of diffusion, the rate of crystal growth, or the mole fraction of a mixture.

Harmonic Mean (Average)

Example 7: The interfacial diffusion coefficient, $D = \frac{\lambda^2}{2\tau}$, where λ is the interfacial distance (1.5 μm) and τ is the diffusion time, is given in Table 5 along with τ. Using data, show that $\bar{D}_A = \frac{\lambda^2}{2\bar{\tau}_H}$.

Table 5: τ and D

τ/ms (x)	D/cm^{2}s^{-1}	(1/x)
7.37	0.153	0.1357
7.28	0.155	0.1374
8.97	0.125	0.1115
6.28	0.179	0.1592
3.89	0.289	0.2571
5.52	0.204	0.1812
8.93	0.126	0.1120
3.46	0.325	0.2890
4.59	0.245	0.2179
3.63	0.310	0.2755

Solution: Since D is inversely related to τ, harmonic mean of τ will give better result for $\bar{D}_A$.

$$\bar{\tau}_H = \frac{n}{\sum \frac{1}{\tau_i}} = \frac{10}{1.8765} = 5.33 \text{ ms}$$

$$\bar{D}_A = \frac{\sum D_i}{n} = \frac{2.111}{10} = 0.2111 \text{ cm}^2\text{s}^{-1}$$

From equation,

$$D = \frac{\lambda^2}{2\bar{\tau}_H} = \frac{\left(1.5\ \mu m \times \frac{1\ cm}{1000\ \mu m}\right)^2}{2 \times 5.33\ ms \times \frac{1\ s}{10^6\ ms}}$$

$$= 0.21107 \text{ cm}^2\text{s}^{-1} \cong 0.2111 \text{ cm}^2\text{s}^{-1}$$

Hence, $D = \frac{\lambda^2}{2\bar{\tau}_H} = \bar{D}_A$

Harmonic Mean (Average)

Example 8: In 0.1 M $CaCl_2$ aqueous solution, diffusion coefficient of Ca^{2+} and Cl^- are 7.9×10^{-10} m^2/s and 20.3×10^{-10} m^2/s, respectively. Find the diffusion coefficient of $CaCl_2$.

Solution:

The diffusion coefficient of a salt in water is given by the harmonic mean of the diffusion coefficients of the components of the salt, weighted by their molal concentration:

$$D_{salt} = \frac{\sum m_i}{\sum (m_i / D_i)}$$

Here, $m_{Ca} = 0.1$ M, $m_{Cl} = 0.2$ M, $D_{Ca} = 7.9\times10^{-10}$ m^2/s

$D_{Cl} = 20.3\times10^{-10}$ m^2/s

Now, $D_{salt} = \frac{0.1+0.2}{\frac{0.1}{7.9\times10^{-10}}+\frac{0.2}{20.3\times10^{-10}}} = \frac{0.3\times10^{-10}}{0.0225} = 13.33\times10^{-10}$ m^2/s

Median & Mode

The **median** is the middle value of a dataset when it is ordered from least to greatest. If the dataset has an even number of observations, the median is the average of the two middle numbers.

Example: For the dataset: 3, 5, 6, 8, 9 The median is 6 (the middle value).

For the dataset: 3, 5, 6, 8 The median is

$$\frac{5+6}{2} = 5.5$$

Mode

The **mode** is the value that appears most frequently in a dataset. A dataset can have more than one mode if multiple values have the same highest frequency.

Example: For the dataset: 2, 4, 4, 6, 6, 6, 8 The mode is 6 (appears most frequently).

Quartiles & Percentiles

Quartiles

Quartiles divide the dataset into four equal parts. The first quartile (Q1) is the median of the lower half, the second quartile (Q2) is the median, and the third quartile (Q3) is the median of the upper half.

Example: For the dataset: 1, 3, 5, 7, 9, 11, 13

- Q1 (first quartile) = 3
- Q2 (median) = 7
- Q3 (third quartile) = 11

Percentiles

Percentiles divide the dataset into 100 equal parts. The nth percentile is the value below which n% of the data falls.

Example: For the dataset: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 The 25th percentile (P25) is 3 (25% of the data is below 3).

Measures of Dispersion

Measures of dispersion describe the spread or variability within a dataset. They help us understand how much the data points differ from the central value.

- **Range**
- **Variance**
- **Standard Deviation**
- **Interquartile Range (IQR)**
- **Mean Absolute Deviation (MAD)**

Range & Variance

Range

The **range** is the difference between the maximum and minimum values in a dataset. It gives a quick sense of the spread.

$$R = x_{max} - x_{min}$$

Example: For the dataset: 4, 8, 6, 5, 3

Range=8−3=5

Variance

The **variance** measures the average squared deviation of each data point from the mean. It provides a sense of how much the data points vary from the mean.

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Example: For the dataset: 4, 8, 6, 5, 3

1.Calculate the mean:

$$\text{Mean} = \frac{4+8+6+5+3}{5} = 5.2$$

Find the squared deviations:

$$(4-5.2)^2, (8-5.2)^2, (6-5.2)^2, (5-5.2)^2, (3-5.2)^2 = 1.44, 7.84, 0.64, 0.04, 4.84$$

Calculate the variance:

$$\text{Variance} = \frac{1.44+7.84+0.64+0.04+4.84}{5} = 2.96$$

Standard Deviation & Interquartile Range (IQR)

Standard Deviation

The **standard deviation** is the square root of the variance. It is the most commonly used measure of dispersion and provides a sense of how spread out the data points are around the mean.

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

Example: For the dataset: 4, 8, 6, 5, 3

1.Calculate the variance: 2.96

2.Calculate the standard deviation: Standard Deviation=$\sqrt{2.96}$≈1.72

Interquartile Range (IQR)

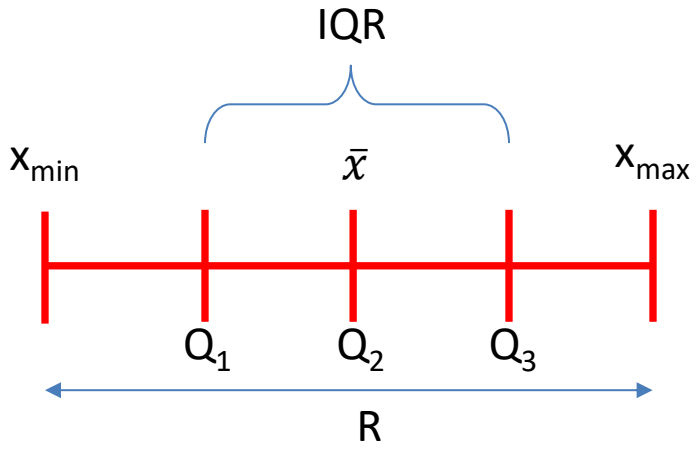
The **interquartile range** is the difference between the first quartile (Q1) and the third quartile (Q3). It measures the spread of the middle 50% of the data and is less affected by outliers.

Example: For the dataset: 1, 3, 5, 7, 9, 11, 13

1.Q1 (first quartile) = 3

2.Q3 (third quartile) = 11

3.Calculate the IQR: IQR=11−3=8



Mean Absolute Deviation (MAD)

The **mean absolute deviation** measures the average distance between each data point and the mean of the dataset. It provides a sense of the variability in the data.

$$MAD = \frac{\sum|x_i - \bar{x}|}{n}$$

Example: For the dataset: 4, 8, 6, 5, 3

1.Calculate the mean: 5.2

Find the absolute deviations:

$$|4 - 5.2|, |8 - 5.2|, |6 - 5.2|, |5 - 5.2|, |3 - 5.2| = 1.2, 2.8, 0.8, 0.2, 2.2$$

Calculate the MAD

$$\text{MAD} = \frac{1.2 + 2.8 + 0.8 + 0.2 + 2.2}{5} = 1.44$$

Are measures of central tendency and dispersion sufficient?

Measures of central tendency and dispersion provide estimates of data, but they fall short in key areas:

- **Quantifying uncertainty**: Assess data unpredictability and variability.
- **Assessing reliability**: Evaluate data consistency and accuracy for valid conclusions.
- **Predicting outcomes**: Use past data to forecast future events.
- **Analyzing distribution**: Examine data for skewness, kurtosis, and patterns.
- **Making inferences**: Draw conclusions, factoring in uncertainty and assumptions.
- **Supporting decisions**: Apply probabilistic models to minimize risks in choices.

A deeper understanding of probability is essential to address these limitations.

Probability

Experiment:

- An experiment is any process or action that produces a set of possible outcomes. Experiments are generally repeatable and controlled, with outcomes that can vary.
    - *Example*: Rolling a six-sided die is an experiment, as it produces one outcome from a set of six possible numbers (1, 2, 3, 4, 5, or 6).

Trial:

- A trial is a single instance or execution of an experiment. Each trial yields one outcome from the possible outcomes of the experiment.
    - *Example*: If we roll the die three times, each roll is a separate trial. Each trial might yield a different outcome, such as rolling a 2 on the first roll, a 5 on the second roll, and a 3 on the third roll.

Event:

- An event is a specific outcome or a set of outcomes from an experiment. Events can be as simple as a single outcome or more complex, involving multiple outcomes.
    - *Example*: For the die roll experiment, an event could be "rolling an even number," which includes the outcomes {2, 4, 6}. Another event could be "rolling a 3."

Probability

Probability is a measure of how likely an event is to occur, quantified betwe en 0 (impossible) and 1 (certain). It's the foundation of statistics and ever yday decision-making.

Example:

Think of flipping a fair coin. There are two possible outcomes: heads or tail s. Each outcome is equally likely. The probability P of getting heads (or tail s) is:

$$P(heads) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{1}{2} = 0.5$$

In this case, the probability is 0.5 or 50%, meaning there's an equal chance of getting heads or tails.

Another Example:

Rolling a six-sided die. The probability of rolling a 4 is:

$$P(\text{rolling a 4}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{1}{6} \approx 0.167$$

This means there's roughly a 16.7% chance of rolling a 4 on a six-sided die.

These examples illustrate how probability helps us understand and predict the likelihood of different outcomes.

Measurement Error & Probability

The measured quantity can be represented as

$$x = \mu_x + \epsilon_x$$

Where,

- x **is the measured quantity** is the observed or recorded value obtained from the measurement process.
- μ_x **is true value** is the actual, precise value of the quantity being measured (which is often unknown or idealized).
- ϵ_x **is error** represents the difference between the measured quantity and the true value. This error can be positive or negative, depending on whether the measured value is above or below the true value.

Measurement Error & Probability

Measured error can be modelled by probability distribution:

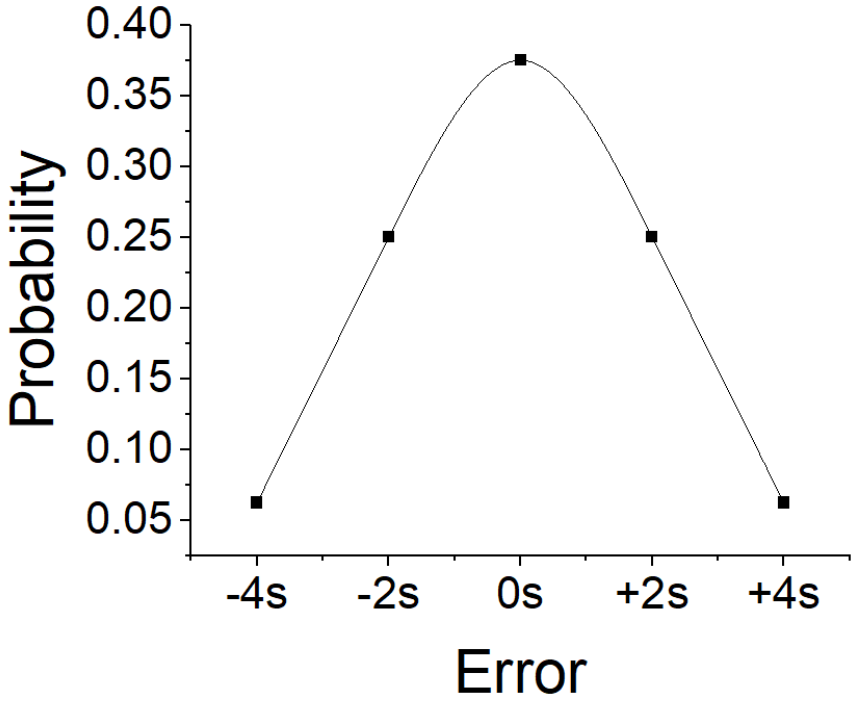
- Measurement error can be modeled with probability to quantify uncertainty in measurements.
- Errors often stem from instrument limitations, observer variability, or environmental factors.
- Imagine a situation in which just four small random errors, s, combine to give an overall error.
- Assume that each error has an equal probability of occurring and that each can cause the final result to be high or low by a fixed uncertainty $\pm U$.

Combinations of Uncertainties	Magnitude of Random Error	Number of Combinations	Relative Frequency
$-U_1-U_2-U_3-U_4$	-4s	1	0.0625
$+U_1-U_2-U_3-U_4$ $-U_1+U_2-U_3-U_4$ $-U_1-U_2+U_3-U_4$ $-U_1-U_2-U_3+U_4$	-2s	4	0.2500
$+U_1+U_2-U_3-U_4$ $-U_1-U_2+U_3+U_4$ $+U_1-U_2+U_3-U_4$ $-U_1+U_2-U_3+U_4$ $-U_1+U_2+U_3-U_4$ $+U_1-U_2-U_3+U_4$	0s	6	0.3750
$-U_1+U_2+U_3+U_4$ $+U_1-U_2+U_3+U_4$ $+U_1+U_2-U_3+U_4$ $+U_1+U_2+U_3-U_4$	+2s	4	0.2500
$+U_1+U_2+U_3+U_4$	+4s	1	0.0625

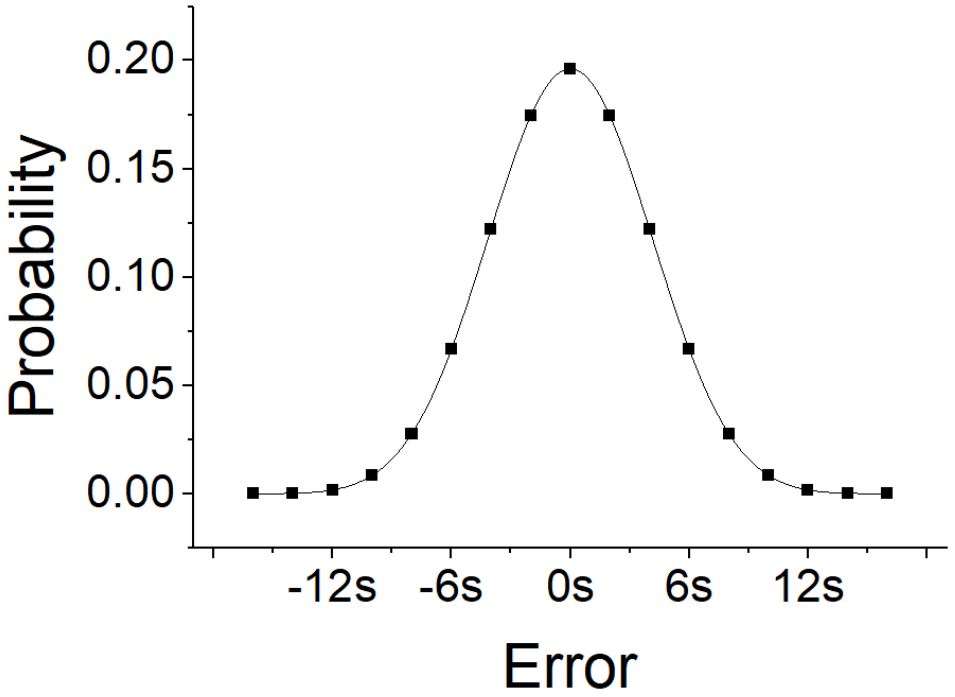
Measurement Error & Probability

Measured error can be modelled by probability distribution:

Combinations of 4 small errors of equal probability



Combinations of 16 small errors of equal probability

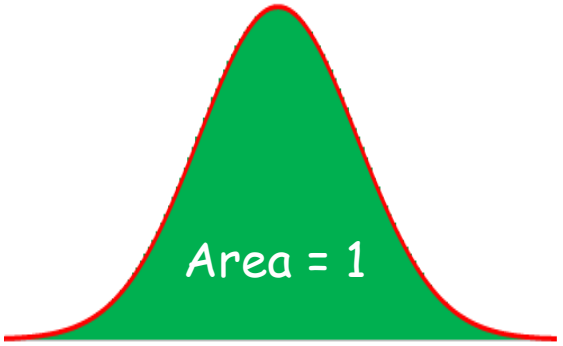


- With increasing the number of measurement, data distribution becomes narrower, smooth, bell-shaped curved.
- The mathematical function that describes the curve is probability distribution function (PDF).
- The PDF that represents bell-shaped curve is normal or Gaussian PDF.

Probability Distribution Function (PDF)

Normal/ Gaussian PDF

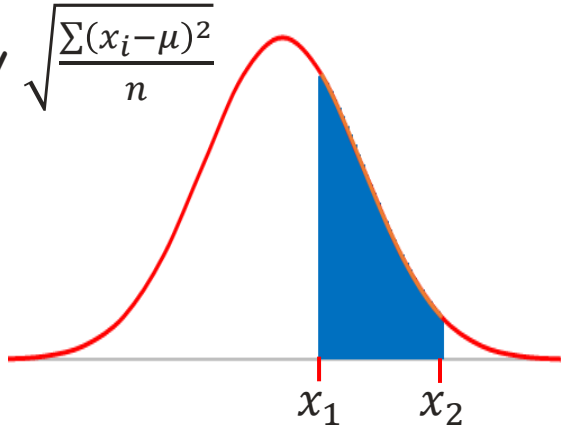
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



where

μ is the population mean, which is equal to $\frac{\sum x_i}{n}$

σ is the population standard deviation given by $\sqrt{\frac{\sum(x_i-\mu)^2}{n}}$



Properties

1) **Symmetric, continuous, bell-shaped**
2) **Area under the curve** $\int_{-\infty}^{\infty} f(x)dx = 1$
3) **Area under the curve represents the total probability which is 1 indicating that probability is 100%**
4) **The probability within the interval x_1 and x_2 is** $\int_{x_1}^{x_2} f(x)dx$

Probability Distribution Function (PDF)

Standard Normal/ Gaussian PDF

Let $z = \frac{x-\mu}{\sigma}$, and

g(z) is the pdf of z

g(z) is given by

$$g(z) = f(x) \cdot \left|\frac{dx}{dx}\right|$$

Now, $x = z\sigma + \mu \Longrightarrow \frac{dx}{dz} = \sigma$

Substituting above information,

$$g(z) = f(z\sigma + \mu) \cdot \sigma$$

$$g(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z\sigma+\mu-\mu)^2}{2\sigma^2}} \cdot \sigma$$

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$ Standard Normal/ Gaussian PDF

Standard Normal/ Gaussian PDF can also be simply derived by putting

$$\boldsymbol{\mu = 0}$$

$$\boldsymbol{\sigma = 1}$$

$$\boldsymbol{z = x}$$

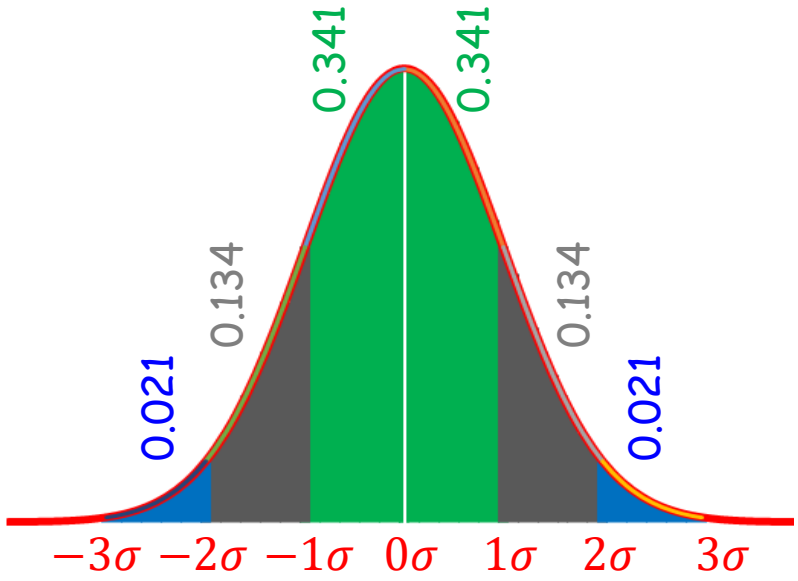
In Normal/ Gaussian PDF

$$\boldsymbol{g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}}$$

Probability Distribution Function (PDF)

Standard Normal/ Gaussian PDF

Range between	Area	Area %
0σ to 1σ	0.341	34.1
-1σ to 0σ	0.341	34.1
-1σ to 1σ	0.682	68.2
1σ to 2σ	0.134	13.4
-2σ to -1σ	0.134	13.4
-2σ to 2σ	0.954	95.4
2σ to 3σ	0.021	02.1
-2σ to -1σ	0.021	02.1
-3σ to 3σ	0.997	99.7



As can be seen that

- Data with 1σ is about 68.2%
- Data with 2σ is about 95.4%
- Data with 3σ is about 99.7%

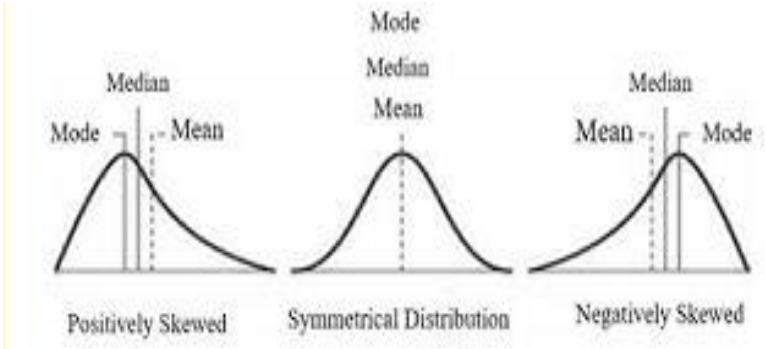
Probability Distribution Function (PDF)

Skewness

Skewness is the asymmetry of a distribution around its mean.

$$\boldsymbol{Skewness, s_k = \frac{3(mean - median)}{\sigma}}$$

- If $s_k > 3$, it is positive (right) skew
- If $s_k < 3$, it is negative (left) skew
- If $s_k = 3$, there is now skew

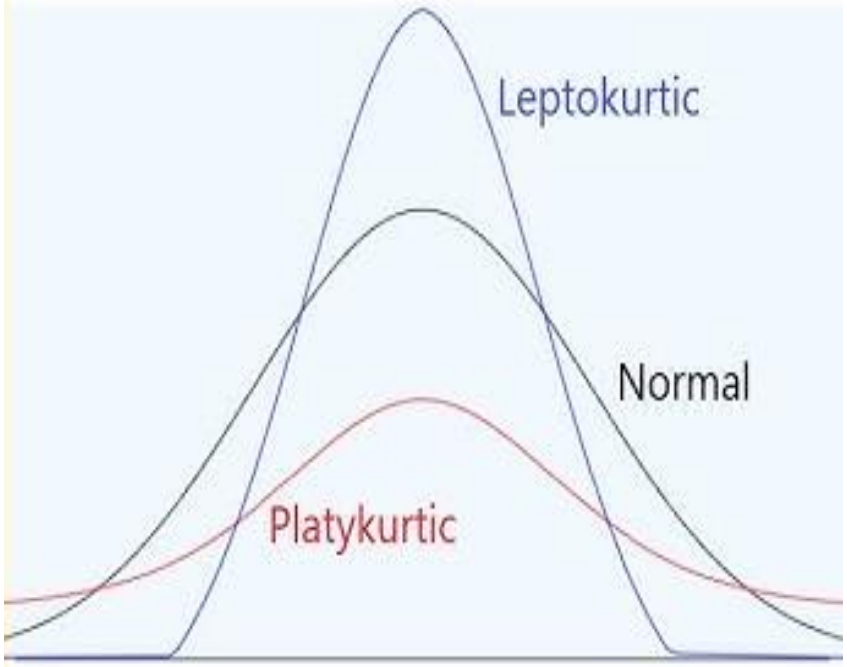


Kurtosis

Kurtosis is the "tailedness" or the sharpness of the peak of a distribution. It helps describe whether data points tend to cluster around the mean or if they produce outliers.

$$\boldsymbol{Kurtosis, \gamma = \frac{(mean - median)^4}{\sigma^2}}$$

- If $\gamma < 3$, it is Platykurtic
- If $\gamma > 3$, it is Leptokurtic
- If $\gamma = 3$, it is Mesokurtic (normal)

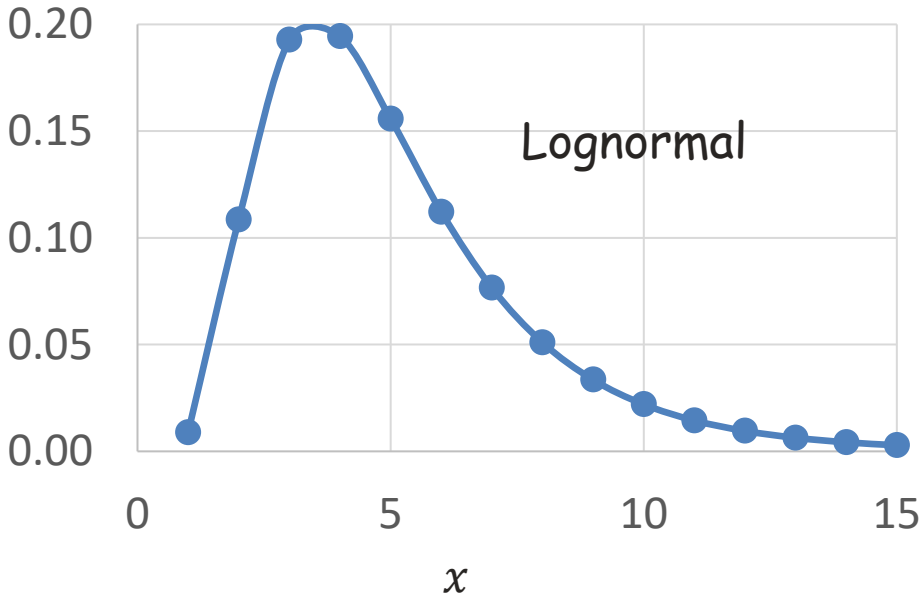


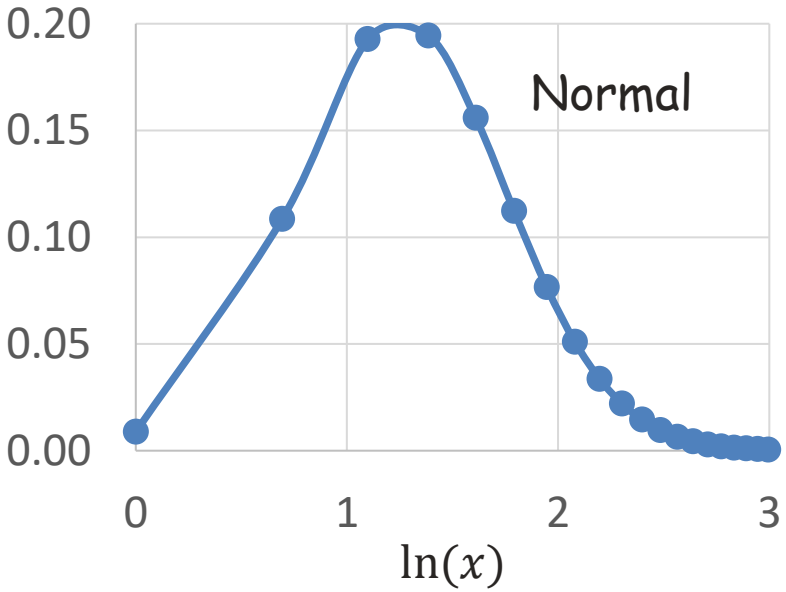
Other Probability Distribution Function (PDF)

Lognormal PDF

The **lognormal probability density function (PDF)** describes a continuous probability distribution of a random variable whose logarithm is normally distributed. If x is lognormally distributed, then $\ln(x)$ follows a normal distribution.

$$\text{Lognormal PDF: } f(x,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \qquad x > 0$$





When does it occur?

The lognormal distribution arises in chemistry when a variable is influenced by **multiplicative processes** or **random growth factors**.

Example: **Reaction Times, Concentration Data, Particle Size Distribution, etc.**

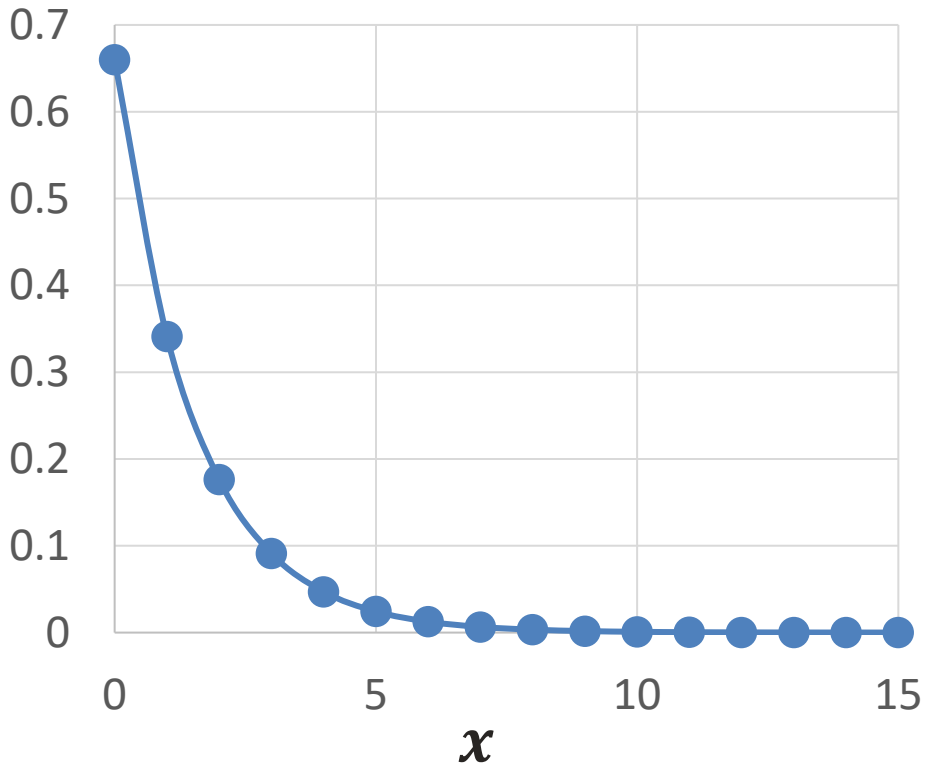
Other Probability Distribution Function (PDF)

Exponential PDF

It is used to model the time or distance between independent events that occur at a constant average rate.

$$\text{Exponential PDF: } f(x,\lambda) = \lambda e^{-\lambda x} \text{ where } x \geq 0$$

Here, λ is the rate parameter, which is the reciprocal of the mean ($\lambda = 1/\mu$).



When does it occur?

Exponential distributions often describe **random events or waiting times** in processes where events are memoryless (i.e., the probability of an event occurring in the future does not depend on past events).

Example:

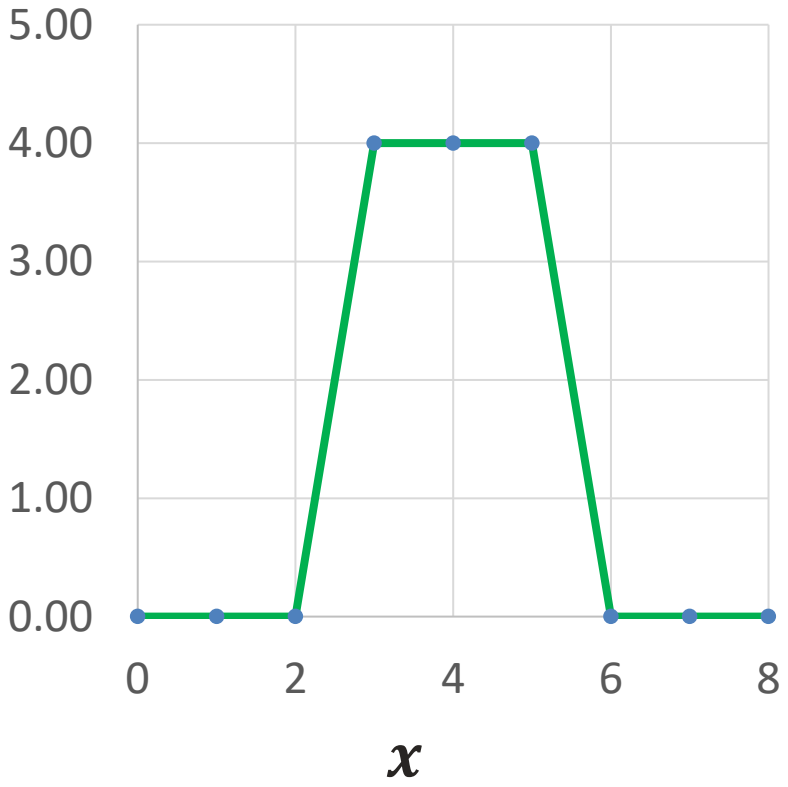
- **Radioactive Decay**:
- **Reaction Times in First-Order Reactions**:
- **Diffusion-related Processes**:

Other Probability Distribution Function (PDF)

Rectangular (Uniform) PDF

It describes a situation where all outcomes within a specific range are equally likely, with zero probability outside that range.

Rectangular PDF: $f(x, a, b) = \frac{1}{b-a}$ where $a \leq x \leq b$



When does it occur?

A rectangular PDF can describe the distribution of data in situations where uniformity is expected within defined bounds

Example:

- **Measurement Noise or Uncertainty** (Instrument specification $\pm 5\%$)
- **Energy Distributions (**Probability of occupation of energy state**)**
- **Random Mixing Scenarios (**equal probability of solute conc. in uniform solution**)**

Other Probability Distribution Function (PDF)

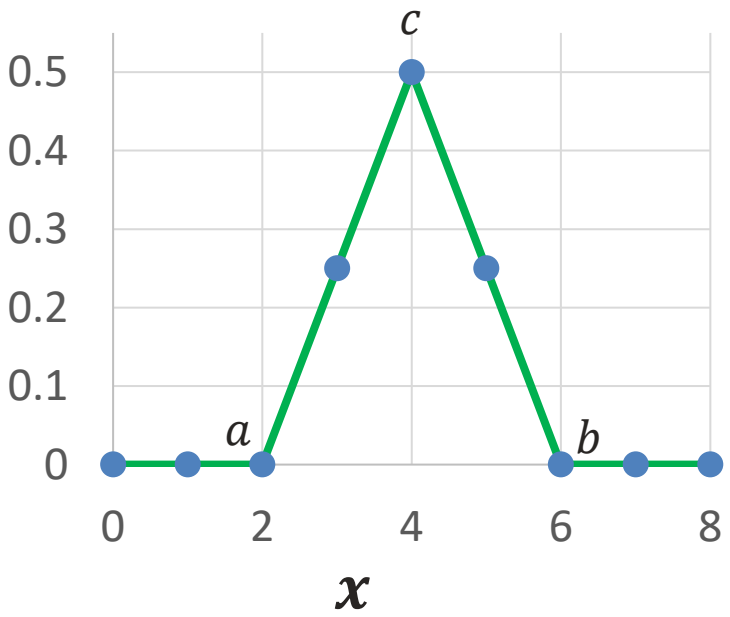
Triangular PDF

A triangular probability density function (PDF) is a distribution shaped like a triangle, defined by three parameters: the minimum value a, the maximum value b, and the mode c (the most probable value).

Triangular PDF:

$$f(x) = \frac{2(x-a)}{(b-a)(c-a)} \text{ where } a \leq x \leq c$$

$$f(x) = \frac{2(b-x)}{(b-a)(b-c)} \text{ where } c \leq x \leq b$$



When does it occur?

The PDF rises linearly from a to c and then decreases linearly from c to b, making it suitable for scenarios where values near the mode are more likely but extremes are possible within a defined range.

Example:

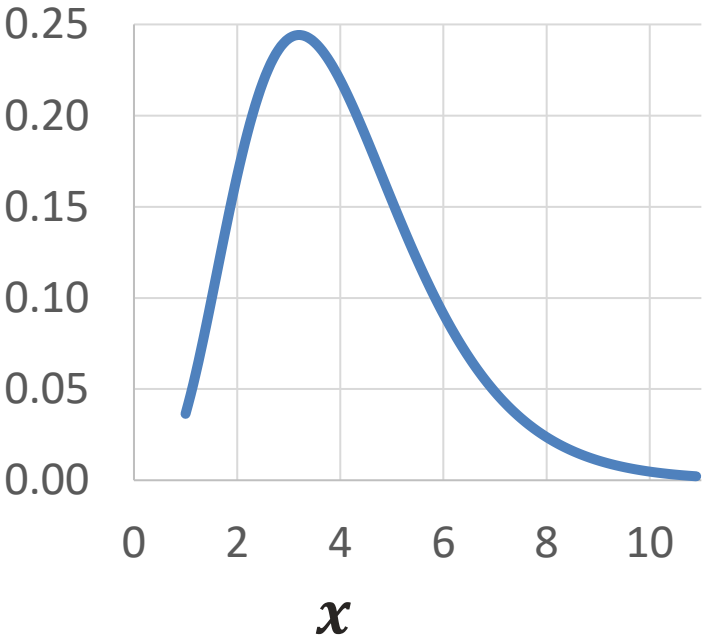
- **Empirical or Limited Data** (Instrument specification $\pm 5\%$)
- **Processes with Natural Limits (**Probability of occupation of energy state**)**
- **Random Mixing Scenarios (**equal probability of solute conc. in uniform solution**)**

Other Probability Distribution Function (PDF)

Gamma Distribution:

It is a continuous probability distribution that models the time required for a specific number of events to occur in a Poisson process.

- Two parameters characterize it:
  - ✓ Shape parameter (k): Determines the skewness.
  - ✓ Scale parameter (θ): Determines the spread.
- The gamma distribution is widely used in scenarios involving waiting times, lifetimes, and queuing systems.



Gamma PDF:

$$f(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)} \quad (x > 0, k > 0, \theta > 0$$

The gamma function, $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt$

Other Probability Distribution Function (PDF)

Gamma Distribution:

When does it occur?

- Consider a multi-step chemical reaction where the formation of a product requires k intermediate steps, and the time for each step follows an exponential distribution with rate parameter λ.
- The total reaction time for the k steps will follow a gamma distribution with:
  - k = number of steps (shape parameter),
  - $\theta = 1/\lambda$ (scale parameter).
- This is useful in studying reaction mechanisms and estimating reaction times under various conditions.
- Example:
  - **Radioactive Decay**
  - **Reaction Kinetics**
  - **Chromatography**
  - **Molecular Events**

Other Probability Distribution Function (PDF)

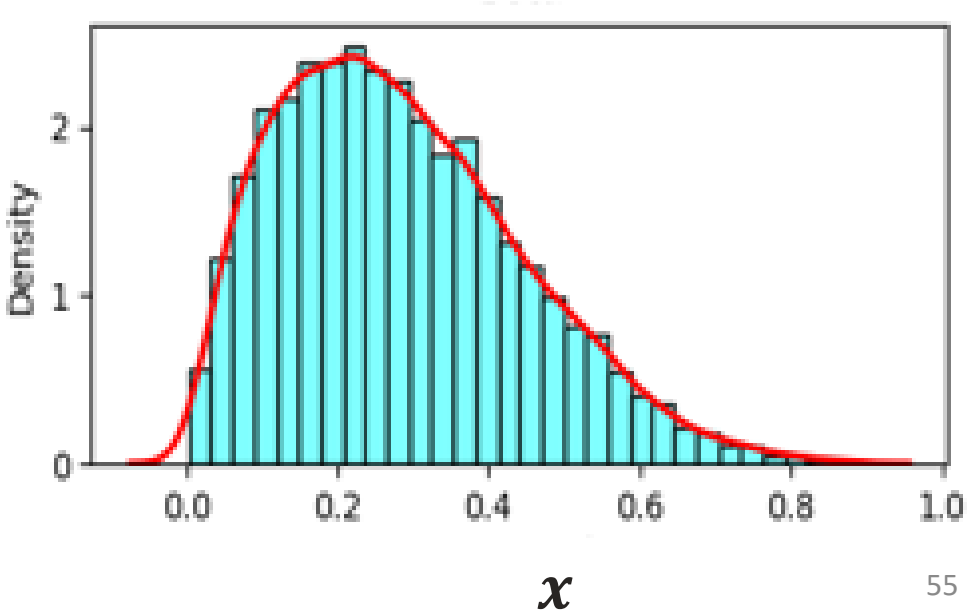
Beta Distribution:

The beta distribution is a continuous probability distribution defined on the interval [0,1]. It is parameterized by two positive shape parameters, α and β, which control the shape of the distribution. Its probability density function (PDF) is given by:.

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad for\ 0 \le x \le 1$$

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- The gamma distribution is widely used in scenarios involving waiting times, lifetimes, and queuing systems.



Other Probability Distribution Function (PDF)

Beta Distribution:

When does it occur?

- Consider the fractional binding of a ligand to a receptor. If the proportion of bound receptors, x, is influenced by two competing factors (e.g., ligand affinity and receptor saturation), the distribution of x might follow a beta distribution. For instance:
  - α: relates to the number of successful ligand-receptor binding events.
  - β: relates to the number of non-binding attempts or receptors remaining unoccupied.
- If measurements of x yield varying probabilities, the beta distribution provides a way to model the range of fractional bindings between 0 (no binding) and 1 (full saturation).
- Example:
  - **Reaction Probabilities**
  - **Molecular Populations**
  - **Spectroscopic Intensities**
  - **Chemical Kinetics**

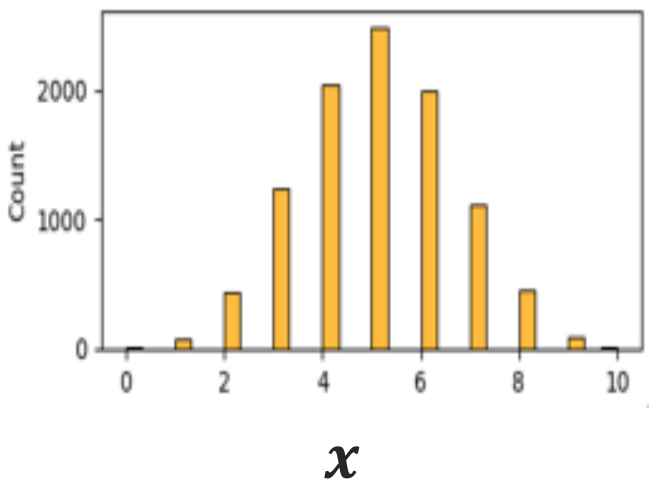
Other Probability Distribution Function (PDF)

Binomial Distribution (discrete):

The binomial distribution describes the probability of obtaining a fixed number of "successes" in a specified number of independent trials, each with the same probability of success.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad where \ \binom{n}{k} = \frac{n!}{n!(n-k)!}$$

- P(X=k) is the probability of exactly k successes in n trials.
- n = number of trials.
- k = number of successes.
- p = probability of success in a single trial.



- The binomial distribution appears in chemistry in scenarios where processes can be considered a series of independent binary outcomes (e.g., success/failure, yes/no, presence/absence).
  - Molecular Interactions
  - Spectroscopy
  - Chemical Kinetics
  - Statistical Thermodynamics

Other Probability Distribution Function (PDF)

Binomial Distribution (discrete):

Example in Chemistry

Consider a reaction where the probability p of a molecule reacting upon collision is 0.4, and there are n=10 collisions. The binomial distribution can help determine the probability of exactly k=5 successful reactions:

$$P(X = 5) = \binom{10}{5} 0.4^5 (1 - 0.4)^{10-5}$$

$$\binom{10}{5} = \frac{10!}{10!\,(10 - 5)!} = 252$$

$$P(X = 5) = 250 \times 0.4^5 \times 0.6^5 = 0.2007$$

Thus, the probability of exactly 5 successful reactions is 20.07%.

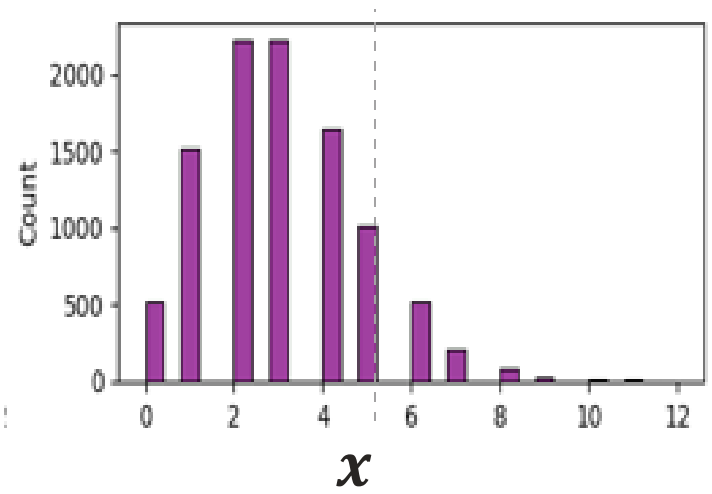
Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring within a fixed interval of time, space, or other domains, provided these events occur independently and at a constant average rate.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- P(X=k): Probability of observing k events.
- λ: Mean number of events per interval (rate parameter).
- k: Number of occurrences (non-negative integer).
- e: Euler's number (≈2.718).



- In chemistry, the Poisson distribution often describes processes where:
- Rare Events: Events occur randomly and rarely, e.g., radioactive decay.
- Constant Probability: Each event has the same likelihood of occurring.
- Independence: Events occur independently of each other.

Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

1. Radioactive Decay:

   The number of radioactive emissions detected in a fixed time interval (e.g., the count rate from a Geiger-Müller counter) often follows a Poisson distribution.

   Example: If a sample emits an average of λ=5 decays per second, the probability of detecting 3 decays in one second is:

$$P(X = 3) = \frac{5^3 e^{-5}}{3!} = 0.1404$$

2. Photon Emission in Spectroscopy:

   The number of photons emitted by a molecule during fluorescence within a fixed time.

3. Molecular Collisions in Gases:

   The number of collisions experienced by a molecule in a small time interval.

4. Chemical Reaction Events:

   In single-molecule reaction studies, the number of reaction events detected in a fixed observation window.

Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

The **t-distribution**, **F-distribution**, and **chi-square (χ^2) distribution** are statistical distributions primarily used in inferential statistics for hypothesis testing and confidence interval estimation. They are classified as **sampling distributions** because they describe the distribution of sample-based statistics under certain conditions. Here's why they are needed and their respective probability density functions (PDFs):

t-Distribution

- **Why needed**: Used when estimating the population mean when the sample size is small and the population standard deviation is unknown. It accounts for additional variability in smaller samples.
- **Sampling distribution?** Yes, because it describes the distribution of the sample mean when the population standard deviation is unknown.
- **PDF** (for $-\infty < t < \infty$):

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where ν is the degrees of freedom, and Γ is the gamma function.

Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

F-Distribution

- **Why needed**: Used to compare variances between two populations or to test the overall significance in ANOVA (Analysis of Variance).
- **Sampling distribution?** Yes, because it represents the ratio of two scaled chi-square distributions (variance estimates from samples).
- **PDF** (for $F > 0$):

$$f(F) = \frac{\left(\frac{d_1}{d_2}\right)^{d_1/2} F^{(d_1/2)-1}}{B(d_1/2, d_2/2)\left(1 + \frac{d_1}{d_2}F\right)^{(d_1+d_2)/2}}$$

where d_1 and d_2 are degrees of freedom for the numerator and denominator, and B is the beta function.

Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

Chi-Square (χ^2) Distribution

- **Why needed**: Used to test independence (e.g., chi-square test of independence), goodness of fit, and the variance of a population.
- **Sampling distribution?** Yes, because it describes the distribution of the sum of squared standard normal variables, often arising from sample variance estimates.
- **PDF** (for $x > 0$):

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2}$$

where ν is the degrees of freedom.

Why called **sampling distributions**?

They describe the probability distributions of statistics derived from a sample (e.g., sample mean, variance ratio) rather than from a population. These distributions are foundational in making inferences about population parameters based on sample data.

Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

Central Limit Theorem (CLT)

The **Central Limit Theorem (CLT)** states that, for a sufficiently large sample size, the sampling distribution of the sample mean (or sum) will approximate a **normal distribution** regardless of the shape of the population distribution, provided the samples are independent and identically distributed (i.i.d.).

This theorem is fundamental to inferential statistics as it justifies the use of normal distribution-based methods (like z-tests) even when the underlying population distribution is not normal.

Key Features of CLT:

1. **Parent Distribution Independence**: The population distribution can be any shape (uniform, exponential, skewed, etc.).
2. **Large Sample Size**: As the sample size n increases, the sampling distribution of the sample mean becomes increasingly normal.
3. **Mean and Variance**:
   - Sampling distribution mean $\mu_{\bar{x}} = \mu$ (population mean).
   - Sampling distribution variance $\sigma_{\bar{x}}^2 = \sigma^2/n$ (population variance divided by sample size).

Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

Example: Demonstrating CLT

To illustrate, we'll simulate random samples from different parent distributions (e.g., uniform, exponential, and bimodal), calculate their sample means, and show that the distribution of the sample means approaches normality.

Procedure:

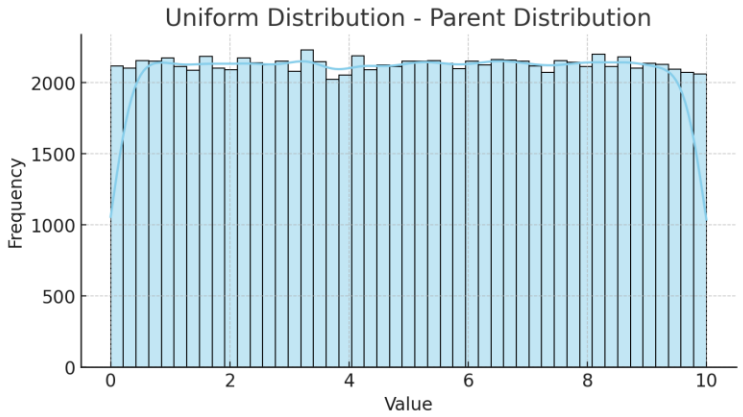
1. Generate a parent population from a specific distribution (e.g., uniform, exponential, etc.).
2. Take multiple random samples of size n from this population.
3. Compute the sample mean for each sample.
4. Plot the distribution of these sample means to show it approximates a normal distribution.

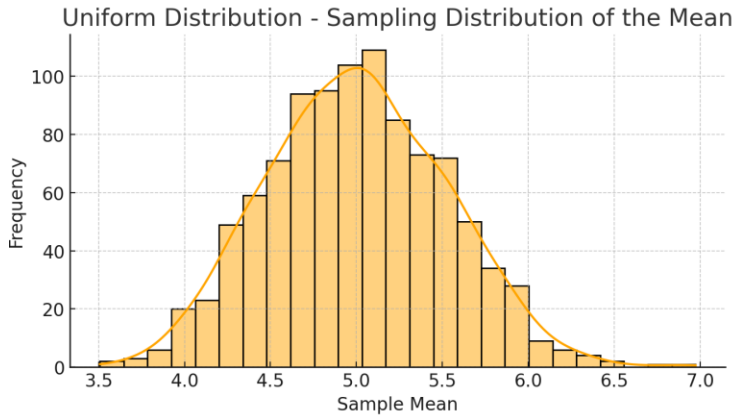
The plots above demonstrate the **Central Limit Theorem (CLT)**:

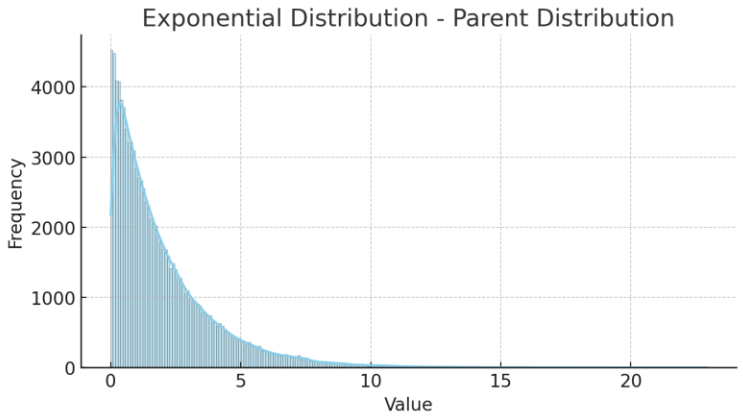
1. **Left Column**: Shows the parent distributions, which are:
   - **Uniform Distribution**: Flat and evenly spread.
   - **Exponential Distribution**: Skewed with a long tail.
   - **Bimodal Distribution**: Two distinct peaks.
2. **Right Column**: Displays the sampling distributions of the sample means (from 1000 samples of size 30 each). Regardless of the shape of the parent distribution:
   - The sample means approximate a **normal distribution**.
   - This highlights the essence of CLT — the sampling distribution of the mean approaches normality as sample size increases. [›-]

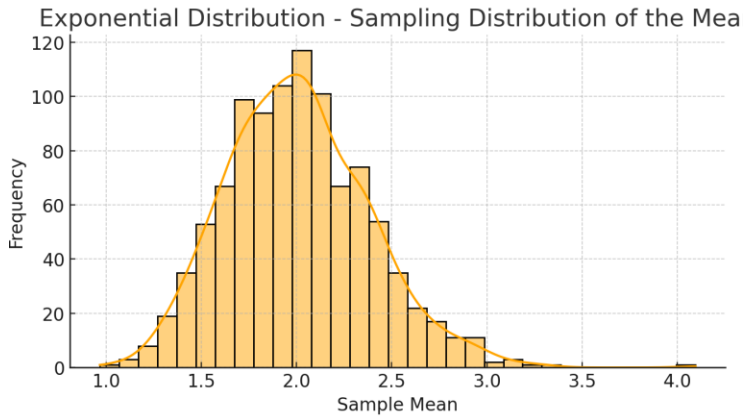
Other Probability Distribution Function (PDF)

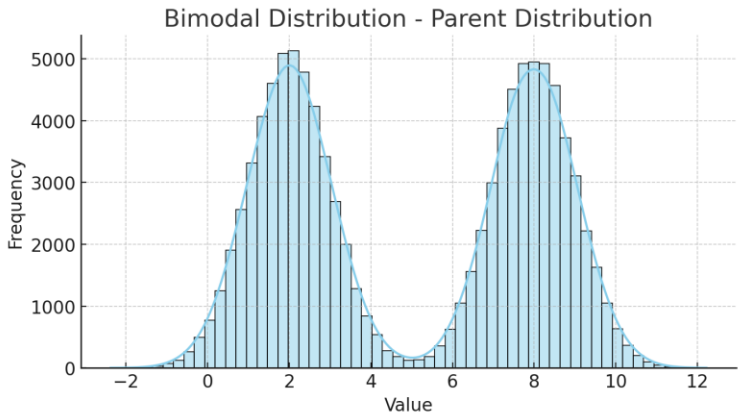
Po

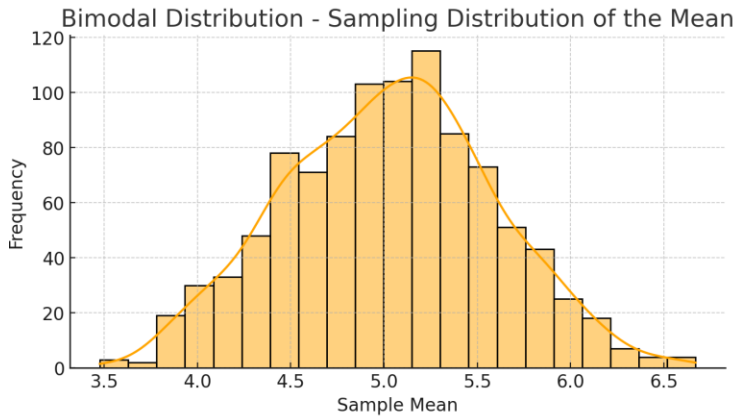












Other Probability Distribution Function (PDF)

Poison Distribution (discrete):

Limitations of CLT

1. **Sample Size Requirement**: CLT requires a sufficiently large sample size; for heavily skewed or complex distributions, larger samples are needed for normality.
2. **Independent and Identically Distributed (i.i.d.) Assumption**: Violations (e.g., dependent samples) can lead to incorrect conclusions.
3. **Finite Variance Assumption**: CLT applies only when the parent distribution has finite variance.
4. **Approximations May Fail**: In small samples or extreme parent distributions, the normal approximation may not hold.

Necessity of Studying Parent Distributions

1. **Understanding Real Data Behavior**: Many phenomena (e.g., financial data, environmental events) follow non-normal distributions like exponential or power-law.
2. **Tail Risks and Extremes**: Parent distributions affect the likelihood of extreme values, important in fields like risk management.
3. **Accurate Modeling**: Theoretical assumptions (e.g., exponential decay, bimodal trends) depend on understanding the actual parent distribution.
4. **Statistical Method Selection**: Methods like non-parametric tests or robust estimators are tailored to specific distribution characteristics.
5. **Beyond Means**: CLT focuses on sample means, but variance, skewness, and other parameters also depend on the parent distribution.