Quantitative Methods in Chemistry

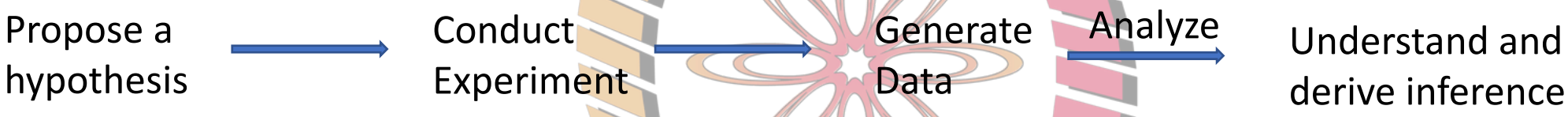
Week 3

Topics:
Hypothesis validation (null hypothesis, confidence levels, confidence intervals, one-tail test, two-tail test, use of statistical tables such as z-table, t-table, F-table, identifying outliers in data with Q-test)

Learning objectives:
Define the concept of hypothesis. Compare and contrast the available statistical tools to develop and verify hypotheses.

Department of Chemistry, https://chm.iiserb.ac.in/

This week we will understand how statistical tools can be employed to analyze scientific data and arrive at an unbiased conclusion or inference.



What is a “hypothesis”?

Dictionary definition is “a supposition or a proposed explanation based on limited evidence as a starting point for further investigation.”

It is a proposition without any assumption of its truth!

In other words, we need to test the hypothesis that is being proposed, and arrive at a conclusion that is not influenced by our personal biases or beliefs.

Examples of testable hypotheses:

(i) A new treatment is significantly better than the previous one in treating a certain disease. (It may or may not be!)

(ii) The water or air quality of the city has deteriorated in the last few years. (Who knows what is the truth?)

(iii) Student X in course 1 has better academic credentials than student Y in course 2. (This also may or may not be true)

However, Human beings (including scientists!) are prone to personal biases.

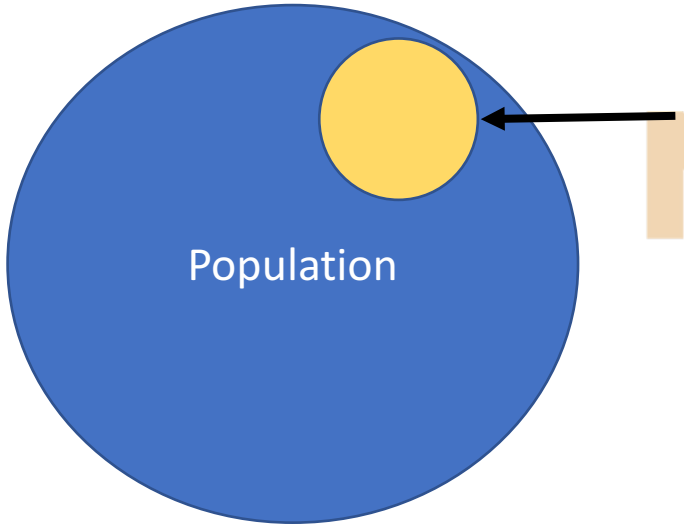
Q. How do we remove these while drawing inferences from scientific data?

A. Statistics provide us scientific tools to analyze data and test hypotheses in unbiased manner.

Quick re-cap of the statistical terms introduced in this course:

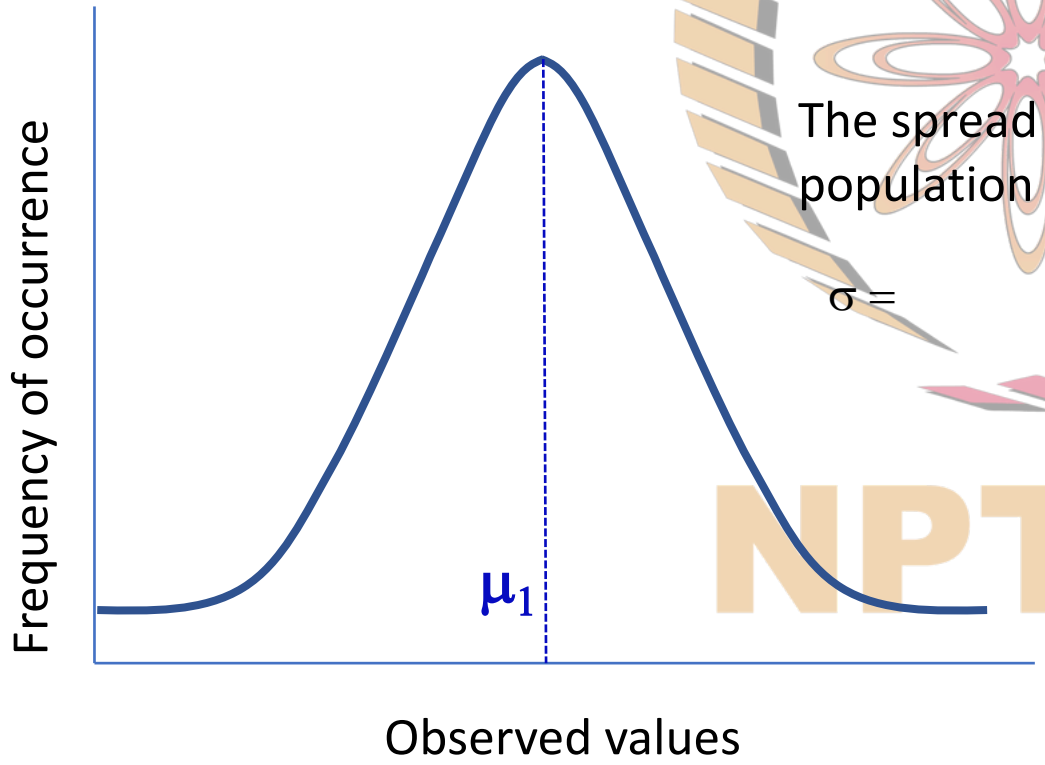
1. Population – A collection of **all** the data points/ measurements of interest to us. E.g. all the cells involved of a particular type, all the items being produced by a company, all the citizens of a city or state or country. A population contains complete the information necessary to make inferences but can be too tedious or impractical to analyze.

2. Sample – A subset of Population from which inferences about the population are to be drawn. It provides approximate information or indication about the population. It does not contain complete information about the population but is more convenient to analyze.



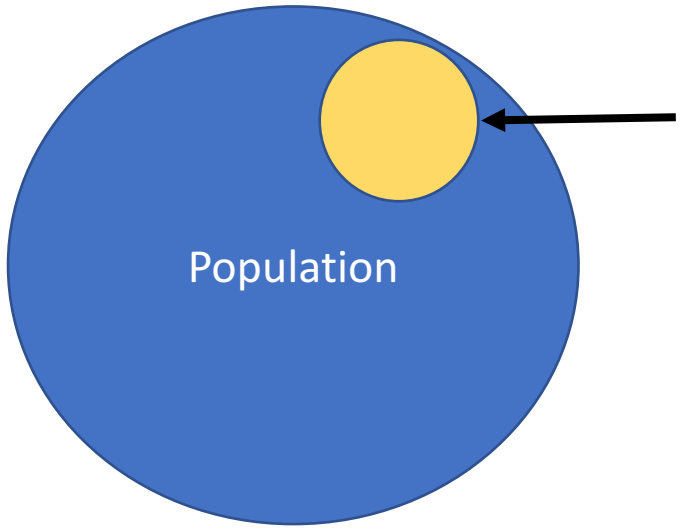
Any randomly accumulated large data set is considered to follow a Gaussian distribution.

It is distributed symmetrically around the mean value (μ).



The spread of the data is given by the population standard deviation (σ)

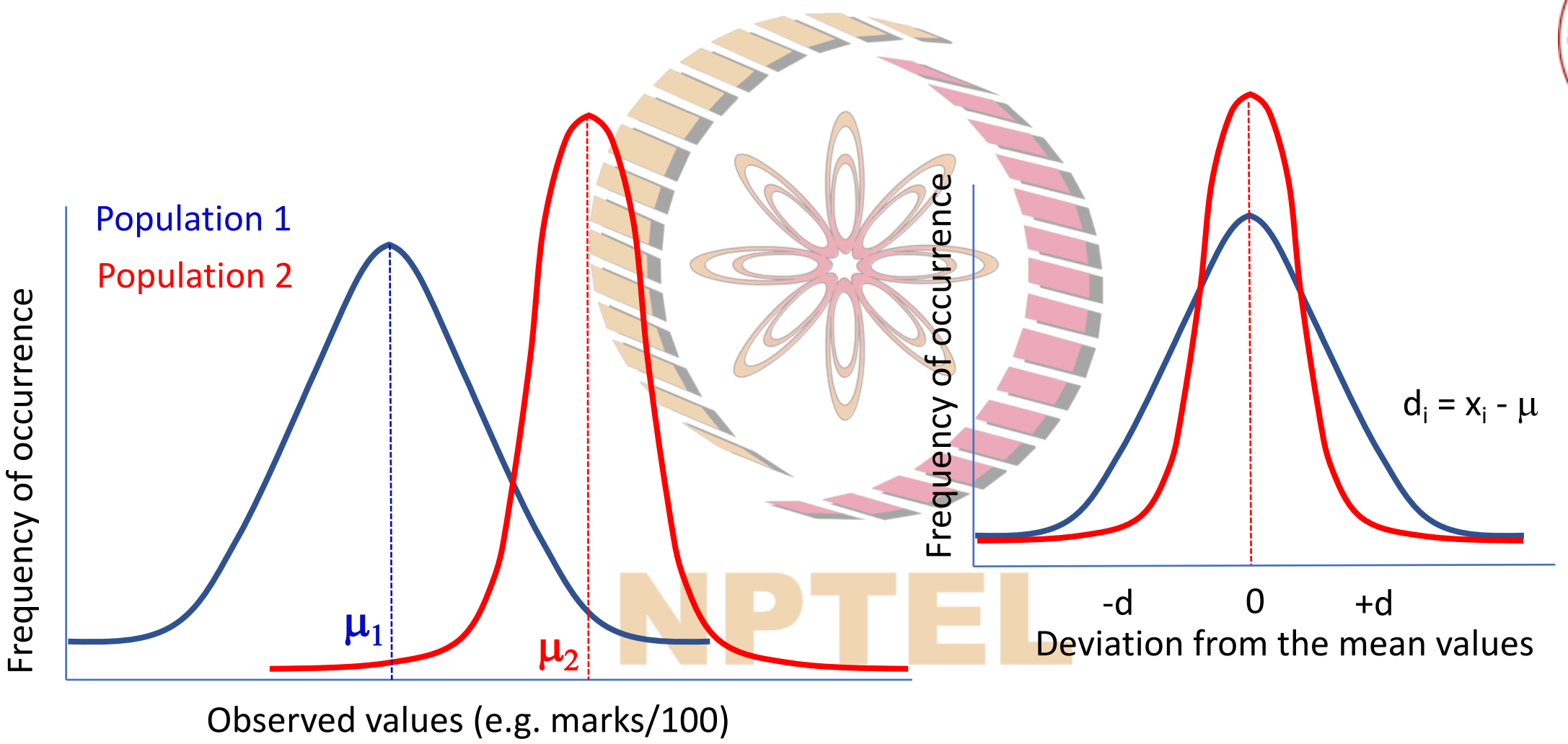
$\sigma =$

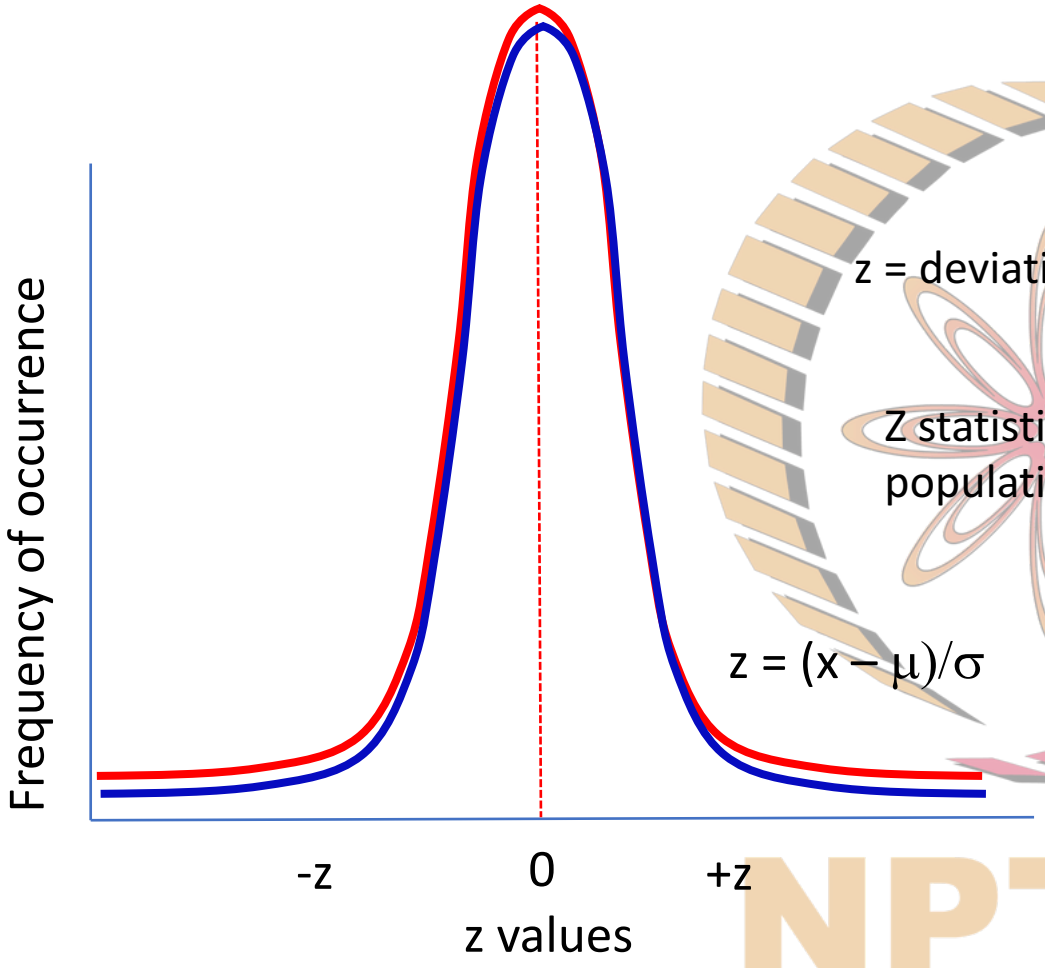


Sample (as an approximation of population)

Sample mean is denoted by

Sample standard deviation is denoted by

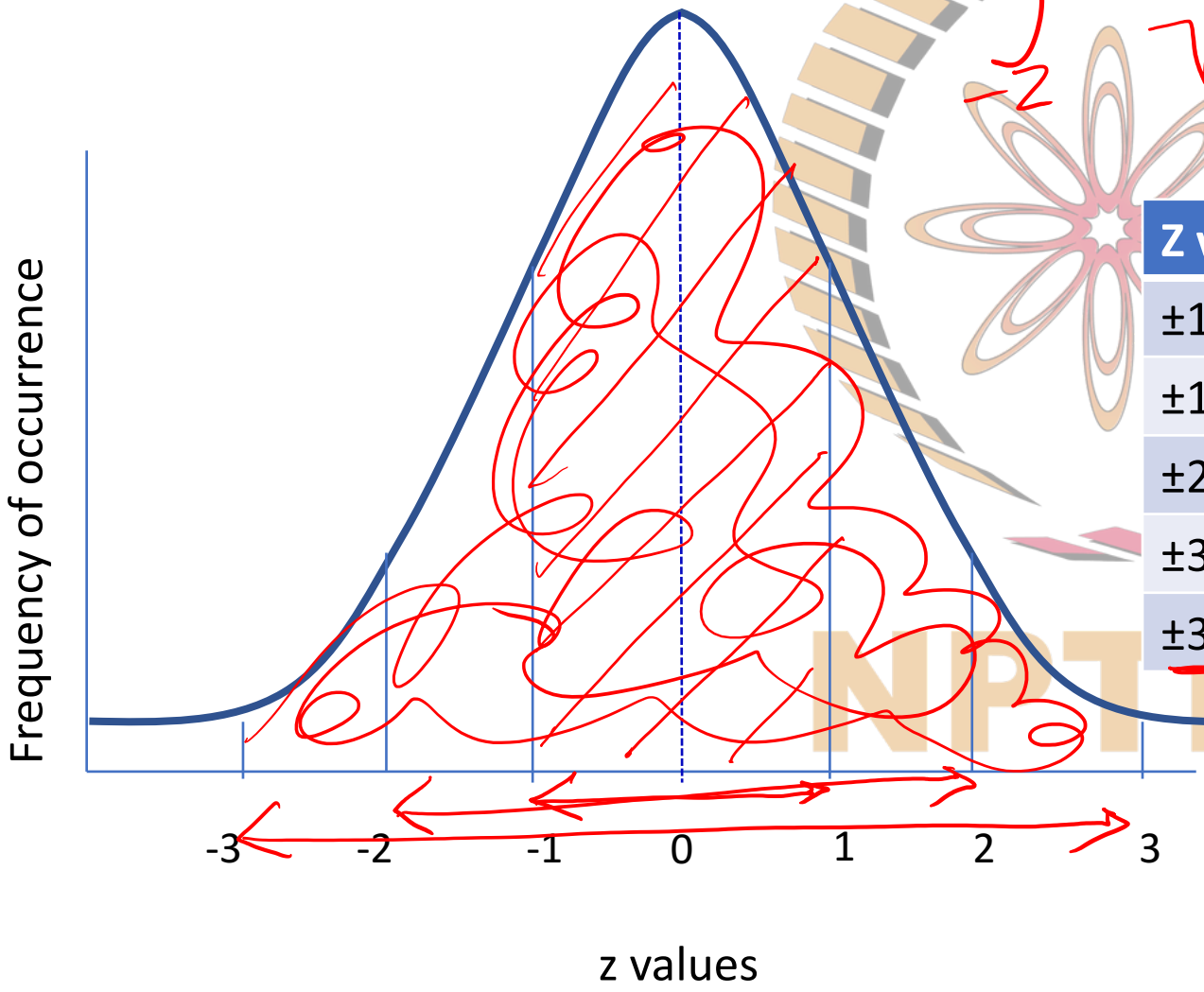

Population 1
Population 2
Frequency of occurrence
μ_1
μ_2
Observed values (e.g. marks/100)
Frequency of occurrence
$d_i = x_i - \mu$
-d
0
+d
Deviation from the mean values



z = deviation (in std. dev. Units)

Z statistics is applied when we know the population mean and population standard deviation

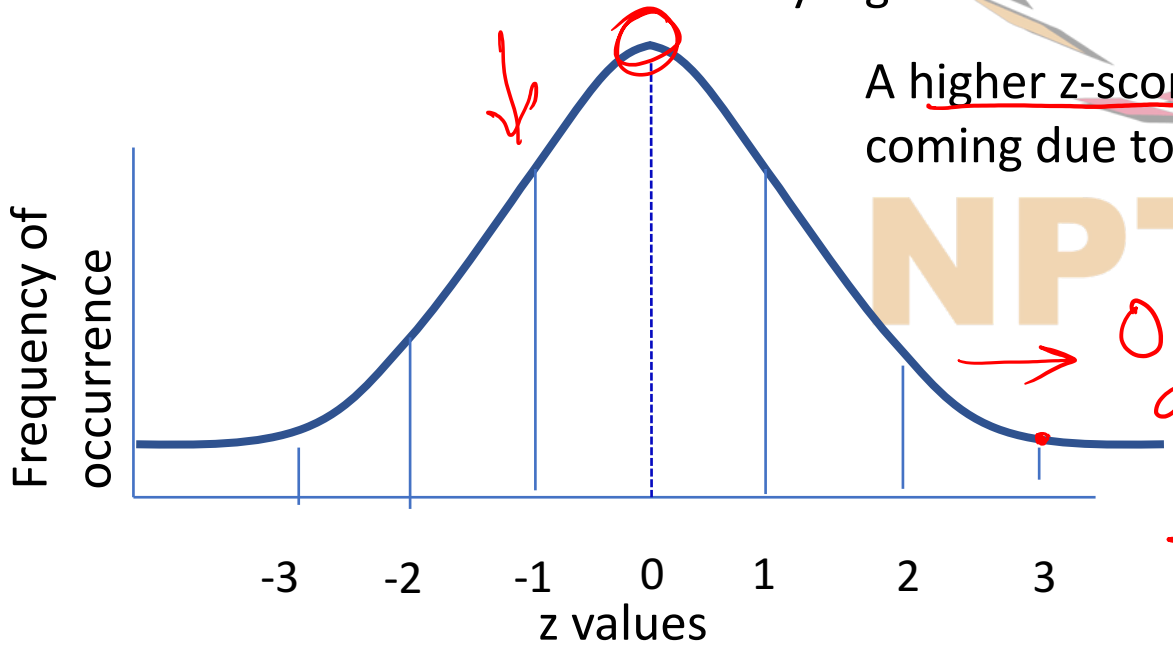
$z = (x - \mu)/\sigma$



$$AUC = \int_{-z}^{+z} \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

Z value	AUC
±1.00	68.3%
±1.96	95%
±2.00	95.4%
±3.00	99.7%
±3.29	99.9%

- Random fluctuations during measurements cause the repeat measurement values to take a Gaussian profile.
- These fluctuations are natural and beyond our control.
- For a data following Gaussian profile, the frequency of occurrence of a value very far from the population mean due to random fluctuations diminishes rapidly, while that near the mean value is very high.



A higher z-score indicates that the probability of that reading coming due to random fluctuations is rather low!

0.3% coming due to random fluctuations

A few Statistical terms to be acquainted with

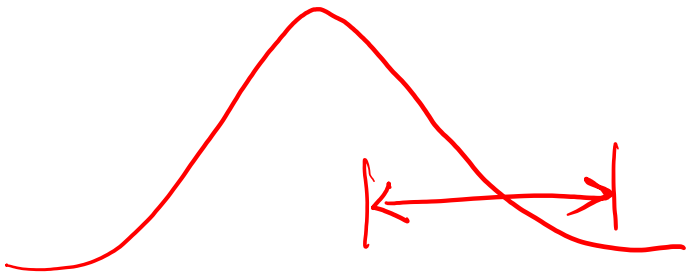
Confidence Interval (C I) – The range of values within which the population mean is expected to occur with a certain probability. The larger the probability, the larger is the interval.

The limits of this interval are known as **Confidence Limits**.

e.g. we can say: There is 95% probability that the population mean for Calcium in blood of human beings lies within 2.4 ± 0.2 mM. This is the Confidence Interval.

The Confidence Limits (or extremities of Confidence Interval) for this data set is 2.2 – 2.6 mM.

Confidence Level (C L) is the probability of finding the population mean within the Confidence Interval. In the above example, Confidence Level is 95%.



C I of the population mean depends on the **C L**

Larger the C L needed, larger will be the Confidence Interval for the mean

e.g. the "Normal" Blood Calcium level = 2.4 ± 0.2 mM for 95% population

Now. **C I for μ (for single reading) = $\bar{x} \pm z\sigma$**

C I for μ (for N readings) = $\bar{x} \pm \frac{z\sigma}{\sqrt{N}}$

So, to halve the C I (at the same C L), we need to take **4N** readings!

Q. How many readings are required to reduce the C I to 1/3rd the original value?

9x

Z-statistics could be applied only when we have good estimate of the population standard deviation, σ
(This usually happens when we have a large number of observations available with us, or, when the sample size is quite large).

What to do if we do not have a good estimate of σ? (e.g. when dealing with small samples?)

William Sealy Gosset

Karl Pearson

Ronald A Fisher

Gosset developed statistical methods to analyze small samples.
He published his findings in 1908 in journal *Biometrika* under the pseudonym "Student".

C I can be expressed in terms of t-statistics (for small samples)

C I for μ (for N readings) = $\bar{x} \pm \frac{ts}{\sqrt{N}}$

z statistics

$\bar{x} \pm \frac{z\sigma}{\sqrt{N}}$

t value to be inserted from the table based on the degrees of freedom and the **significance level**

t Distribution: Critical Values of *t*

Degrees of freedom	*Two-tailed test:* / *One-tailed test:*	*Significance level* 10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
∞	→	1.645	1.960	2.326	2.576	3.090	3.291

Significance level (α) = 100 − CL %

5% α ⇒ 95% CL

Key take-aways from the statistical t-table

- t-values are larger than the z-values at same confidence level or significance level
- As a result, the C I for μ is larger when t-statistics is used instead of the z-statistics
- As $N \rightarrow \infty$, $t \rightarrow z$

Confidence Level and Significance Level

Confidence Level is the probability of finding the population mean with a certain interval of sample mean.
This interval is known as the confidence interval.

Significance Level (denoted by α, and expressed as %) is the probability of a result being outside the confidence interval.

In other words, Significance Level tells us about the "**Statistical Significance**" of a particular measurement value.

Often, $\alpha = 100 - CL$