Chapter 1– Basics and Statistics of Analytical Biochemistry

Biochemistry and Molecular Biology (BMB)

1.1 Biochemical Studies

1.2 Units of Measurements

1.3 Weak Electrolytes

1.4 Buffer Solution

1.6 Quantitative Biochemical Measurements

1.7.1-1.7.2 Principle of Clinical Biochemical Analysis

Others:

- Receiver Operating Characteristic Curve
- Diagnosis Sensitivity and Specificity

Basic principles

- **Molarity** : Number of moles of the substances in 1 dm^3 of solution.

- **One mole**: equal to molecular mass of the substance

- **Molecular mass**:

  Da: *daltons*

  kDa: *Kilodaltons* =1000 Da

  M_r: *no unit*

  Relative molecular mass

  = the molecular mass of a substance relative to 1/12 of the atomic mass of the ^{12}C .

Units for Different Concentrations

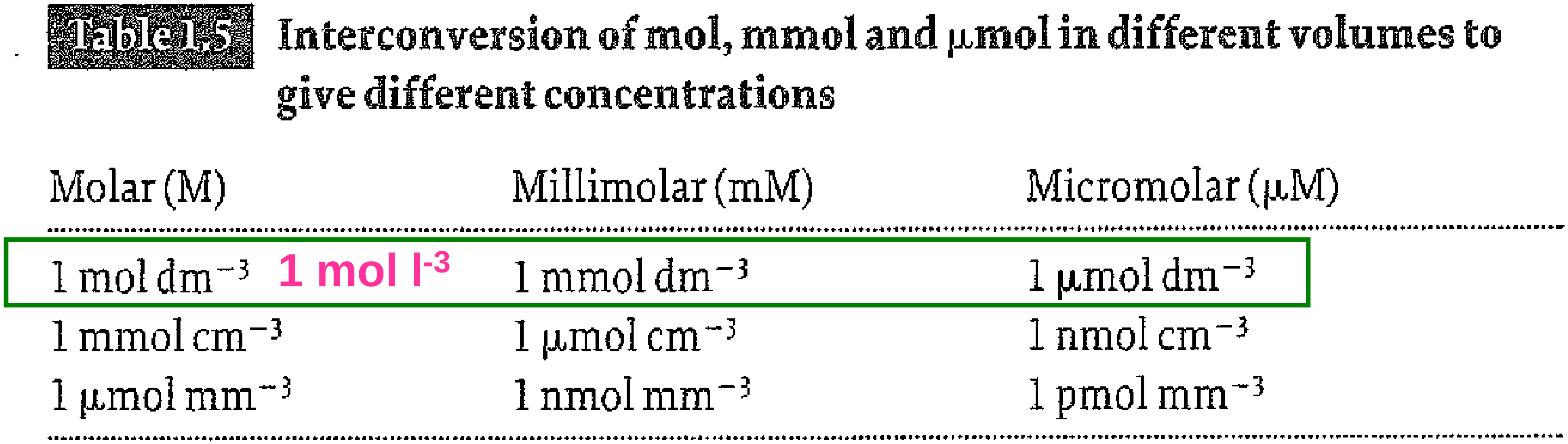
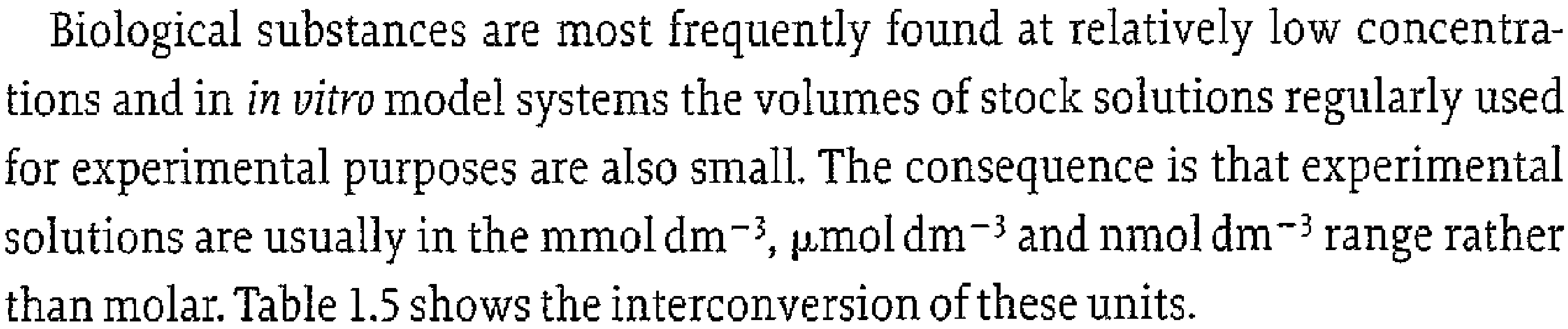


Table 1.5 **Interconversion of mol, mmol and μmol in different volumes to give different concentrations**

Molar (M)	Millimolar (mM)	Micromolar (μM)
1 mol dm^{-3} 1 mol l^{-3}	1 mmol dm^{-3}	1 μmol dm^{-3}
1 mmol cm^{-3}	1 μmol cm^{-3}	1 nmol cm^{-3}
1 μmol mm^{-3}	1 nmol mm^{-3}	1 pmol mm^{-3}



Biological substances are most frequently found at relatively low concentrations and in *in vitro* model systems the volumes of stock solutions regularly used for experimental purposes are also small. The consequence is that experimental solutions are usually in the mmol dm^{-3}, μmol dm^{-3} and nmol dm^{-3} range rather than molar. Table 1.5 shows the interconversion of these units.

Ion Strengths

Reason of deviation:

Presence of electrolytes will result in electrostatic interaction with other ions and solvents

Total ion charge in solution

$$M=1/2 *(c_1z_1^2 + c_1z_1^2 +....+ c_nz_n^2)$$

c_1, c_2, …c_n: concentrations of each ion in *molarity*

z_1, z_2, …z_n: charge on the individual ion

Example 2 CALCULATION OF IONIC STRENGTHS

Question

Calculate the ionic strength of (i) 0. 1 M NaCl, (ii) 0. 1 M NaCl + 0.05 M KNO_3 + 0.01 M Na_2SO_4.

Answer

Ionic strength can be calculated using the equation $\mu = \frac{1}{2}\Sigma cz^2$.

(i) Calculating cz^2 for each ion:

$$Na^+ = 0.1 \times (+1)^2 = 0.1\ M$$

$$Cl^- = 0.1 \times (-1)^2 = 0.1\ M$$

Hence

$$\frac{1}{2}\Sigma cz^2 = 0.2/2 = 0.1\ M$$

(ii)

$$\begin{aligned}
Na^+ &= 0.1 \times (+1)^2 + 0.02 \times (+1)^2 &&= 0.12\ M \\
Cl^- &= 0.1 \times (-1)^2 &&= 0.10\ M \\
K^+ &= 0.05 \times (+1)^2 &&= 0.05\ M \\
NO_3^- &= 0.05 \times (-1)^2 &&= 0.05\ M \\
SO_4^{2-} &= 0.01 \times (-2)^2 &&= 0.04\ M
\end{aligned}$$

Hence

$$\frac{1}{2}\Sigma cz^2 = \frac{1}{2}(0.36) = 0.18\ M$$

Activity and Activity Coefficients

Activity : the effective concentration in solution

A_x = [Concentration] γ_x

γ_x : Activity coefficient

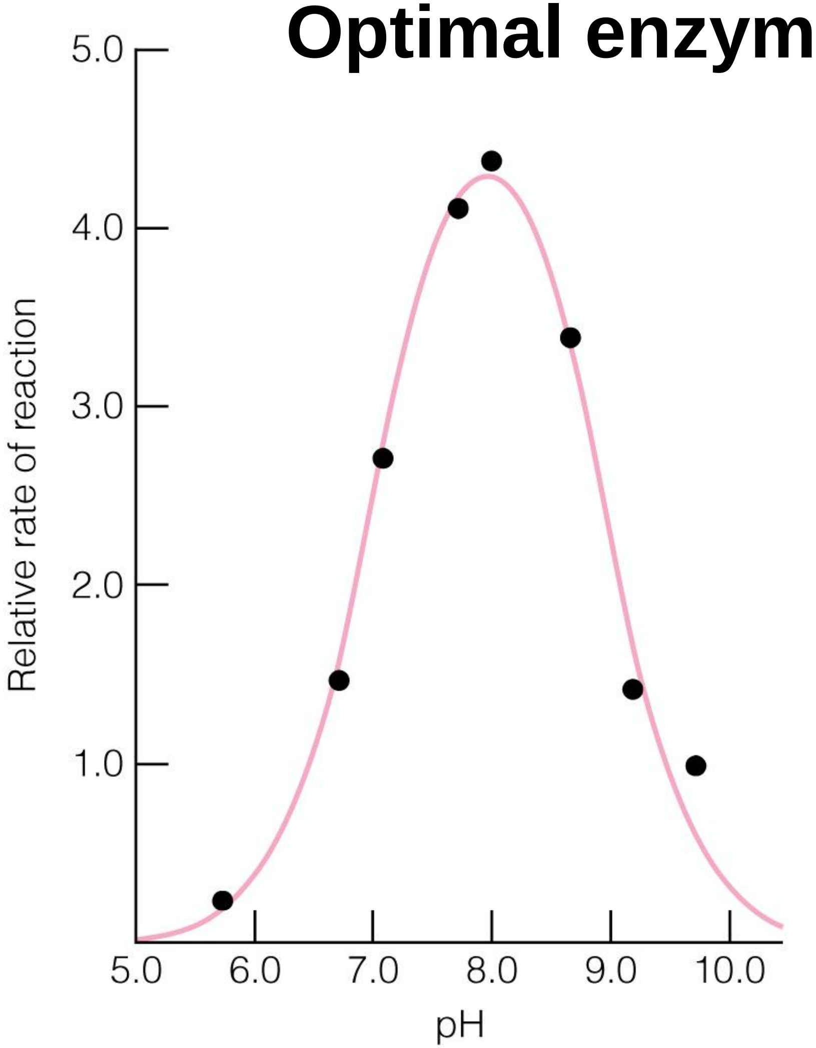
- The coefficient establish the relationship between activity and concentration.
- It will **decrease** when the **ionic strength increases** (include concentration, charge and ion mobility)

e.g. 0.001 M Mg^{2+} 0.872
Fe^{3+} 0.738

Except for very diluted solution, the effective concentrations are usually less than the actual concentration

Preparation of Buffer Solution

Optimal enzyme activity pH 8



α-Chymotrypsin: catalyzed cleavage of the C-N bond

Henderson-Hasselbalch Equation

For a weak acid, which dissociates as follows:

$$HA \rightarrow H^{+} + A^{-}$$

$$\text{equilibrium constant} = K_{eq} = K_{a} = \frac{[H^{+}] \times [A^{-}]}{[HA]}$$

$$\log10Ka = \log10[H+] + \log10[A-\ ] - \log10[HA]$$

$$-\log10[H+] = -\log10Ka + \log10[A-] - \log10[HA]$$

$$pH = pK_{a} + \log_{10}\left(\frac{[A^{-}]}{[HA]}\right)$$

$$pH = pK_{a} + \log_{10}\left(\frac{[\text{conjugate base}]}{[\text{conjugate acid}]}\right) = pK_{a} + \log_{10}\left(\frac{[\text{proton acceptor}]}{[\text{proton donor}]}\right)$$

Why is pKa useful?

$$pH = pK_a + \log_{10}\left(\frac{[A^-]}{[HA]}\right)$$

Perhaps it is useful to look at this in another way: if we consider the situation where the acid is one half dissociated, in other words **where [A-] is equal to [HA],** then, substituting in the Henderson-Hasselbalch Equation

$$pH = pKa + \log10(1)$$

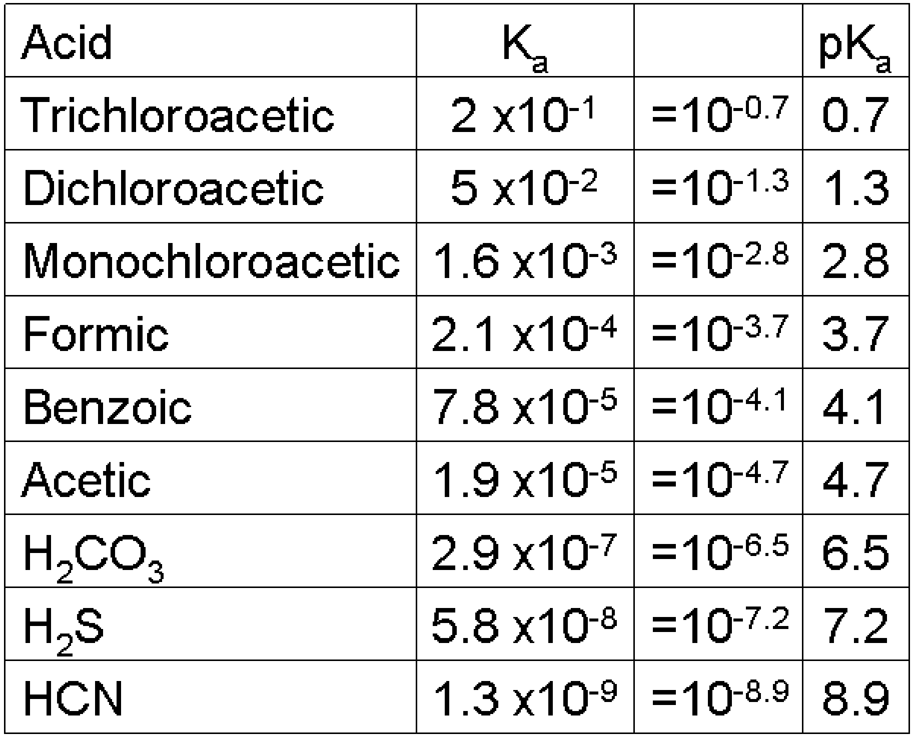
$$pH = pKa + 0$$

$$pH = pKa$$

This means that an acid is half dissociated when the pH of the solution is numerically equal to the pKa of the acid.

$$pH = pK_a + \log_{10}\left(\frac{[A^-]}{[HA]}\right)$$

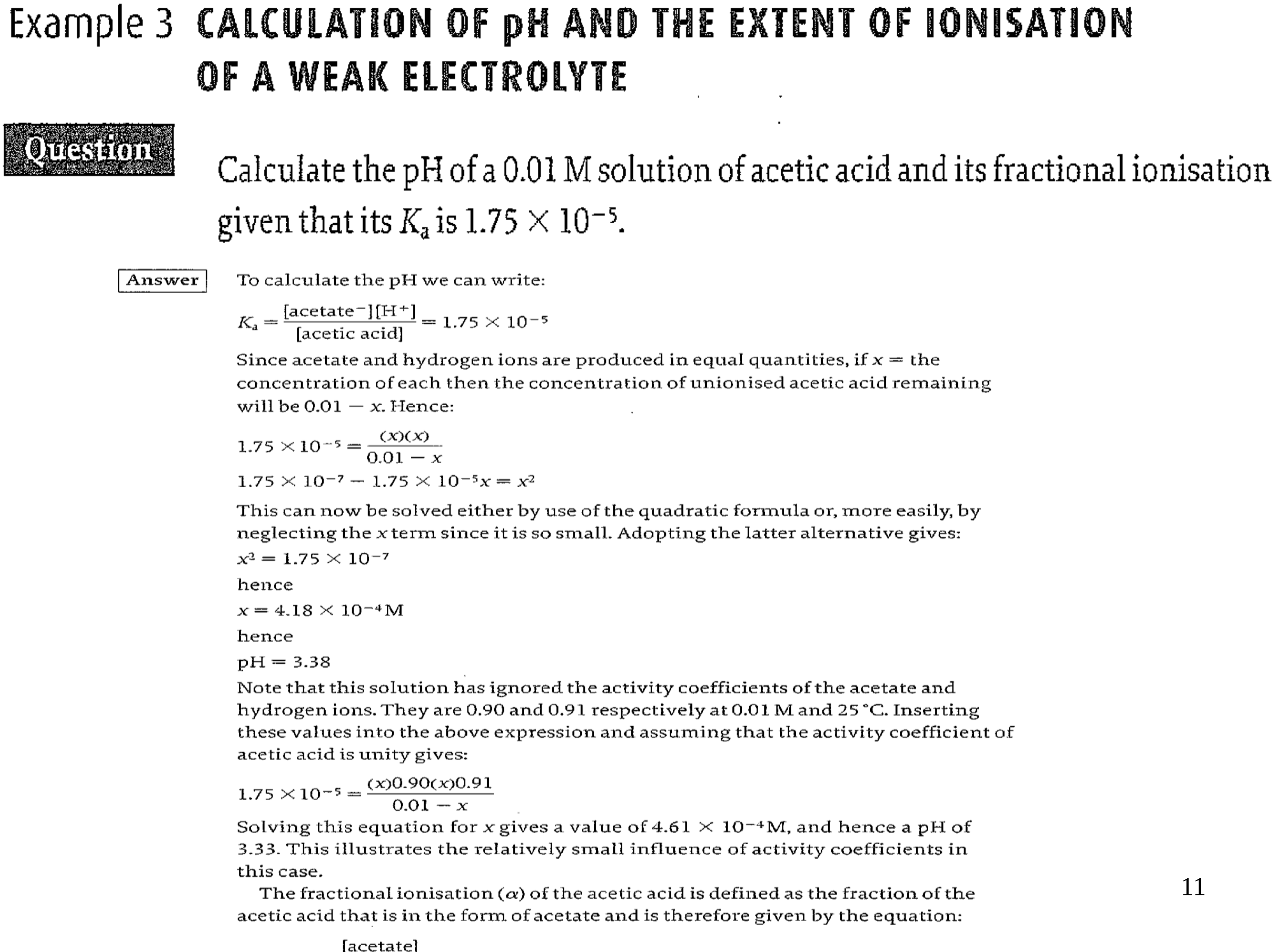
$HA \rightarrow H^+ + A^-$



Acid	K_a		pK_a
Trichloroacetic	2 x10^{-1}	=$10^{-0.7}$	0.7
Dichloroacetic	5 x10^{-2}	=$10^{-1.3}$	1.3
Monochloroacetic	1.6 x10^{-3}	=$10^{-2.8}$	2.8
Formic	2.1 x10^{-4}	=$10^{-3.7}$	3.7
Benzoic	7.8 x10^{-5}	=$10^{-4.1}$	4.1
Acetic	1.9 x10^{-5}	=$10^{-4.7}$	4.7
H_2CO_3	2.9 x10^{-7}	=$10^{-6.5}$	6.5
H_2S	5.8 x10^{-8}	=$10^{-7.2}$	7.2
HCN	1.3 x10^{-9}	=$10^{-8.9}$	8.9

Acids with the lowest pKa values are able to dissociate in solutions of low pH, i.e. even where the hydrogen ion concentration is high.

Acids with higher pKa values dissociate only in solutions of high (more alkaline) pH.



Example 3 CALCULATION OF pH AND THE EXTENT OF IONISATION OF A WEAK ELECTROLYTE

Question

Calculate the pH of a 0.01 M solution of acetic acid and its fractional ionisation given that its K_a is 1.75×10^{-5}.

Answer

To calculate the pH we can write:

$$K_a = \frac{[\text{acetate}^-][\text{H}^+]}{[\text{acetic acid}]} = 1.75 \times 10^{-5}$$

Since acetate and hydrogen ions are produced in equal quantities, if $x =$ the concentration of each then the concentration of unionised acetic acid remaining will be $0.01 - x$. Hence:

$$1.75 \times 10^{-5} = \frac{(x)(x)}{0.01 - x}$$

$$1.75 \times 10^{-7} - 1.75 \times 10^{-5}x = x^2$$

This can now be solved either by use of the quadratic formula or, more easily, by neglecting the x term since it is so small. Adopting the latter alternative gives:

$$x^2 = 1.75 \times 10^{-7}$$

hence

$$x = 4.18 \times 10^{-4}\,\text{M}$$

hence

$$\text{pH} = 3.38$$

Note that this solution has ignored the activity coefficients of the acetate and hydrogen ions. They are 0.90 and 0.91 respectively at 0.01 M and 25 °C. Inserting these values into the above expression and assuming that the activity coefficient of acetic acid is unity gives:

$$1.75 \times 10^{-5} = \frac{(x)0.90(x)0.91}{0.01 - x}$$

Solving this equation for x gives a value of 4.61×10^{-4}M, and hence a pH of 3.33. This illustrates the relatively small influence of activity coefficients in this case.

The fractional ionisation (α) of the acetic acid is defined as the fraction of the acetic acid that is in the form of acetate and is therefore given by the equation:

[acetate]

Quantitative Biochemical Measurements

What to study?	**Model**
How to study	**Method**
Is the results correct?	**Performance**
How to interpret results?	**Report**

Quantitative Biochemical Measurements

- Analytical Considerations:

(I) Test Model :

in vivo v.s. ***in vitro***

Material: urine, serum/plasma/blood

Matrix v.s Analyte

Sampling v.s population

in vivo v.s. ***in vitro***

In vivo: In a living cell or organism

In vitro: Biological or chemical work done in the test tube

(in glass)

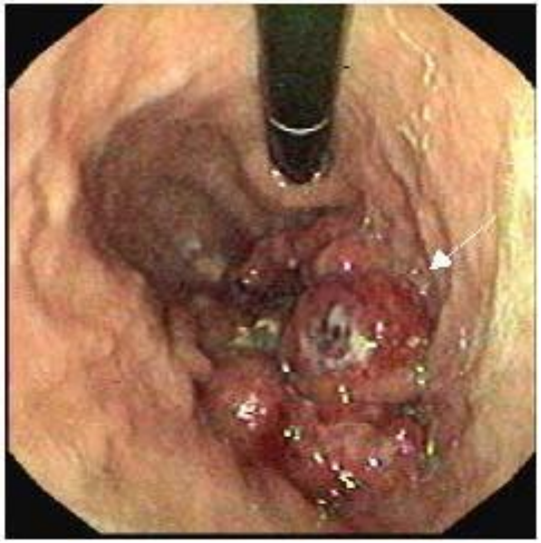
Sampling v.s Population

Population: Representative portion of analyte

Heterogeneous v.s Homogeneous

Extraction Methods:

- Liquid extraction
- Solid-phase extraction
- Laser microdisection (cancer cell)
- ……….etc



Quantitative Biochemical Measurements

(II) Selection of Analytical Methods

- Qualitative v.s Quantitative analysis
- Chemical and physical properties of analyte
- Precision, accuracy and detection limit
- Interference from matrix
- Cost and value
- Possible hazard and risk

NOTE

Precision v.s. Accuracy for Quantitative or Numerical data

Accuracy— a measure of rightness.

Accuracy can be defined how closely a measured value agrees with the correct value.

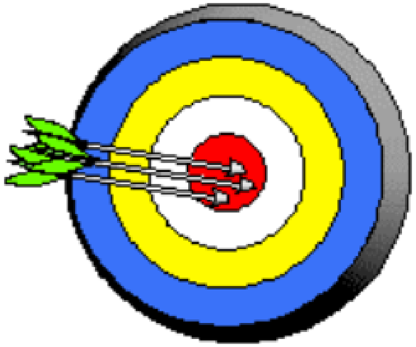
Accuracy is determined by comparing a number to a known or accepted value.

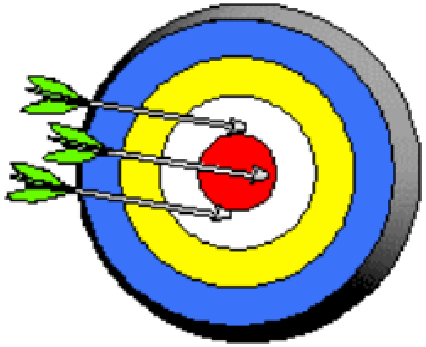
Precision — a measure of exactness.

Precision can be defined how closely individual measurements agree with each other.
It is sometimes defined as reproducibility

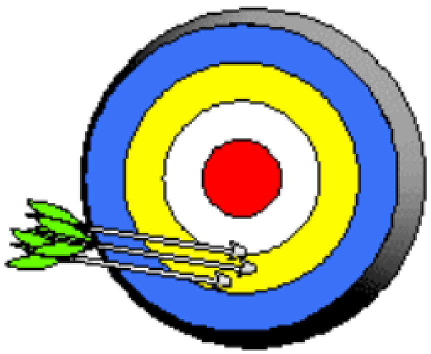
Accuracy	Precision
√	√

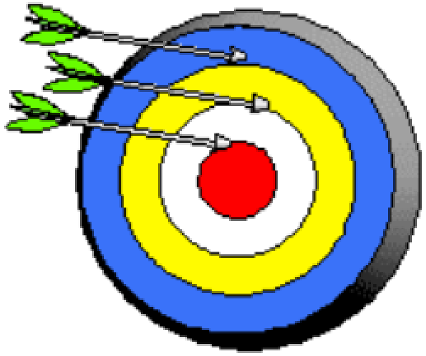
Accuracy	Precision
√	X





The average is close to the center but the individual values are not similar





Accuracy	Precision
X	√

Accuracy	Precision
X	X

Physical Basis of Analytical Methods

Physical properties that can be measured with some degree of precision	Examples of properties used in the		
	Protein	Lead	Oxygen
Extensive			
Mass	+	+	
Volume			+
Mechanical			
Specific gravity	+		
Viscosity	+		
Surface tension	+		
Spectral			
Absorption	+	+	
Emission			
Fluorescence			
Turbidity	+		
Rotation			
Electrical			
Conductivity			
Cuurent/voltage			+
Half-cell potential			+
Nuclear			
Radioactivity	+		

Major manipulative steps in a generalized method of analysis

Purification of the test substance
↓
Development of a physical characteristic by the formation of a derivative
↓
Detection of an inherent or induced physical characteristic
↓
Signal amplification
↓
Signal measurement
↓
Computation
↓
Presentation of result

Quantitative Biochemical Measurements

(III) Experimental Errors

- Systematic error
- Random error

⇩

Standard Operation Procedures (SOP)

Systematic Error

- Constant or proportional (**Bias)**
- Also called

Overestimation /underestimation

(1) **Analyst error**: pipette, calibration, solution preparation, method design

(2) **Instrumental error**: contamination of instrument, power fluctuation, variation in T, pH, electronic noise

(3) **Method error**: side reaction, incomplete reaction

Identification of Systematic Errors

- Blank sample
- Standard reference sample
- Alternative methods
- External quality assessment sample

Random Error

- Variable, either positive or negative
- also called

  Indeterminate error

(1) **Instrumental error**: random electric noise

Standard Operating Procedures (SOP)

Detailed, written instructions to achieve uniformity of the performance of a specific process;

Include:

- Quantity/quality of reagent
- Preparation of standard solution
- Calibration of instrument
- Methodology of actual analytical procedures

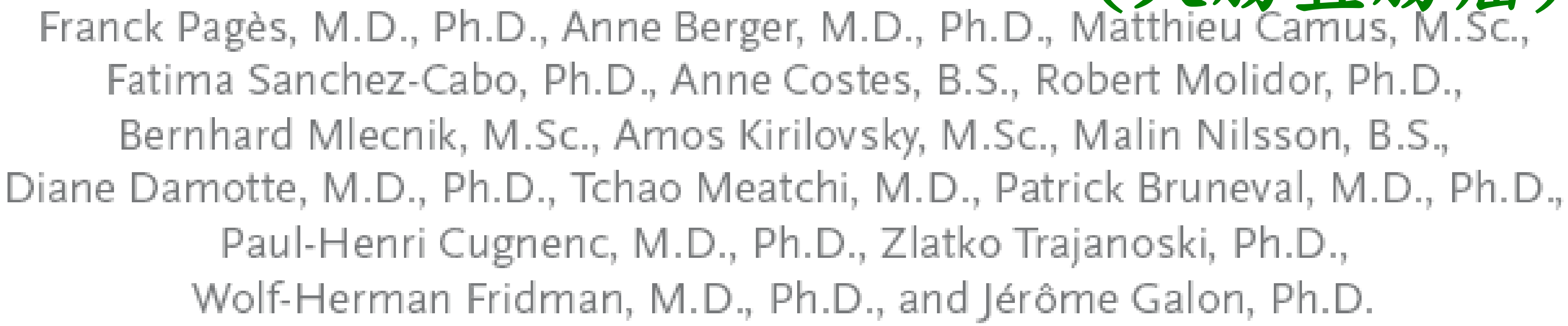
Assessment of Performance of Analytical Method

The NEW ENGLAND JOURNAL of MEDICINE

Question:

1. What is the correlation of the **memory of immune cell** and cancer metastasis?
2. Will it affect the survival rate?

（大腸直腸癌）



Franck Pagès, M.D., Ph.D., Anne Berger, M.D., Ph.D., Matthieu Camus, M.Sc.,
Fatima Sanchez-Cabo, Ph.D., Anne Costes, B.S., Robert Molidor, Ph.D.,
Bernhard Mlecnik, M.Sc., Amos Kirilovsky, M.Sc., Malin Nilsson, B.S.,
Diane Damotte, M.D., Ph.D., Tchao Meatchi, M.D., Patrick Bruneval, M.D., Ph.D.,
Paul-Henri Cugnenc, M.D., Ph.D., Zlatko Trajanoski, Ph.D.,
Wolf-Herman Fridman, M.D., Ph.D., and Jérôme Galon, Ph.D.

NEJM, 353, 2654-2666, 2005

Background

The role of tumor-infiltrating (浸潤) immune cells in the early metastatic invasion (轉移性侵犯) of colorectal cancer（直腸癌）is unknown.

Methods

We studied pathological signs of early metastatic invasion (venous emboli 靜脈栓塞 and lymphatic 淋巴 and perineural invasion(神經旁間隙) in 959 specimens of resected colorectal cancer. The local immune response within the tumor was studied by flow cytometry (39 tumors), low density-array real-time polymerase-chain-reaction assay (75 tumors), and tissue microarrays (415 tumors).

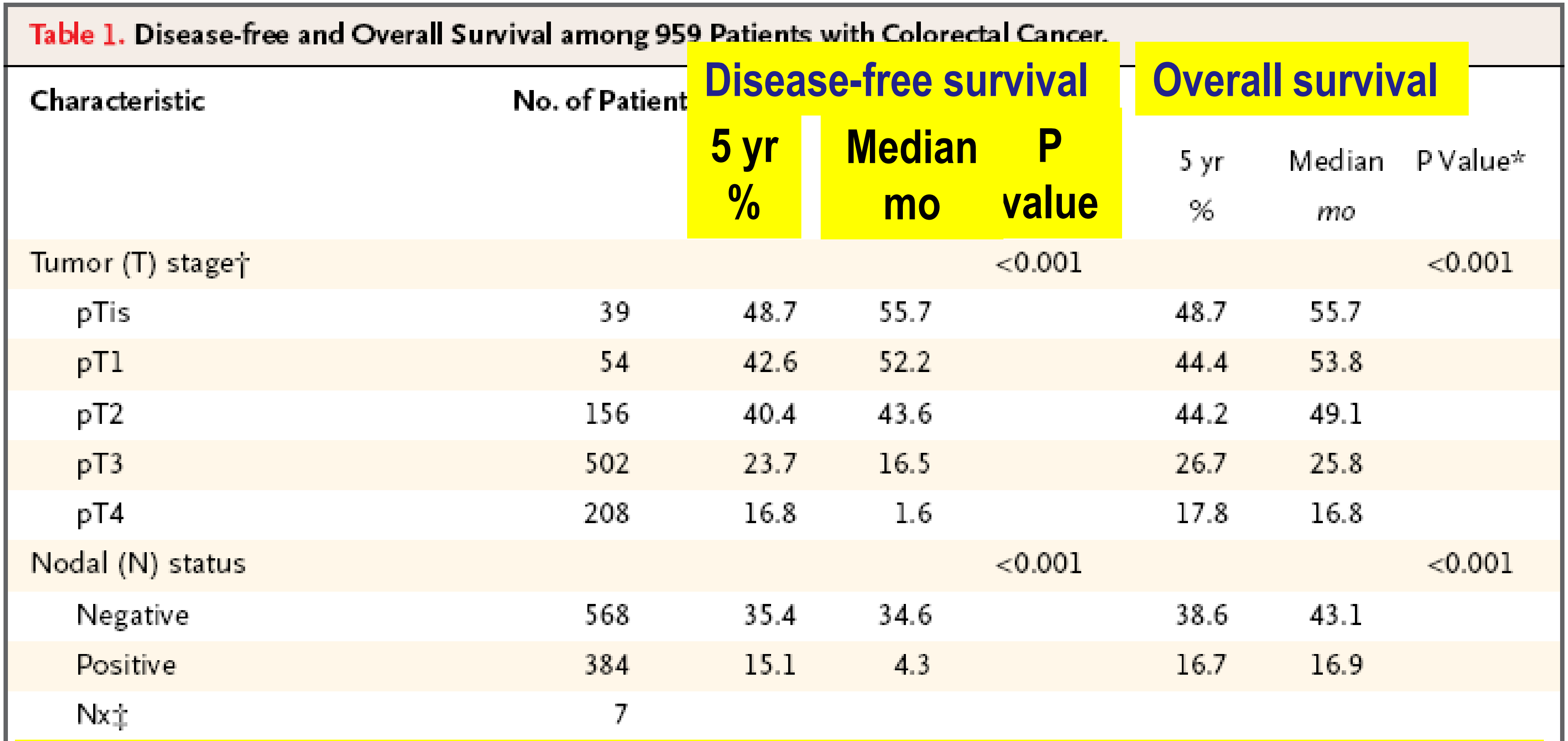


Table 1. Disease-free and Overall Survival among 959 Patients with Colorectal Cancer.

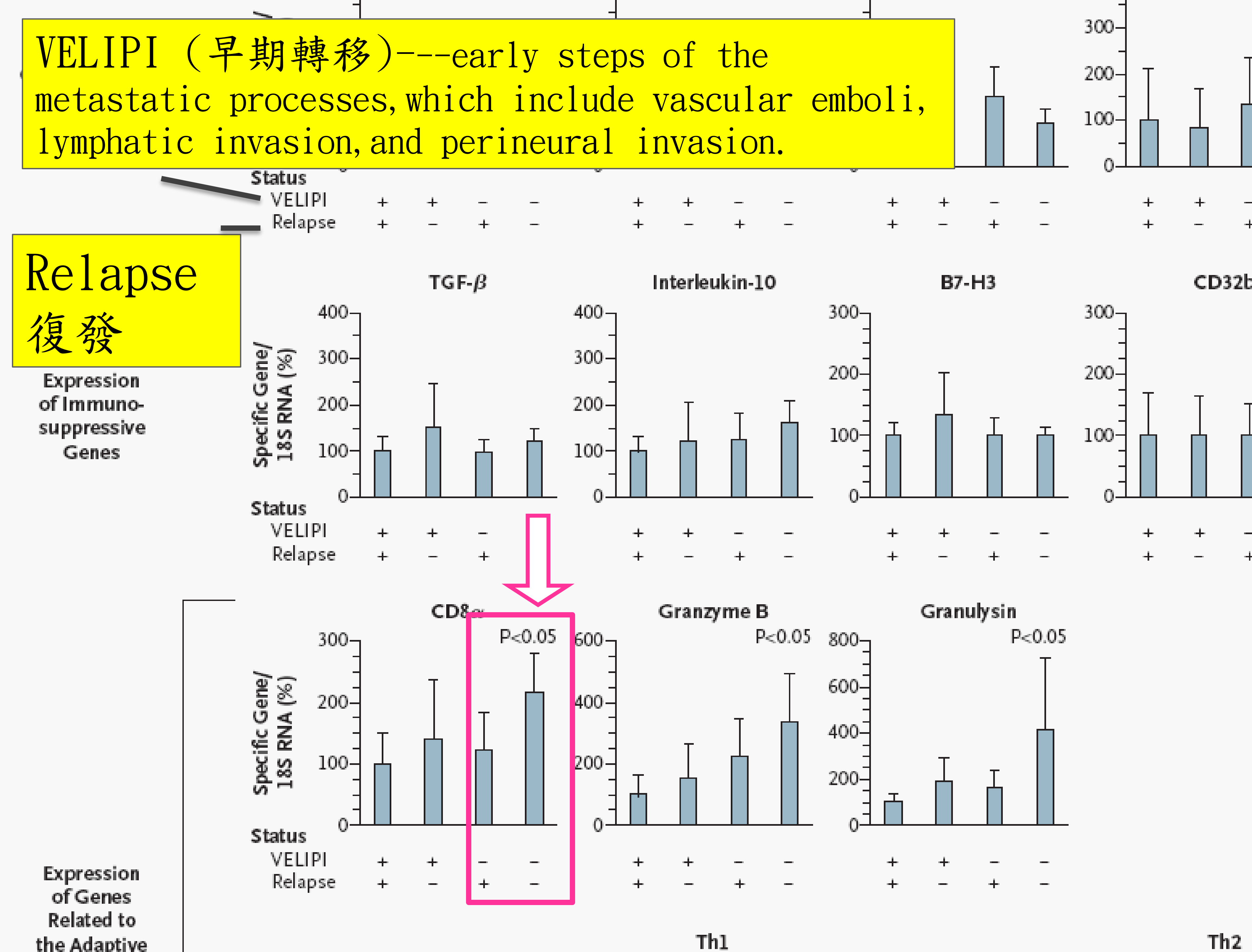
Characteristic	No. of Patient	Disease-free survival			Overall survival		
		5 yr %	Median mo	P value	5 yr %	Median mo	P Value*
Tumor (T) stage†				<0.001			<0.001
pTis	39	48.7	55.7		48.7	55.7	
pT1	54	42.6	52.2		44.4	53.8	
pT2	156	40.4	43.6		44.2	49.1	
pT3	502	23.7	16.5		26.7	25.8	
pT4	208	16.8	1.6		17.8	16.8	
Nodal (N) status				<0.001			<0.001
Negative	568	35.4	34.6		38.6	43.1	
Positive	384	15.1	4.3		16.7	16.9	
Nx‡	7						

- **Disease-free survival** (DFS) denotes the **chances of staying free of disease** after a particular treatment for a group of individuals suffering from a cancer.

- **Overall survival** is a term that denotes the **chances of staying alive** for a group of individuals suffering from a cancer.

VELIPI（早期轉移）---early steps of the metastatic processes, which include vascular emboli, lymphatic invasion, and perineural invasion.

Relapse
復發



Interpretation of Quantitative Data

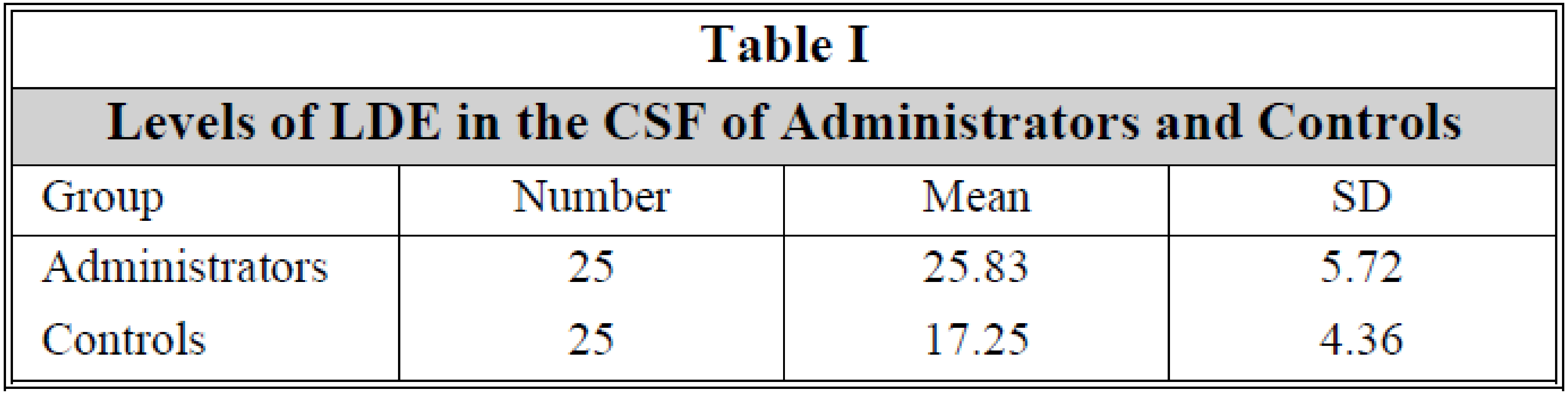


Table I

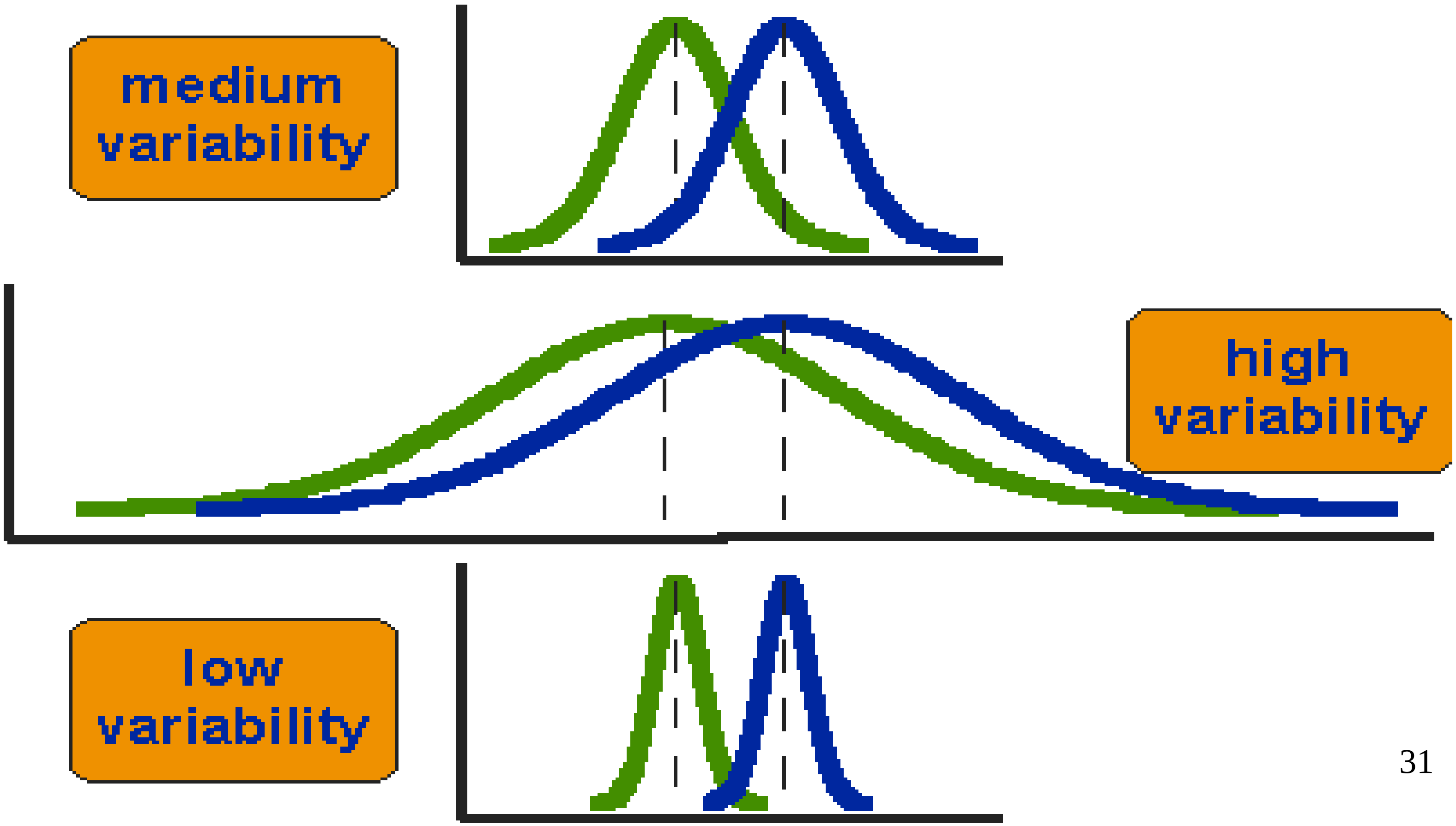
Levels of LDE in the CSF of Administrators and Controls

Group	Number	Mean	SD
Administrators	25	25.83	5.72
Controls	25	17.25	4.36

Is the difference of measured mean values from the two groups significantly different ?

How do we evaluate the data ?
Are the two groups different?

Normal control (健康) 52 54 Cancer Patient (癌症)



Normal v.s Patient?

A. Discrimination - Comparison of Data Groups

1. 2 groups with equal variances
2. 2 groups with unique variances

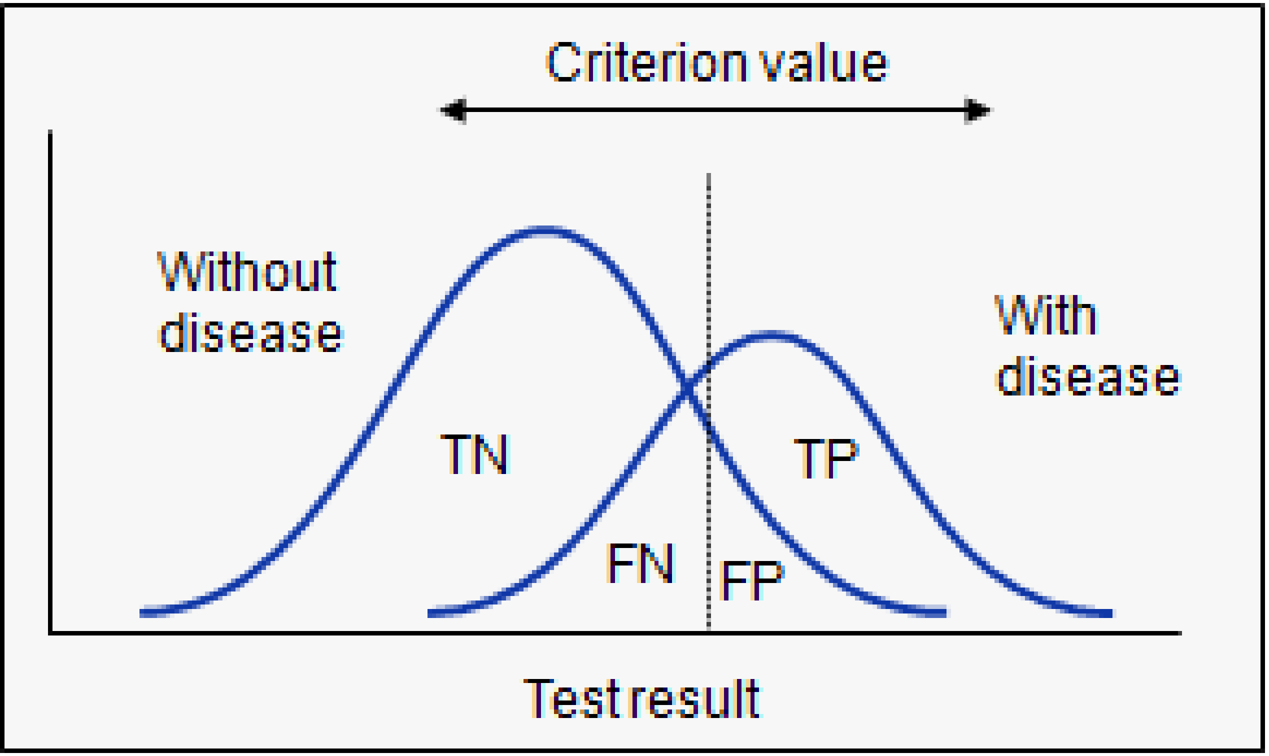
B. Receiving Operating Characteristic (ROC) curve

1. 2 X 2 contingency table
2. sensitivity & specificity
3. plotting ROC curve
4. uses of ROC curve

When the two study groups do have statistically significant difference, how do we evaluate the correlation of any new data with the two groups?

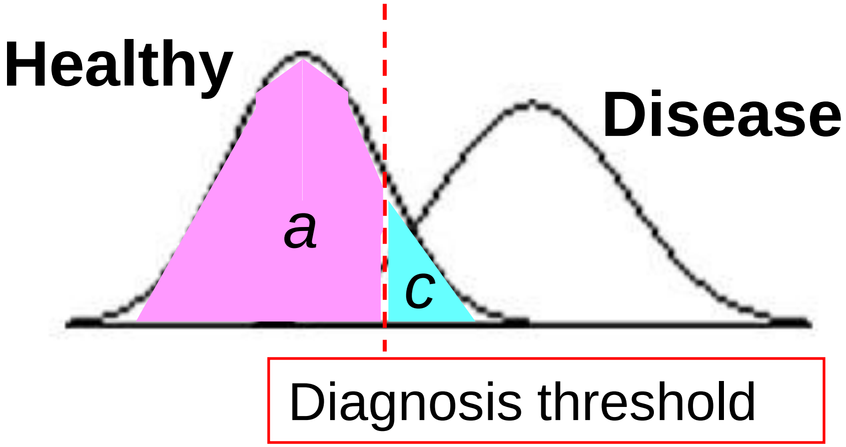
Receiver Operating Characteristics Curve (ROC curve analysis)

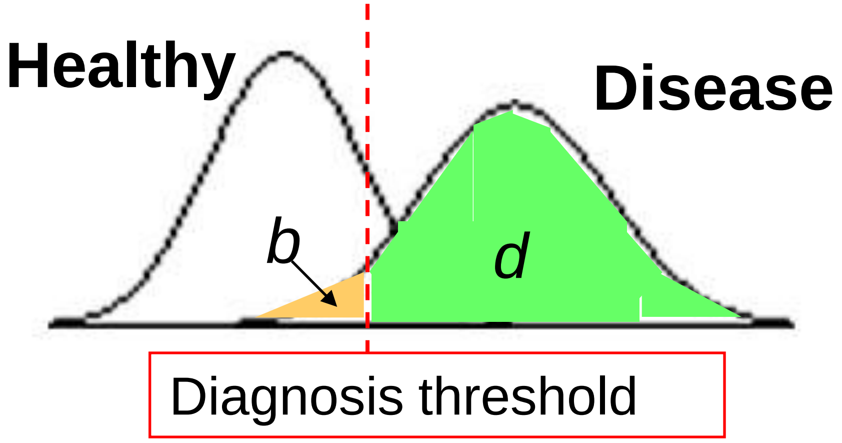
The diagnostic performance of a test, or the accuracy of a test to discriminate diseased cases from normal cases is evaluated using Receiver Operating Characteristic (ROC) curve analysis



TN: true negative
FN: false negative
TP: true positive
FP: false positive

2 x 2 Contingency Table

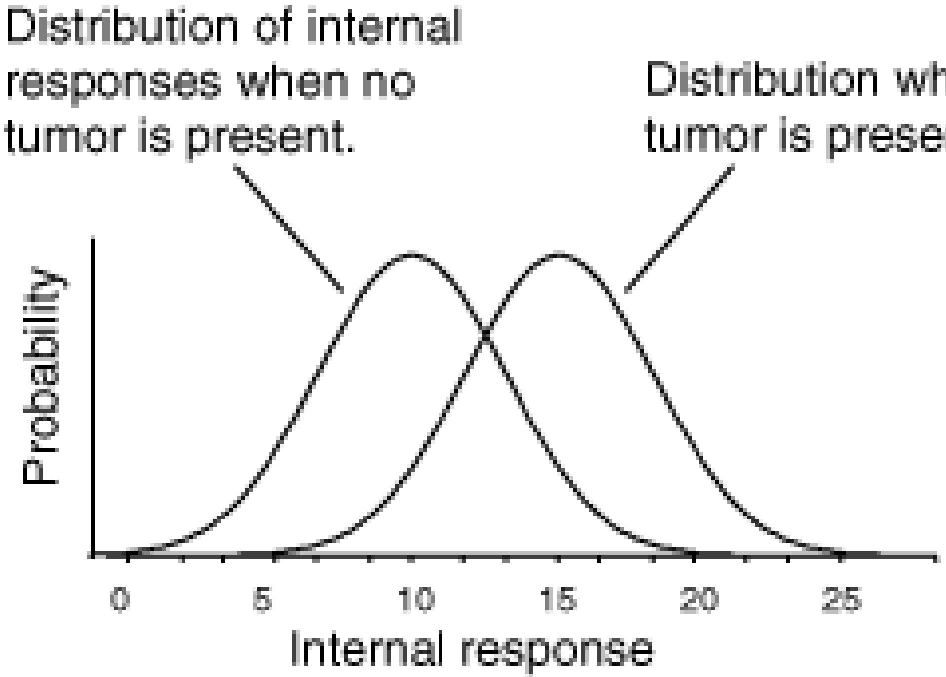


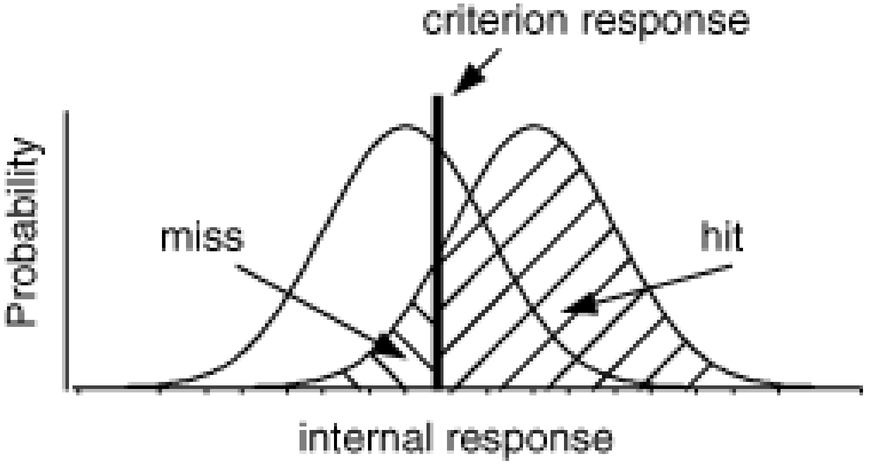


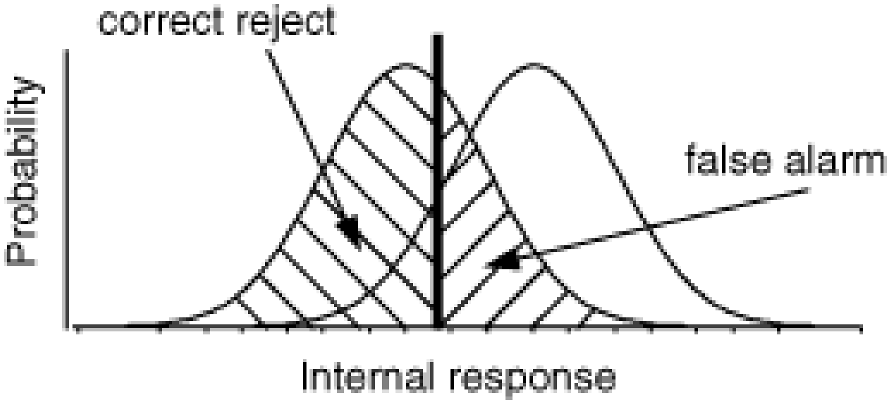
Result	Disease (true)		Total
	Absent	Present	
Normal (negative)	*a*	*b*	*a+b*
Disease (positive)	*c*	*d*	*c+d*
total	*a+c*	*b+d*	*a+b+c+d*

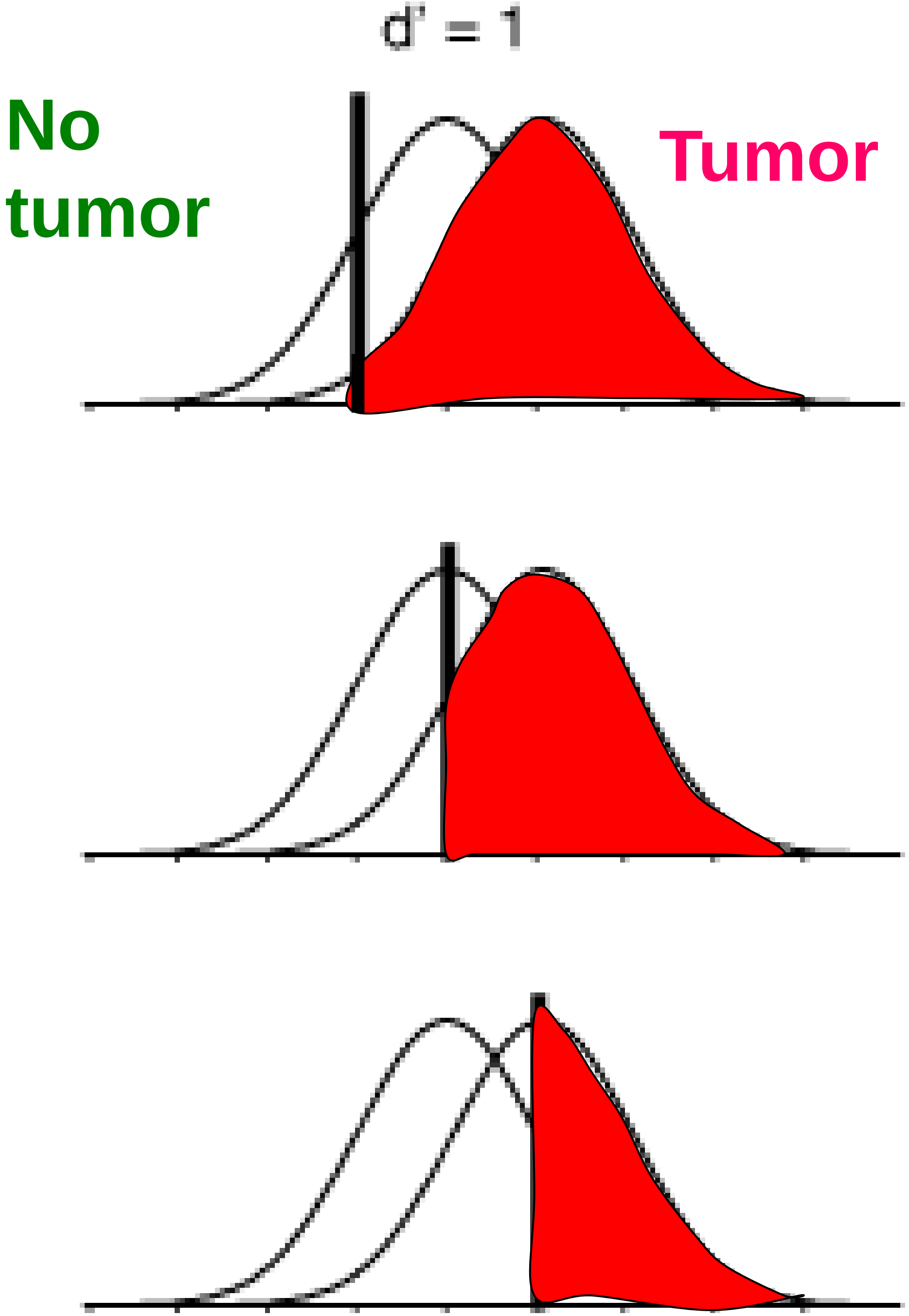
🟪 🟩 Correct

🟦 🟧 Wrong


Distribution of internal responses when no tumor is present.
Distribution when tumor is present.
Probability
0
5
10
15
20
25
Internal response


criterion response
Probability
miss
hit
internal response


correct reject
Probability
false alarm
Internal response

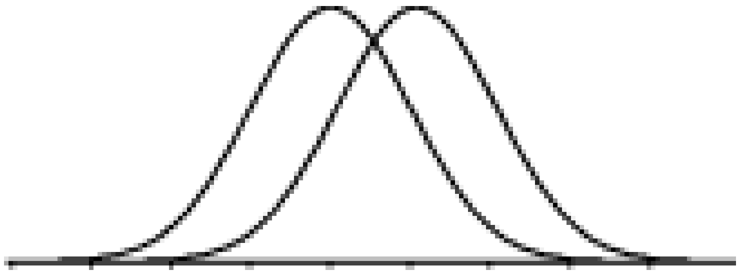


Hits = 97.5%
False alarms = 84%

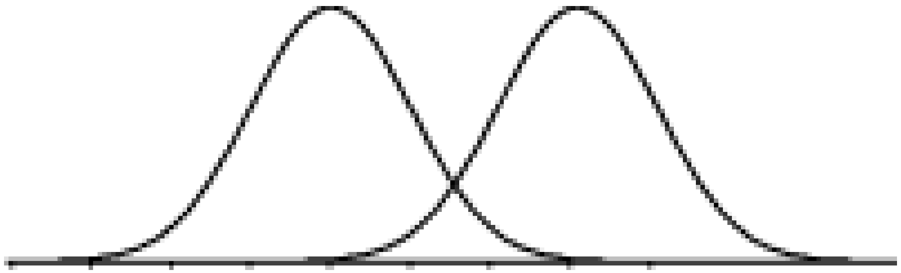
Hits = 84%
False alarms = 50%

Hits = 50%
False alarms = 16%

Receiver Operating Characteristics (ROC) Curve



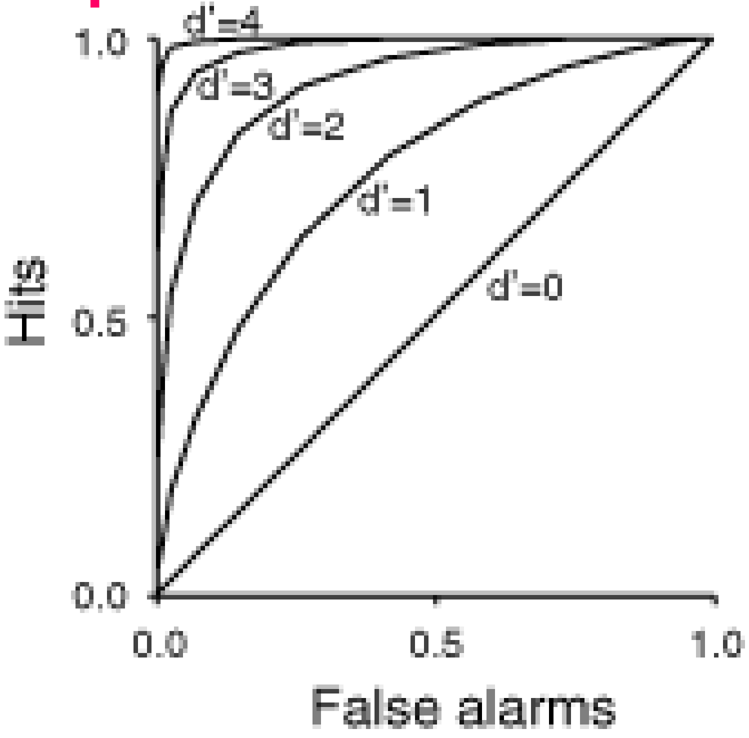
d' = 1 (lots of overlap)



d' = 3 (not much overlap)

High noise,
Lots of overlap

Low noise,
Not much overlap



ROC curves

Sensitivity & Specificity

Sensitivity

- probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage).

Sensitivity = P(disease positive｜**disease**）

= *d* / (b+*d*)

– **True Positive**

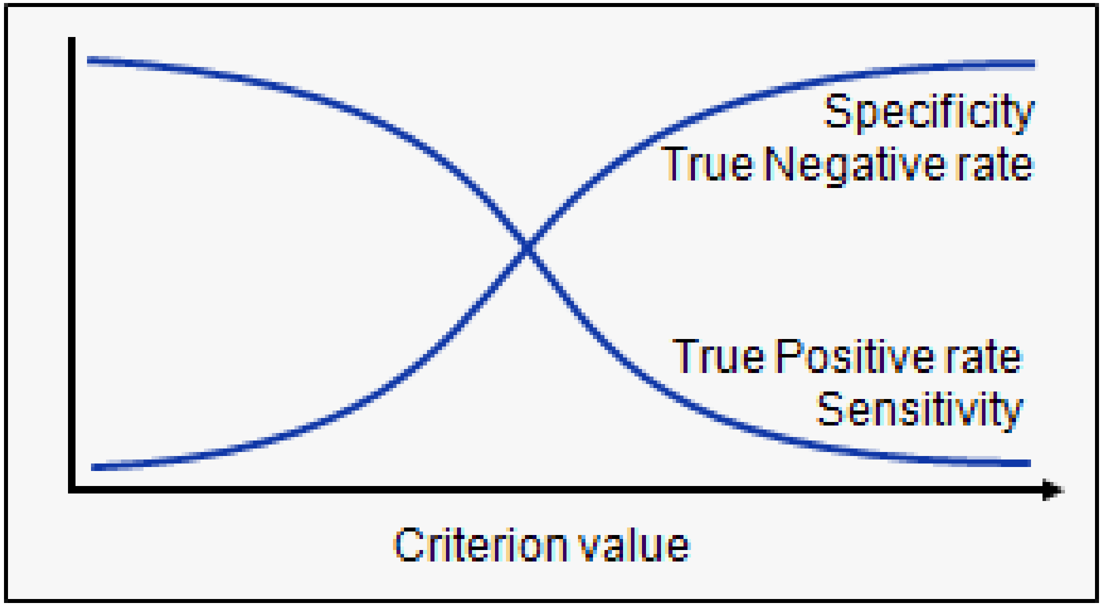
(1-sensitivity) : False Negative

Sensitivity & Specificity

- **Specificity**
- probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage)
- Specificity = P(disease negative | **noraml**)
- = a / (a+c)
  - **True negative**

(1-specificity) : False positive

Sensitivity and Specificity versus Criterion Value



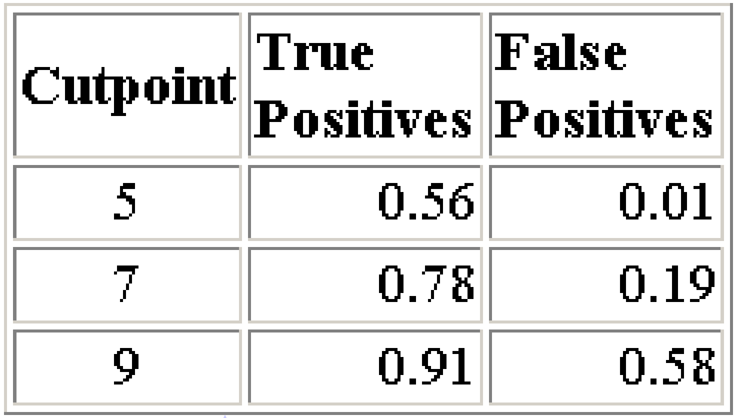
When you select a higher criterion value, the false positive fraction will decrease with increased specificity but on the other hand the true positive fraction and sensitivity will decrease.

When you select a lower criterion value, then the true positive fraction and sensitivity will increase. On the other hand the false positive fraction will also increase, and therefore the true negative fraction and specificity will decrease.

Plotting ROC Curve
Receiver Operating Characteristics Curve

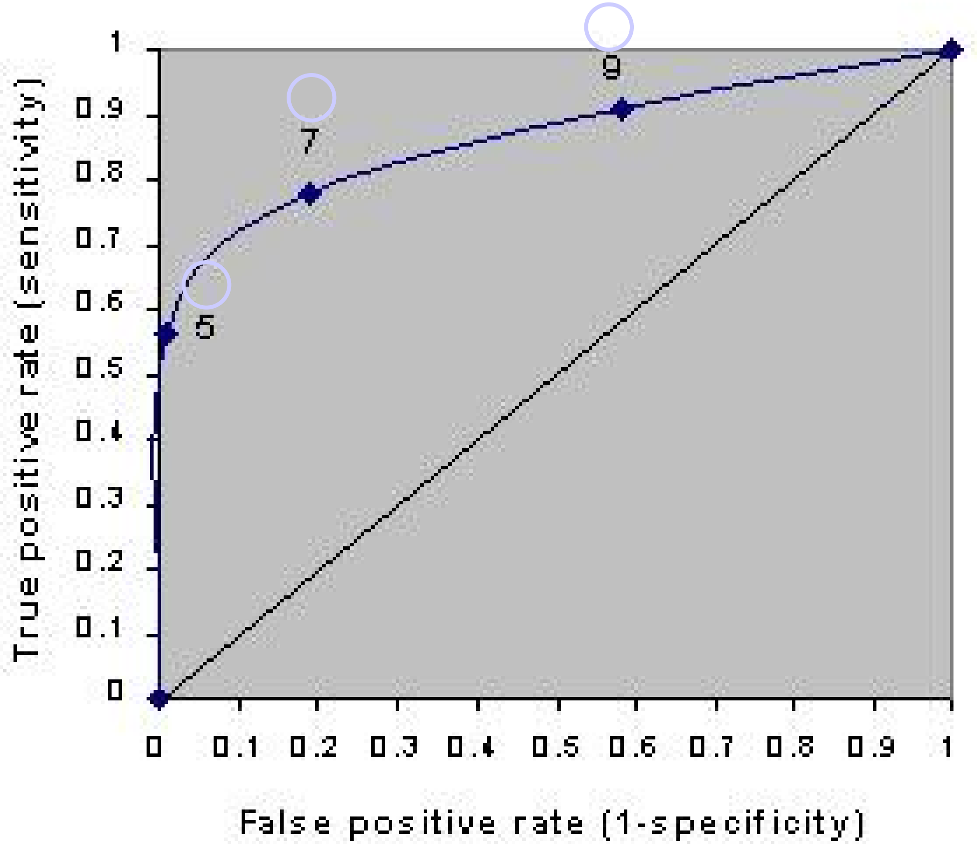
- **Y軸**：Sensitivity (true positive)
- **X軸**（1-specificity）(false positive)

(normal, but wrong diagnosis)



Cutpoint	True Positives	False Positives
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58

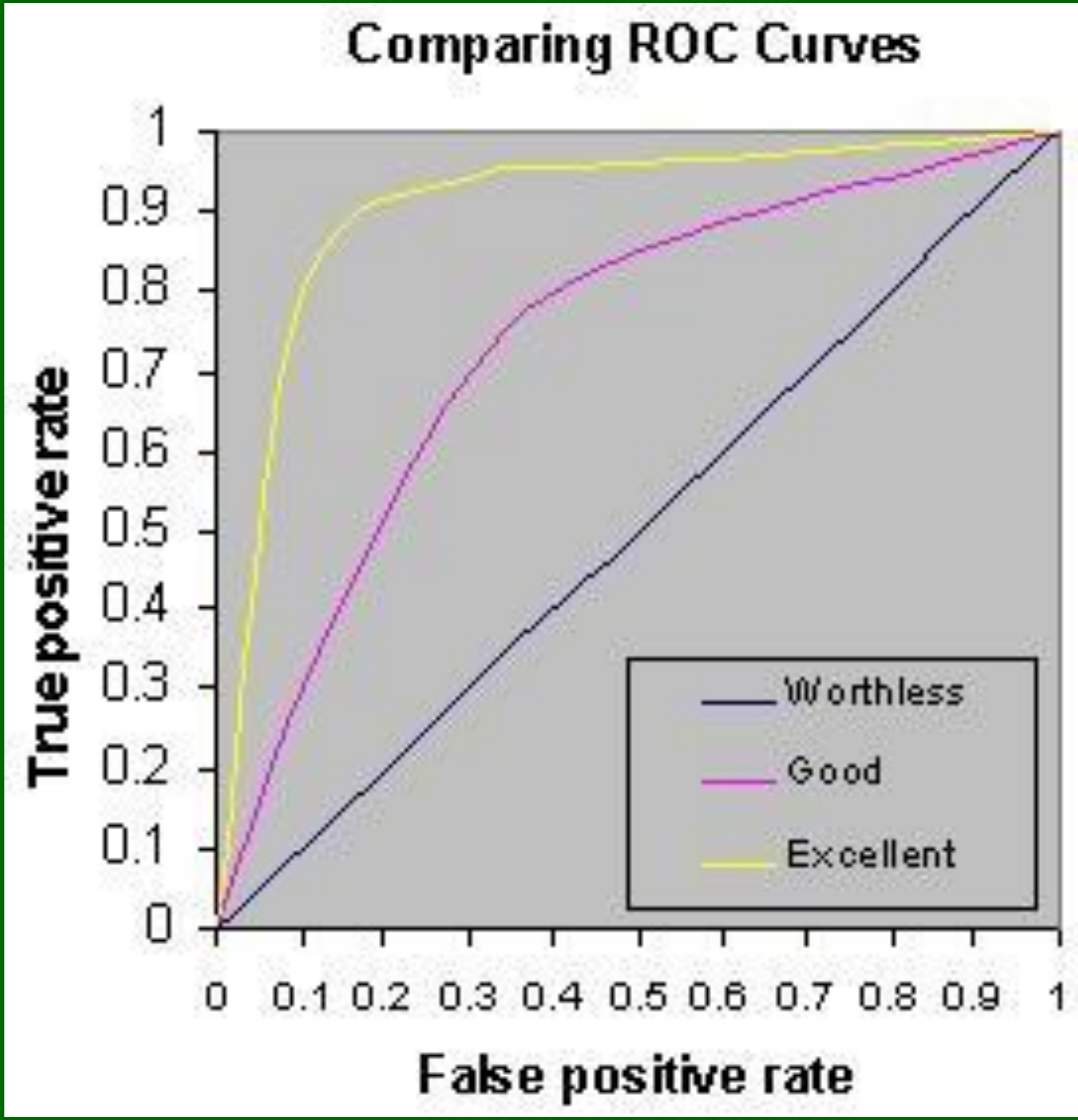
不同判定標準

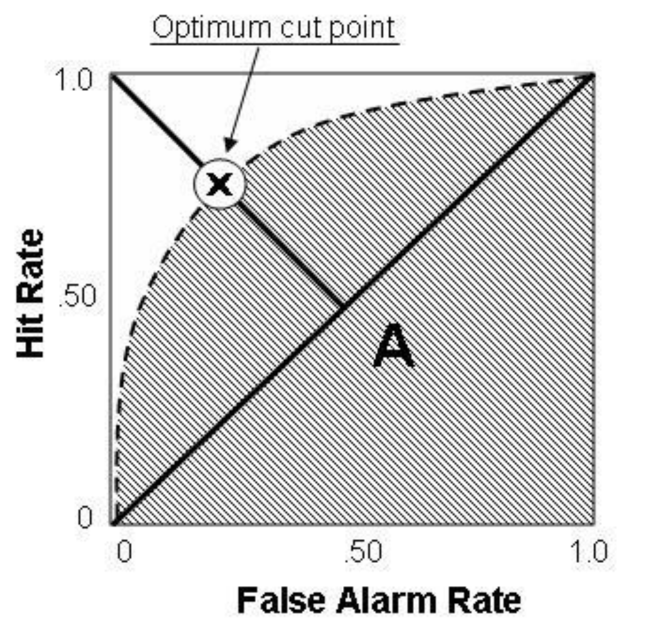


Uses of ROC curve to Determine Diagnosis Threshold

- **Area under Curve (AUC)**
  - 0.9 ~ 1.0: excellent
  - 0.8 ~ 0.9: good
  - 0.7 ~ 0.8: fair
  - 0.6 ~ 0.7: poor

ss

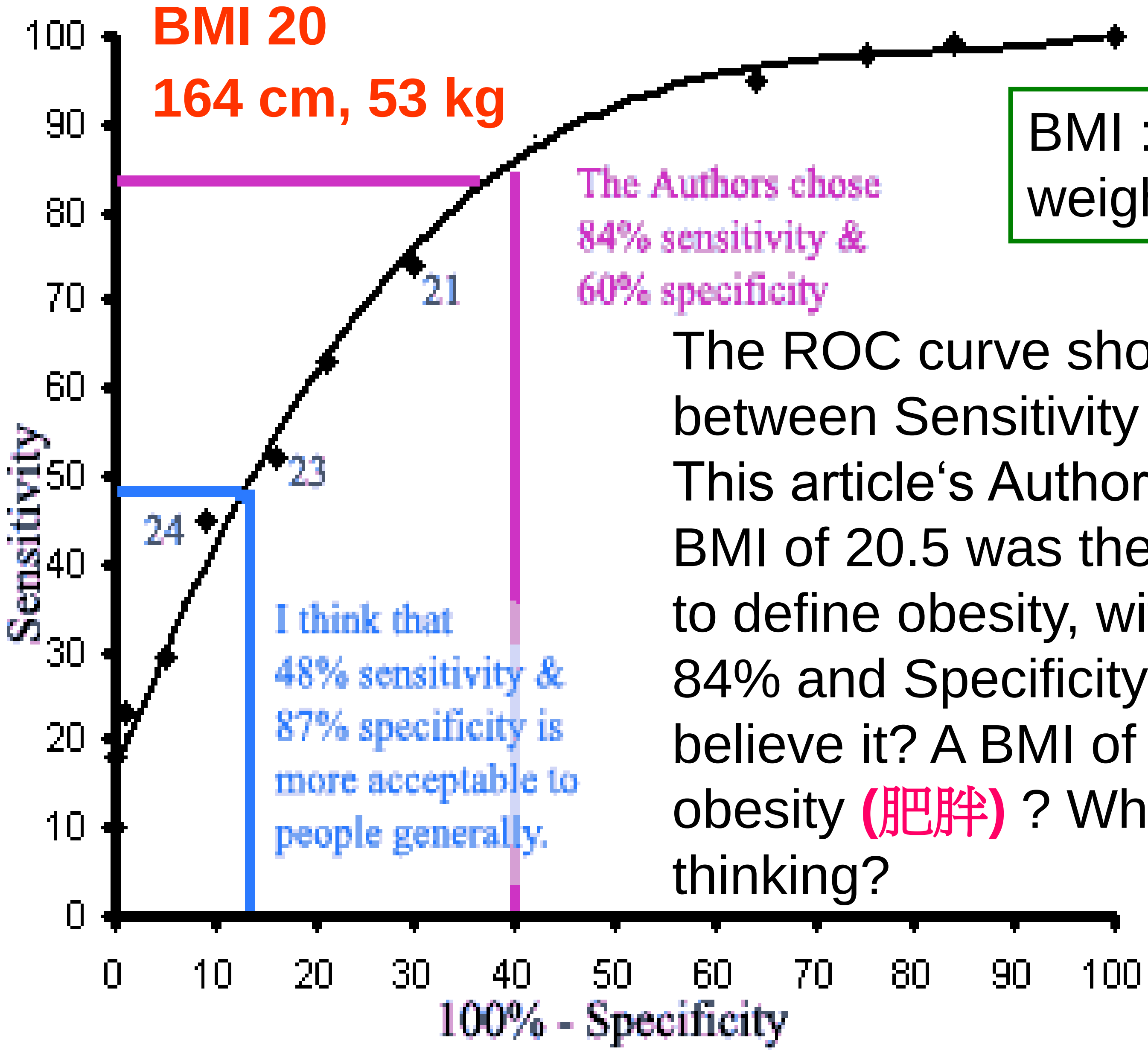




J Clin Epidemiol, Jul 1997;50(7):837-43

BMI 20

164 cm, 53 kg



The Authors chose 84% sensitivity & 60% specificity

I think that 48% sensitivity & 87% specificity is more acceptable to people generally.

BMI : weight (Kg)/Height (m^2)

The ROC curve shows the trade-offs between Sensitivity and Specificity. This article‘s Authors believed that a BMI of 20.5 was the optimum threshold to define obesity, with a Sensitivity of 84% and Specificity of 60%. Can you believe it? A BMI of 20.5 to define obesity **(肥胖)** ? What were they thinking?

Assessment of the Performance of a Method

(BMB 1.6.2)

Summary Statistics

- **Measures of Central Tendency**

– Mean, Median, Mode

- **Spread**

–Range

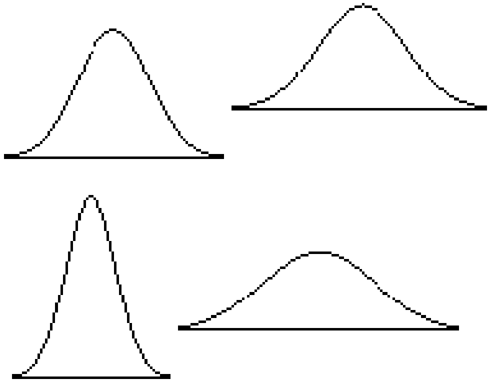
–Variance

–Standard deviation

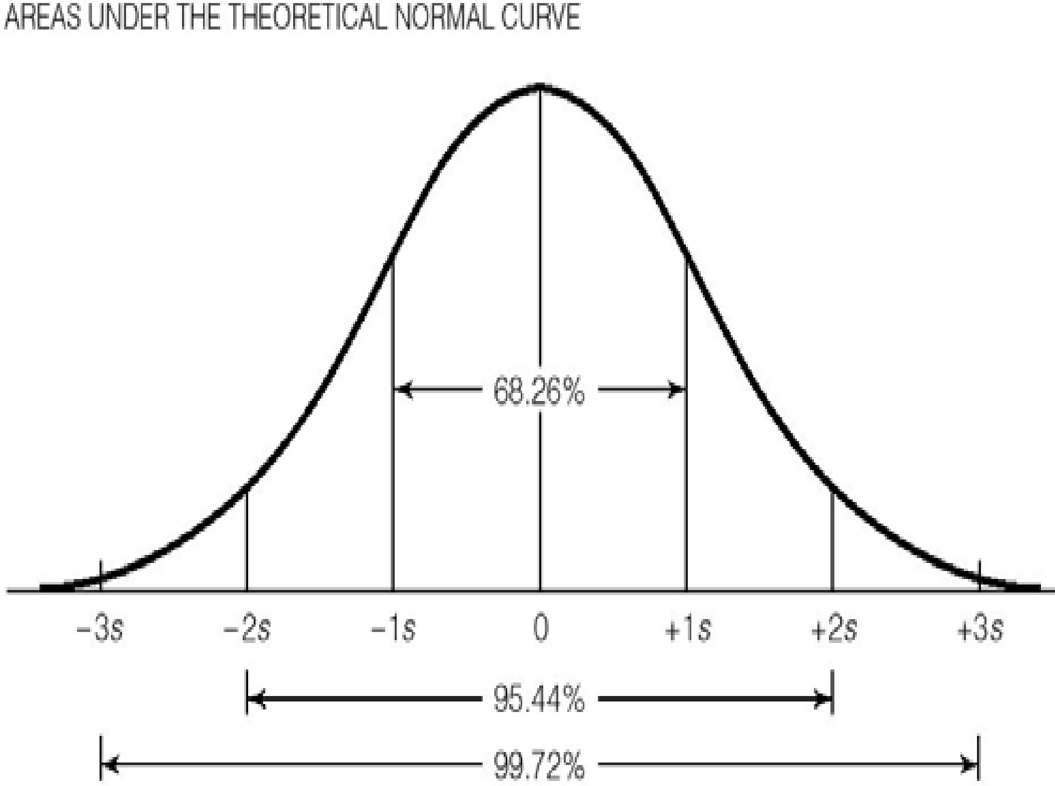
–Stander error

- **Shape**

Data Follows Normal Distribution



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

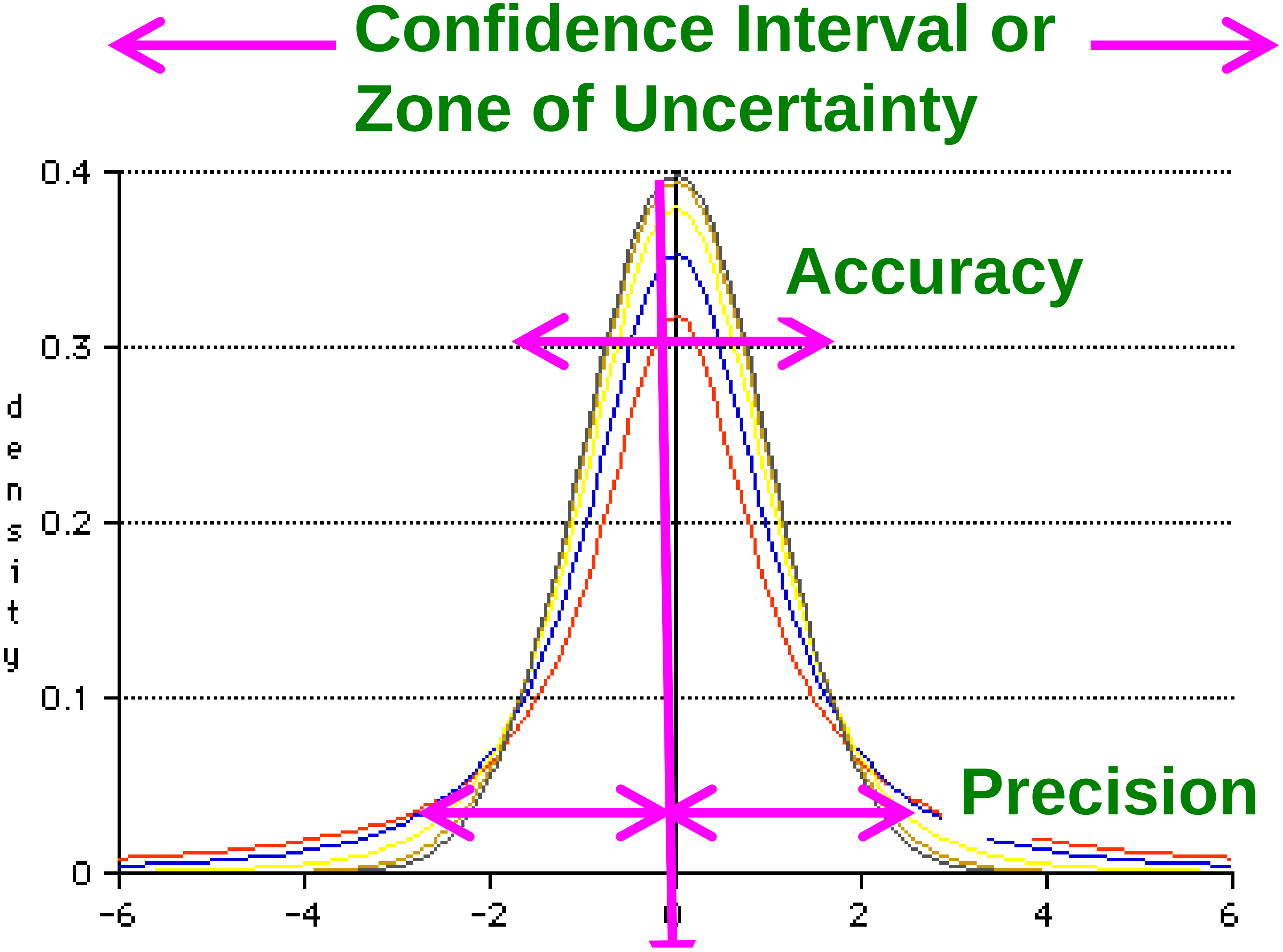


- The ***x-axis*** represents the values of a particular variable

- The ***y-axis*** represents the proportion of members of the population that have each value of the variable

- The area under the curve represents probability – i.e. area under the curve between two values on the x-axis represents the probability of an individual having a value in that range

Real-World Quantification



Confidence Interval or
Zone of Uncertainty

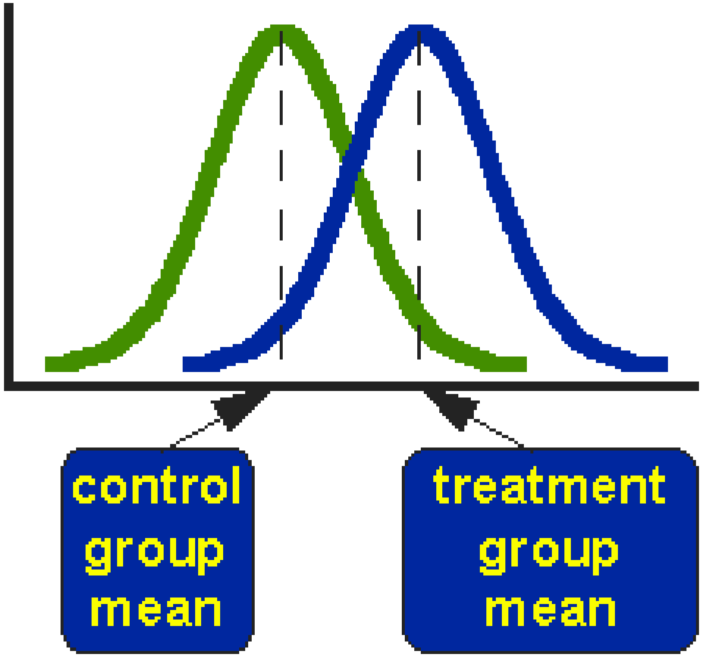
Accuracy

Precision

True mean Value

'Student's' t Test

The *t*-test compares the actual difference between two means in relation to the variation in the data



http://www.socialresearchmethods.net/kb/stat_t.htm

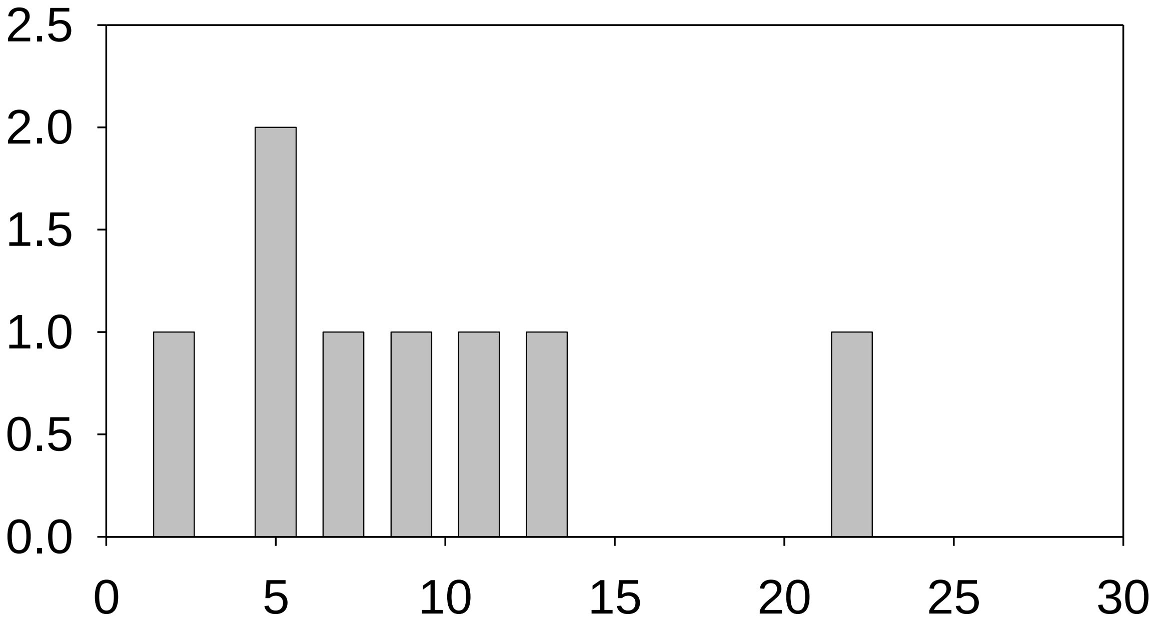
'Student's' t Test

- **One-sample t-test**: know the mean difference between the sample and the known value of the population mean.
- **Unpaired t-test**: compare two population means
- **Paired t-test**: compare the values of means from two related samples, for example in a ‘before and after’ scenario.

When $t_{calc} > t_{table}$ ，the two value are not the same (within the confidence intervals)

Measures of Central Tendency

- –Mode
- –Median
- –Mean



e.g. 2, 5, 5, 7, 9, 11, 13, 22

mode = 5 (greatest frequency)
median= (7+9)/2=8
mean=(2+5+5+7+9+11+13+22)/8

Median=50%
Odd：中間數值
Even：中間兩數之平均

Spread -----Variance

- **Variance** (變異數)：

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

- **Standard Deviation, S.D.** (標準差)= gives the dispersion of numerical data around the mean value ：

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

N-1: degree of freedom

= [Number of observation — 1]

Q: Why do we divide by (n-1) and not by (n)?

- Use of n as a divisor will give a sample standard deviation which tends to underestimate the population standard deviation, whereas the use of **(n-1)** gives what is known as an”unbiased estimator”
- Score deviates less from their own mean than from any other number. So, the calculation subtracting each score from the sample mean will be smaller than subtracting form the population mean------ ***underestimate*** the SD

(n-1)

Statistics for Analytical Chemists, by R. Caulcutt and R. Boddy

Spread -----Coefficient of Variance

Coefficient of Variation (變異係數)：
Relative standard deviation

$$CV = \frac{s}{\bar{x}} \times 100\%$$

e.g. A: 2.00± 0.10 mM, CV=5.0%

B: 8.00± 0.41 mM, CV=5.0%

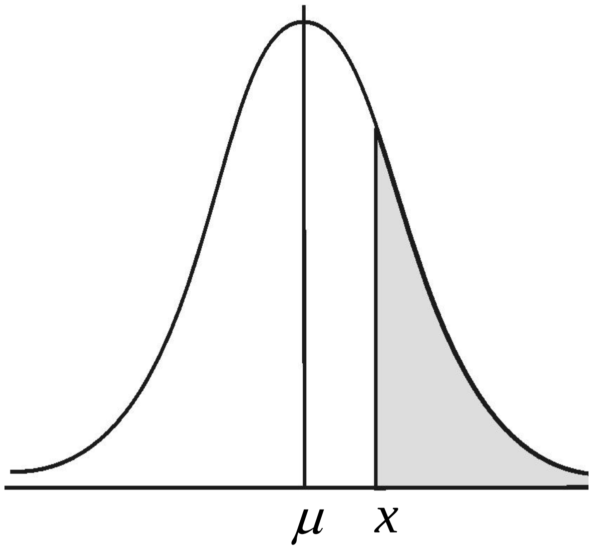
Spread -----Coefficient of Variance

- Possibility of occurrence

80% Rainy

98% Rainy, Cloudy, Sunny

- ***p*-value**



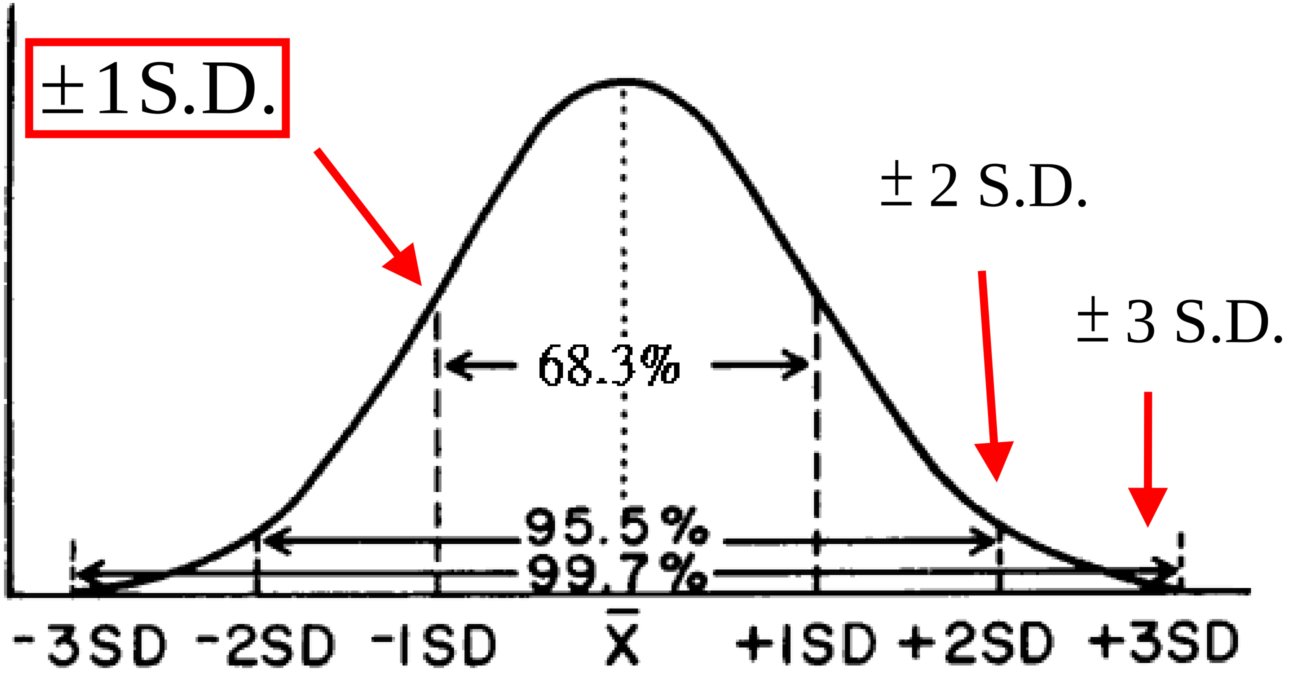
P=0.05 (=95% confidence)
-----***Statistically significant***

Define the spread or distribution of the data

68.3% data will be within the range of $\bar{X} \pm 1\text{S.D.}$

The possibility of a data point within the range of $\bar{X} \pm 1\text{S.D.}$ is 68.3%.

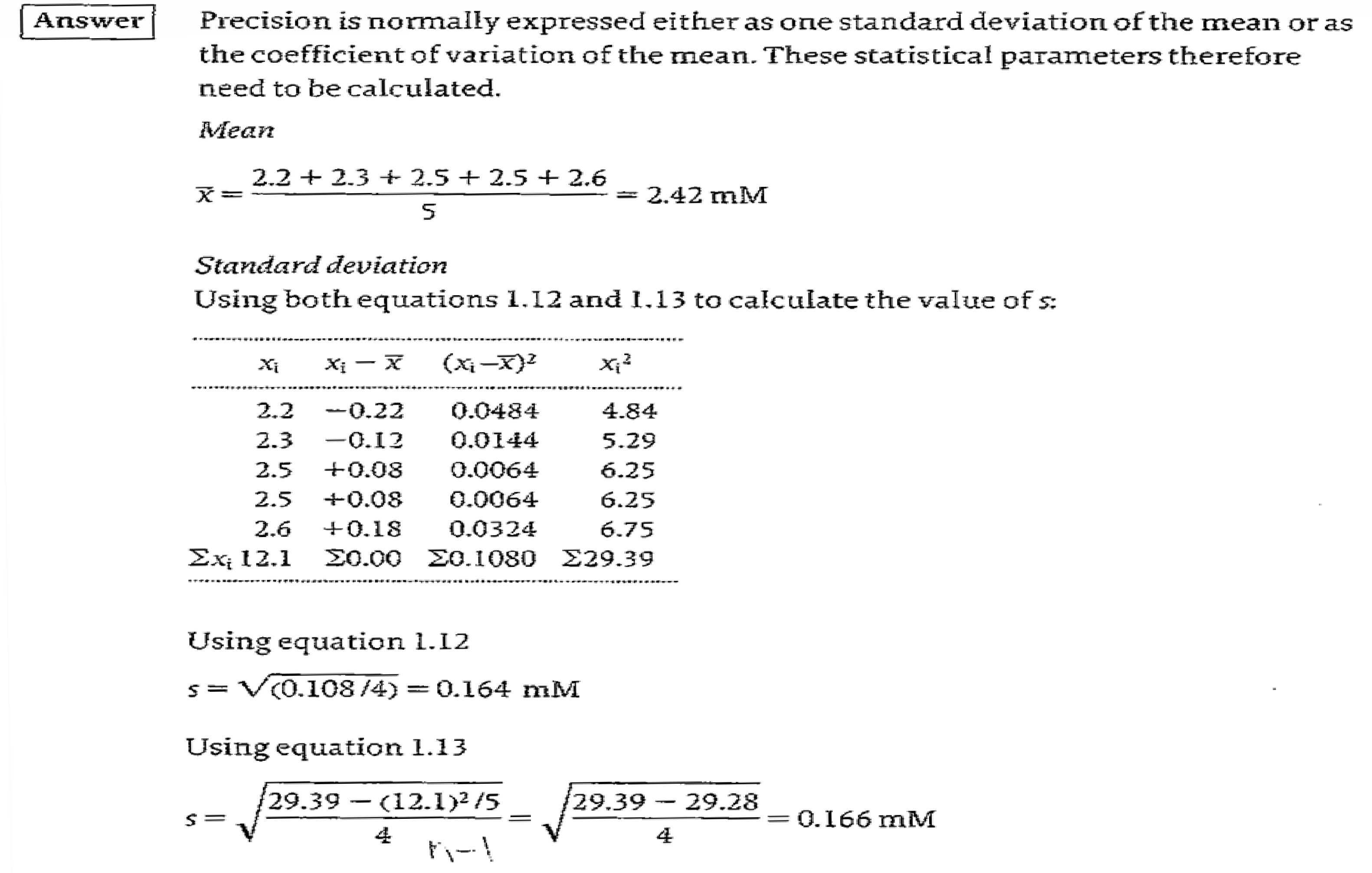
Frequency of occurrence of a measurement



Gaussian Distribution/Normal Distribution

ASSESSMENT OF THE PRECISION OF AN ANALYTICAL DATA SET

Five measurements of the fasting serum glucose concentration were made on the same sample taken from a diabetic patient. The values obtained were 2.3, 2.5, 2.2, 2.6 and 2.5 mM. Calculate the precision of the data set.



Answer

Precision is normally expressed either as one standard deviation of the mean or as the coefficient of variation of the mean. These statistical parameters therefore need to be calculated.

Mean

$$\bar{x} = \frac{2.2 + 2.3 + 2.5 + 2.5 + 2.6}{5} = 2.42\ \text{mM}$$

Standard deviation

Using both equations 1.12 and 1.13 to calculate the value of *s*:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
2.2	−0.22	0.0484	4.84
2.3	−0.12	0.0144	5.29
2.5	+0.08	0.0064	6.25
2.5	+0.08	0.0064	6.25
2.6	+0.18	0.0324	6.75
Σx_i 12.1	Σ0.00	Σ0.1080	Σ29.39

Using equation 1.12

$$s = \sqrt{(0.108/4)} = 0.164\ \text{mM}$$

Using equation 1.13

$$s = \sqrt{\frac{29.39 - (12.1)^2/5}{\underset{n-1}{4}}} = \sqrt{\frac{29.39 - 29.28}{4}} = 0.166\ \text{mM}$$

Accuracy (Bias, Inaccuracy)

Differences between “mean” and “true” value

① When the number of sampling approaches infinity, “mean” is equal to the “population mean ”

② If the “uncertainty” (SD) is close to 0,

Then，n much approach infinity

(Eg：when SD is 1/2→n has to increase to 4-fold)

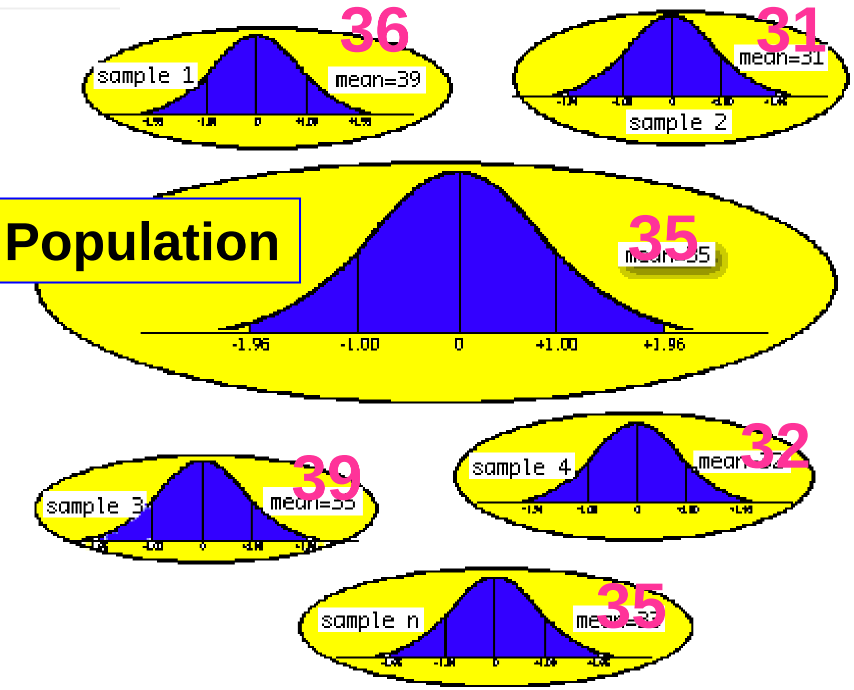
$$s = \left(\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2}$$

How do we evaluate the difference of measured “mean” and “true mean” of the population?

In practical experimental design, it is not possible to sample EVERY analyte from the population.

Animal model, cancer vs healthy group….etc

Standard Error (of the mean, S.E.)



S.D **variability of original data**

The absolute value of S.D. can not tell the difference of mean and popuation mean

S.E $$\frac{SD}{\sqrt{N}}$$ **variability of mean**

…..Why does the denominator read $N^{1/2}$ instead of just N? Because we are really dividing the variance, which is SD^2, by N, but we end up again with squared units, so we take the square root of everything…….

Can J Psychiatry, Vol 41, October 1996

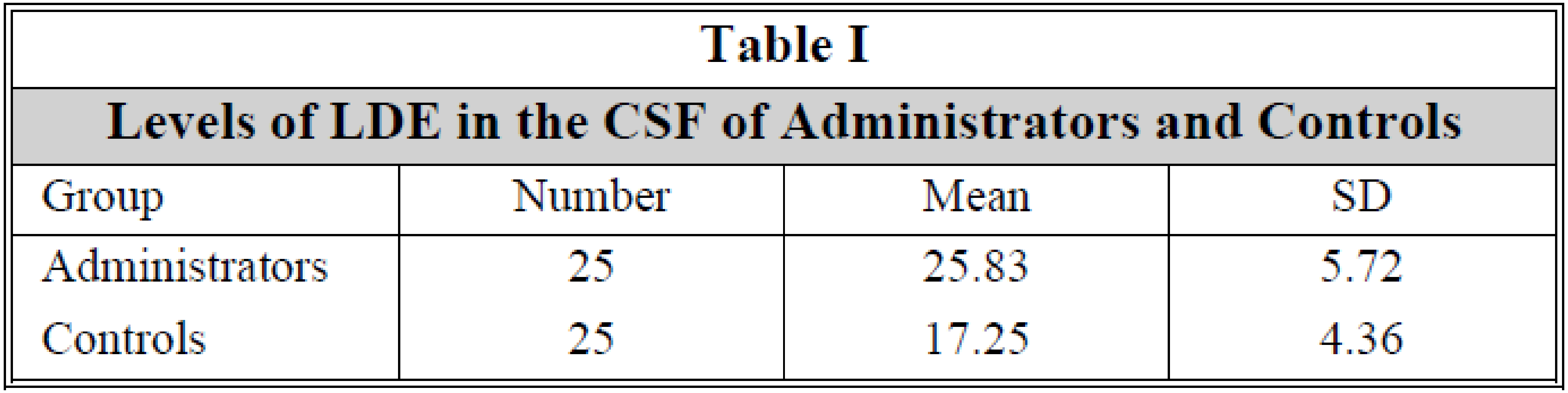


Table I

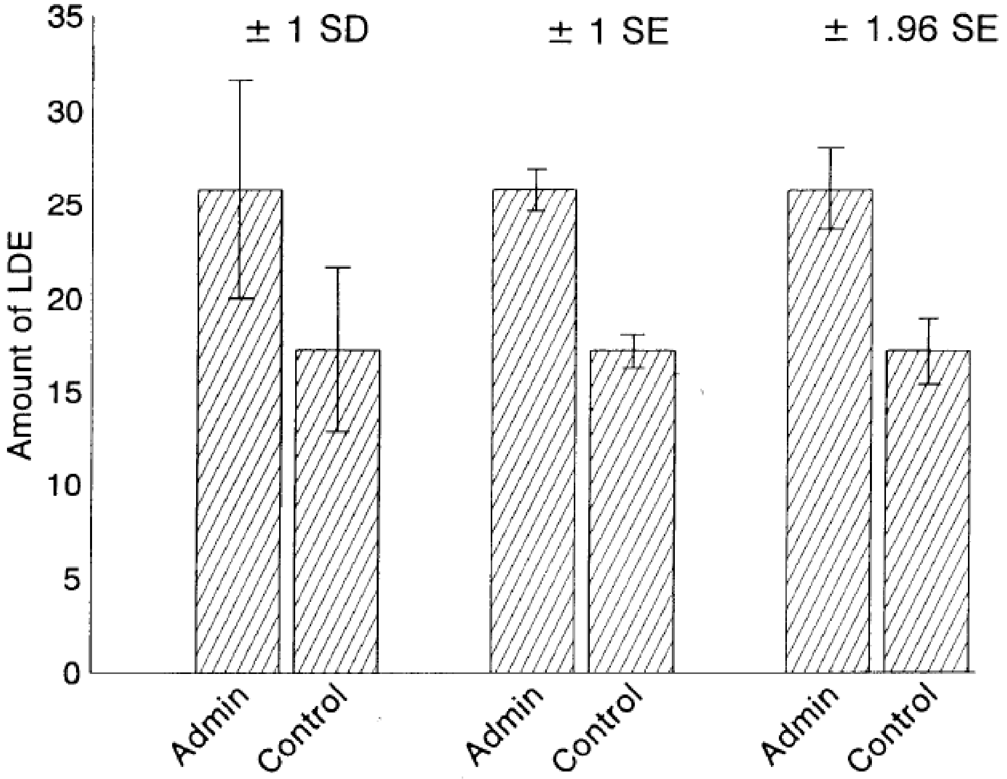
Levels of LDE in the CSF of Administrators and Controls

Group	Number	Mean	SD
Administrators	25	25.83	5.72
Controls	25	17.25	4.36

SD v.s. SE

$$s = \left(\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{1/2}$$

$$\text{SE} = \frac{\text{SD}}{\sqrt{\text{N}}}$$



Spread--Confidence Interval

Gives a range of values about the sample mean within a given probability

- for normal distribution

$$P(-1.96 \le z \le 1.96) = 0.95\text{, and } z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\Rightarrow P(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data

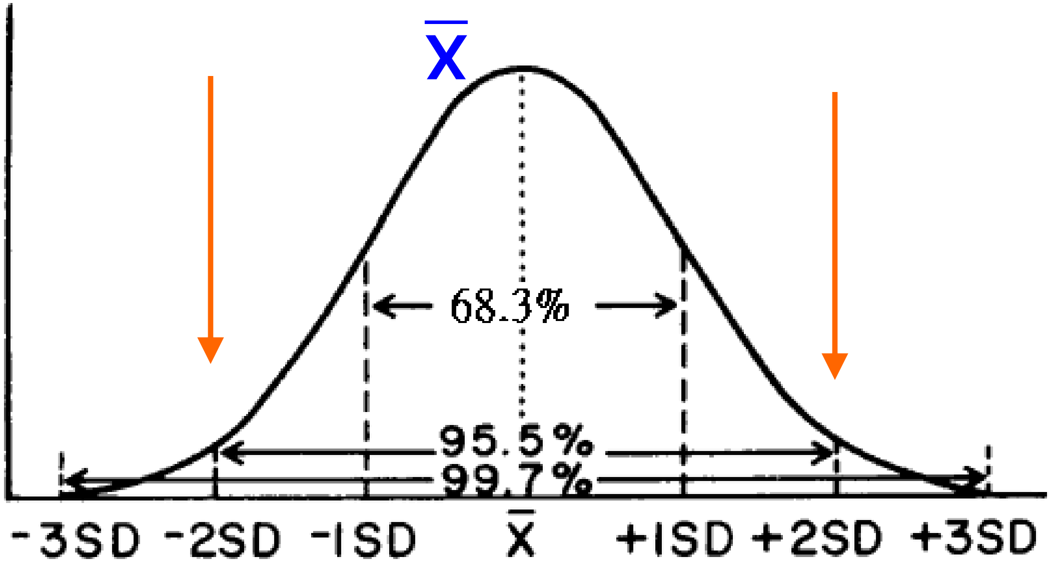
Spread---Confidence Interval

The lower and upper boundaries / values of a confidence interval, that is, the values which define the range of a confidence interval

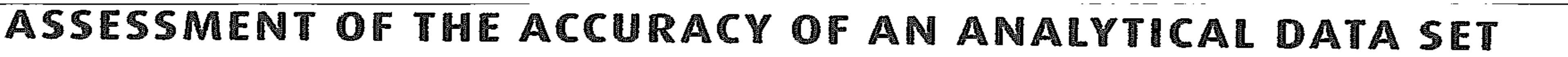
$$\left[\overline{X} - (t) \cdot \frac{SD}{\sqrt{n}}\right] \le M \le \left[\overline{X} + (t) \cdot \frac{SD}{\sqrt{n}}\right]$$

Confidence Limit

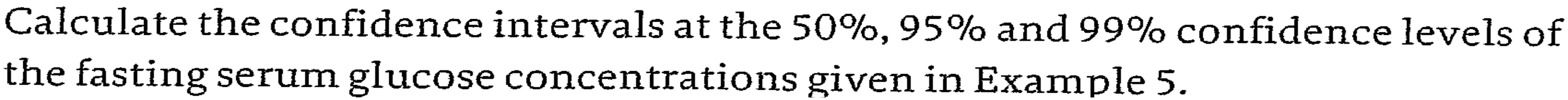
t：student‘s factor (Table 1.9)



Example 6

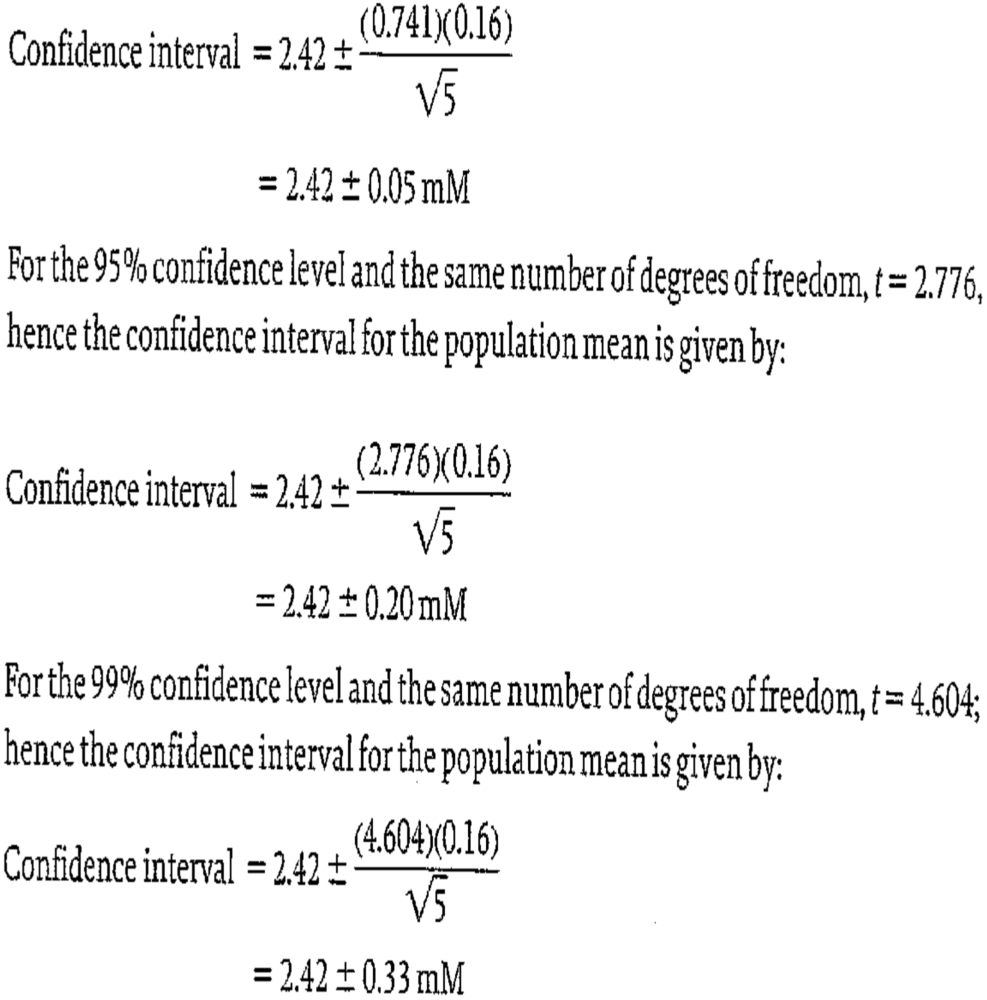


ASSESSMENT OF THE ACCURACY OF AN ANALYTICAL DATA SET



Calculate the confidence intervals at the 50%, 95% and 99% confidence levels of the fasting serum glucose concentrations given in Example 5.

$$\left[\overline{X} - (t)\cdot\frac{SD}{\sqrt{n}}\right] \le M \le \left[\overline{X} + (t)\cdot\frac{SD}{\sqrt{n}}\right]$$



$$\text{Confidence interval} = 2.42 \pm \frac{(0.741)(0.16)}{\sqrt{5}}$$

$$= 2.42 \pm 0.05\ \text{mM}$$

For the 95% confidence level and the same number of degrees of freedom, $t = 2.776$, hence the confidence interval for the population mean is given by:

$$\text{Confidence interval} = 2.42 \pm \frac{(2.776)(0.16)}{\sqrt{5}}$$

$$= 2.42 \pm 0.20\ \text{mM}$$

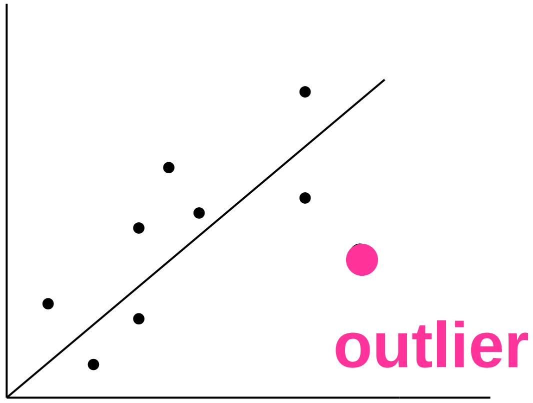
For the 99% confidence level and the same number of degrees of freedom, $t = 4.604$; hence the confidence interval for the population mean is given by:

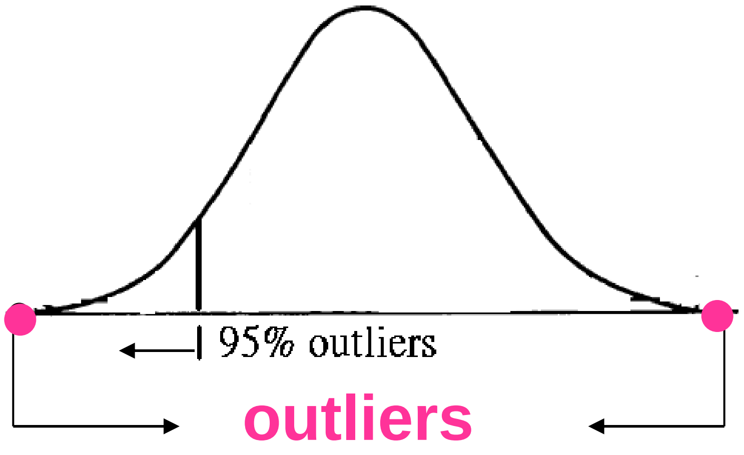
$$\text{Confidence interval} = 2.42 \pm \frac{(4.604)(0.16)}{\sqrt{5}}$$

$$= 2.42 \pm 0.33\ \text{mM}$$

Outlier

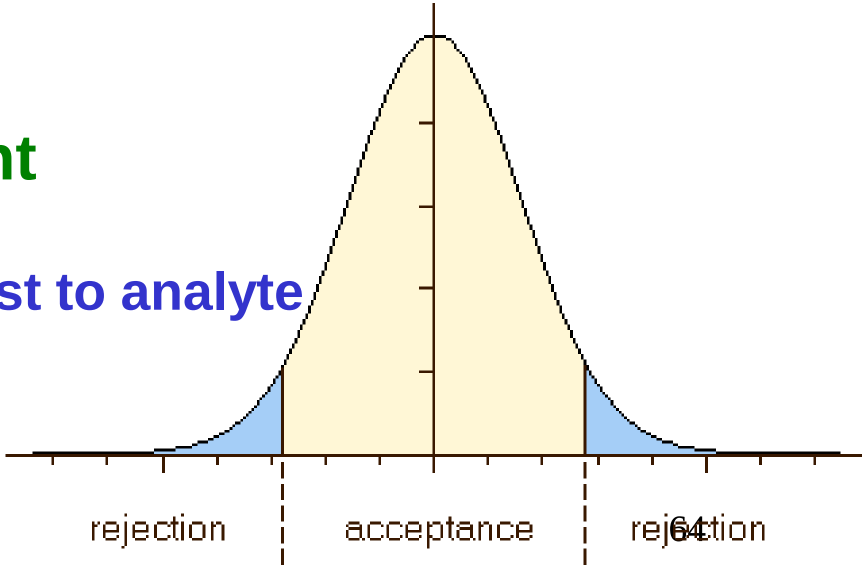
Rejection of outlier experimental data





Q exp (Dixon’s Q-test)

Experimental rejection quotient



The data point closest to analyte

$$Q_{exp} = \frac{X_n - X_{n-1}}{X_n - X_1} = \frac{\text{gap}}{\text{range}}$$

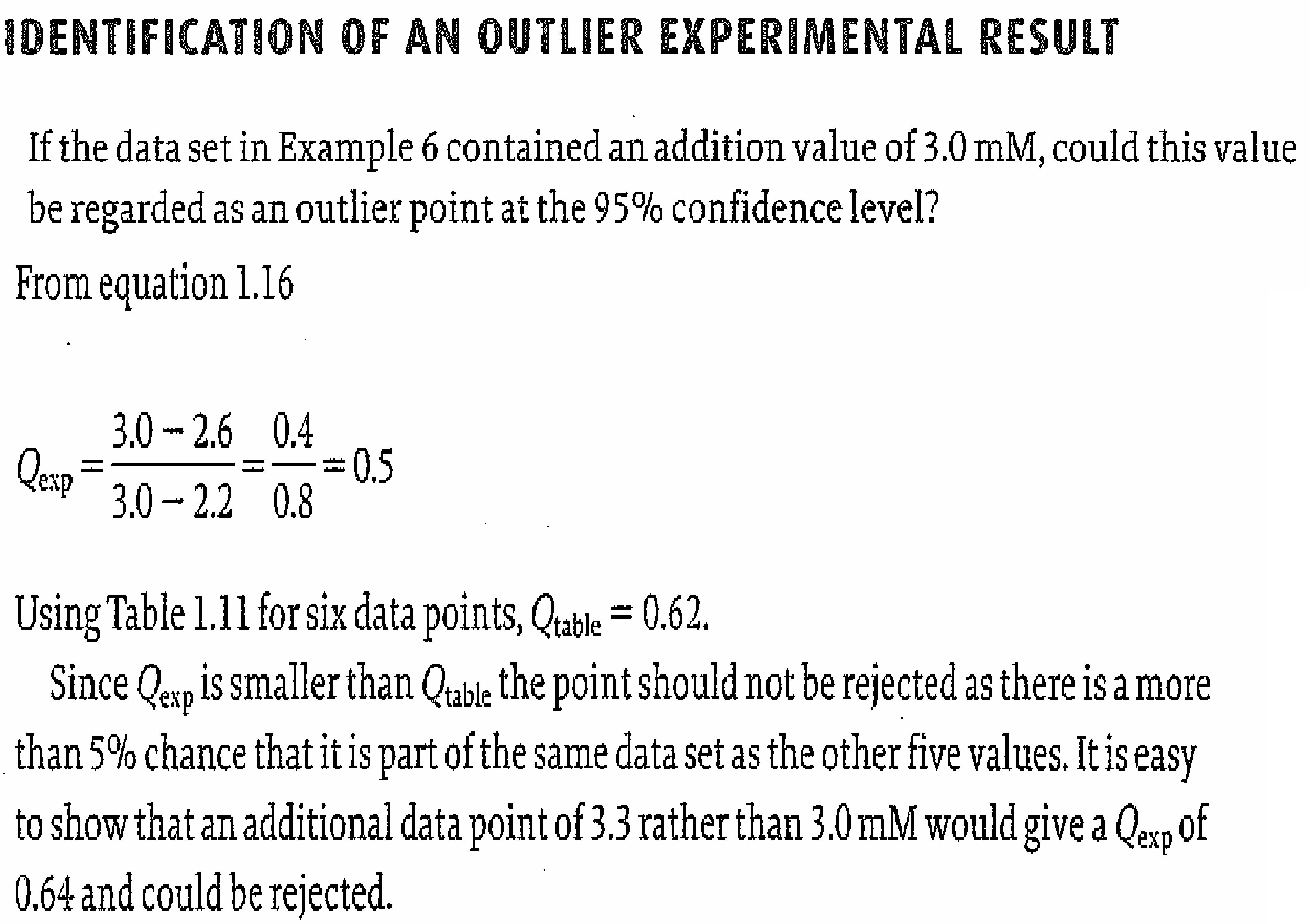
Outlier– **Q values**

Table.1.1	Values of Q for the rejection of outliers	
Number of observations		Q (95% confildence)
4		0.83
5		0.72
6		0.62
7		0.57
8		0.52

$Q_{exp} < Q_{\text{Table 1.10}}$ ------ Accept the datapoint

$Q_{exp} > Q_{\text{Table 1.10}}$ ------ Reject the datapoint

Example 7



IDENTIFICATION OF AN OUTLIER EXPERIMENTAL RESULT

If the data set in Example 6 contained an addition value of 3.0 mM, could this value be regarded as an outlier point at the 95% confidence level?

From equation 1.16

$$Q_{\mathrm{exp}} = \frac{3.0 - 2.6}{3.0 - 2.2} = \frac{0.4}{0.8} = 0.5$$

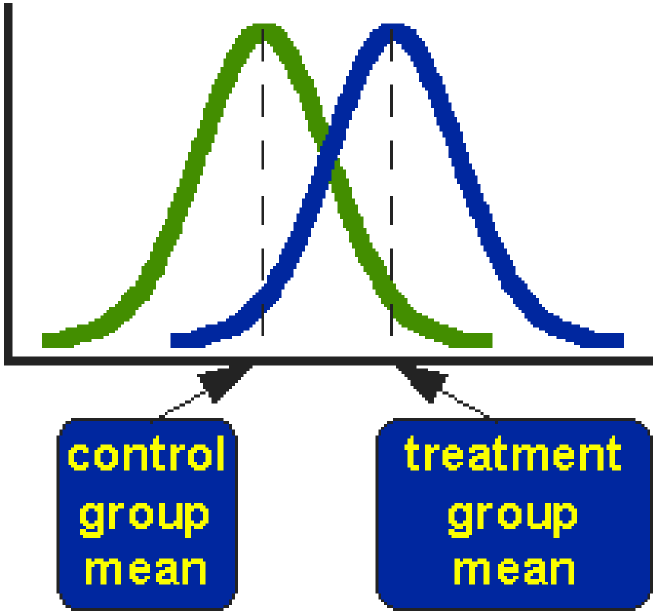
Using Table 1.11 for six data points, $Q_{\mathrm{table}} = 0.62$.

Since Q_{exp} is smaller than Q_{table} the point should not be rejected as there is a more than 5% chance that it is part of the same data set as the other five values. It is easy to show that an additional data point of 3.3 rather than 3.0 mM would give a Q_{exp} of 0.64 and could be rejected.

‘Student’s‘ t Test— Test of Difference (檢驗差異是否具有統計意義)

The ***t*-test** compares the actual difference between two means relative to the variation in the data —

sample mean v.s.true mean

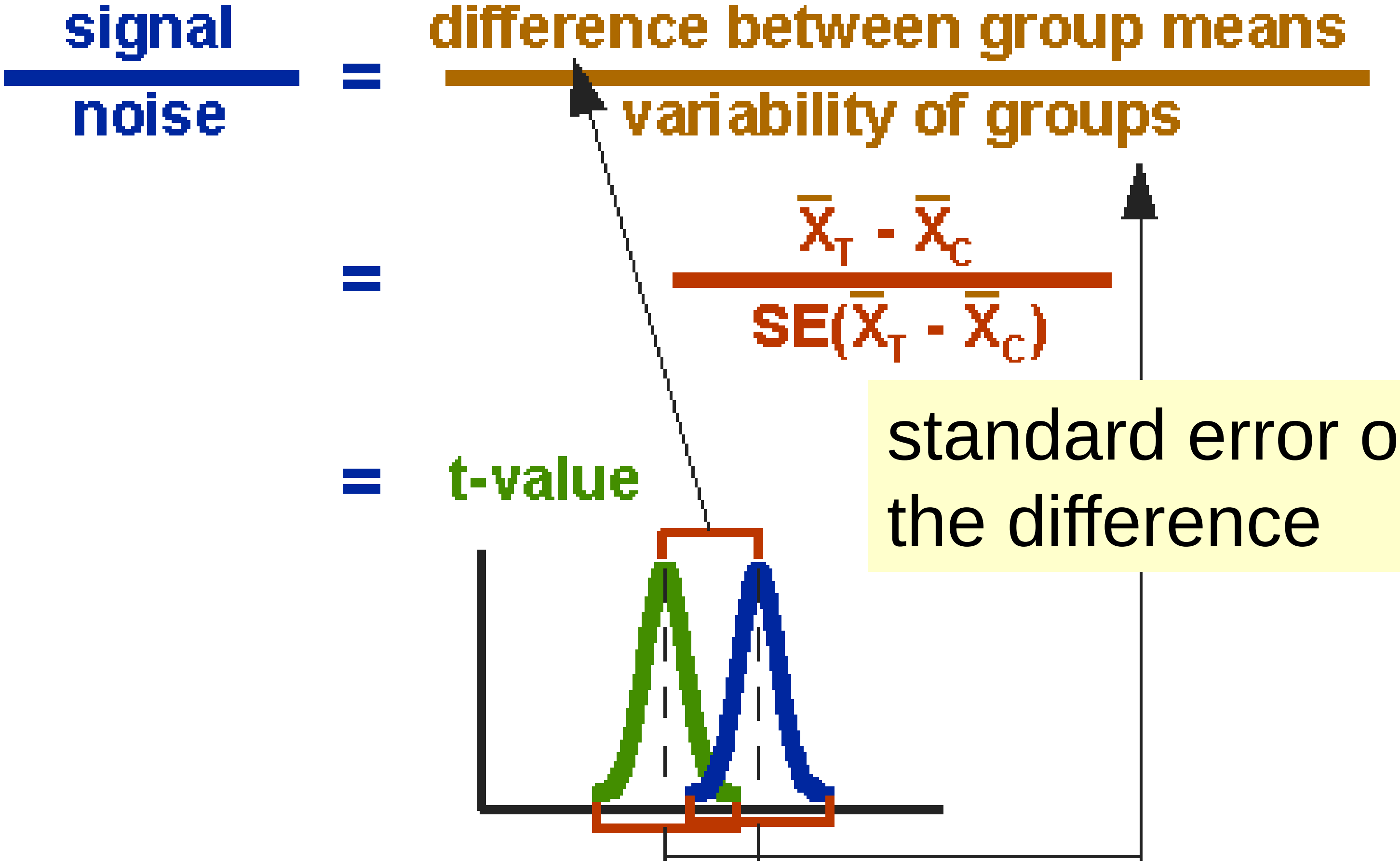


Determine whether a significant difference exist between two mean or whether the two population means are equal.

http://www.socialresearchmethods.net/kb/stat_t.htm

t value:

calculated by integrating the distribution between confident limits



$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

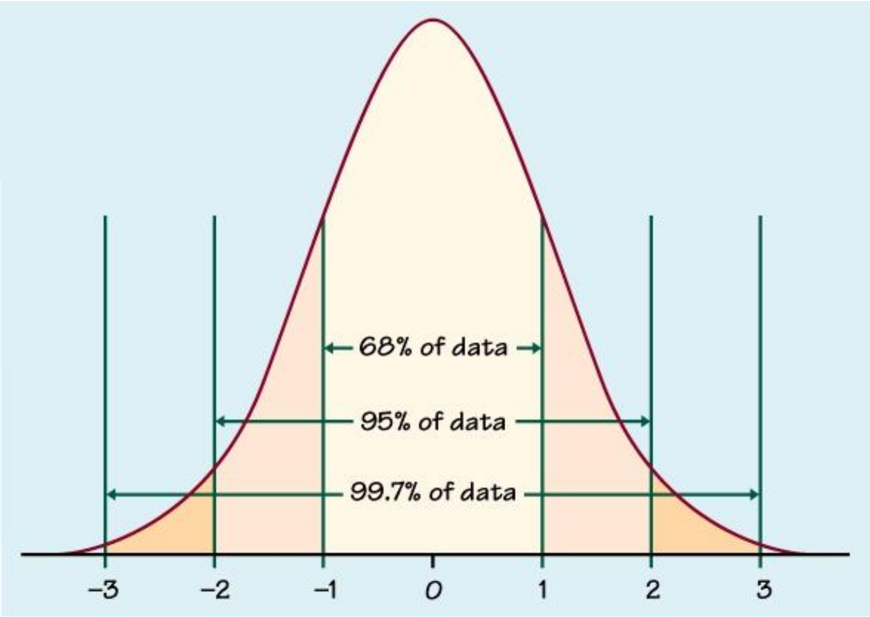
$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

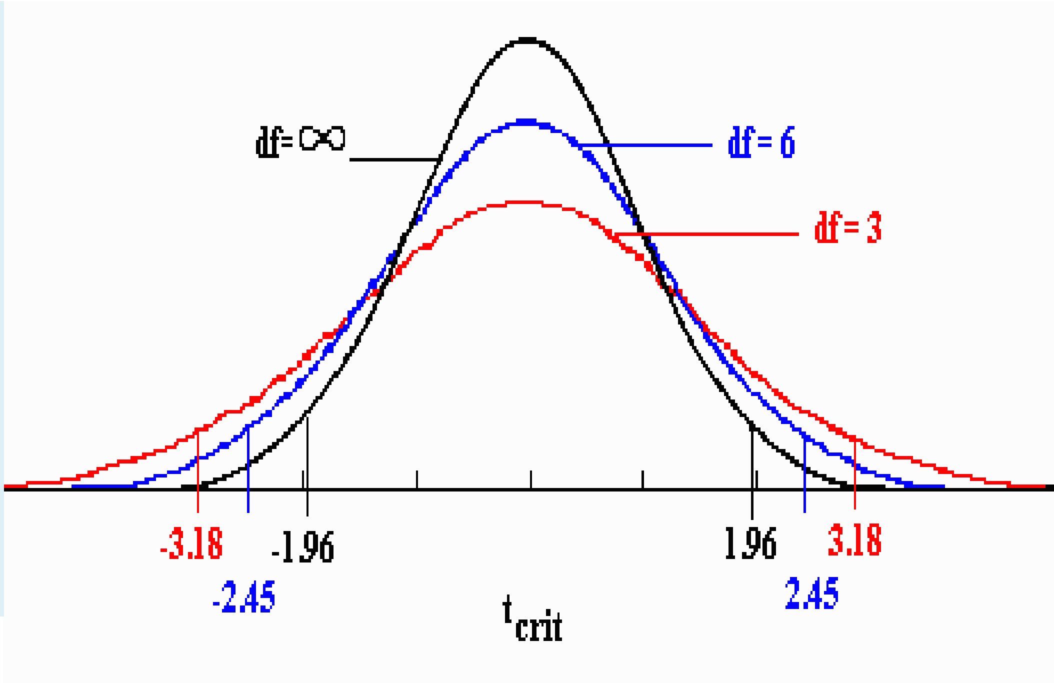
$$= \text{t-value}$$

standard error of the difference

The t-distribution

- In fact we have many t-distributions, each one is calculated in reference to the number of degrees of freedom (*d.f.*) also know as variables (*v*)





Normal distribution

t-distribution

Student’s t Values (Table 1.9)

Table 1.9 **Values for Student's t**
MBM, p38

	Confidence Level (%)					
Degree of Freedom	50	90	95	98	99	99.9
N-1						
2	0.816	2.92	4.303	6.965	9.925	
3	0.765	2.353	3.182	4.541	5.841	
4	0.741	2.132	2.776	3.747	4.604	
5	0.727	2.015	2.571	3.365	4.032	
6	0.718	1.943	2.447	3.143	3.707	

One-Sample t-test

- Compare the result with the **known value** of the solution

often test whether the mean of a variable is less than, greater than, or equal to a specific value.

$$M = \bar{X} \pm \frac{\bar{t}S}{\sqrt{n}} \quad \Rightarrow \quad t_{calc} = \frac{(M - \bar{X})\sqrt{n}}{S}$$

Known value

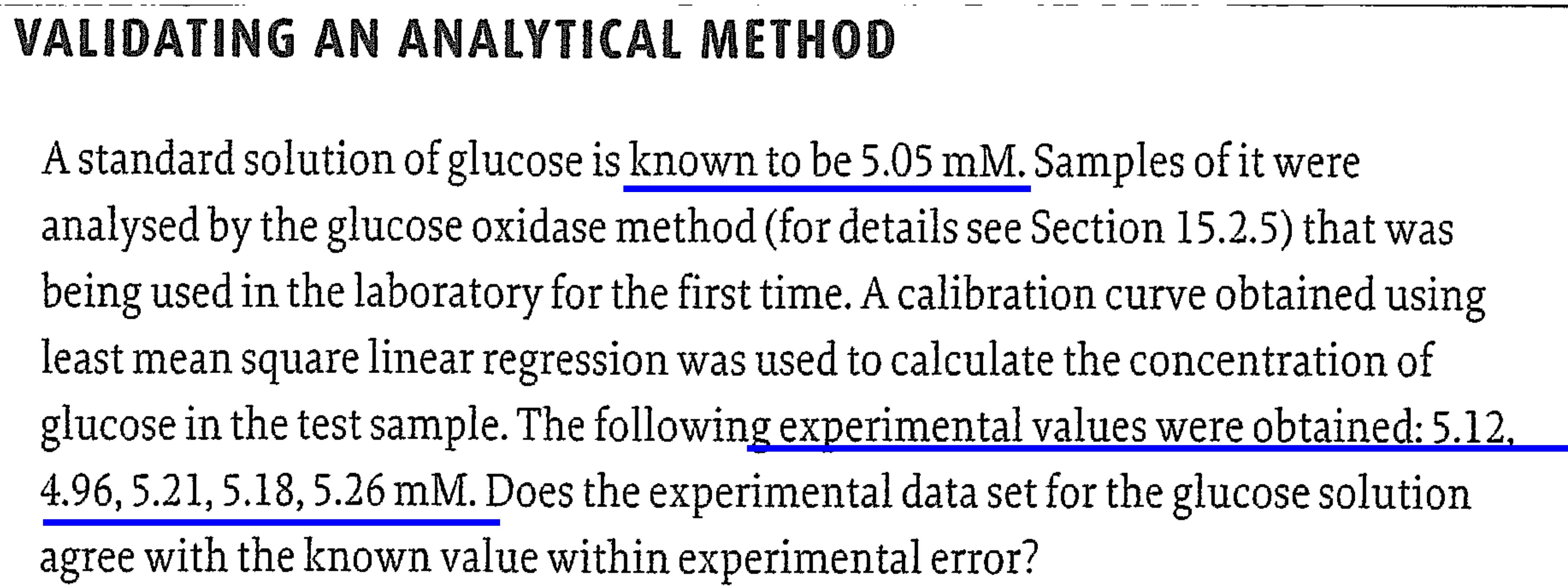
When $t_{calc} \rangle t_{table(1.9)}$ →if S.E is not in the range of $\bar{X} \pm XX$

The result is not within the range of population

$$t_{calc} < t_{table(1.9)}$$

The result is consistent with the population

Example 8



VALIDATING AN ANALYTICAL METHOD

A standard solution of glucose is known to be 5.05 mM. Samples of it were analysed by the glucose oxidase method (for details see Section 15.2.5) that was being used in the laboratory for the first time. A calibration curve obtained using least mean square linear regression was used to calculate the concentration of glucose in the test sample. The following experimental values were obtained: 5.12, 4.96, 5.21, 5.18, 5.26 mM. Does the experimental data set for the glucose solution agree with the known value within experimental error?

It is first necessary to calculate the mean and standard deviation for the set and then to use it to calculate a value for Student's *t*.

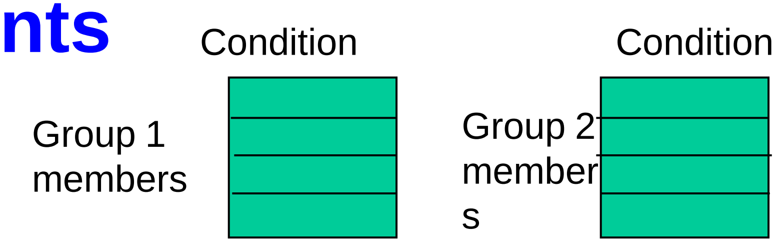
Applying equations 1.12 and 1.13 to the data set gives $\bar{x} = 5.15$ mM and $s = \pm 0.1$ mM.

Now applying equation 1.17 to give t_{calc}:

$$t_{\text{calc}} = \frac{(5.05 - 5.15)\sqrt{5}}{0.1} = 2.236$$

Note that the negative difference between the two mean values in this calculation is ignored. From Table 1.9 at the 95% confidence level with 4 degrees of freedom, $t_{\text{table}} = 2.776$. t_{calc} is therefore less than t_{table} and the conclusion can be drawn that measured mean value does agree with the known value. Using equation 1.14, the coefficient of variation for the measured values can be calculated to be 1.96%.

Unpaired Statistical Experiments



- **Overall setting:** 2 groups of 4 individuals each
  - Group1: TIGP students
  - Group2: NTU students
- **Experiment 1:**
  - We measure the height of all students
  - We want to establish if members of one group are consistently (or on average) taller than members of the other, and if the measured difference is significant
- **Experiment 2:**
  - We measure the weight of all students
  - We want to establish if members of one group are consistently (or on average) heavier than the other, and if the measured difference is significant
- **Experiment 3:**
  - ………

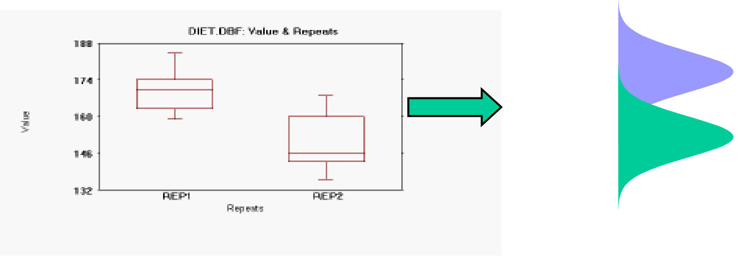
Unpaired Statistical Experiments

- In unpaired experiments, you typically have two groups of people that are not related to one another, and measure some property for each member of each group

- e.g. you want to test whether a new drug is effective or not, you divide similar patients in two groups:
  - One groups takes the drug
  - Another groups takes a placebo
  - You measure (quantify) effect of both groups some time later

- You want to establish whether there is a significant difference between both groups at that later point

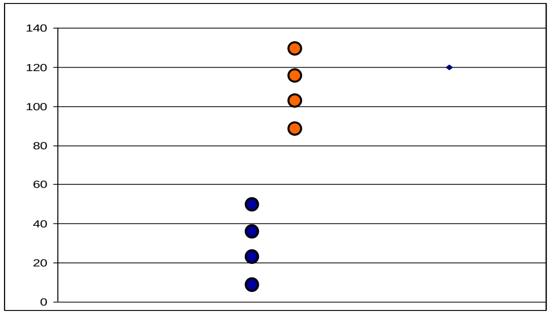
Unpaired Statistical Experiments



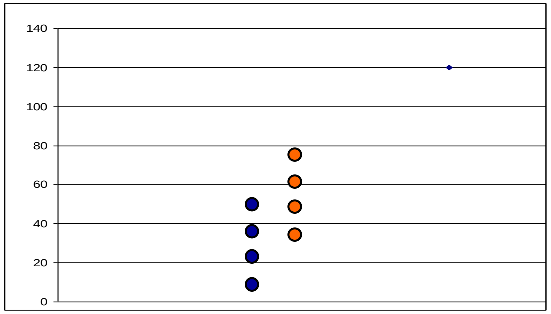
1. How do we address the problem?
2. Compare two sets of results (alternatively calculate mean for each group and compare means)

1. Graphically:
    1. Scatter Plots
    2. Box plots, etc

2. Compare Statistically
    1. Use unpaired t-test



Are these two series significantly different?



Are these two series significantly different?

Unpaired t-test：

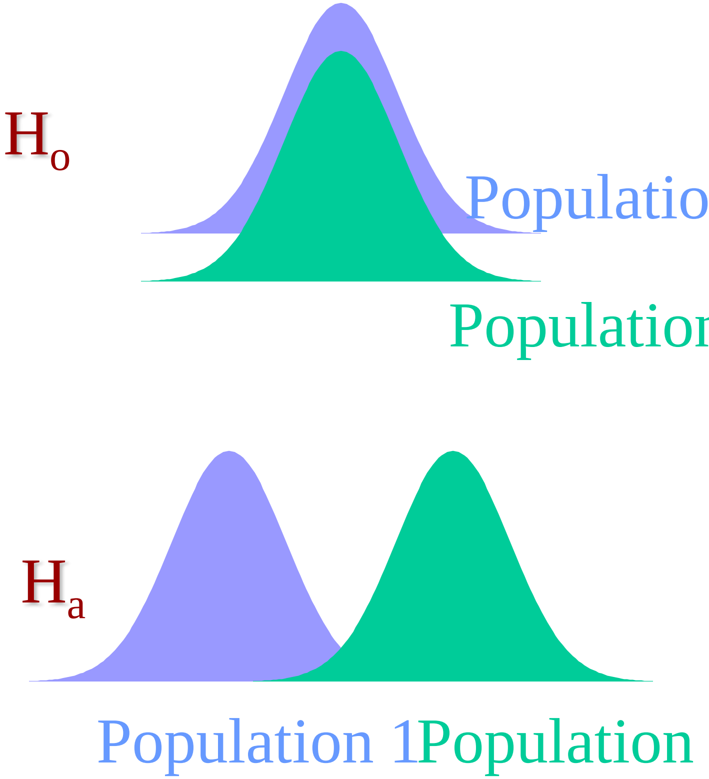
Are two data sets different ??

- applied to two independent groups

  e.g. diabetic patients versus non-diabetics

- sample size from the two groups may or may not be equal

- in addition to the assumption that the data is from a normal distribution, there is also the assumption that the **standard deviation (SD)s** is approximately the same in both



首先要比較2方法的標準差 (Similar ? Different?)

F-test $$F = \frac{S_1^{\ 2}}{S_2^{\ 2}} \approx \frac{\text{largest variation}}{\text{smallest variation}}$$

例：Group A mean：50 mg/l ， n=5 ， S=2.0mg/l

B mean：45 mg/l ， n=6 ， S=1.5mg/l

$F= 2^2 / 1.5^2$ =1.78

Degree of freedom at 95%:

(n-1)+(n2-1)=(5-1)+(6-1)=9

Table1.1

$$F_{table} = 7.39 \rangle F_{calc} = 1.78$$

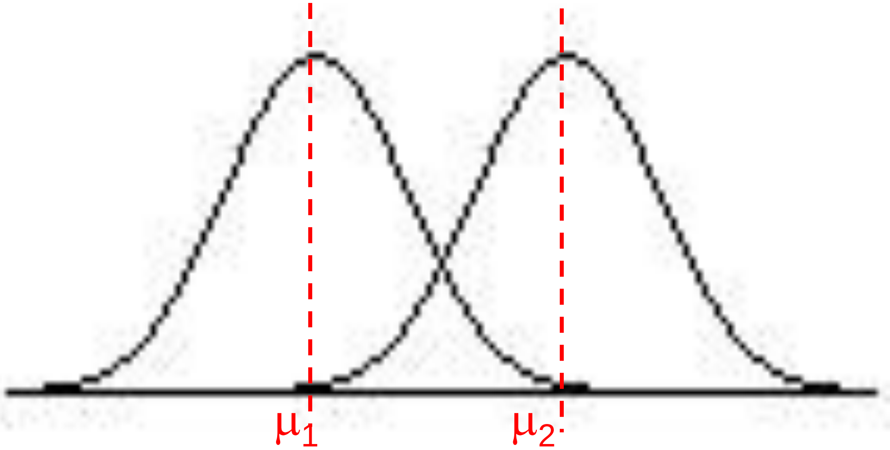
→ the variance values are the same

→ **the mean really differs**

Similar S (With Equal Variances)

- equation 1.18、1.19

$$t_{\mathrm{calc}} = \frac{\bar{x}_1 - \bar{x}_2}{S_{\mathrm{pooled}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$



S pooled

an estimator of the common standard deviation of the two samples:

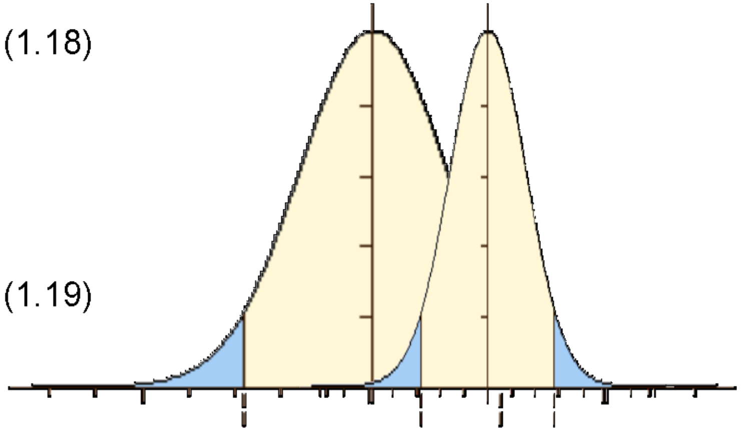
$$S_{\mathrm{pooled}} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

$$\text{degree of freedom} = n_1 + n_2 - 2$$

Different S (With Unequal Variances)

This test is used only when the two sample sizes are unequal and the variance is assumed to be different.

- equation 1.20、1.21

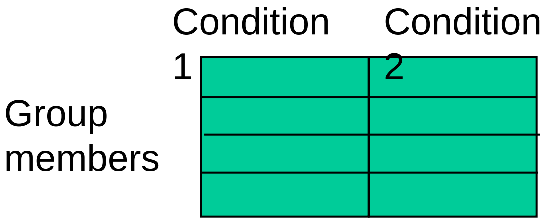


$$t_{\text{calc}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}}$$

$$\text{degree of freedom} = \left\{ \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{[(s_1^2 / n_1)^2 / (n_1 + 1)] + [(s_2^2 / n_2)^2 / (n_2 + 1)]} \right\} - 2$$

或 $$\frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{[(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)]}$$

Paired statistical experiments



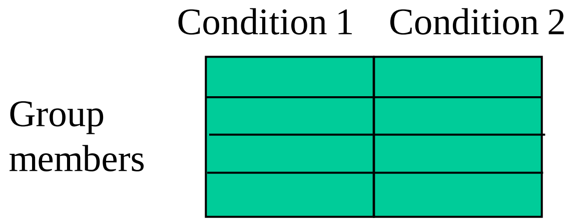
Condition 1

Condition 2

Group members

- **Overall setting:** 1 groups of 4 individuals each
  - Group1: TIGP students
  - We make measurements for each student in two situations
- **Experiment 1:**
  - We measure the height of all students before Bioinformatics course and after Bioinformatics course
  - We want to establish if Bioinformatics course consistently (or on average) affects students’ heights
- **Experiment 2:**
  - We measure the weight of all students before Bioinformatics course and after Bioinformatics
  - We want to establish if Bioinformatics course consistently (or on average) affects students’ weights
- **Experiment 3:**
  - ………

Paired statistical experiments



- In paired experiments, you typically have one group of people, you typically measure some property for each member before and after a particular event (so measurement come in pairs of before and after)

- e.g. you want to test the effectiveness of a new cream for tanning
  - You measure the tan in each individual before the cream is applied
  - You measure the tan in each individual after the cream is applied

- You want to establish whether the there is a significant difference between measurements before and after applying the cream for the group as a whole

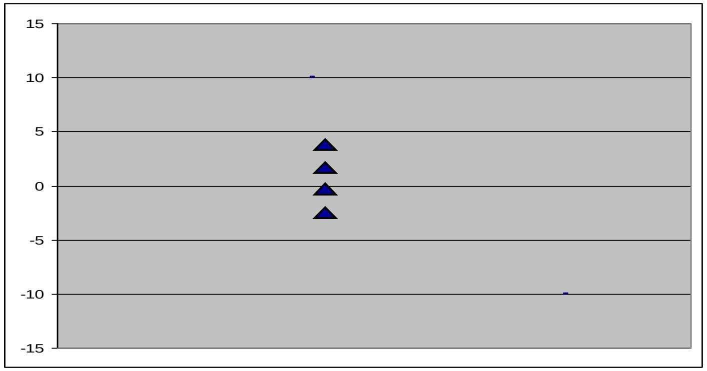
Paired statistical experiments

- The WT/KO example is a paired experiment if the rats in the experiments are the same!

Experiments for Gene 96608_at		
Rat #	WT gene expression	KO gene expression
Rat1	100	200
Rat2	100	300
Rat3	200	400
Rat4	300	500

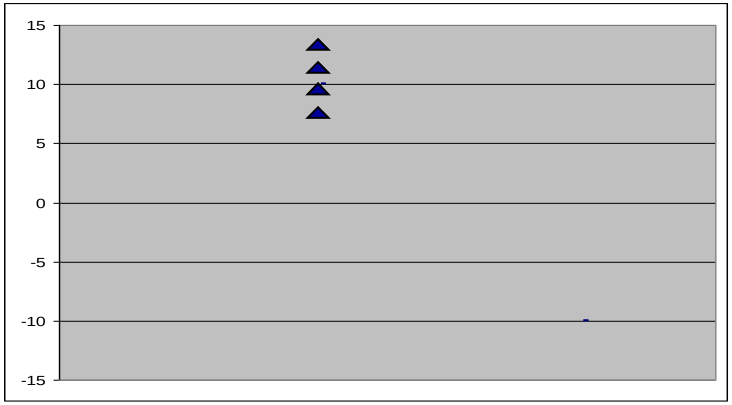
Paired statistical experiments

1. How do we address the problem?
2. Calculate difference for each pair
3. Compare differences to zero
4. Alternatively (compare average difference to zero)



5. Graphically:
   1. Scatter Plot of difference
   2. Box plots, etc
6. Statistically
   1. Use unpaired t-test

Are differences close to Zero?



Paired t-test

- Data is derived from study subjects who have been measured at two time points (so each individual has two measurements). The two measurements generally are before and after a treatment intervention

  Eg: control versus treated sample

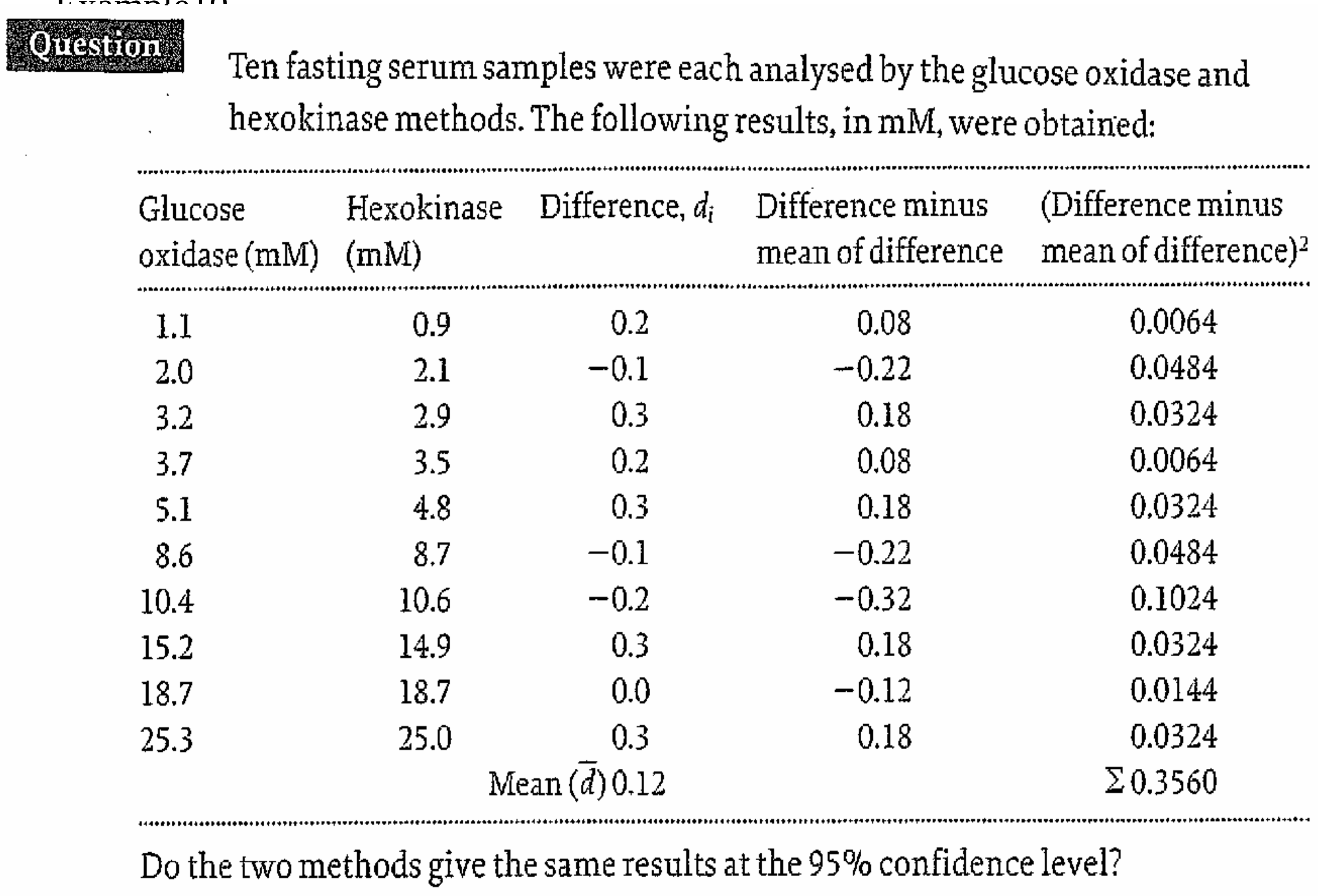
- 95% confidence interval is derived from the difference between the two sets of paired observations

equation 1.22、1.23

$$t_{\text{calc}} = \frac{\bar{d}}{s_{\text{d}}}\sqrt{n} \qquad s_{\text{d}} = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

COMPARISON OF TWO ANALYTICAL METHODS USING DIFFERENT TEST SAMPLES

Example10



Question

Ten fasting serum samples were each analysed by the glucose oxidase and hexokinase methods. The following results, in mM, were obtained:

Glucose oxidase (mM)	Hexokinase (mM)	Difference, d_i	Difference minus mean of difference	(Difference minus mean of difference)^2
1.1	0.9	0.2	0.08	0.0064
2.0	2.1	−0.1	−0.22	0.0484
3.2	2.9	0.3	0.18	0.0324
3.7	3.5	0.2	0.08	0.0064
5.1	4.8	0.3	0.18	0.0324
8.6	8.7	−0.1	−0.22	0.0484
10.4	10.6	−0.2	−0.32	0.1024
15.2	14.9	0.3	0.18	0.0324
18.7	18.7	0.0	−0.12	0.0144
25.3	25.0	0.3	0.18	0.0324
		Mean ($\overline{d}$) 0.12		Σ 0.3560

Do the two methods give the same results at the 95% confidence level?

answer

Before addressing the main question, note that the 10 samples analysed by the two methods were chosen to cover the whole analytical range for the methods. To assess whether or not the two methods have given the same result at the chosen confidence level, it is necessary to calculate a value for t_{calc} and to compare it with t_{table} for the 9 degrees of freedom in the study. To calculate t_{calc}, it is first necessary to calculate the value of s_d in equation 1.23. The appropriate calculations are shown in the table above.

$$s_d = \sqrt{[\Sigma(d_i - \bar{d})^2]/(n-1)}$$

$$= \sqrt{(0.356/9)}$$

$$= 0.199$$

From equation 1.22

$$t_{\text{calc}} = \frac{\bar{d}\sqrt{n}}{s_d}$$

$$= (0.12\sqrt{10})/0.199$$

$$= 1.907$$

Using Table 1.9, t_{table} at the 95% confidence level and for 9 degrees of freedom is 2.262. Since t_{calc} is smaller than t_{table} the two methods do give the same results at the 95% confidence level. Inspection of the two data sets shows that the glucose oxidase method gave a slightly high value for 7 of the 10 samples analysed.

An alternative approach to the comparison of the two methods is to plot the two data sets as an x/y plot and to carry out a regression analysis of the data. If this is done using the glucose oxidase data as the y variable, the following results are obtained:

Slope: 1.0016, intercept: 0.1057, correlation coefficient r: 0.9997

The slope of very nearly 1 confirms the similarity of the two data sets, whilst the small positive intercept on the y-axis confirms that the glucose oxidase method gives a slightly higher, but insignificantly different, value from that of the hexokinase method.

$$t_{\text{calc}} = \frac{(\text{known value} - \bar{x})\sqrt{n}}{s} \qquad (1.17)$$

$$t_{\text{calc}} = \frac{\bar{x}_1 - \bar{x}_2}{S_{\text{pooled}}}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad (1.18)$$

$$S_{\text{pooled}} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \qquad (1.19)$$

$$t_{\text{calc}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}} \qquad (1.20)$$

$$\text{Degree of freedom} = \left\{ \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{[(s_1^2 / n_1)^2 / (n_1 + 1)] + [(s_2^2 / n_2)^2 / (n_2 + 1)]} \right\} - 2 \qquad (1.21)$$

$$t_{\text{calc}} = \frac{\bar{d}}{s_{\text{d}}}\sqrt{n} \qquad (1.22)$$

$$s_{\text{d}} = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} \qquad (1.23)$$