- The **standard deviation of the mean**, σ_n

→ a measure of the uncertainty of the mean of n measurements.

$$\sigma_n = \frac{\sigma}{\sqrt{n}}$$

Sample number	Method 1 (μg/L)	Method 2 (μg/L)
1	17.2	14.2
2	23.1	27.9
3	28.5	21.2
4	15.3	15.9
5	23.1	32.1
6	32.5	22.0
7	39.5	37.0
8	38.7	41.5
9	52.5	42.6
10	42.6	42.8
11	52.7	41.1

. . . . .

- Uncertainty decreases

→ by a factor of 2 by making four times as many measurements

→ by a factor of 10 by making 100 times as many measurements.

4-2. Confidence Intervals

Calculating Confidence Intervals

- From a limited number of measurements,

  → we cannot find the true population mean, μ, or the true standard deviation, σ

  → what we can determine are $\bar{x}$ and s,

  the sample mean and the sample standard deviation.

- The **confidence interval** is an expression stating that

→ at some level of confidence, a range of values that include the true population mean.

- The confidence interval of μ is given by

Confidence interval: $$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

- where

→ s is the measured standard deviation,

→ n is the number of observations,

→ and t is Student's t, taken from Table 4-4.

- Student's t is a statistical tool used most frequently

  i) to find confidence intervals

  and ii) to compare mean values measured by different methods.

- The Student's t table is used to look up "t-values" according to degrees of freedom and confidence levels.

See Table 4-4

Example: Calculating Confidence Intervals

- The carbohydrate content of a glycoprotein (a protein with sugars attached to it) is determined to be 12.6, 11.9, 13.0, 12.7, and 12.5 g of carbohydrate per 100 g of protein in replicate analyses. Find the 50% and 90% confidence intervals for the carbohydrate content.

Solution First calculate $\bar{x}$ ($= 12.5_4$) and s ($= 0.4_0$) for the five measurements. For the 50% confidence interval, look up t in Table 4-2 under 50 and across from *four* degrees of freedom (degrees of freedom $= n - 1$.) The value of t is 0.741, so the 50% confidence interval is

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}} = 12.5_4 \pm \frac{(0.741)(0.4_0)}{\sqrt{5}} = 12.5_4 \pm 0.1_3$$

The 90% confidence interval is

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}} = 12.5_4 \pm \frac{(2.132)(0.4_0)}{\sqrt{5}} = 12.5_4 \pm 0.3_8$$

There is a 50% chance that the true mean, μ, lies within the range $12.5_4 \pm 0.1_3$ (12.4_1 to 12.6_7). There is a 90% chance that μ lies within the range $12.5_4 \pm 0.3_8$ (12.1_6 to 12.9_2).

- The 50% confidence interval is

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}} = 12.5_4 \pm 0.1_3$$

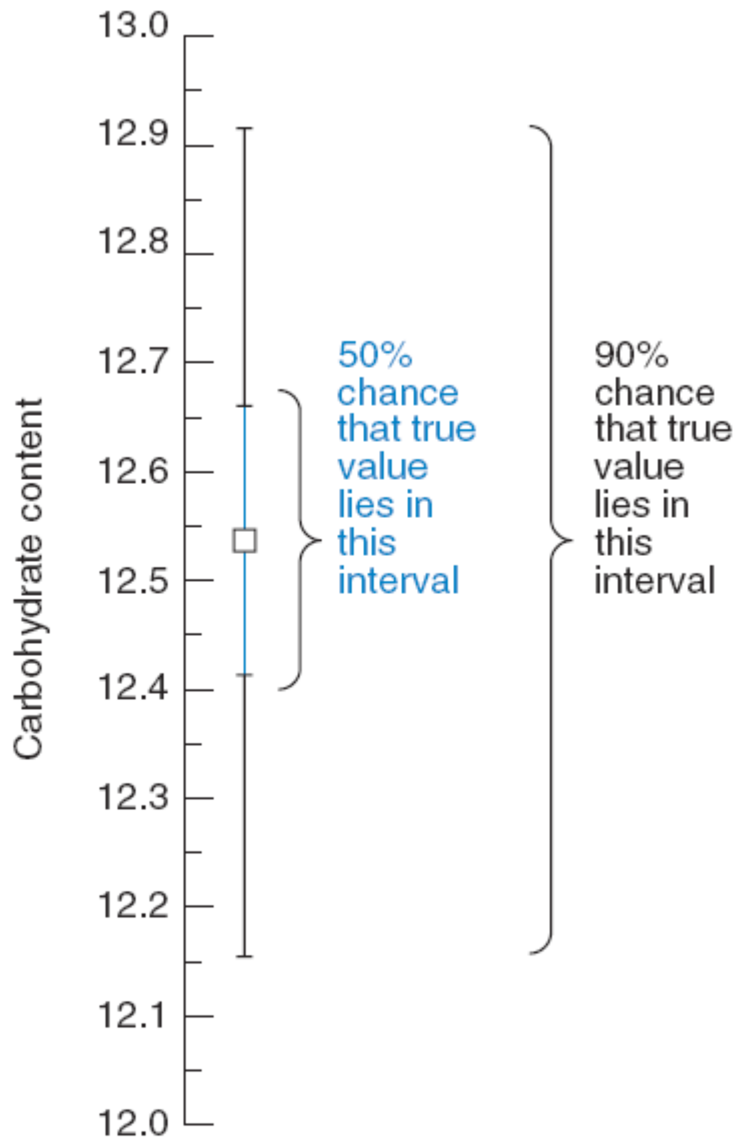
- The 90% confidence interval is

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}} = 12.5_4 \pm 0.3_8$$

- If you repeated sets of five measurements many times,

→ half of 50 % confidence intervals are expected to include the true mean, μ

→ nine tenths of 90 % confidence intervals are expected to include the true mean, μ



The Meaning of a Confidence Interval

- A computer chose numbers **at random**

  → from a Gaussian population with a population mean of 10 000 and a population standard deviation of 1 000

- In trial 1,

  → four numbers were chosen,

  → their mean (9526) and standard deviation were calculated

  → then, the 50% confidence interval was calculated using t = 0.765 from Table 4-4

  (50% confidence, 3 degrees of freedom → t = 0.765).

- This trial is plotted as the first point at the left in Figure 4-5a;

See Fig 4-5

- The square is centered at the mean value of 9 526,
- The error bar extends from the lower limit to the upper limit of the 50% confidence interval
- The experiment was repeated 100 times to produce the points in Figure 4-5a.

- In Figure 4-5a , the experiment was performed 100 times,

  → 45 of the error bars (open square) in Figure 4-5a pass through the horizontal line at 10 000.

See Fig 4-5

- The 50% confidence interval is defined such that,

  → if we repeated this experiment an infinite number of times, 50% of the error bars in Figure 4-5a would include the true population mean of 10 000.

- Figure 4-5b shows the same experiment with the same set of random numbers,

  → but this time the 90% confidence interval was calculated.
- For an infinite number of experiments,

  → we would expect 90% of the confidence intervals to include the population mean of 10 000.
- In Figure 4-5b, 89 of the 100 error bars cross the horizontal line at 10 000.

See Fig 4-5

Comparison of Mean with Student's t

- If you make two sets of measurements of the same quantity,

  → because of small, random variations in the measurements,

  the mean value from one set will generally not be equal to the mean value from the other set

- We use a ***t* test** to compare one mean value with another

  → to decide whether there is a statistically significant difference between the two.

  → That is, do the two means agree "within experimental error"?

- In inferential statistics, the term "**null hypothesis**" is a general statement

  → that there is no relationship between two measured phenomena.

- Rejecting the null hypothesis corresponds to

  → concluding that there is a relationship between two phenomena

- Until evidence indicates otherwise,

  → the null hypothesis is generally assumed to be true

- The **null hypothesis** in statistics regarding comparison of means
  → states that the mean values from two sets of measurements are not different.

- Statistics gives us a probability
  → that the observed difference between two means arises from random measurement error.

- If there is less than a 5% chance that that the observed difference arises from random variations
  → We customarily reject the null hypothesis

- With this criterion, we have a 95% chance that our conclusion is correct.
  → One time out of 20 when we conclude that two means are not different
  : we will be wrong.

For example,

- Measure a quantity several times, obtaining an average value and standard deviation.

- Compare our answer with an accepted answer.

- If the average is not exactly the same as the accepted answer,
  → Does our measured answer agree with the accepted answer "within experimental error"?

- You purchased a Standard Reference Material coal sample certified by the National Institute of Standards and Technology to contain 3.19 wt% sulfur.

- You are testing a new analytical method
  → to see whether it can reproduce the known value.

- The measured values are 3.29, 3.22, 3.30, and 3.23 wt% sulfur, giving a mean of $\bar{x} = 3.26_0$ and a standard deviation of $s = 0.04_1$.

- Does your answer agree with the known answer?
  → To find out,
    1) compute the 95% confidence interval for your answer
    2) see if that range includes the known answer.
  → If the known answer is not within your 95% confidence interval, then the results do not agree.

- For four measurements,

  → there are 3 degrees of freedom and $t_{95\%} = 3.182$ in Table 4-4.

- The 95% confidence interval is

$$95\% \text{ confidence interval} = \bar{x} \pm \frac{ts}{\sqrt{n}} = 3.26_0 \pm \frac{(3.182)(0.04_1)}{\sqrt{4}} = 3.26_0 \pm 0.06_5$$

$$95\% \text{ confidence interval} = 3.19_5 \text{ to } 3.32_5 \text{ wt\%}$$

- The known answer (3.19 wt%) is just outside the 95% confidence interval.

- Therefore we conclude that

  → there is less than a 5% chance that our method agrees with the known answer.

  → We conclude that our method gives a “different” result from the known result.

Is My Red Blood Cell Count High Today?

- At the opening of this chapter,

  → red cell counts on five "normal" days were 5.1, 5.3, 4.8, 5.4, and 5.2 × 10^6 cells/L.

  → The question was whether today's count of 5.6 × 10^6 cells/L is "significantly" higher than normal?

- Disregarding the factor of 10^6,

  → the mean of the normal values is $\bar{x} = 5.16$

  → the standard deviation is s = 0.23.

$$95\% \text{ confidence interval} = \bar{x} \pm \frac{ts}{\sqrt{n}} = 5.16 \pm \frac{2.776 \cdot 0.23}{\sqrt{5}} = 5.16 \pm 0.26$$

- Today's value is 5.6×10^6

- Today's red cell count lies in the upper tail of the curve containing less than 2.5% of the area of the curve.

  → There is less than a 5% probability of observing a count of 5.6×10^6 cells/L on "normal" days.

- It is reasonable to conclude that today's count is elevated.

See Fig 4-9

x-axis : *t*-value

98% confidence interval $= \bar{x} \pm \frac{ts}{\sqrt{n}} = 5.16 \pm \frac{3.747 \cdot 0.23}{\sqrt{5}} = 5.16 \pm 0.39$

99% confidence interval $= \bar{x} \pm \frac{ts}{\sqrt{n}} = 5.16 \pm \frac{4.604 \cdot 0.23}{\sqrt{5}} = 5.16 \pm 0.47$

- We see that 5.6 lies in 99% confidence levels.

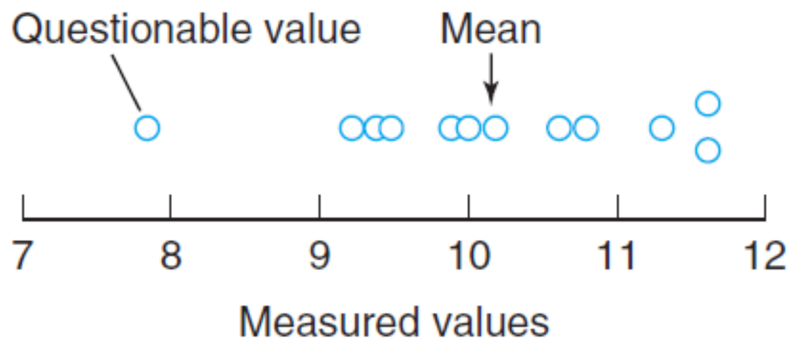
- More specifically,
  → There is less than a 2% probability of observing a count of 5.6×10^6 cells/L on "normal" days.

Grubbs Test for an Outlier

- To tell how much of zinc was included in the nail, students
  1) dissolved zinc from a galvanized nail
  2) and measured the mass lost by the nail

- Here are 12 results in mass loss (%):
  → 10.2, 10.8, 11.6, 9.9, 9.4, 7.8, 10.0, 9.2, 11.3, 9.5, 10.6, 11.6

- The value 7.8 appears out of line.



→ A datum that is far from other points is called an **outlier**.

- Should 7.8 be discarded before averaging the rest of the data or should 7.8 be retained?

- We answer this question with the **Grubbs test**.

 1) First compute

  → the average ($\bar{x} = 10.16$)

  → and the standard deviation (s = 1.11)

 of the complete data set (all 12 points in this example).

 2) Then compute the Grubbs statistic ***G***, defined as

Grubbs test:

$$G_{\text{calculated}} = \frac{|\text{questionable value} - \bar{x}|}{s}$$

Grubbs test:

$$G_{\text{calculated}} = \frac{|\text{questionable value} - \bar{x}|}{s}$$

- where the numerator is the absolute value of the difference between the suspected outlier and the mean value.

- If $G_{calculated}$ is greater than G_{table} in Table 4-6,

  → the value in question is out of the 95% confidence interval

  → the value in question can be rejected with 95% confidence.

  → the questionable point should be discarded.

See Table 4-6

- In our example,

$$G_{calculated} = \frac{|7.8 - 10.16|}{1.11} = 2.13$$

- In Table 4-6,

  G_{table} = 2.285 for 12 observations.

See Table 4-6

- Because $G_{calculated} < G_{table}$,

  → the questionable point should be retained.

The Method of Least Squares

- For most chemical analyses,

  → the response of the procedure must be evaluated for known quantities of analyte (called standards)

  → the response to an unknown quantity can be interpreted.

- For this purpose, we commonly prepare a calibration curve,

  → such as the one for caffeine in Figure 0-7.

- Most often, we work in a region

  → where the calibration curve is a straight line.

- We use the method of least squares to draw the “best” straight line.

Finding the Equation of the Line

Assumptions)

1) The procedure we use assumes that the errors in the y values are substantially greater than the errors in the x values.

- This condition is often true in a calibration curve
  → in which the experimental response (y values) is less certain than the quantity of analyte (x values).

2) A second assumption is that uncertainties (standard deviations) in all y values are similar.

- Suppose we seek to draw the best straight line through the points in Figure 4-11 by minimizing the vertical deviations between the points and the line.

See Fig 4-11

- The Gaussian curve drawn over the point (3,3) is a schematic indication of the fact that each value of y_i is normally distributed about the straight line.
- That is, the most probable value of y will fall on the line, but there is a finite probability of measuring y some distance from the line.

See Fig 4-11

- We minimize only the vertical deviations because we assume that uncertainties in y values are much greater than uncertainties in x values.

- Let the equation of the line be

Equation of straight line: $$\boxed{y = mx + b}$$

Equation of straight line: $$y = mx + b$$

- in which m is the slope and b is the y-intercept.

- The vertical deviation for the point (x_i, y_i) in Figure 4-11 is $y_i - y$,

  → where y is the ordinate of the straight line when $x = x_i$.

$$\text{Vertical deviation} = d_i = y_i - y = y_i - (mx_i + b)$$

- Some of the deviations are positive and some are negative.
- Because we wish to minimize the magnitude of the deviations irrespective of their signs,

  → we square all the deviations so that we are dealing only with positive numbers:

$$d_i^2 = (y_i - y)^2 = (y_i - mx_i - b)^2$$

- Because we minimize the squares of the deviations,

  → this is called **the method of least squares**.

- Finding values of m and b that minimize the sum of the squares of the vertical deviations involves some calculus, which we omit.

- We express the final solution for slope and intercept in terms of determinants, which summarize certain arithmetic operations.

- The determinant $\begin{vmatrix} e & f \\ g & h \end{vmatrix}$

  → represents the value eh – fg.

- For example, $\begin{vmatrix} 6 & 5 \\ 4 & 3 \end{vmatrix} = (6 \times 3) - (5 \times 4) = -2$

- The slope and the intercept of the "best" straight line are found to be

$$\text{Least-squares "best" line} \begin{cases} \text{Slope:} \quad m = \begin{vmatrix} \Sigma(x_i y_i) & \Sigma x_i \\ \Sigma y_i & n \end{vmatrix} \div D & \text{(4-16)} \\ \text{Intercept:} \quad b = \begin{vmatrix} \Sigma(x_i^2) & \Sigma(x_i y_i) \\ \Sigma x_i & \Sigma y_i \end{vmatrix} \div D & \text{(4-17)} \end{cases}$$

: where D is

$$D = \begin{vmatrix} \Sigma(x_i^2) & \Sigma x_i \\ \Sigma x_i & n \end{vmatrix}$$

: n is the number of points.

$$m = \frac{n\Sigma(x_i y_i) - \Sigma x_i \Sigma y_i}{n\Sigma(x_i^2) - (\Sigma x_i)^2}$$

$$b = \frac{\Sigma(x_i^2)\Sigma y_i - \Sigma(x_i y_i)\Sigma x_i}{n\Sigma(x_i^2) - (\Sigma x_i)^2}$$

- Let's use these equations to find the slope and intercept of the best straight line through the four points in Figure 4-11.
  → The work is set out in Table 4-7.

See Table 4-7

- Noting that n = 4 and putting the various sums into the determinants in Equations 4-16, 4-17, and 4-18 gives

$$m = \begin{vmatrix} 57 & 14 \\ 14 & 4 \end{vmatrix} \div \begin{vmatrix} 62 & 14 \\ 14 & 4 \end{vmatrix} = \frac{(57 \times 4) - (14 \times 14)}{(62 \times 4) - (14 \times 14)} = \frac{32}{52} = 0.615\,38$$

$$b = \begin{vmatrix} 62 & 57 \\ 14 & 14 \end{vmatrix} \div \begin{vmatrix} 62 & 14 \\ 14 & 4 \end{vmatrix} = \frac{(62 \times 14) - (57 \times 14)}{(62 \times 4) - (14 \times 14)} = \frac{70}{52} = 1.346\,15$$

- The equation of the best straight line through the points in Figure 4-11 is therefore

$$y = 0.615\,38x + 1.346\,15$$

How Reliable Are Least-Squares Parameters?

- To estimate the uncertainties (expressed as standard deviations) in the slope and intercept,

  → an uncertainty analysis must be performed on Equations 4-16 and 4-17.

$$\text{Least-squares “best” line} \begin{cases} \text{Slope:} \quad m = \begin{vmatrix} \Sigma(x_i y_i) & \Sigma x_i \\ \Sigma y_i & n \end{vmatrix} \div D & \text{(4-16)} \\ \text{Intercept:} \quad b = \begin{vmatrix} \Sigma(x_i^2) & \Sigma(x_i y_i) \\ \Sigma x_i & \Sigma y_i \end{vmatrix} \div D & \text{(4-17)} \end{cases}$$

- Because the uncertainties in m and b are related to the uncertainty in measuring each value of y,
  → we first estimate the standard deviation that describes the population of y values.

- This standard deviation, σ_y, characterizes the little Gaussian curve inscribed in Figure 4-11

See Fig 4-11

- We estimate σ_y, the population standard deviation of all y values, by calculating s_y, the standard deviation, for the four measured values of y.

- The deviation of each value of y_i from the center of its Gaussian curve is → $d_i = y_i - y = y_i - (mx_i + b)$.

- The standard deviation of these vertical deviations is

$$\sigma_y \approx s_y = \sqrt{\frac{\sum(d_i - \bar{d})^2}{(\text{degrees of freedom})}} \qquad (4\text{-}19)$$

- But the average deviation, $\bar{d}$, is 0 for the best straight line,
  → so the numerator of Equation 4-19 reduces to

$$\Sigma(d_i^2)$$

- The degrees of freedom is the number of independent pieces of information available.
  → For n data points, there are n degrees of freedom.

- If you were calculating the standard deviation of n points,
  → you would first find the average to use in Equation 4-2.
  → This calculation leaves n – 1 degrees of freedom in Equation 4-2 because only n – 1 pieces of information are available in addition to the average.

- If you know n – 1 values and you also know their average,
  → then the nth value is fixed and you can calculate it.