Technical Assistance
HYDROLOGY PROJECT

World Bank & Government of The Netherlands funded

Training module # WQ - 48

Applied Statistics

New Delhi, September 2000

CSMRS Building, 4th Floor, Olof Palme Marg, Hauz Khas,
New Delhi – 11 00 16 India
Tel: 68 61 681 / 84 Fax: (+ 91 11) 68 61 685
E-Mail: dhvdelft@del2.vsnl.net.in

DHV Consultants BV & DELFT HYDRAULICS

with
HALCROW, TAHAL, CES, ORG & JPS

Table of contents

1. Module context

This module discusses statistical procedures, which are commonly used by chemists to evaluate the precision and accuracy of results of analyses. Prior training in 'Basic Statistic', module no. 47, or equivalent is necessary to complete this module satisfactorily.

While designing a training course, the relationship between this module and the others, would be maintained by keeping them close together in the syllabus and place them in a logical sequence. The actual selection of the topics and the depth of training would, of course, depend on the training needs of the participants, i.e. their knowledge level and skills performance upon the start of the course.

2. Module profile

Title	:	Applied Statistics
Target group	:	HIS function(s): Q2, Q3, Q5, Q6
Duration	:	one session of 90 min
Objectives	:	After the training the participants will be able to*:* • Apply common statistical tests for evaluation of the precision of data
Key concepts	:	• confidence interval • regression analyses
Training methods	:	Lecture, exercises, OHS
Training tools required	:	Board, flipchart
Handouts	:	As provided in this module
Further reading and references	:	• 'Analytical Chemistry', Douglas A. Skoog and Donald M. West, Saunders College Publishing, 1986 (Chapter 3) • 'Statistical Procedures for analysis of Environmenrtal monitoring Data and Risk Assessment', Edward A. Mc Bean and Frank A. Rovers, Prentice Hall, 1998.

3. Session plan

No	Activities	Time	Tools
1	***Preparations***		
2	***Introduction***: • Need for statistical procedures • Scope of the module	10 min	OHS
3	***Confidence interval*** • Definition • Significance of standard deviation and its correct estimation • When population SD is known • When population SD is not known	30 min	OHS
4	***Rejection of Data*** • Caution in rejection • Procedures	20 min	OHS
5	***Regression analysis***	20 min	OHS
6	***Wrap up***	10 min	

4. Overhead/flipchart master

OHS format guidelines

Type of text	Style	Setting
Headings:	OHS-Title	Arial 30-36, with bottom border line (not: underline)
Text:	OHS-lev1 OHS-lev2	Arial 24-26, maximum two levels
Case:		Sentence case. Avoid full text in UPPERCASE.
Italics:		Use occasionally and in a consistent way
Listings:	OHS-lev1 OHS-lev1-Numbered	Big bullets. Numbers for definite series of steps. Avoid roman numbers and letters.
Colours:		None, as these get lost in photocopying and some colours do not reproduce at all.
Formulas/Equations	OHS-Equation	Use of a table will ease horizontal alignment over more lines (columns) Use equation editor for advanced formatting only

Applied Statistics

- Confidence intervals
- Number of replicate observations
- Rejection of data
- Regression analysis

Confidence Intervals (1)

- Population mean μ is always unknown
- Sample mean $\bar{x}$
  - *Only for large number of replicate* $\bar{x} \longrightarrow \mu$
- Estimation of limits for μ from $\bar{x}$

  $(\bar{x} - \mathbf{a}) < \mu < (\bar{x} + \mathbf{a})$
- Value of **a** depend
  - *confidence level (probability of occurrence)*
  - *standard deviation*
  - *difference between limits is the interval*

Confidence Intervals

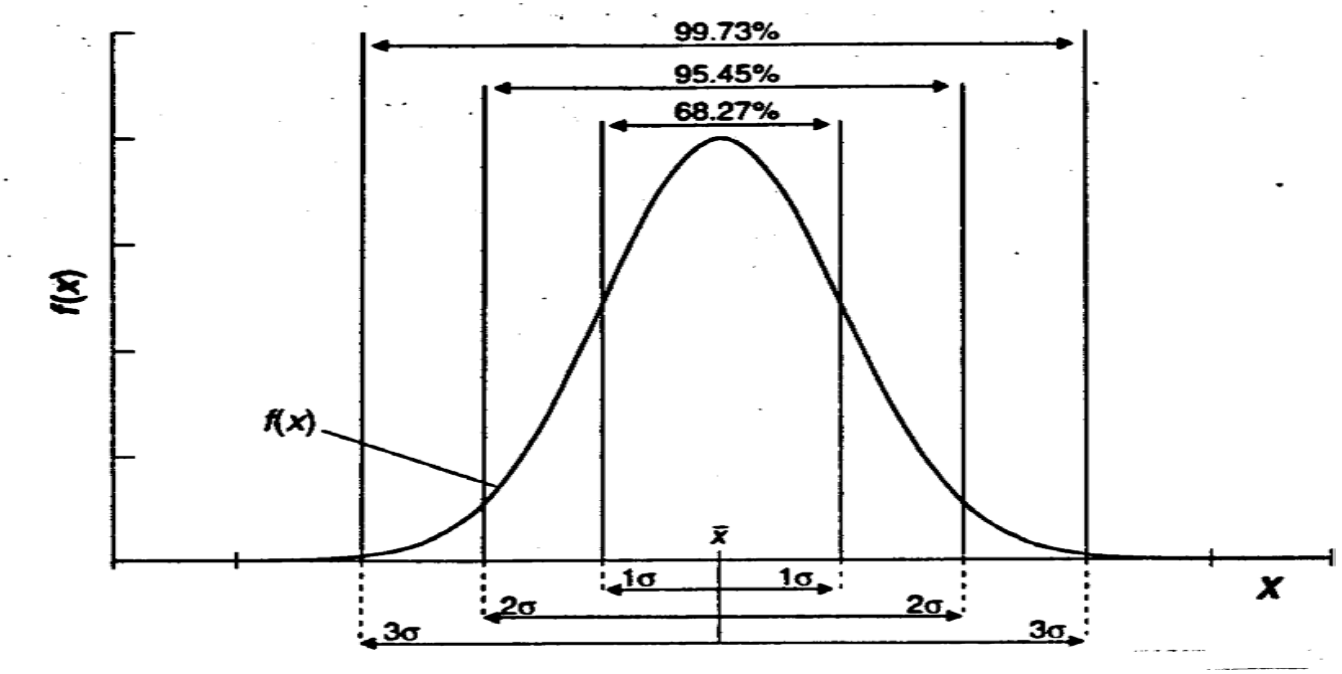


Figure 1. Normal distribution curve and areas under curve for different distances

Confidence Intervals (2)

- When s is good estimate of σ
  - *for single observation,* $(x - z\sigma) < \mu < (x - z\sigma)$

Confidence level	**50**	**90**	**95**	**99**	**99.7**
Z	**0.67**	**1.64**	**1.96**	**2.58**	**3**

- as z or interval increases confidence level also increases

Example (1)

- Mercury concentration in fish was determined as 1.8 μg/kg. Calculate limits for μ for 95% confidence level if σ = 0.1 μg/kg
  - *For 95% confidence level z = 1.96*
  - *Limits for mean*

$$1.8 - 1.96 \times 0.1 < \mu < 1.8 + 1.96 \times 0.1$$

$$1.6 < \mu < 2.0$$

- What are the limits for 99.7% confidence level?

Confidence Intervals (3)

- Confidence interval when mean, $\bar{x}$, of n replicates is available

$$\bar{x} - z\sigma/\sqrt{n} < \mu < \bar{x} + z\sigma/\sqrt{n}$$

- Confidence interval is reduced compared to single measurement as n increases

n	1	2	3	4	9	16
$\sqrt{n}$	1	1.4	1.7	2	3	4

- Analyse 2 to 4 replicates, more replicates give diminishing return

Example (2)

- Average mercury conc. from 3 replicates was 1.67 μg/ kg
- Calculate limits for 95% confidence level, σ = 0.1 μg/ kg

$1.67 - 1.96 \times 0.1/\sqrt{3} < \mu < 1.67 + 1.96 \times 0.1/\sqrt{3}$

$1.56 < \mu < 1.78$

Example (3)

- How Many replicate analyses will be required to decrease the 95% confidence interval to $\pm$ 0.07, σ = 0.1

  Interval = $\pm z\sigma/\sqrt{n}$

  $\pm$ 0.07 = $\pm$ 1.96 x 0.1/$\sqrt{n}$

  n = 7.8

- 8 measurements will give slightly better than 95% chance for μ to be with $\bar{x} \pm 0.07$

Confidence Interval (4)

- When σ is unknown
- Sample s based on n measurements is available

$$\bar{x} - t\,s/\sqrt{n} < \mu < \bar{x} + t\,s/\sqrt{n}$$

Degrees of Freedom	1	4	10	α
T_{95}	12.7	2.78	2.23	1.96
T_{99}	63.7	4.60	3.17	2.58

- Note $T \longrightarrow z$ as degrees of freedom increases to $\propto$

Example (4)

- Groundwater samples were analysed for TOC, n = 5, $\bar{x}$ = 11.7 mg/L s = 3.2 mg/L
- Find the 95% confidence limit for the true mean
  - *Degrees of freedom = n – 1 = 5 – 1 = 4*
  - T_{95} *= 2.78 (from table)*

$$11.7 - 2.78 \times 3.2 / \sqrt{5} < \mu < 11.7 + 2.78 \times 3.2 / \sqrt{5}$$

$$7 < \mu < 15.74$$

Detection of Outliers

- Extreme values, very high or low, will influence calculated statistics
- Rejection
  - *sound basis*
  - *may be due to an unrecorded event*
  - *errors in transcription, reading of instruments, inconsistent methodology, etc.*
  - *otherwise apply statistical yardsticks*

Q Test

- $Q_{exp} = \dfrac{\text{Difference between suspect and neighbour}}{\text{Spread of entire test}}$

- Reject if $Q_{exp} = Q_{crit}$

Total No. of observations	3	5	7	10
Q_{crit} 90%	0.94	0.64	0.51	0.41
Q_{crit} 96%	0.98	0.73	0.59	0.48

- Other criteria are also available

Regression Analysis

- Quantification of relationships between variables
  - *used for predictions*
  - *calibration of instruments*
- Least-square method
  - *deviations due to random errors*

    $y = a + bx$
  - *linear relationship*

Calibration of GC

	Pesticide conc., μg/L (x_i)	Peak area, cm^2 (y_i)	x_i^2	y_i^2	$x_i y_i$
	0.352	1.09	0.12390	1.1881	0.3868
	0.803	1.78	0.64481	3.1684	1.42934
	1.08	2.60	1.16640	6.7600	2.80800
	1.38	3.03	1.90140	9.1809	4.18140
	1.75	4.01	3.06250	16.0801	7.01750
$\sum$	5.365	12.51	6.90201	36.3775	15.81992

$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 1.145365$

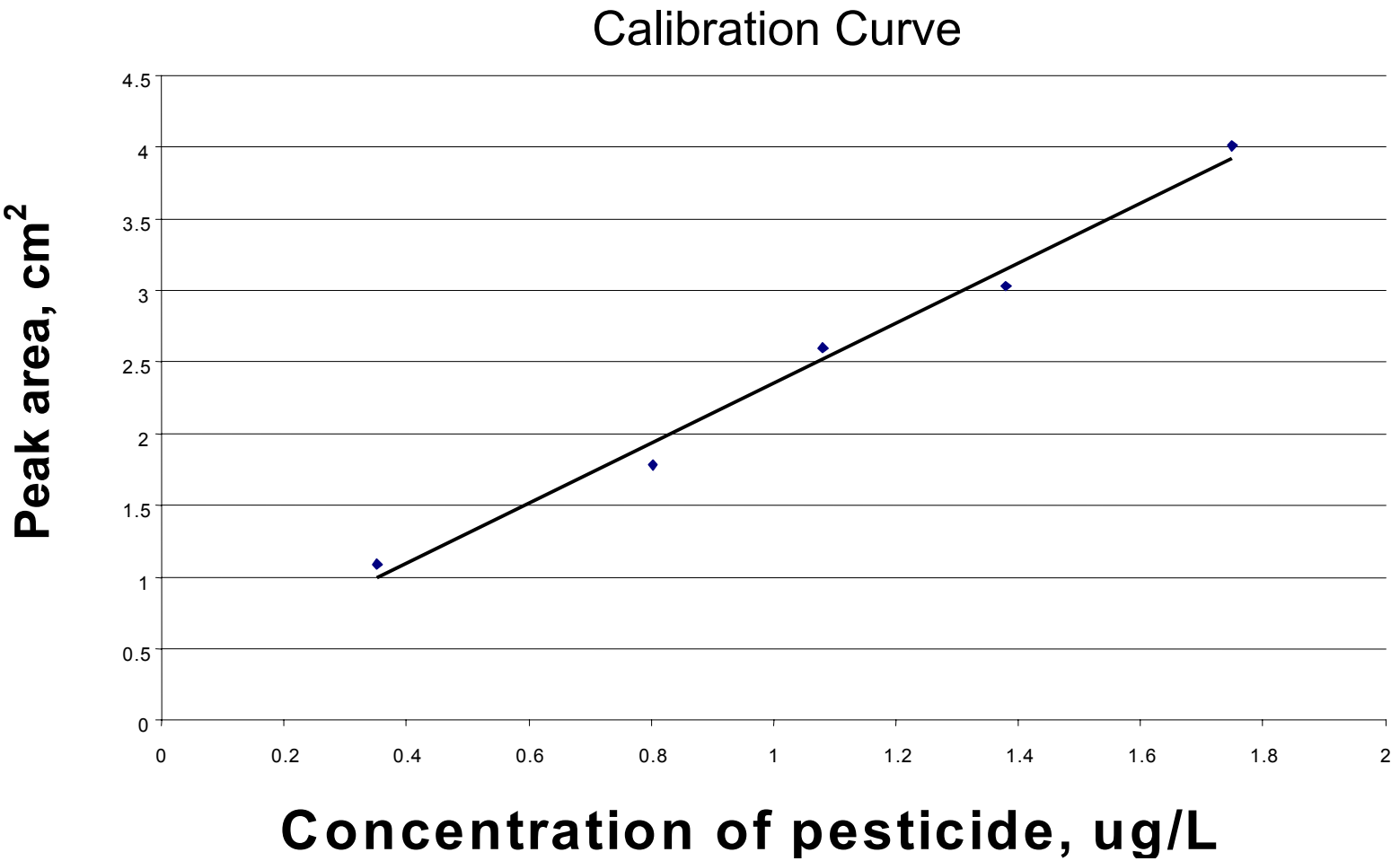
$S_{yy} = \sum y_i^2 - (\sum y_i)^2/n = 5.07748$

$S_{xy} = \sum x_i y_i - \sum x_i \sum y_i /n = 2.39669$

$b = S_{xy} / S_{xx} = 2.09$

$a = \bar{y} - b\ \bar{x} = 0.26$

Figure 2 - Calibration Curve



5. Evaluation sheets

6. Handout

Applied Statistics

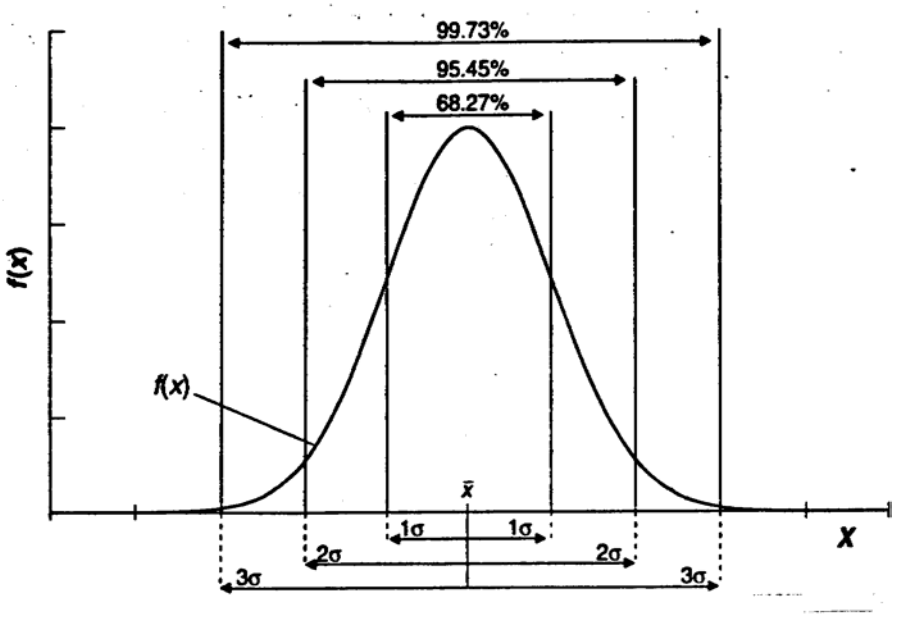
- Confidence intervals
- Number of replicate observations
- Rejection of data
- Regression analysis

Confidence Intervals (1)

- Population mean μ is always unknown
- Sample mean $\bar{x}$
  - *Only for large number of replicate* $\bar{x} \longrightarrow \mu$
- Estimation of limits for μ from $\bar{x}$
  $(\bar{x} - \mathbf{a}) < \mu < (\bar{x} + \mathbf{a})$
- Value of **a** depend
  - *confidence level (probability of occurrence)*
  - *standard deviation*
  - *difference between limits is the interval*

Confidence Intervals

Figure 1. Normal distribution curve and areas under curve for different distances



Confidence Intervals (2)

- When s is good estimate of σ
  - *for single observation, $(x - z\sigma) < \mu < (x - z\sigma)$*

Confidence level	**50**	**90**	**95**	**99**	**99.7**
Z	**0.67**	**1.64**	**1.96**	**2.58**	**3**

  - *as z or interval increases confidence level also increases*

Example (1)

- Mercury concentration in fish was determined as 1.8 μg/kg.
- Calculate limits for μ for 95% confidence level if σ = 0.1 μg/k
  - *For 95% confidence level z = 1.96*
  - *Limits for mean*

  $1.8 - 1.96 \times 0.1 < \mu < 1.8 + 1.96 \times 0.1$

  $1.6 < \mu < 2.0$
- What are the limits for 99.7% confidence level?

Confidence Intervals (3)

- Confidence interval when mean, $\bar{x}$, of n replicates is available

  $\bar{x} - z\sigma/\sqrt{n} < \mu < \bar{x} + z\sigma/\sqrt{n}$
- Confidence interval is reduced compared to single measurement as n increases

n	1	2	3	4	9	16
$\sqrt{n}$	1	1.4	1.7	2	3	4

- Analyse 2 to 4 replicates, more replicates give diminishing return

Example (2)

- Average mercury conc. from 3 replicates was 1.67 μg/ kg.
- Calculate limits for 95% confidence level, σ = 0.1 μg/ k

  $1.67 - 1.96 \times 0.1/\sqrt{3} < \mu < 1.67 + 1.96 \times 0.1/\sqrt{3}$

  $1.56 < \mu < 1.78$

Example (3)

- How Many replicate analyses will be required to decrease the 95% confidence interval to $\pm$ 0.07, σ = 0.1

  Interval = $\pm z\sigma/\sqrt{n}$

  $\pm$ 0.07 = $\pm$ 1.96 x 0.1/ $\sqrt{n}$

  n = 7.8

- 8 measurements will give slightly better than 95% chance for μ to be with $\bar{x} \pm$ 0.07

Confidence Intervals (4)

- When σ is unknown
- Sample s based on n measurements is available
  $\bar{x} - t\,s/\sqrt{n} < \mu < \bar{x} + t\,s/\sqrt{n}$

Degrees of Freedom	**1**	**4**	**10**	α
T_{95}	12.7	2.78	2.23	1.96
T_{99}	63.7	4.60	3.17	2.58

- Note T $\longrightarrow$ z as degrees of freedom increases to α

Example (4)

- Groundwater samples were analysed for TOC n = 5, $\bar{x}$ = 11.7 mg/L s = 3.2 mg/L
- Find the 95% confidence limit for the true mean
  - *Degrees of freedom = n – 1 = 5 – 1 = 4*
  - *T_{95} = 2.78*
    11.7 – 2.78 x 3.2 /$\sqrt{5}$ < μ < 11.7 + 2.78 x 3.2 /$\sqrt{5}$
    7 < μ < 15.74

Detection of Outliers

- Extreme values, very high or low, will influence calculated statistics

- Rejection
  - *sound basis*
  - *may be due to an unrecorded event*
  - *errors in transcription, reading of instruments, inconsistent methodology, etc.*
  - *otherwise apply statistical yardsticks*

Q Test

- $Q_{exp} = \dfrac{\text{Difference between suspect and neighbour}}{\text{Spread of entire test}}$

- Reject if $Q_{exp} = Q_{crit}$

Total No. of observations	3	5	7	10
Q_{crit} 90%	0.94	0.64	0.51	0.41
Q_{crit} 96%	0.98	0.73	0.59	0.48

- Other criteria are also available

Regression Analysis

- Quantification of relationships between variables
  - *used for predictions*
  - *calibration of instruments*

- Least-square method
  - *deviations due to random errors*
    y = a + bx
  - *linear relationship*

Calibration of GC

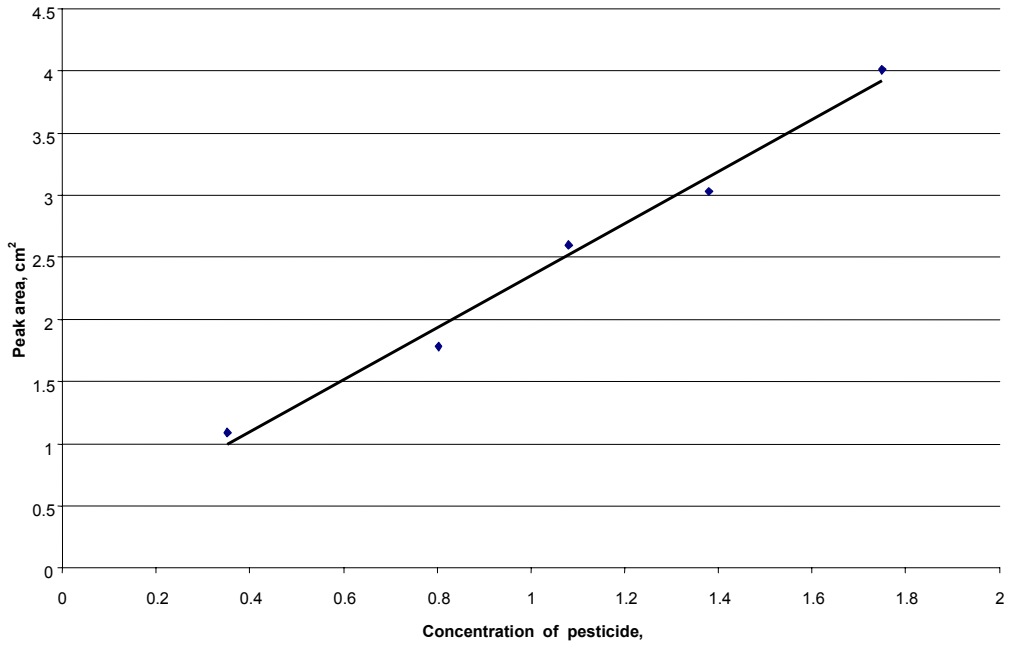
	Pesticide conc., μg/L (x_i)	Peak area, cm^2 (y_i)	x_i^2	y_i^2	$x_i y_i$
	0.352	1.09	0.12390	1.1881	0.3868
	0.803	1.78	0.64481	3.1684	1.42934
	1.08	2.60	1.16640	6.7600	2.80800
	1.38	3.03	1.90140	9.1809	4.18140
	1.75	4.01	3.06250	16.0801	7.01750
$\sum$	5.365	12.51	6.90201	36.3775	15.81992

$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n$ =1.145365
$S_{yy} = \sum y_i^2 - (\sum y_i)^2/n$ =5.07748
$S_{xy} = \sum x_i y_i - \sum x_i \sum y_i /n$ =2.39669

$b = S_{xy} / S_{xx} = 2.09$

$a = \bar{y} - b\ \bar{x} = 0.26$

Figure 2 – Calibration Curve



Add copy of Main text in chapter 8, for all participants.

7. Additional handout

These handouts are distributed during delivery and contain test questions, answers to questions, special worksheets, optional information, and other matters you would not like to be seen in the regular handouts.

It is a good practice to pre-punch these additional handouts, so the participants can easily insert them in the main handout folder.

8. Main text

Contents

Applied Statistics

For small data sets, there is always the question as to how well the data represent the population and what useful information can be inferred regarding the population characteristics. Statistical calculations are used for this purpose. This module deals with four such applications:

1. Definition of a population around the mean of a data set within which the true mean can be expected to be found with a particular degree of probability.
2. Determination of the number of measurements needed to obtain the true mean within a predetermined interval around the experimental mean with a particular degree of probability.
3. Determining if an outlying value in a set of replicate measurements should be retained in calculating the mean for the set.
4. The fitting of a straight line to a set of experimental points.

1. Confidence Intervals

The true mean of population, μ, is always unknown. However, limits can be set about the experimentally determined mean, $\bar{x}$, within which the true mean may be expected to occur with a given degree of probability. These bounds are called *confidence limits* and the interval between these limits, the *confidence interval*.

For a given set of data, if the confidence interval about the mean is large, the probability of an observation falling within the limits also becomes large. On the other hand, in the case when the limits are set close to the mean, the probability that the observed value falls within the limits also becomes smaller or, in other words, a larger fraction of observations are expected to fall out side the limits. Usually confidence intervals are calculated such that 95% of the observations are likely to fall within the limits.

For a given probability of occurrence, the confidence limits depend on the value of the standard deviation, s, of the observed data and the certainty with which this quantity can be taken to represent the true standard deviation, σ, of the population.

1.1 Confidence limits where σ is known

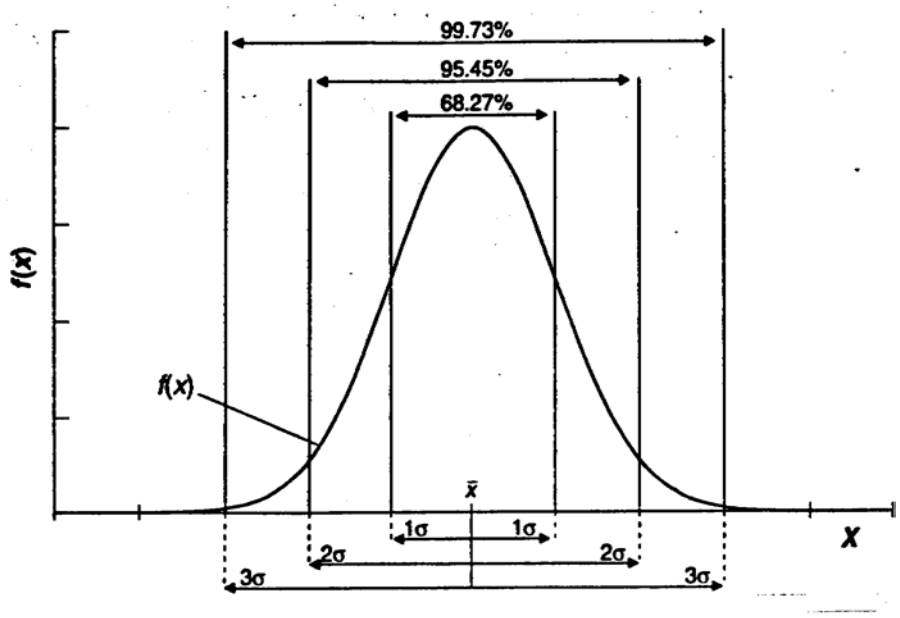
When the number of repetitive observations is 20 or more, the standard deviation of the observed set of data, s, can be taken to be a good approximation of the standard deviation of the population, σ.

However, it may not always be possible to perform such a large number of repetitive analyses, particularly when costly and time consuming extraction and analytical procedures are involved. In such cases, data from previous records or different laboratories may be pooled, provided that identical precautions and analytical steps are followed in each case. Further, care should be taken that the medium sampled is also similar, for example, analysis results of groundwater samples having high TDS should not be pooled with the results of surface waters, which usually have low TDS.

As discussed in the previous module, most of the randomly varying data can be approximated to a normal distribution curve. For normal distribution, 68% of the observations lie between the limits $\mu \pm \sigma$, 96% between the limits $\mu \pm 2\sigma$, 99.7% between the limits $\mu \pm$

3σ, etc., Figure 1. For a single observation, x_i, the confidence limits for the population mean are given by:

Figure 1. Normal distribution curve and areas under curve for different distances



$$\text{confidence limit for } \mu = x_i \pm z\sigma \qquad (1)$$

where z assumes different values depending upon the desired confidence level as given in Table 1.

Table 1: Values of z for various confidence levels

Confidence level, %	z
50	0.67
68	1.00
80	1.29
90	1.64
95	1.96
96	2.00
99	2.58
99.7	3.00
99.9	3.29

Example 1:
Mercury concentration in the sample of a fish was determined to be 1.80 μg/kg. Calculate the 50% and 95% confidence limits for this observation. Based on previous analysis records, it is known that the standard deviation of such observations, following similar analysis procedures, is 0.1 μg/kg and it closely represents the population standard deviation, σ.

From Table 1, it is seen that z = 0.67 and 1.96 for the two confidence limits in question. Upon substitution in Equation 1, we find that

50% confidence limit = $1.80 \pm 0.67 \times 0.1 = 1.8 \pm 0.07$ μg/kg
95% confidence limit = $1.80 \pm 1.96 \times 0.1 = 1.8 \pm 0.2$ μg/kg

Therefore if 100 replicate analyses are made, the results of 50 analysis will lie between the limits 1.73 and 1.87 and 95 results are expected to be within an enlarged limit of 1.6 and 2.0

Equation 1 applies to the result of a single measurement. In case a number of observations, n, is made and an average of the replicate samples is taken, the confidence interval decreases. In such a case the limits are given by:

$$\text{confidence limit for } \mu = \bar{x} \pm z\sigma/\sqrt{n} \qquad (2)$$

Example 2:
Calculate the confidence limits for the problem of Example 1, if three samples of the fish were analysed yielding an average of 1.67 μg/kg Hg.

Substitution in Equation 2 gives:

50% confidence limit = $1.67 \pm 0.67 \times 0.1/\sqrt{3} = 1.67 \pm 0.04$ μg/kg
95% confidence limit = $1.67 \pm 1.96 \times 0.1/\sqrt{3} = 1.67 \pm 0.11$ μg/kg

For the same odds the confidence limits are now substantially smaller and the result can be said to be more accurate and probably more useful.

Note that Equation 2 indicates that the confidence interval can be halved by increasing the number of analyses to 4 ($\sqrt{4} = 2$) compared to a single observation. Increasing the number of measurements beyond 4 does not decrease the confidence interval proportionately. To narrow the interval by one fourth, 16 measurements would be required ($\sqrt{16} = 4$), thus giving diminishing return. Consequently, 2 to 4 replicate measurements are made in most cases.

Equation 2 can also be used to find the number of replicate measurements required such that with a given probability the true mean would be found within a predetermined interval. This is illustrated in Example 3.

Example 3:
How many replicate measurements of the specimen in Example 1 would be needed to decrease the 95% confidence interval to ±0.07.

Substituting for confidence interval in Equation 2:

$0.07 = 1.96 \times 0.1/\sqrt{n}$
$\sqrt{n} = 1.96 \times 0.1/0.07, \quad n = 7.8$

Thus, 8 measurements will provide slightly better than 95% chance of he true mean lying within ±0.07 of the experimental mean.

1.2 Confidence limits where σ is unknown

When the number of individual measurements in a set of data are small the reproducibility of the calculated value of the standard deviation, s, is decreased. Therefore, for a given probability, the confidence interval must be larger under these circumstances.

To account for the potential variability in s, the confidence limits are calculated using the statistical parameter t:

$$\text{confidence limit for } \mu = x \pm ts/\sqrt{n} \qquad (3)$$

In contrast to z in Equation 2, t depends not only on the desired confidence level, but also upon the number of degrees of freedom available in the calculation of s. Table 2 provides values for t for various degrees of freedom and confidence levels. Note that the values of t become equal to those for z (Table 1) as the number of degrees of freedom becomes infinite.

Table 2: Values of t for various confidence levels and degrees of freedom

Degrees of Freedom	Confidence Level		
	90	**95**	**99**
1	6.31	12.7	63.7
2	2.92	4.30	9.92
3	2.35	3.18	5.84
4	2.13	2.78	4.60
5	2.02	2.57	4.03
6	1.94	2.45	3.71
8	1.86	2.31	3.36
10	1.81	2.23	3.17
12	1.78	2.18	3.11
12	1.78	2.16	3.06
14	1.76	2.14	2.98
∞	1.64	1.96	2.58

Example 4:
A chemist obtained the following data for the concentration of total organic carbon (TOC) in groundwater samples: 13.55, 6.39, 13.81, 11.20, 13.88 mg/L. Calculate the 95% confidence limits for the mean of the data.

Thus, from the given data, n = 5, $\bar{x}$ = 11.77 and s = 3.2. For degrees of freedom = 5 – 1 = 4 and 95% confidence limit, t from Table 2 = 2.78. Therefore, from Equation 3:

$$\text{confidence limit for } \mu = x \pm ts/\sqrt{n} = 11.77 \pm 2.78 \times 3.2/\sqrt{5}$$
$$= 11.77 \pm 3.97$$

2. Detection of Data Outliers

Data outliers are extreme (high or low) values that diverge widely from the main body of the data set. The presence of one or more outliers may greatly influence any calculated statistics and yield biased results. However, there is also the possibility that the outlier is a legitimate member of the data set. Outlier detection tests are to determine whether there is sufficient statistical evidence to conclude that an observation appears extreme and does not belong to the data set and should be rejected.

Data outliers may result from faulty instruments, error in transcription, misreading of instruments, inconsistent methodology of sampling and analysis and so on. These aspects should be investigated and if any of such reasons can be pegged to an outlier, the value may be safely deleted from consideration. However, this is to be kept in mind that the suspect data may rightfully belong to the set and may be the consequence of an unrecorded event, such as, a short rainfall, intrusion of sea water or a spill.

Table 3: Critical values for rejection quotient

Number of observations	Q_{crit} (reject if $Q_{exp} > Q_{crit}$)	
	90% confidence	96% confidence
3	0.94	0.98
4	0.76	0.85
5	0.64	0.73
6	0.56	0.64
7	0.51	0.59
8	0.47	0.54
9	0.44	0.51
10	0.41	0.48

Of the numerous statistical criteria available for detection of outliers, the Q test, which is commonly used, will be discussed here. To apply the Q test, the difference between the questionable result and its closest neighbour is divided by the spread of the entire set. The resulting ratio Q_{exp} is compared with the rejection values, Q_{crit} given in Table3, that are critical for a particular degree of confidence. If Q_{exp} is larger, a statistical basis for rejection exists. The table shows only some selected values. Any standard statistical analysis book may be consulted for a complete set.

Example 5:
Concentration measurements for fluoride in a well were measured as 2.77, 2.80, 2.90, 2.92, 3.45, 3.95, 4.44, 4.61, 5.21, 7.46. Use the Q test to examine whether the highest value is an outlier.

$Q_{exp} = (7.46 – 5.21) / (7.46 – 2.77) = 0.51$
Since 0.51 is larger than 0.48, the Q_{crit} value for 96% confidence, there is a basis for excluding the value.

Other criteria that can be used to evaluate an apparent outlier are:

- Plotting of a scatter diagram indicating inter-parameter relationship between two constituents. If there is a correlation between the two constituents, an outlier would lie a significant distance from the general trend.

- Applying test for normal distribution by calculating the mean and standard deviation of all the data and determining if the extreme value is outside the mean ± 3 times the standard deviation limit. If so, it is indeed an unusual value. (A cumulative normal probability plot may also be used for this test.)

- When the data set is not large, calculating and comparing the standard deviation, with and without the suspect observations. A suspect value which has a considerable influence on the calculated standard deviation suggests an outlier.

3. Regression Analysis

In assessing environmental quality , it is often of interest to quantify relationship between two or more variables. This may allow filling of missing data for one constituent and also may help in predicting future levels of a constituent.

Regression analysis is focused on determining the degree to which one or more variable(s) is dependant on an other variable, the independent variable. Thus regression is a means of calibrating coefficients of a predictive equation. In *correlation* analysis neither of the variables is identified as more important than the other. Correlation is not causation. It provides a measure of the goodness of fit.

In the calibration step, in most analytical procedures where a 'best' straight line is fitted to the observed response of the detector system when known amounts of analyte (standards) are analysed. This section discusses the regression analysis procedure employed for this purpose.

Least-squares method is the most straightforward regression procedure. Application of the method requires two assumptions; a linear relationship exists between the amount of analyte (x) and the magnitude of the measured response (y) and that any deviation of individual points from the straight line is entirely the consequence of indeterminate error, that is, no significant error exists in the composition of the standards.

The line generated is of the form:

$$y = a + bx \tag{4}$$

where a is the value of y when x is zero (the intercept) and b is the slope. The method minimises the squares of the vertical displacements of data points from the best fit line.

For convenience, the following three quantities are defined:

$$S_{xx} = \sum (x_i - x)^2 = \sum x_i^2 - (\sum x_i)^2/n$$
$$S_{yy} = \sum (y_i - y)^2 = \sum y_i^2 - (\sum y_i)^2/n$$
$$S_{xy} = \sum (x_i - x)\,(y_i - y) = \sum x_i y_i - \sum x_i \sum y_i /n$$

Calculating these quantities permits the determination of the following:

1. The slope of the line b:

$$b = S_{xy}/S_{xx} \tag{5}$$

2. The intercept a:

$$a = \bar{y} - b\,\bar{x} \tag{6}$$

3. The standard deviation about the regression line, s_r:

$$s_r = \sqrt{\{(S_{yy} - b^2 S_{xx})/(n-2)\}} \quad (7)$$

4. The standard deviation of the slope s_b:

$$s_b = \sqrt{(s_r^2/S_{xx})} \quad (8)$$

5. The standard deviation of the results based on the calibration curve s_c:

$$s_c = (s_r/b) \times \sqrt{\{(1/m) + (1/n) + (y_c - y)^2/b^2 S_{xx}\}} \quad (9)$$

where y_c is the mean of m replicate measurements made using the calibration curve.

Example 6:
The following table gives calibration data for chromatographic analysis of a pesticide and computations for fitting a straight line according to the least-square method.

	Pesticide conc., μg/L (x_i)	Peak area, cm^2 (y_i)	x_i^2	y_i^2	$x_i y_i$
	0.352	1.09	0.12390	1.1881	0.3868
	0.803	1.78	0.64481	3.1684	1.42934
	1.08	2.60	1.16640	6.7600	2.80800
	1.38	3.03	1.90140	9.1809	4.18140
	1.75	4.01	3.06250	16.0801	7.01750
∑	5.365	12.51	6.90201	36.3775	15.81992

Therefore

$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n$ =1.145365
$S_{yy} = \sum y_i^2 - (\sum y_i)^2/n$ =5.07748
$S_{xy} = \sum x_i y_i - \sum x_i \sum y_i /n$ =2.39669

and from Equations 5 & 6

b = 2.0925 = 2.09 and a = 0.2567 = 0.26

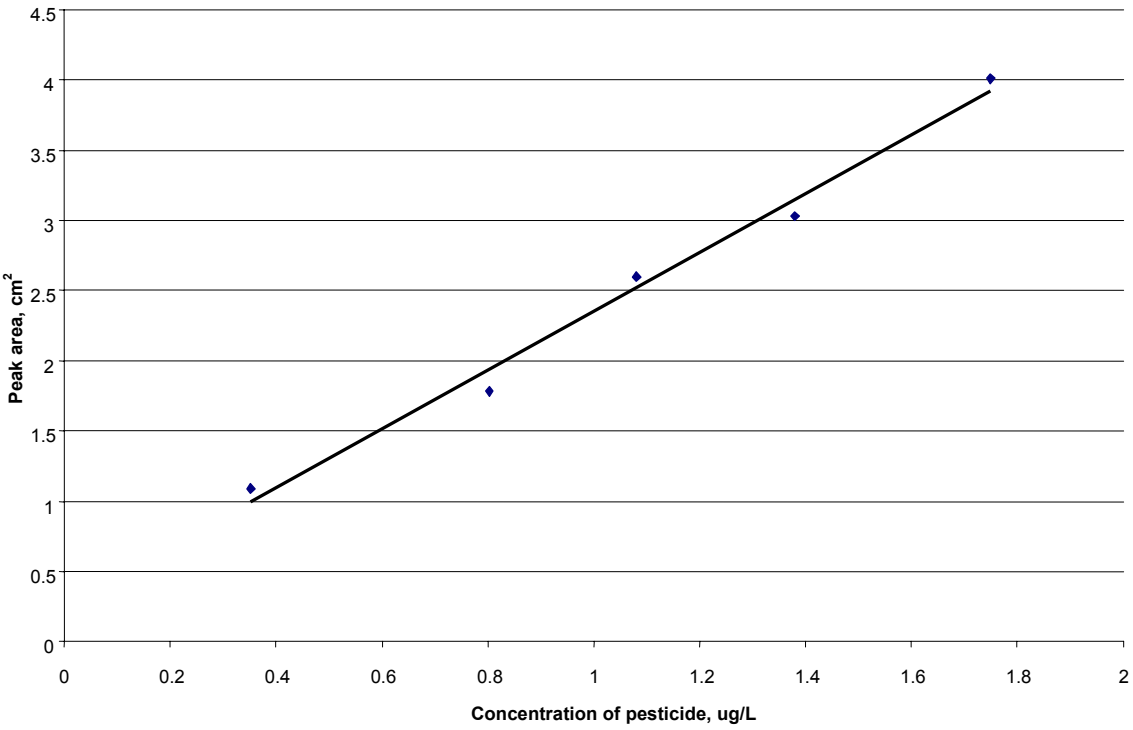
Thus the equation for the least-square line is

y = 0.26 + 2.09x

Note that rounding should not be performed until the end in order to avoid rounding errors.

The data points and the equation is plotted in Figure2.

Figure 2: Calibration Curve



Example 7:
Using the relationship derived in Example 6 calculate

1. concentration of pesticide in sample if the peak area of 2.65 was obtained.
2. the standard deviation of the result based on the single measurement of 2.65.
3. the standard deviation of the measurement if 2.65 represents the average of 4 replicates.

1. Substituting in the equation derived in the previous example

   2.65 = 0.26 + 2.09x
   x = 1.14 μg/L

2. Substituting in Equation 9 for m = 1

   $s_c = (0.144/2.09) \times \sqrt{\{(1/1) + (1/5) + (2.65 - 12.51/5)^2/2.09^2 \times 1.145\}} = \pm 0.08$

3. Substituting in Equation 9 for m = 4

   $s_c = (0.144/2.09) \times \sqrt{\{(1/4) + (1/5) + (2.65 - 12.51/5)^2/2.09^2 \times 1.145\}} = \pm 0.05$