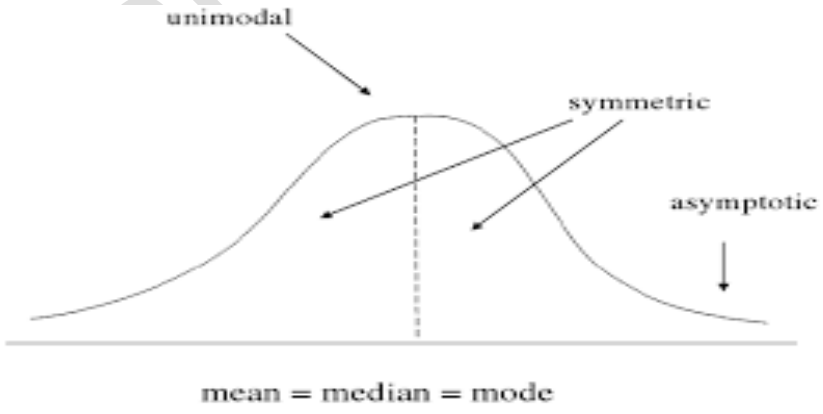
IT Skills for Chemist
(Statistical analysis)

Statistical analysis: Statistical analysis is the collection and interpretation of data in order to uncover patterns and trends. It is a component of data analytics. Statistical analysis can be used in situations like gathering research interpretations, statistical modeling or designing surveys and studies. It can also be useful for business intelligence organizations that have to work with large data volumes.

p-value in statistics mean: The P value is defined as **the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.** The P stands for probability and measures how likely it is that any observed difference between groups is due to chance.

a p-value of 0.05 mean: A p-value less than 0.05 is typically considered to be **statistically significant, in which case the null hypothesis should be rejected**. A p-value greater than 0.05 means that deviation from the null hypothesis is not statistically significant, and the null hypothesis is not rejected.

Gaussian distribution: Normal distributions have key characteristics that are easy to spot in graphs: The mean, median and mode are exactly the same. The distribution is symmetric about the mean—half the values fall below the mean and half above the mean. The distribution can be described by two values: the mean and the standard deviation.



Gaussian distribution

IT Skills for Chemist
(Statistical analysis)

Characteristics:

1. Ball shaped
2. symmetry about mean
3.mean, median, mode are equal
4. continuous distribution
5.never touch x-axix
6.area under curve is 1

Random vs. Systematic Error

In scientific research, **measurement error** is the difference between an observed value and the true value of something. It's also called observation error or experimental error.

There are two main types of measurement error:

- **Random error** is a chance difference between the observed and true values of something (e.g., a researcher misreading a weighing scale records an incorrect measurement).
- **Systematic error** is a consistent or proportional difference between the observed and true values of something (e.g., a miscalibrated scale consistently registers weights as higher than they actually are).

In research, systematic errors are generally a bigger problem than random errors.

Random error isn't necessarily a mistake, but rather a natural part of measurement. There is always some variability in measurements, even when you measure the same thing repeatedly, because of fluctuations in the environment, the instrument, or your own interpretations.

But variability can be a problem when it affects your ability to draw valid conclusions about relationships between variables. This is more likely to occur as a result of systematic error.

Dr Sandip Kr Rajak

IT Skills for Chemist
(Statistical analysis)

Precision vs accuracy

Random error mainly affects **precision**, which is how reproducible the same measurement is under equivalent circumstances. In contrast, systematic error affects the **accuracy** of a measurement, or how close the observed value is to the true value.

Taking measurements is similar to hitting a central target on a dartboard. For accurate measurements, you aim to get your dart (your observations) as close to the target (the true values) as you possibly can. For precise measurements, you aim to get repeated observations as close to each other as possible.

Random error introduces variability between different measurements of the same thing, while systematic error skews your measurement away from the true value in a specific direction.

When you only have random error, if you measure the same thing multiple times, your measurements will tend to cluster or vary around the true value. Some values will be higher than the true score, while others will be lower. When you average out these measurements, you'll get very close to the true score.

For this reason, random error isn't considered a big problem when you're collecting data from a large sample—the errors in different directions will cancel each other out when you calculate descriptive statistics. But it could affect the precision of your dataset when you have a small sample.

Systematic errors are much more problematic than random errors because they can skew your data to lead you to false conclusions. If you have systematic error, your measurements will be biased away from the true values. Ultimately, you might make a false positive or a false negative conclusion (a Type I or II error) about the relationship between the variables you're studying.

Statistical significance testing: The t test. The F test

The t-test and F-test are both statistical tests that help researchers determine if there are differences between groups or variables. They are often used in hypothesis testing to help researchers decide whether to accept or reject the null hypothesis. The main difference between the two tests is the number of groups they compare:

T-test

Compares the means of two groups, such as two independent groups, or the effects of treatment on the same group over time. It can also be used to compare the relationship between a categorical variable and a continuous variable. To determine if a t-test result is statistically significant, you can compare the calculated t-score to a table value. If the t-score is less than 0.05, then the result is statistically significant.

F-test

Compares the means of two or more groups, or the variances between samples. It's also called the F-ratio and is often used in analysis of variance (ANOVA). To determine if an F-test result is statistically significant, you can compare the F-statistic to a table value. If the F-statistic is larger, then the result is significant, which indicates greater differences between the sample averages.

IT Skills for Chemist
(Statistical analysis)

Independent sample t-test

Number of words recalled

Group 1	Group 2 (Imagery)
21	22
19	25
18	27
18	24
23	26
17	24
19	28
16	26
21	30
18	28
mean = 19	mean = 26
std = sqrt(40)	std = sqrt(50)

$df = (n_1-1) + (n_2-1) = 18$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}} = \frac{19-26}{1} = -7$$

$$t_{(0.05,18)} = \pm 2.101$$

$$t > t_{(0.05,18)}$$

→ Reject H_0

Dr Sand

……******End*******……