Module 5 : Data Presentation

In this module, various presentation methods for different types of data are introduced. For example, standard curve of protein concentration, chromatogram, comparisons of different sets of samples using ANOVA, etc.

There are quizzes for learning analysis of gel photo and data presentation methods. You may try the challenges in this module.

5a. Introduction

Any data collected by a research (known as raw data) are always in an unorganized form and need to be organized and presented in a meaningful and readily comprehensible form in order to facilitate further data analysis and/or biochemical investigations. We can present the collected data in following ways:

1. Tabulation
2. Diagrammatic Presentation
3. Graphical Presentation

5a(i). Tabulation

Tabulation is the process of summarizing data in the form of a table. A table is a systematic arrangement of classified data in columns and rows. It facilitates comparison and often reveals certain patterns in data, which may not be obvious. Besides, tabulation facilitates computation of various statistical measures like average, standard deviation, correlation etc. Moreover, it is easier to present the information in the form of graphs and diagrams through tabulated data.

An ideal table should consist of the following main parts:

1. Title of the table;
2. Captions or column headings, with appropriate units;
3. Proper arrangement of data in the table body in accordance to the description of captions.

Tips:

In PowerPoint, you can click on the insert tab and then on table to insert a table. You can state the number of rows and columns and can manage text alignment, colors etc by right clicking on the table. If you have a prepared excel chart, please click on insert object tab and then you can insert the excel chart itself into the slide.

Select Sample:

i. Protein content in a sample [e.g. Milk Protein]

i. Protein content in a sample [e.g. Milk Protein]

Concentration(mg/ml)	1st Trial	2nd Trail	Average
0	0	0	0
0.2	0.127	0.141	0.134
0.4	0.302	0.261	0.282
0.6	0.46	0.465	0.463
0.8	0.597	0.572	0.585
1	0.688	0.672	0.68

ii. Enzymatic assay (e.g. ester substrate)

ii. Enzymatic assay (e.g. ester substrate)

Enzymatic and non-enzymatic hydrolysis of ester substrate

Time (min)	Absorbance at 400nm					
	A. Non-enzymatic hydrolysis			B. Enzymatic hydrolysis		
	sample 1	sample 2	Average	sample 3	sample 4	Average
0.0	0.034	0.029	0.032	0.187	0.195	0.191
0.5	0.035	0.036	0.036	0.221	0.225	0.223
1.0	0.035	0.037	0.036	0.236	0.240	0.238
1.5	0.035	0.038	0.037	0.248	0.255	0.252
2.0	0.035	0.039	0.037	0.258	0.267	0.263
3.0	0.035	0.039	0.037	0.284	0.294	0.289
4.0	0.035	0.039	0.037	0.313	0.322	0.318
5.0	0.036	0.039	0.038	0.339	0.352	0.346
7.5	0.036	0.040	0.038	0.411	0.428	0.420
10.0	0.037	0.041	0.039	0.486	0.509	0.498
12.5	0.038	0.041	0.040	0.561	0.590	0.576
15.0	0.039	0.042	0.041	0.640	0.650	0.645
17.5	0.041	0.042	0.042	0.721	0.761	0.741
20.0	0.043	0.042	0.043	0.803	0.846	0.825
25.0	0.047	0.045	0.046	0.964	1.012	0.988
30.0	0.054	0.047	0.051	1.086	1.098	1.092
4 days	0.859	0.772	0.816	1.166	0.970	1.068
7 days	1.008	0.934	0.971	1.174	0.975	1.075

iii. Chemical inhibition experiment (e.g. Inhibition of Red blood cells (RBC) hemolysis by green tea) Results:

iii. Chemical inhibition experiment
(e.g. Inhibition of Red blood cells (RBC) hemolysis by green tea)

Results:

	Trail 1	Trail 2	Trail 3	Inhibition % (1)	Inhibition % (2)	Inhibition % (3)	Mean Inhibition %	S.D.
AAPH (control)	0.699	0.778	0.794	0.000	0.000	0.000	0.000	0.000
Chinese Green Tea	0.317	0.315	0.314	0.546	0.595	0.605	0.582	0.031
Japanese Green Tea	0.572	0.568	0.555	0.182	0.270	0.301	0.251	0.062
Organic Jasmine Green Tea	0.488	0.488	0.499	0.302	0.373	0.372	0.349	0.041
Rickshaw Green Tea	0.611	0.578	0.548	0.126	0.257	0.310	0.231	0.095
Luk Yu Jasmine Tea	0.499	0.587	0.49	0.286	0.246	0.383	0.305	0.071

iv. Inhibition of RBC hemolysis by green tea and black tea

iv. Inhibition of RBC hemolysis by green tea and black tea

	Tea Concentration (mg)	Trial 1	Trial 2	Trail 3	Inhibition % (1)	Inhibition % (2)	Inhibition % (3)	Mean Inhibition	S.D.	t
AAPH (control)		0.699	0.778	0.794	0	0	0	0	0	
Green Tea (0.25mg)	0.25	0.416	0.458	0.419	0.405	0.411	0.472	0.429	0.037	
Green Tea (0.5mg)	0.5	0.317	0.315	0.314	0.546	0.595	0.605	0.582	0.031	
Green Tea (1.5mg)	1.5	0.483	0.465	0.466	0.309	0.402	0.413	0.375	0.057	
Black Tea (0.25mg)	0.25	0.494	0.498	0.529	0.293	0.36	0.334	0.329	0.034	0.025920713
Black Tea (0.5mg)	0.5	0.495	0.489	0.468	0.292	0.371	0.411	0.358	0.061	0.010784171
Black Tea (1.5mg)	1.5	0.44	0.44	0.321	0.371	0.434	0.596	0.467	0.116	0.307679479

5a(ii). Diagrammatic Presentation

It is used to present data classified in accordance to categories and geographical aspects.

Selection of a suitable diagram should consider:

a. The nature of data
b. Purpose of the diagrammatic presentation
c. The type of audience for whose understanding the diagrammatic presentation is made.

Tips: In PowerPoint, you can click on the 'insert chart' tab to insert a chart. To change chart type, please right click on the chart and select type of chart from menu.

Select Sample:

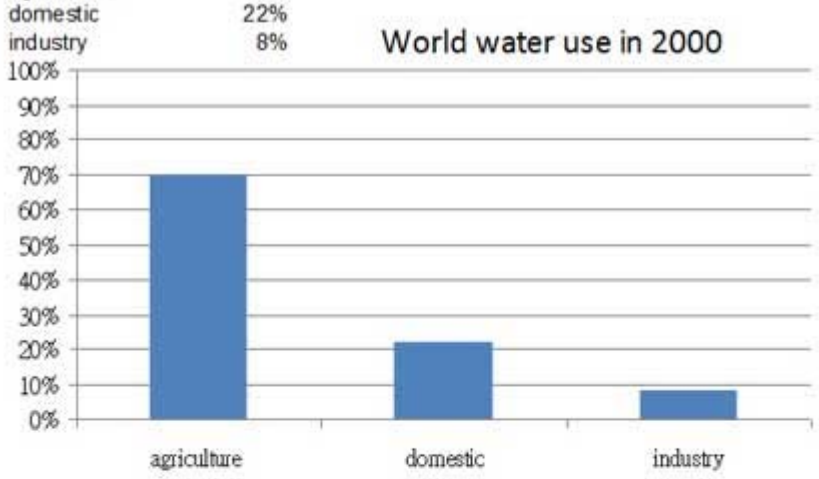
(See next page)

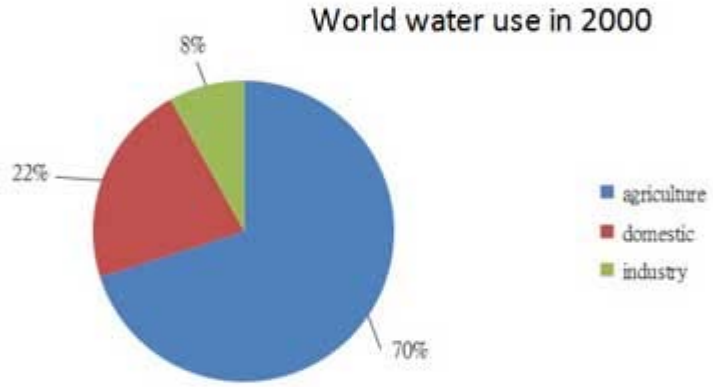
i. Pie Chart/ Bar Chart

i. Pie Chart/ Bar Chart

The world water use in 2000

agriculture	70%
domestic	22%
industry	8%

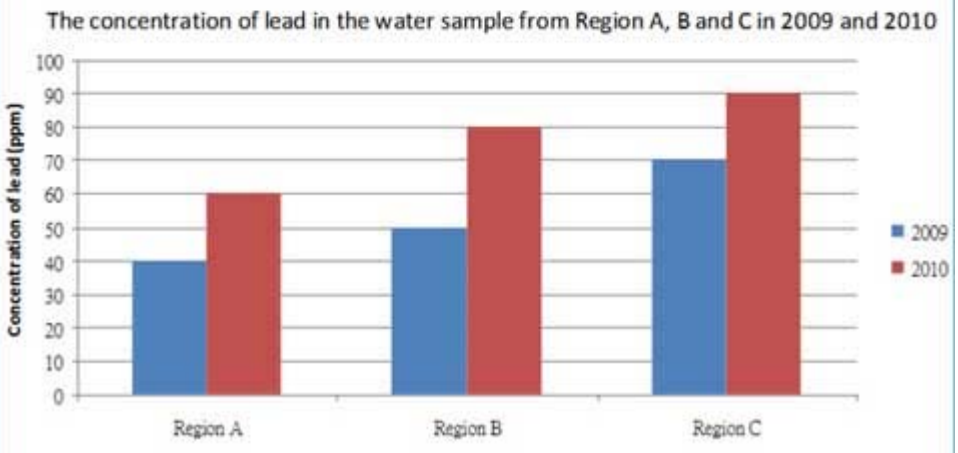


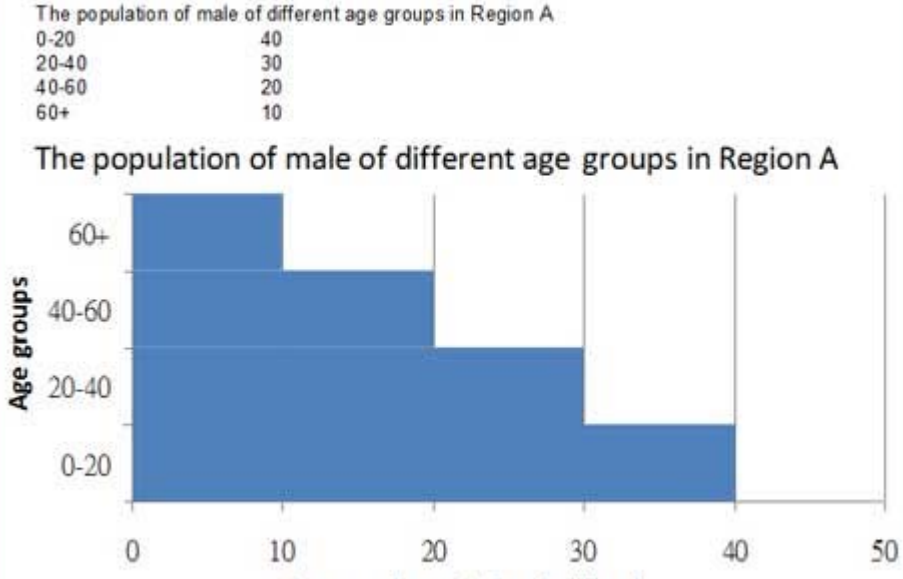


The concentration of lead in the water sample from Region A, B and C in 2009 and 2010

	2009	2010
Region A	40	60
Region B	50	80
Region C	70	90

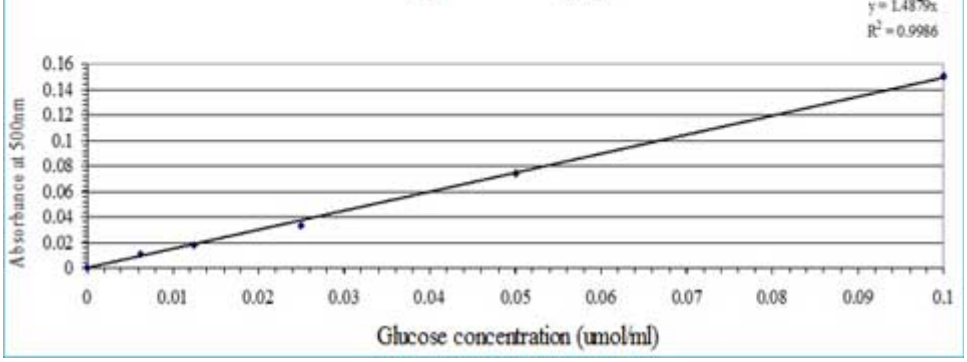
(Multiple Bars)

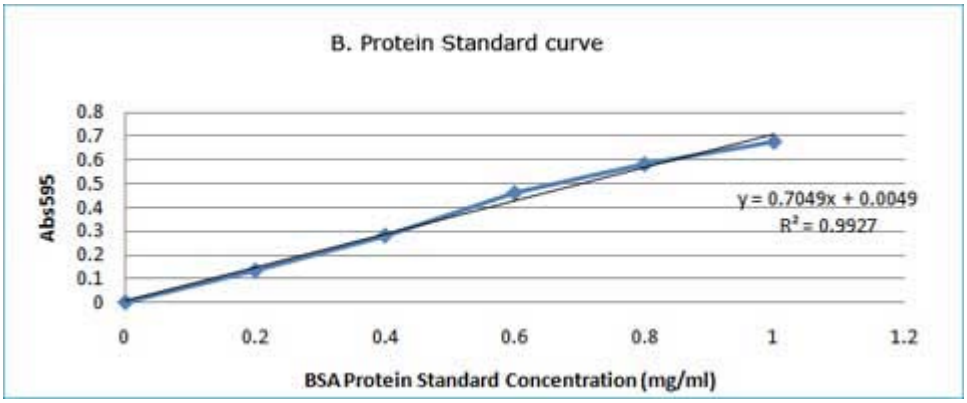


ii. Pyramid Diagram	ii. Pyramid Diagram The population of male of different age groups in Region A 0-20 40 20-40 30 40-60 20 60+ 10 

5a(iii). Graphical Presentation

It is used to observe some functional relationship between the values of two variables. The dependent variable is conventionally shown on the y-axis and the independent variable (e.g. time) is shown on the x-axis.

Select Sample:
i. Line graph It is used to observe some functional relationship between the values of two variables. The dependent variable is conventionally shown on the y-axis and the independent variable (e.g. time) is shown on the x-axis. A. Glucose standard curve showing the relationship of the absorbance of oxidized O-dianisine at 500nm at different glucose concentration by using glucose oxidase method Glucose conc (umol/ml) / Absorbance (500nm) 0 / 0 0.00625 / 0.011 0.0125 / 0.018 0.025 / 0.033 0.05 / 0.074 0.1 / 0.15   



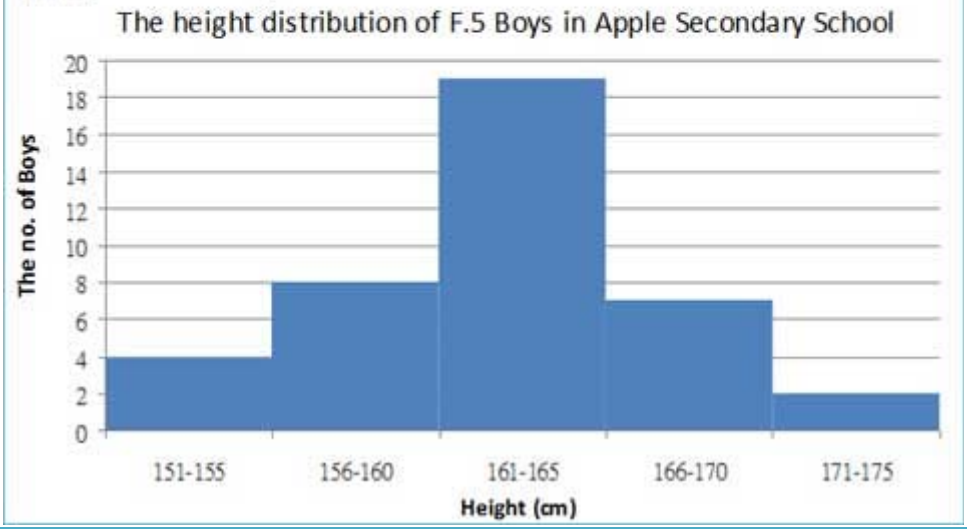
This type of curve is linear curve, and error bars can also be incorporated to show the degree of deviation of data.

ii. Histogram

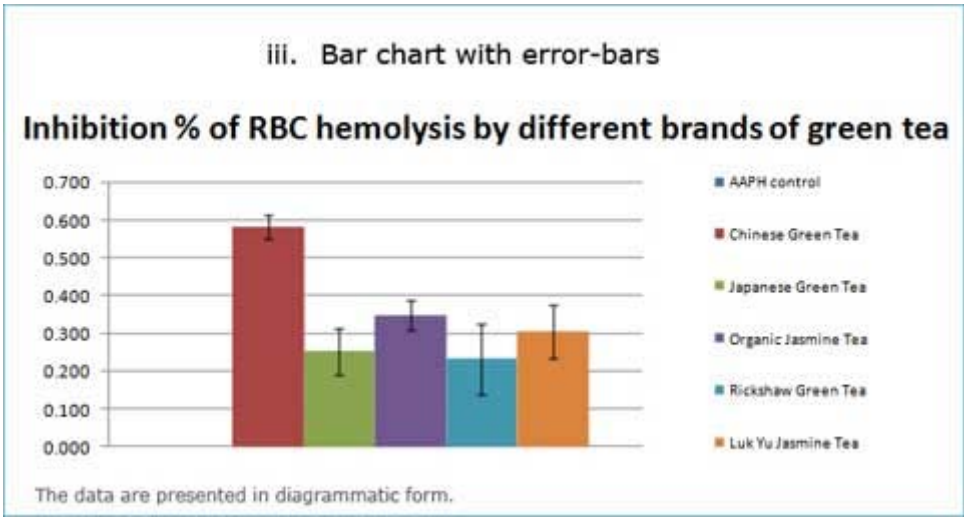
Represented by rectangles adjacent to each other
Example:

The height distribution of F.5 Boys in Apple Secondary School

151-155	4
156-160	8
161-165	19
166-170	7
171-175	2



iii. Bar chart with error-bars



Error bars can be incorporated to represent the deviation of data. Furthermore, ANOVA is calculated as below:

One way ANOVA
(單因子變異數分析)

Summary

組	個數	總和	平均	變異數 (Variance)
列 1	3	1.746145	0.582048	0.00097
列 2	3	0.752619	0.250873	0.003831
列 3	3	1.046147	0.348716	0.001647
列 4	3	0.692787	0.230929	0.00897
列 5	3	0.914496	0.304832	0.00498

ANOVA

Source	SS(平方和)	d.f.(自由度)	MS(平均平方和)	F(檢定統計量)	p-value	critical value*
Treatments	0.239039	4	0.05976	14.64786	0.000347	3.478049691
Error	0.040798	10	0.00408			
Total	0.279837	14				

*Critical value can be checked at
http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm#ONE-05-1-10

(v1 is the degree of freedom (df) of treatments, while v2 is the df of error)

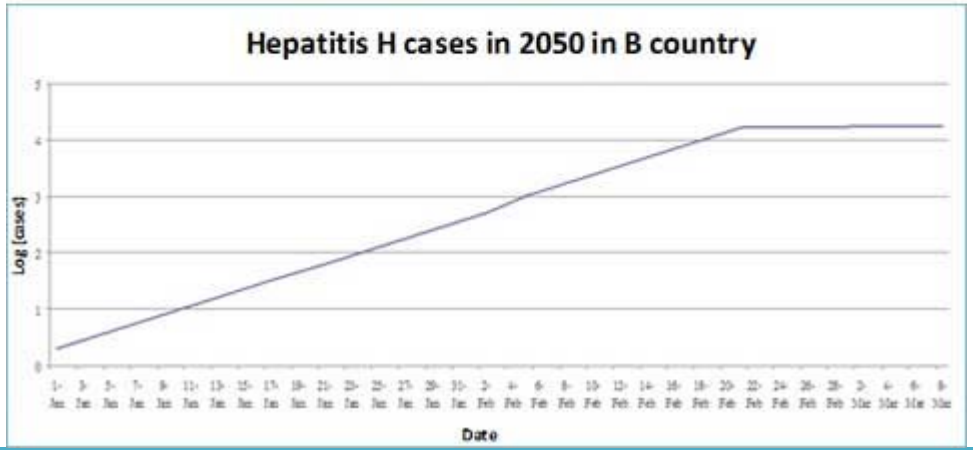
Other statistical analysis will be elaborated at the "Linear Regression" section

iv. Semi-log graph

A semi-log graph or semi-log plot is a way of presenting data that are changing with an exponential relationship. One axis is plotted on a logarithmic scale. This kind of plot is useful when values of one of the variables cover a large range while other has only a restricted range. The merit of this plot can show the features of data that could not be observed by linear plot easily.

(See next page)

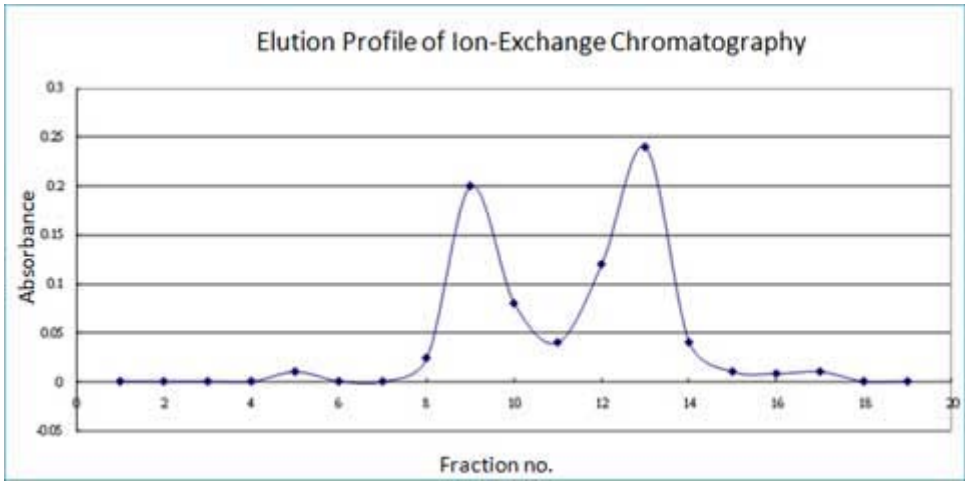
Hepatitis H cases in 2050 in B country								
Date	1-Jan	5-Jan	9-Jan	13-Jan	17-Jan	21-Jan	25-Jan	29-Jan
Log(cases)	0.30103	0.60206	0.90309	1.20412	1.50515	1.80618	2.10721	2.40824
Cases	2	4	8	16	32	64	128	256
Date	2-Feb	5-Feb	9-Feb	13-Feb	17-Feb	21-Feb	25-Feb	
Log(cases)	2.70927	3.0103	3.31133	3.61236	3.91339	4.21442	4.21748	
Cases	512	1024	2048	4096	8192	16384	16500	
Date	1-Mar	5-Mar	8-Mar					
Log(cases)	4.23045	4.23805	4.24304					
Cases	17000	17300	17500					



v. Chromatogram

Chromatogram is the visual output of the chromatograph. In the case of an optimal separation (e.g. Gel filtration), different peaks or patterns on the chromatogram correspond to different components of the separated mixture.

Values on x-axis represents the retention time or fraction number, while values on y-axis is a signal (e.g. absorbance values obtained by a spectrophotometer,) corresponding to the response generated by the analytes exiting the system. In the case of an optimal system the signal amplitude is proportional to the concentration of the specific analyte separated.



This type of curve is chromatogram. The peak(s) indicates certain compound(s) of interest during separation. It also indicates the quality of separation.

5b. Some common calculations in biochemical experiments

i. Cell counting in cell culture

Calculation for cells number (by Hemocytometer)

Using a hemocytometer to count cells is still a widely used method, The hemocytometer consists of two chambers, each of which is divided into nine squares with the dimension of 1x1 mm. A cover glass is supported 0.1 mm over these squares so that the total volume over each square is 1.0 mm x 0.1 mm or 0.1 mm3, or 10-4 cm3. So in total each square has volume of 0.0001 ml.

In general, the number of cells is expressed as cells/ml.

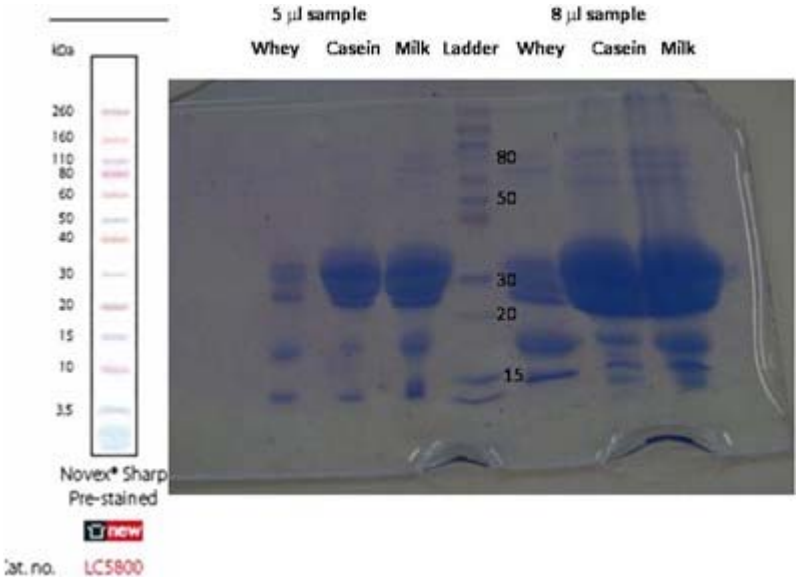
Cell number is calculated by:

Total number of cells counted in 9 squares x dilution factor x (1 X 10^4)

Total number of squares counted (9 squares)

ii. Determine the size and quantities of DNA samples in agarose gel electrophoresis

Determine the sizes and quantities of DNA samples by comparison with DNA marker. This technique can also be applied in protein size determination, such as in SDS PAGE experiments.



Reference information:
Protein M.W.

Casein(80%): ~20-25kDa
as1: 23.5kDa (199AA)
as2 : 25.2kDa(207AA)
b : 24kDa(209AA)
k : 19kDa(169AA)

Whey(20%)
b-lactoglobulin: 18kDa
a-lactalbumin: 14kDa
Serum albumin: ~66kDa
Lactoferrin: ~80kDa

Immunoglobuline-G: ~150kDa

Quiz: Could you suggest the factors that can affect the quality of bands?
Suggested answer: Samples quantities, dilution factors, running condition

iii. Protein determination by absorbance method (280 nm)

1. Pipette 0.9 ml distilled water and 0.1 ml BSA/ Lysozyme into a microfuge tube.

2. Take absorbance using the Spectrophotometer at 280 nm.

Calculation:

Protein concentration can be calculated by the following equation:

$A = \varepsilon\lambda \times C \times L$
A = Absorbance
$\varepsilon\lambda$ = Extinction coefficient (cm^{-1} $(mg/ml)^{-1}$)
C = Protein concentration (mg/ml)
L = Path length. (Path length for spectrometers is 1 cm)

Extinction coefficients of two proteins:
• BSA : 0.667

• Lysozyme : 2.65

5c. Basic quantitative analysis of biochemical data

Biochemistry is an experimental science that usually involves analytical and quantitative techniques. Methods for the analysis of experimental data are essential tools for manipulating the results of many biochemical studies. In this section, we describe introductory concepts of some common analytical techniques in biochemical field, including linear regression and an application in the field of enzyme kinetics.

5c(i). Fitting data by the method of least squares → Linear regression

Linear Regression is one of the classical procedures in general regression analysis of several data points. For data pairs in the form of x, y, where y is a function of x, the linear equation: y= a + bx that minimizes the sum of errors squared (SSD)* is shown in the following equation **

This method allows the equation of the best straight line fitting the experimental data to be calculated directly:

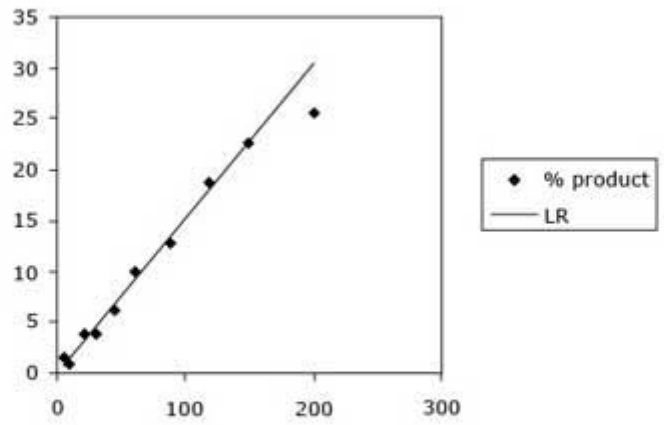
Equation: $y = a + bx$

**Slope $b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

y Intercept $a = \frac{(\Sigma x^2)(\Sigma y) - (\Sigma x)(\Sigma xy)}{n\Sigma x^2 - (\Sigma x)^2}$

*SSD represents the deviation between theoretical and experimental data

$$SSD = \sum_i (S_{i,exp} - S_{i,theo})^2$$



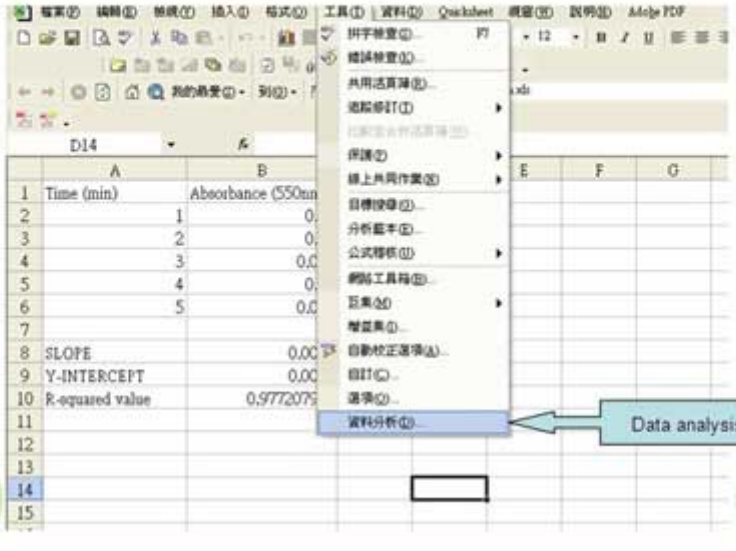
At least five points are required for the linear region.
Linear Regression implies that all data points have the same standard deviation and accuracy.

Tips:
To perform linear regression in Excel, the slope, y-intercept and R-squared values (coefficient of determination) can be obtained as shown in following example:

ColumnA	Column B
Time (min)	Absorbance (550nm)
1	0.01
2	0.02
3	0.035
4	0.04
5	0.049
SLOPE	0.0098
Y-INTERCEPT	0.0014
R-squared value	0.977207977

=SLOPE(B2:B6, A2:A6)
=INTERCEPT(B2:B6, A2:A6)
=RSQ(B2:B6, A2:A6)

Besides, we can also use Excel's regression analysis package to perform linear regression:
If "Data analysis" does not exist, please check the "Analysis Toolpak" in the Tools/Add-Ins (增益集) menu.



Choose Regression(迴歸):

For 95% confidence interval, the results show that :

-0.0077<Y-intercept<0.011
0.007<Slope<0.0126

*Remarks:
SS: Sum of square
MS: Mean of SS

$$\text{MS of Regression} = \frac{\text{SS of Regression}}{\text{k-1 (df)}}$$ where k is the number of data group (k=2 in this example)

$$\text{MS of Residual} = \frac{\text{SS of Residual}}{\text{n-k (df)}}$$ where n is the number of data (n=5 in this example)

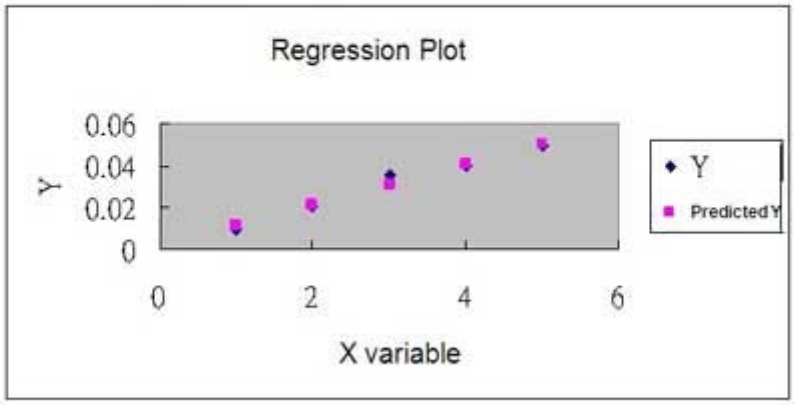
F value: MS of Regression
MS of Residual

The larger the F value, the better in data fitting.

Significance of F - This indicates the probability that the Regression output could have been obtained by chance. A small Significance of F confirms the validity of the Regression output. For example, if Significance of F = 0.050, there is only a 5% chance that the Regression output was only a chance occurrence.

P-value of each coefficient and the Y-intercept - The P-Values of each of these provide the likelihood that they are real results and did not occur by chance. The lower the P-Value, the higher the likelihood that coefficient or Y-Intercept is valid. For example, a P-Value of 0.023 for a regression coefficient indicates that there is only a 2.3% chance that the result occurred merely as a result of chance.

Plot of "Observed Y" and "Predicted Y"



5c(ii). Error Estimation

How to estimate the accuracy of experimental data is among the most difficult tasks in every day research activities.

Typical considerations are:

i. The reliability of data;

ii. The accuracy of my results;

* In data analysis, we could consider:

- Any data points can be removed from the analysis. In general, we could delete one of these points, fit again, observe the differences in the results, and then, put point back and delete another one.
- Provides appropriate qualitative estimate of the reliability of results. Please check if individual data points have a sound effect on the result of the fit.

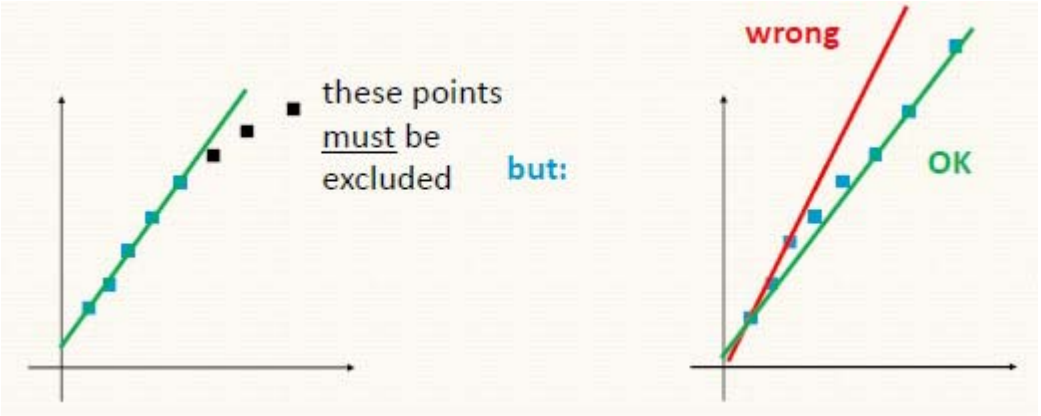
Introduction to error analysis

Methods for error analysis:

- Repeat identical experiment, and collect more than one data set;
- Perform checking the accuracy of individual experiment; (like described in *)
- Apply statistical methods to analysis the results (mean, standard deviation etc.)
- Errors can also be caused by apparatus or procedures. So apparatus checking (e.g. equipment calibration, solution checking) and procedure verification are essential to minimise the risk of these errors.

Treatment of outliers

- Identification of outlieres:
  - Data points that differ significantly from all other data
  - Data points that probably caused by the experimental errors.
- Outliers can be act as stimulus to repeat experiments and improve experimental practice. Besides, least-squares estimates are highly sensitive to the effect of outliers.
  - In term of statistics, outlier is defined by a probability of occurrence < significance level (0.01 for today)



Only analysis the initial part of the curve (initial curve)

Please use Linear regression line for Michaelis-Menten Kinetics analysis

5c(iii). Applications

i. Linear Regression

As mentioned in section 1, linear regression is one of the classical procedures in general regression analysis. This method allows the equation of the best straight line fitting (least-square fit) a set of experimental data to be calculated directly (please refer to section 1).

ii. Michaelis-Menten kinetics

In a simple single substrate – single product enzymatic reaction, in general, it is considered that the initial binding is reversible, and that the back reaction from product can be neglected. At the start of the reaction, the product concentration (P) is assumed zero.

$$S + E \underset{k_2}{\overset{k1}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} E + P$$

Where
ES = Enzyme-substrate complex
k_1= association rate constant for formation of the ES complex
k_2 = dissociation rate constant for breakdown of the ES complex
k_{cat} = enzyme catalytic constant

When the concentration of substrate (S) is much higher than that of enzyme (E) , after a short pre-steady state, the rate of product turnover is constant. In this steady state status, the concentration of enzyme-substrate complex is constant, and the rate of reaction can be illustrated as:

$$\frac{dc_p}{dt} = \frac{k_{cat} \cdot Cs \cdot C_{Etotal}}{Cs + K_M}$$

Where
C_{Etotal} = total concentration of enzyme
C_s = concentration of substrate
K_M = Michaelis constant, it is the substrate concentration that gives half of the maximum rate, i.e. $V_{(Cs)} = V_{max}/2$
V_{max} = Maximum reaction rate
$V_{(Cs)}$ = steady state (or initial) velocity

C_{Etotal} . k_{cat} represents the maximum turnover rate V_{max}. At steady state, dc_P/dt represents the rate of reaction $V_{(Cs)}$, and the equation can be illustrated as Michaelis-Menton equation:

$$V_{(Cs)} = k_{cat} . C_{E.total} \frac{Cs}{Cs + K_M} = V_{max} \frac{Cs}{Cs + K_M}$$

The function is a rectangular hyperbola.

The above equation can be presented in Lineweaver-Burk Plot as shown below ($1/V_{(Cs)}$ is plotted against 1/Cs)

$$\frac{1}{V_{(Cs)}} = \frac{1}{V_{max}} + \frac{K_M}{k_{cat}} \frac{1}{Cs}$$

The values of V_{max} and K_M can be determined from the intercepts on the abscissa $1/V_{max}$ and ordinate $(-1/K_M)$ respectively.

As Michaelis-Menten analysis is directly coupled to the linear regression, and the fit is performed with the original data, therefore the errors are relatively lower than that of least square method which may involve operator subjectivity.

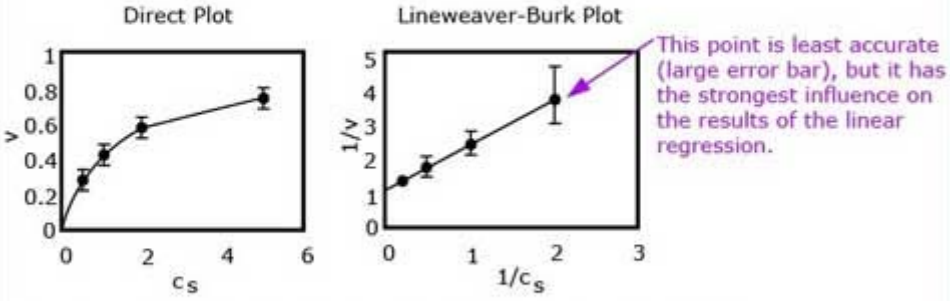


Figure: Error propagation in the direct analysis and Lineweaver-Burk analysis of Michaelis-Menten kinetics.

5d. Common statistical methods

Univariate descriptive		
Statistical Analyses	**Description**	**Example**
Mode	The most commonly occurring value	5 people with ages 21, 23, 25, 27, 27 – mode = 27
Median	Described as the numerical value separating the higher half of a sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values.	8 people with ages 21, 22. 23, 25, 27, 30, 36, 40 Median = 26
Mean	The mathematical average The formula is ΣX/N	3 people with ages 30, 36 and 40. The mean age is 35.3
Variance	The average of the squared differences from the Mean. It is a measure of its statistical dispersion, indicating how far from the expected value its values typically are. It is the square root of its variance. The formula is: $\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$ Where *x* is a random variable, μ is the expected value (mean)	5 students with weight(kg): 50, 62, 45, 55, 59. The mean weight is 54.2kg Therefore the variance is: $\sigma^2 = [(-4.2)^2 + 7.82 + (-9.2)^2 + 0.82 + 4.82] /5$ = 37.36
Standard Deviation	It is a measure of the dispersion of a set of data from its mean. The more spread apart the data, the higher the deviation. It is the square root of its variance. The formula is $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$ Where x is a random variable, μ is the expected value (mean)	5 students with weight(kg): 50, 62, 45, 55, 59. The standard deviation =σ= 6.11
Bi- and Multivariate Inferential Statistical Tests		
Statistical Analyses	**Description**	**Example**
Chi Square	It is a quantitative measure used to determine whether a relationship exists between two categorical variables. It measures the difference between the expected and observed frequencies and is thus a quantitative measure of this relationship. The formula: $\sum_{i}\frac{(O_i - E_i)^2}{E_i}$ where O_i is the observed frequency in a category and E_i is the	Whether a relationship exists between gender and voting behavior.

	expected frequency in the corresponding category. The frequency associated with these rates when no relationship exists is named expected frequencies. Then the greater the relationship, the greater the value of chi-square. Hypothesis testing will use Chi square result to determine whether the relationship exists between the two variables.	
t-Test	t-Test or 'Student's' t Test looks at difference between 2 groups on some variable of interest. It is one of the most commonly used techniques for testing a hypothesis on the basis of a difference between sample means. In general, the t-Test determines a probability that two populations are the same with respect to the variable of interest. It assumes the data sets are in normal distribution and the underlying variances are equal. The Null hypothesis is that there is no significant difference between the sample means. To calculate the t-value, the steps involves: (assume equal sample size and variances): $t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$ where $S_{X_1X_2} = \sqrt{\frac{1}{2}(S_{X_1}^2 + S_{X_2}^2)}$ X_1 and X_2 are the means of the two populations, while S_{X1} and S_{X2} are the variances of the two populations. n is the number of participants of each population. For significance testing, the degree of freedom (df) of this test is 2n-2. For a given df, if the t-value is larger than the value in the table, the null hypothesis (no significance difference) should be rejected.	Whether male and female patients differ in the period of recovery after medical treatment in a given age group.
ANOVA	Analysis of variance (ANOVA) tests the significance of group differences between two or more groups(means). In an experiment, ANOVA is adopted to test the hypothesis that the variation is no greater than that due to normal variation of individuals' characteristics and error in their measurement. The tests in an ANOVA are based on the F-ratio, i.e. The variation caused by an experimental treatment or effect The variation caused by experimental error Or Variance between groups Variance within groups In mathematical terms, F = MST / MSE, where MST = SST / DFT MSE = SSE / DFE (MST and MSE are Mean squares of treatments and Mean squares of errors respectively; SST and SSE are Sum of squares of Treatments and Sum of Squares of Errors respectively; DFT and DFE are Degree of freedom for Treatment and Degree of freedom for Errors respectively.) The null hypothesis is this ratio equals to 1.0, or the treatment effect is	Does the occurrence of diabetic cases differ for poor, developing, and developed countries?

	the same as the experimental error. The hypothesis is rejected if the F-ratio or the test statistics is larger than the critical value in F distribution table (ratio of two *Chi* square distribution) under certain level of confidence and degree of freedom (DFT= k-1 where k is the number of groups, and DFE= N-k where N is the total number of measurements). Typically, however, the one-way ANOVA (one independent variable) is used to test for differences among at least three groups, since the two-group case can be covered by a t-test. ANOVA only determines that there is a difference between groups, but doesn't tell which is different.	
Correlation	Correlation is a statistical measurement of the relationship between two variables. It is a single number that describes the degree of relationship between two variables. The number is range from +1 to –1. A zero correlation indicates that there is no relationship between the variables. A correlation of –1 indicates a perfect negative correlation or inverse association, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that both variables move in the same direction together. Suppose we have two variables X and Y, with means μX and μY respectively, and standard deviation S_x and S_y respectively, and n measurements in each variable. The correlation(or Pearson correlation) is computed as: $$r = \frac{\sum^{n} i\text{-}1\ (X_i\text{-}\mu_X)(Y_i\text{-}\mu_Y)}{(n\text{-}1)\ S_X \text{ and } S_Y}$$ Correlation does not distinguish between independent and dependent variables. After determined the r value, significance test can be conducted. The mutually exclusive hypotheses can be tested. For Null hypothesis, the r =0 , and Alternative hypothesis, the r <> 0. Under certain significance level (e.g. alpha = 0.05), and degree of freedom (n-2), if the calculated r value is greater than the critical value [shown in the Table of critical value of r], then we can conclude that the correlation is “statistically significant”. The null hypothesis can be rejected and accept the alternative.	Does the degree of Inhibition of Red blood cells hemolysis correlate to the concentration of green tea?
Multiple Regression	In multiple regression, more than one variable is used to predict the criterion. It applies several independent variables and one dependent variable, and identifies the best set of predictor variables. Predicted scores from multiple regression are linear combinations of the predictor variables. The general form of a prediction equation is: $Y'=b_1X_1+b_2X_2+...b_kX_k+A$ Where Y' is the predicted score, $X_1$is the score on the first predictor variable, X_2 is the second and so on. A is the Y-intercept and b_1, b_2… are regression coefficients, that are analogous to the slope in simple	<u>Dependent variable:</u> Occurrence of diabetic cases <u>Independent variables:</u> Gender, Geographical regions, family

regression.
The multiple correlation coefficient(R) is the Pearson correlation between the predicted scores and the observed scores (Y' and Y). As r^2 is the proportion of the sum of squares applied in one-variable regression, similarly, R^2 is the proportion of the sum of squares explained in multiple regression.

R^2 = 1-[Residual SS/ Total SS]
Where Residual SS(or error) =$\sum(Y-Y')^2$
Total SS =$\sum(Y-Y_M)^2$
(SS = Sum of Square ;Y' = value of Y predicted from the regression line; Y_M = mean of Y)

Like ANOVA, F statistics can be calculated as: history.

$$F = \frac{\text{MS of Regression}}{\text{MS of Residual(or error)}}$$

= [Regression SS/(k-1)]/[Residual SS/(n-k)]

Where n is the number of sample and k is the number of independent (predictor) variable including the intercept.
(Regression SS= $\sum(Y'-Y_M)^2$)
With reference to the F distribution table (for df=k-1, n-k), we can then determine whether the null hypothesis can be rejected.

MS Excel can be applied to compute these statistical data, please refer to here.

Reference:
1. A. Pingoud, C. Urbanke, J. Hoggett and A. Jeltsch (2002) Biochemical Methods A concise guide for students and researchers. Wiley-VCH
2. http://teamwork.jacobs-university.de:8080/confluence/display/cs09s520331QAnalysisBCExp/Lectures
3. http://www.statsoft.com/textbook/distribution-tables/
4. http://www.gifted.uconn.edu/siegle/research/Correlation/corrchrt.htm