Chemometrics

Application of mathematical, statistical, graphical or symbolic methods to maximize chemical information.

-However, this definition can be expanded to include:

biology (biometrics),

environmental science (environmetrics),

economics (econometrics).

-Two lines of development:

- experimental design: planning and performing experiments in a way that the resulting data contains the maximum information about stated questions.
- multivariate data analysis: utilizing all available data in the best possible way.

Chemometrics

1. Errors in Quantitative Analysis & Descriptive Statistics
2. Signal Processing & Time-Series Analysis
3. Experimental Design & Optimization
4. Factorial Designs & Analysis
5. Fractional Factorial Designs & Analysis
6. Univariate Calibration & Least Squares
7. Linear –vs- Multivariate Regression
8. Principal Component Regression
9. Quality Assurance & Good Laboratory Practice

Errors in Quantitative Analysis

- <u>Student A</u>

  Results have two characteristics:

  1. All very close to each other (10.08-10.12 mL).
  2. All too high (10.00 mL exactly).

  Two types of errors have occurred:

  1. Random – cause replicate results to differ from one another → results fall on both sides of the mean (10.10 mL in student's A case).

     → effect the precision or reproducibility of the experiment.

     → Student A – small random errors (precise).
  2. Systematic – cause all the results to be in error in the same sense (High).

     → Total systematic error is termed bias.

  Hence: Student A has precise, but biased results.

Errors in Quantitative Analysis

- No analysis is free of error!
- <u>Types of errors</u>
  1. Gross – readily described, errors that are obvious.

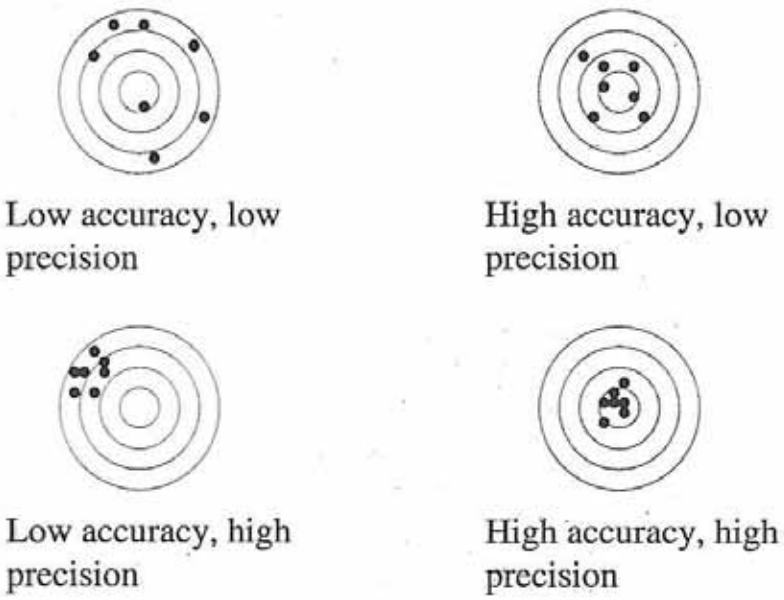
     Instrument breakdown, dropping a sample, contamination (gross).
  2. Random
  3. Systematic

Student	Results (mL)					Comments
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise, unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.18	10.02	9.97	10.04	Precise, unbiased

Titration – each student performs an analysis in which exactly 10.00 mL of exactly 0.1 M NaOH is titrated with exactly 0.1M HCl.

Errors in Quantitative Analysis

- Accuracy – How far a result is from the true value.
- Precision – How close multiple determinations are to each other.



Descriptive Statistics

1. Basic Statistics → Descriptive

Spectrophotometric measurement (Abs) of a sample solution from 15 replicate measurements.

Measurement	Value	Measurement	Value
1	0.3410	9	0.3430
2	0.3350	10	0.3420
3	0.3470	11	0.3560
4	0.3590	12	0.3500
5	0.3530	13	0.3630
6	0.3460	14	0.3530
7	0.3470	15	0.3480
8	0.3460		

Descriptive Statistics

Descriptive statistics for the spectrophotometric measurements.

Parameter	**Value**
Sample #, n	15
Mean	0.3486
Median	0.347
Std Dev	0.00731
RSD %	2.096
Std error	0.00189
Max value	0.363
Min value	0.335

Statistical Tests – Student's *t*-test, *F*-test, tests for outliers
Distributions – Gaussian, Poisson, binominal

Descriptive Statistics

1. Statistics of Repeated Measurements

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

$$Standard\ deviation\ \ s = \sqrt{\sum (x - \bar{x})^2 / (n-1)}$$

Student A

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	10.08	-0.02	0.0004
	10.11	0.01	0.0001
	10.09	-0.01	0.0001
	10.10	0.00	0.0000
	10.12	0.02	0.0010
Total	50.50	0	0.0010

Descriptive Statistics

$$\bar{x} = \frac{\sum x_i}{n} = \frac{50.50}{5} = 10.1 mL$$

$$s = \sqrt{\sum (x - \bar{x})^2 / (n-1)} = \sqrt{0.001/4} = 0.0158 mL$$

Descriptive Statistics

2. Distribution of Repeated measurements
   1. Gaussian distribution
      - **Bell-shaped curve for the frequency of the measurements**

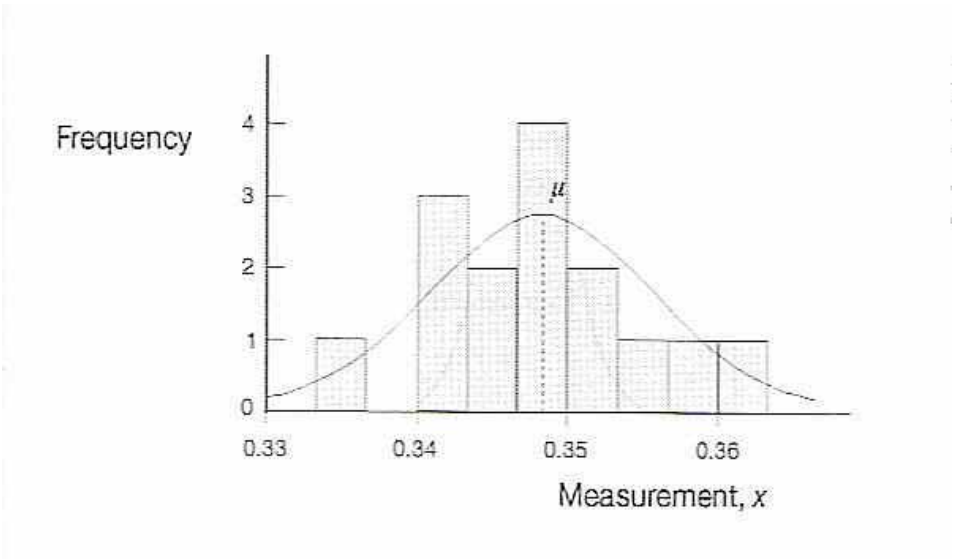


Fig. 1 – Histogram for the measurements of spectrophotometric data. Theoretical distribution with the Gaussian curve in the solid line.

Descriptive Statistics

A. Gaussian Distribution

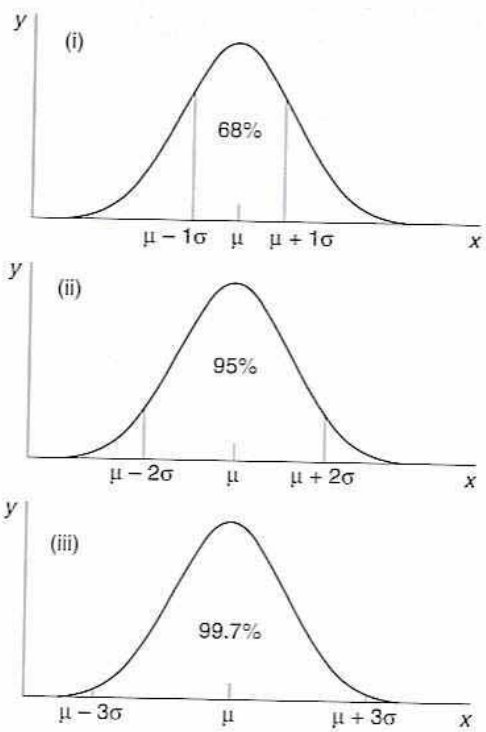


Figure 2.4 Properties of the normal distribution: (i) approximately 68% of values lie within $\pm 1\sigma$ of the mean; (ii) approximately 95% of values lie within $\pm 2\sigma$ of the mean; (iii) approximately 99.7% of values lie within $\pm 3\sigma$ of the mean.

Descriptive Statistics

3. Significance Tests in Analytical Measurements

- Testing the truth of the hypothesis (null hypothesis, H_o)
- Null = implies that no difference exists between the observed and known values $\bar{x}$

Assuming H_o is true, stats can be used to calculate the probabilty that the difference between and true value, μ, arises solely as a result of <u>random errors</u>.

Is the difference significant?

$$t = \frac{|\bar{x} - \mu|}{s}\sqrt{n} \qquad \text{one variable} \quad t-\text{test} \quad (\text{student's } t)$$

where s = estimate of the standard deviation

n = number of parallel measurements

Descriptive Statistics

3. Statistical tests

   A. Student's t

   - If $|t|$ exceeds a certain critical value, then the H_o is rejected,

Table A.2 The t-distribution

Value of t for a confidence interval of	*90%*	*95%*	*98%*	*99%*
Critical value of |t| for P values of number of degrees of freedom	*0.10*	*0.05*	*0.02*	*0.01*
1	6.31	12.71	31.82	63.66
2	2.92	4.30	6.96	9.92
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.89	2.36	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
12	1.78	2.18	2.68	3.05
14	1.76	2.14	2.62	2.98
16	1.75	2.12	2.58	2.92
18	1.73	2.10	2.55	2.88
20	1.72	2.09	2.53	2.85
30	1.70	2.04	2.46	2.75
50	1.68	2.01	2.40	2.68
∞	1.64	1.96	2.33	2.58

The critical values of $|t|$ are appropriate for a *two*-tailed test. For a *one*-tailed test the value is taken from the column for *twice* the desired P-value, e.g. for a one-tailed test, $P = 0.05$, 5 degrees of freedom, the critical value is read from the $P = 0.10$ column and is equal to 2.02.

Descriptive Statistics

3. Statistical tests

A. Student's *t*

Ex: The following results were obtained in the determination of Fe^{3+} in water samples.

504 ppm 50.7 ppm 49.1 ppm 49.0 ppm 51.1 ppm

Is there any evidence of systematic error?

$\bar{x} = 50.06$ $s = 0.956$

H_o → no systematic error , i.e., $\mu = 50$ and using equations

Table: critical value is t_4 = 2.78 (p=0.05).

<u>NO</u> significant difference since the observed $|t|$ is less than the critical value (H_o retained)

Descriptive Statistics

3. Statistical tests

B. Two-sided *t*-test

- A comparison of two sample means: $\bar{x}_1 - \bar{x}_2$

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s_d}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where : n_1, n_2 = # of parallel determinations for $\bar{x}_1$ *and* $\bar{x}_2$

s_d = weighted averaged standard deviation:

$$s_d = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Note: The null hypothesis is accepted if $\bar{x}_1$ *and* $\bar{x}_2$ are different only randomly at risk level *P*, i.e. if the calculated *t*-value is lower than the tabulated value for *t*.

Descriptive Statistics

3. Statistical test

C. *F*- test

- Used to compare the standard deviations of two random samples

$$F = \frac{s_1^2}{s_2^2} \qquad (\text{where } s_1^2 > s_2^2)$$

Note : The null hypothesis is accepted if s_1^2 *and* s_2^2 differ randomly, i.e., if the calculated *F*-value is lower than *F*-distribution from the Table with $v_1 + v_2$ degrees of freedom.

Descriptive Statistics

3. Statistical test
   F- test

Table A.3 Critical values of F for a one-tailed test ($P = 0.05$)

ν_2	ν_1												
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*12*	*15*	*20*
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.703	8.660
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858	5.803
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.619	4.558
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.938	3.874
7	5 591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511	3.445
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.218	3.150
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.006	2.936
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.845	2.774
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.719	2.646
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617	2.544
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.533	2.459
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.463	2.388
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.403	2.328
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.352	2.276
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.308	2.230
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.269	2.191
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.234	2.155
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.203	2.124

ν_1 = number of degrees of freedom of the numerator and ν_2 = number of degrees of freedom of the denominator.

Descriptive Statistics

3. Statistical test

Ex. Determination of titanium content (absolute %) by two laboratories.

Lab 1	Lab2
0.470	0.529
0.448	0.490
0.463	0.489
0.449	0.521
0.482	0.486
0.454	0.502
0.477	
0.409	

Descriptive Statistics

Compare the standard deviations between the two laboratories:

$$F = \frac{s_1^2}{s_2^2} = \frac{0.0229^2}{0.0182^2} = 1.58 \quad \text{Critical Value from the table} = 6.85$$

Calculated value is lower than the tabulated, hence the test result is not significant (only random differences).

Descriptive Statistics

3. Significance Tests

D. Dixon's *Q*-test

→ Test for the determination of outliers in data sets

$$Q = \frac{|x_2 - x_1|}{|x_n - x_1|} \quad and \quad Q_n = \frac{|x_n - x_{n-1}|}{|x_n - x_1|}$$

→ The H_o , i.e., that no outlier exists, is accepted if the quantity $Q<Q(1\text{-}P;n)$

Descriptive Statistics

Table: Critical values for the *Q*-test at H_o at the 1% risk level

n	***Q(0.99; n)***
3	0.99
4	0.89
5	0.76
6	0.70
7	0.64
8	0.59
9	0.56
10	0.53
11	0.50
12	0.48
13	0.47
14	0.45
15	0.44
20	0.39

Descriptive Statistics

3. Significance Tests

Ex: Trace analysis of PAH's in soil reveals benzo [a] pyrene in the following amounts (mg kg^{-1})

5.30 5.00 5.10 5.20 5.10 6.20 5.15

→ Use the Q test to determine if the largest & smallest are outliers

$$Q = \frac{|5.10 - 5.00|}{|6.20 - 5.00|} = 0.083 \qquad Q_n = \frac{|6.20 - 5.30|}{|6.20 - 5.00|} \; 0.75$$

Critical values from the Table, Q(1-P = 0.99; n = 7) = 0.64

For the smallest, $Q_1 < Q_{calc}$, thus the value cannot be an outlier

For the largest, $Q_1 > Q_{calc}$, then the outlier exists

Why is this determination important?

Descriptive Statistics

3. Significance Tests

E. Analysis of variance (ANOVA)

→ Separate and estimate the causes of variation, more than one source of random error. Example?

Note: There are numerous chemometric tools which can be used with ANOVA.

→ We will cover them in this course!

Descriptive Statistics

E. ANOVA

→ Within sample variation

Conditions	Replicate Measurements	Mean
A. Freshly prepared	102, 100, 101	101
B. Stored for 1 hr in the dark	101, 101, 104	102
C. Stored for 1 hr in the subdued light	97, 95, 99	97
D. Stored for 1 hr in bright light	90, 92, 94	97

For each sample, the variance can be calculated: $\sum \frac{(x_i - \bar{x})^2}{(n-1)}$

Descriptive Statistics

E. ANOVA

Sample A = $$\frac{(102-101)^2+(100-101)^2+(101-101)^2}{3-1}=1$$

Sample B = $$\frac{(101-102)^2+(101-102)^2+(104-102)^2}{3-1}=3$$

C + D = 4

Sample Mean variance

$$\frac{(101-98)^2+(102-98)^2+(97-98)^2+(92-98)^2}{3-1}=62/3$$

Note: This estimate has 3 degrees of freedom since it is calculated from 4 samples

Descriptive Statistics

E. ANOVA

Summarizing our calculations:

within – sample mean square = 3 with 8 d.f.

between – sample mean sample = 62 with 3 d.f.

$$F = 62/3 = 20.7$$

From Table – the critical value of F= 4.066 (P=0.05)

- since the calculated value of F is greater than this, the null hypothesis is rejected: the sample means <u>DO</u> differ significantly.

Descriptive Statistics

E. The Arithmetic of ANOVA calculations

Source of Variation	Sum squares	d.f.
Between – sample	$\sum_i \frac{Ti^2}{n} - \frac{T^2}{N}$	h-1
within – sample	by subtraction	by subtraction
Total	$\sum_i \sum_j x_{ij}^2 - \frac{T^2}{N}$	N-1

where N = nh = total # of measurements

Ti = sum of the measurements in the *i* th sample

T = sum of all the measurements

- The test statistic is F = between sample mean square / within sample mean square and the critical value is $F_{h-t,\ N-h}$

Descriptive Statistics

Test whether the samples in previous table were drawn from population, with equal means. (All values have had 100 subtracted from them)

				Ti	Ti^2
A	2	0	1	3	9
B	1	1	4	6	36
C	-3	-5	-1	-9	81
D	-10	-8	-6	-24	576
					$\sum_i T^2{}_i = 702$

n=3, h=4, N=12, $\sum_i \sum_j x_{ij}^{L} = 258$

Descriptive Statistics

Source of Variation	Sum of Squares	d.f.	Mean Square
Between sample	702/3-(-24)/12=186	3	186/3 = 62
Within sample	By subtraction =24	8	24/8=3
Total	258-(-24)2/12 = 210	11	
	F= 62/3 = 20.7 → Significant difference		

Descriptive Statistics, Errors and Calibration

Determining Concentration from Calibration Curve

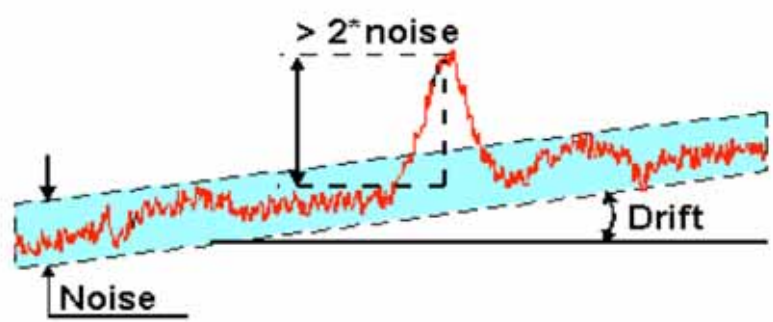
Basic steps:
(1) Make a series of dilutions of known concentration for the analyte.
(2) Analyze the known samples and record the results.
(3) Determine if the data is linear.
(4) Draw a line through the data and determine the line's slope and intercept.
(5) Test the unknown sample in duplicate or triplicate. Use the line equation to determine the concentration of the analyte: y = mx + b

$$\text{Conc}_{\text{analyte}} = \frac{\text{reading} - \text{intercept}}{\text{slope}}$$

Calibration Cont…

<u>Limit of Detection (LOD)</u> - lowest amount of analyte in a sample which can be detected but not necessarily quantitated as an exact value.

- mean of the blank sample plus 2 or 3 times the SD obtained on the blank sample (i.e., LOD = $mean_{blk} + Zs_{blk}$)



LOD calculation - alternative
Data required:
(1) calibration sensitivity = slope of line through the signals of the concentration standards including blank solution
(2) standard deviation for the analytical signal given by the blank solution

$$\text{LOD} = \frac{3x\ SD\ blank\ signals}{\text{slope of signals for std's}}$$

Outliers

Treatment of Outliers

1. Re-examine for Gross Errors
2. Estimate Precision to be Expected
3. Repeat Analysis if Time and Sufficient Sample is Available
4. If Analysis can not be Repeated, Perform a Q-Test
5. If Q-Test Indicates Retention of Value, Consider Reporting the Median