Basic statistical tools for analytical laboratories

Chapter 1
Chemical Measurements & Errors

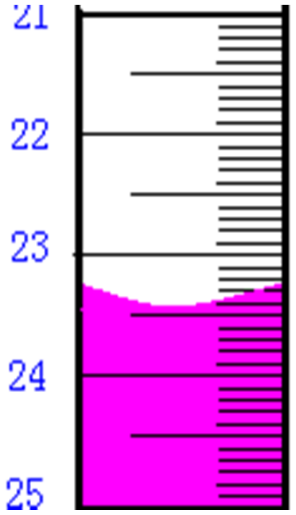
Measurement Errors

- There is always an element of **error** in all measurements, because of many uncontrollable experimental factors like instrumental signals, environmental conditions, etc.
- All test methods *attempt* to determine the '**true value**' of an analyte (or measurand).
- Error = | measured value – true value |

Measurement Errors

- **Where can we find true value?**
- We do not know the 'true value' of the analyte that we want to measure in a sample,
- But we can have one or more of the following *similar* samples for analysis to provide us a 'true value' or reference value for test comparison:
  - Assigned, known or given value
  - Certified reference value by a third party
  - Spiked value by adding known amount of analyte to a matrix
  - Consensus value of proficiency testing sample reached by acceptable participating laboratories
- *But, these 'true values' also have their associated uncertainties.*

Measurement Errors



- **Three types of error**:
- **Gross error** : no excuse
- **Systematic error** β : bias (personal, instrumental and methodological) may be *minimized* or *corrected*
- **Random error** ε : cannot be avoided due to uncontrollable experimental factors
- Random error can be estimated through repeated measurements

Measurement Errors

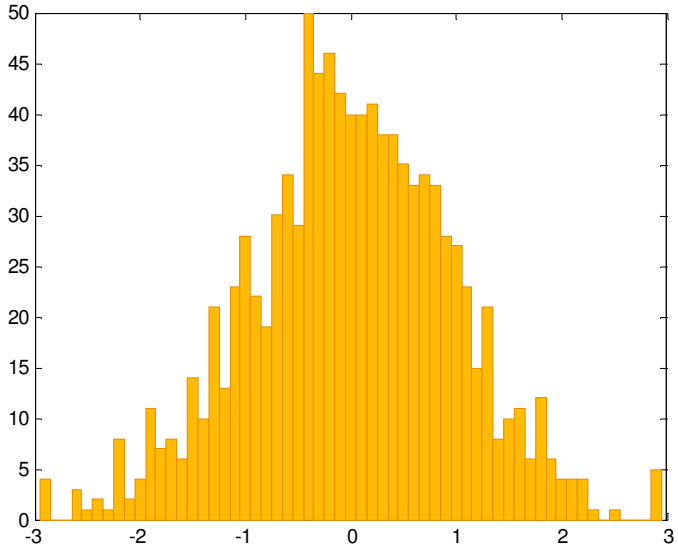
- Measured value, q
- $= TV(X) + \beta + \varepsilon$
- total measurement error δ
- $= \beta + \varepsilon$
- We can try our very best to eliminate β or find a correction factor to rectify the systematic error, but have to deal with the presence of random error, ε.
- We can *estimate* such random error.

Measurement Errors

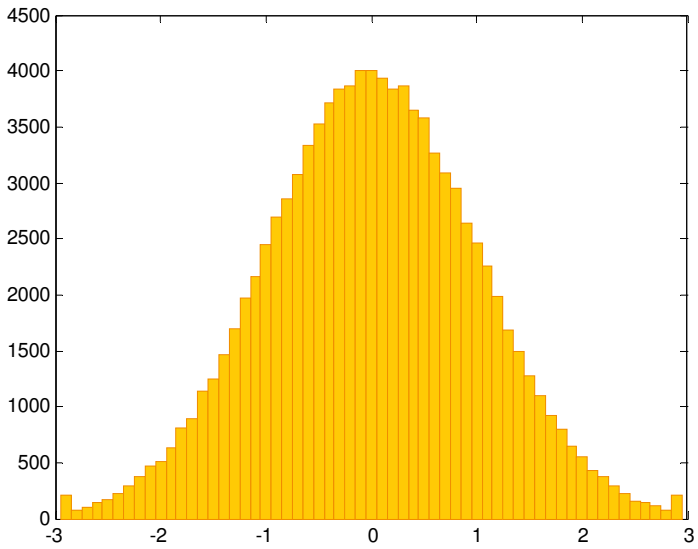
- Upon repeating the analysis for several times, we obtain many test results close to the average of the figures we get, and the larger or smaller results tend to spread away in lesser occurrence.
- This can be reflected in a **histogram**, and a rather smooth bell-shaped curve can be drawn upon the results when the number of results obtained is *very large.*

Histograms

- 1000 data



- 10000 data



Estimation of Random Error

- If repeated independent measurements on a sample are $x_1, x_2, x_3, .., x_i, ..., x_n$
- **Sample Mean (Average)** – $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Median** – the middle of a set of figures when arranged in an ascending order
- **Mode** – the figures which occur the most number of times.
- *In a perfect experiment, all the above three figures are the same.*

Estimation of Random Error

- Data dispersion is best estimated by **variance** and its square root, the **standard deviation**
- Both statistics measure how much each observation deviates from a central point represented by the mean.
- Hence, **deviation** $d = (x_i - \overline{x})$
- and sum of deviations

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

Estimation of Random Error

- **Sum of squares SS =** $\sum_{i=1}^{n} (x_i - \bar{x})^2$
- **Sample standard deviation,**

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

- **Variance** = s^2
- **Coefficient of variation,** $CV = s / \bar{x}$
- **Relative standard deviation,**

$$\%\,RSD = 100 * CV$$

Estimation of Random Error

- Alternative expression of sample standard deviation derived from the last equation:

- $$s = \sqrt{\frac{[\sum x_i^2 - (\sum x_i)^2 / n]}{n-1}}$$

Estimation of Random Error

- **Another important formula for standard deviation**:
- In a set of *independent* **duplicated** measurements, standard deviation, *s* can be expressed as:
- $s = \sqrt{(\Sigma \Delta^2 / 2k)}$ where
  - Δ = *difference between the duplicated results*
  - *k = number of sets of duplicated results*

Example

Copper content in a liquid sample (mg/L)

(duplicate analyses)

k	a	b	$Diff=(a-b)$	$Diff^2$
1	22.5	23.8	-1.30	1.69
2	21.9	22.5	-0.60	0.36
3	23.6	22.3	1.30	1.69
4	22.1	23.6	-1.50	2.25
5	24.0	22.9	1.10	1.21
$s = \sqrt{(\Sigma D^2 / 2k)}$			Sum D^2	7.20
			SD =	0.849

Note:

s (duplicates) > s (n repeats at one time - repeatability)

Pooling of Standard Deviations

- Estimates of the standard deviation obtained at several times may be combined (*pooled*) to obtain a better estimate based upon more degrees of freedom:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2+(n_3-1)s_3^2+\ldots+(n_k-1)s_k^2}{(n_1-1)+(n_2-1)+(n_3-1)+\ldots+(n_k-1)}}$$

- where s_p is based on

$$(n_1-1)+(n_2-1)+(n_3-1)+\ldots+(n_k-1)$$

degrees of freedom

Example

No. of replicates, n

Day #	1	2	3	4	5	Mean	Std Dev
1	96.4	98.3	97.3	98.1	98.8	97.8	0.942
2	99.3	98.1	97.8			98.4	0.794
3	98.8	97.2	99.0			98.3	0.987
4	98.6	99.5	99.8	97.9		99.0	0.866

$$Sp = \sqrt{\frac{(5-1)\times 0.942^2 + (3-1)\times 0.794^2 + (3-1)\times 0.987^2 + (4-1)\times 0.866^2}{(4+2+2+3)}}$$

Pooled

Sp = 0.905

Comparisons of Std Deviation

Trial #	Analyte,%
1	**2.58**
2	**2.69**
3	**2.57**
4	**2.60**
5	**2.52**
6	**2.66**
Mean =	**2.60**
Std Dev =	**0.062**
n =	6

Set #	Dup a,%	Dup b,%	Diff (D)	D^2
1	**2.58**	**2.69**	-0.11	0.0121
2	**2.57**	**2.60**	-0.03	0.0009
3	**2.52**	**2.66**	-0.14	0.0196
			Sum(D^2) =	0.0326
			No. of Set k =	3
			Std Dev =	**0.074**

Repeatability & Reproducibility

- *For chemistry*
- **Repeatability, *r*,** is defined as measurement precision under repeatability (*Normal distribution*) conditions,
- i.e. conditions where independent test results are obtained with
  - the same method on
  - identical test items
  - in the *same* laboratory
  - by the *same* operator
  - using the same equipment within short intervals of time (ISO 3534-2).

Repeatability & Reproducibility

- ***Repeatability standard uncertainty,*** s_r is the **standard deviation** of test results obtained under repeatability conditions *within* a laboratory.
- The relationship between r and s_r is given as:
- $r = 2*\sqrt{2}*s_r$ *or* $r = 2.8\, s_r$

$r = 2.8 s_r$ **Why?**

- Let a duplicated analysis on a sample gives results a and b with associated standard deviation s_a and s_b, respectively
- Here, sample std deviation $s_r = s_a = s_b$
- Therefore, combined std deviation of this duplicated analysis is

$$s_r = \sqrt{s_a^{\,2} + s_b^{\,2}} = \sqrt{s^2 + s^2} = \sqrt{2s^2} = \sqrt{2}s$$

- A coverage factor of 2 is then used in a normal distribution with 95% confidence to be obtain r, i.e. $r = 2 \times \sqrt{2} s_r = 2.8 s_r$

Repeatability & Reproducibility

- **Reproducibility, R,** is precision under reproducibility conditions,
- i.e. conditions where independent test results are obtained
  - with the same method
  - on identical test samples
  - in *different* laboratories
  - by *different* operators using different equipment (ISO 3534-2).

Repeatability & Reproducibility

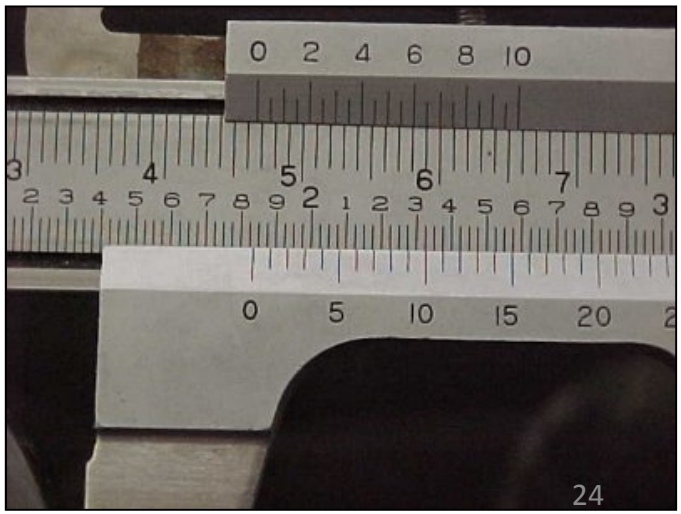
- ***Reproducibility standard uncertainty,*** s_R is the **standard deviation** of test results obtained under reproducibility conditions and is a measure of dispersion of test results under reproducibility conditions.
- The relationship between R and s_R is similarly given as:
- $$R = 2*\sqrt{2}*s_R \quad or \quad R = 2.8\, s_R$$

Accuracy & Precision

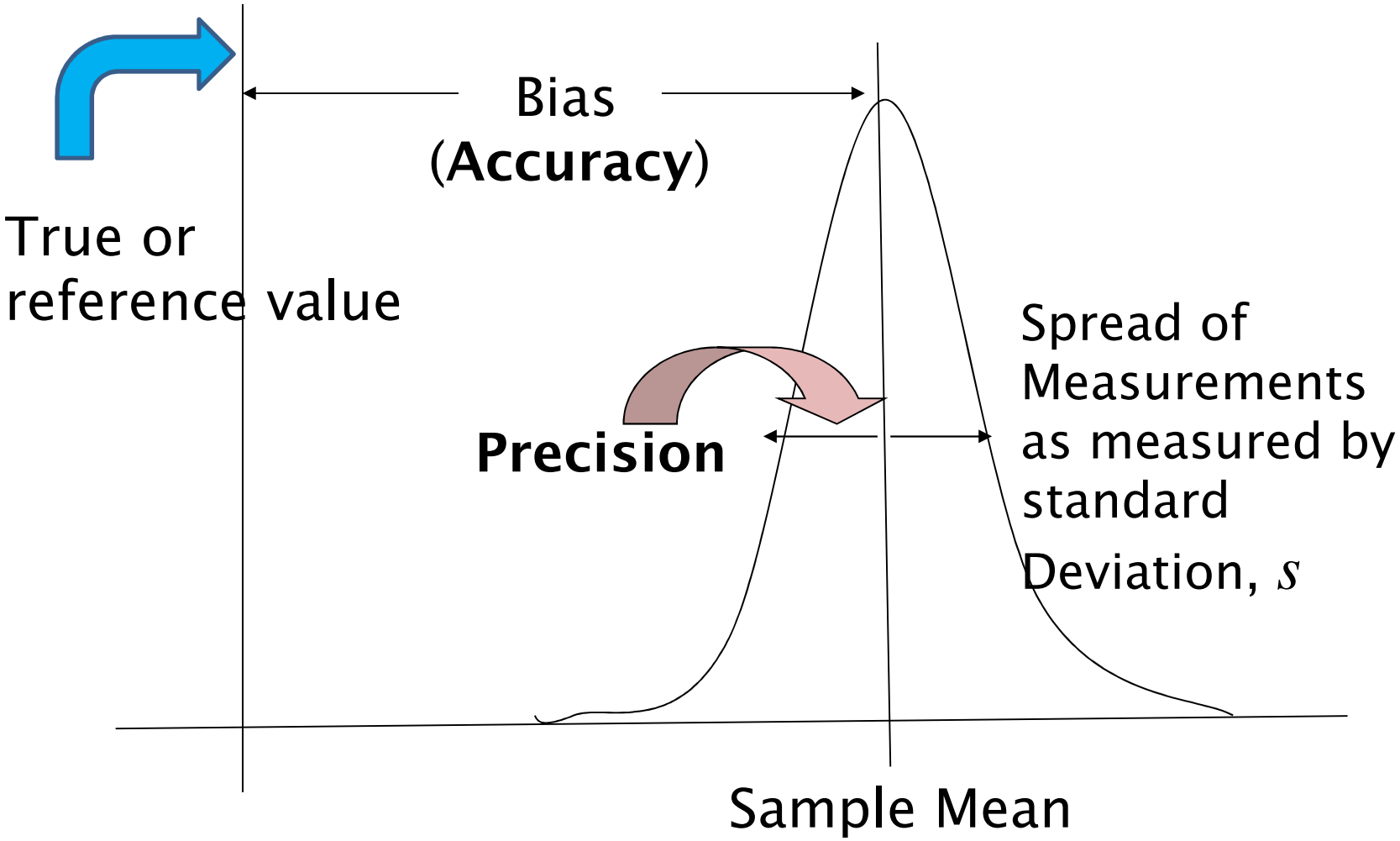
- **Accuracy**
- The closeness of agreement between a test result and a true value of the analyte, i.e. *how **close** the measurement data is to the "true" value.*

Accuracy & Precision

- **Precision**
- The closeness of agreement between independent test results obtained under stipulated conditions; *how much the results are **spread around**, or **deviate from** the "true"" value.*

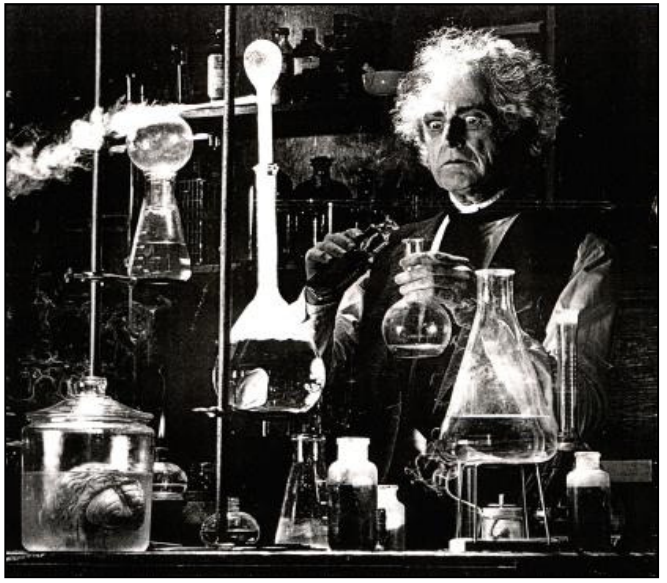


Accuracy & Precision



Accuracy & Precision

- Laboratory analysts always want to produce the most ***precise*** and ***accurate*** experimental data.
- Measurements can be:
  - Precise but not accurate
  - Not precise and not accurate
  - Precise and accurate
  - Not precise and accurate



Accuracy & Precision