Chapter 4
Statistics

Overview

4-1 Gaussian Distribution

4-2 Comparison on Standard Deviations with the *F* Test

4-3 Confidence Intervals

4-4 Comparison of Means with Student's *t*

4-5 *t* Tests with a Spreadsheet

4-6 Grubbs Test for an Outlier

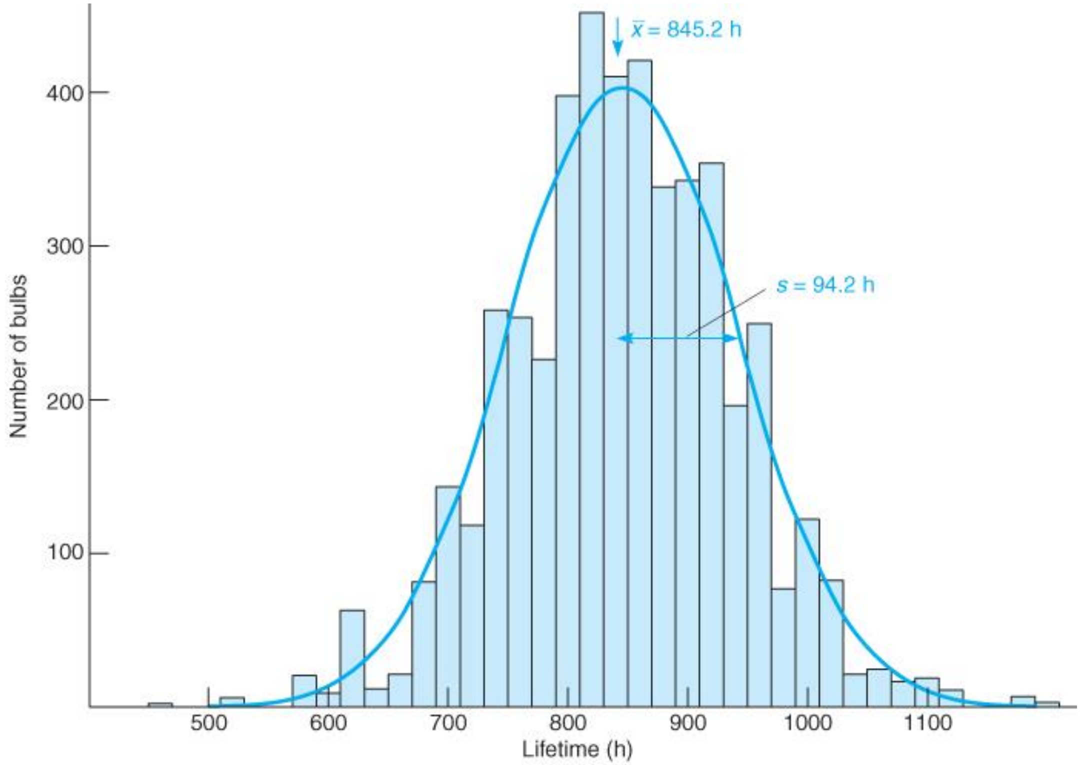
4-7 The Method of Least Squares

4-8 Calibration Curves

4-9 A Spreadsheet for Least Squares

4-1: Gaussian Distribution

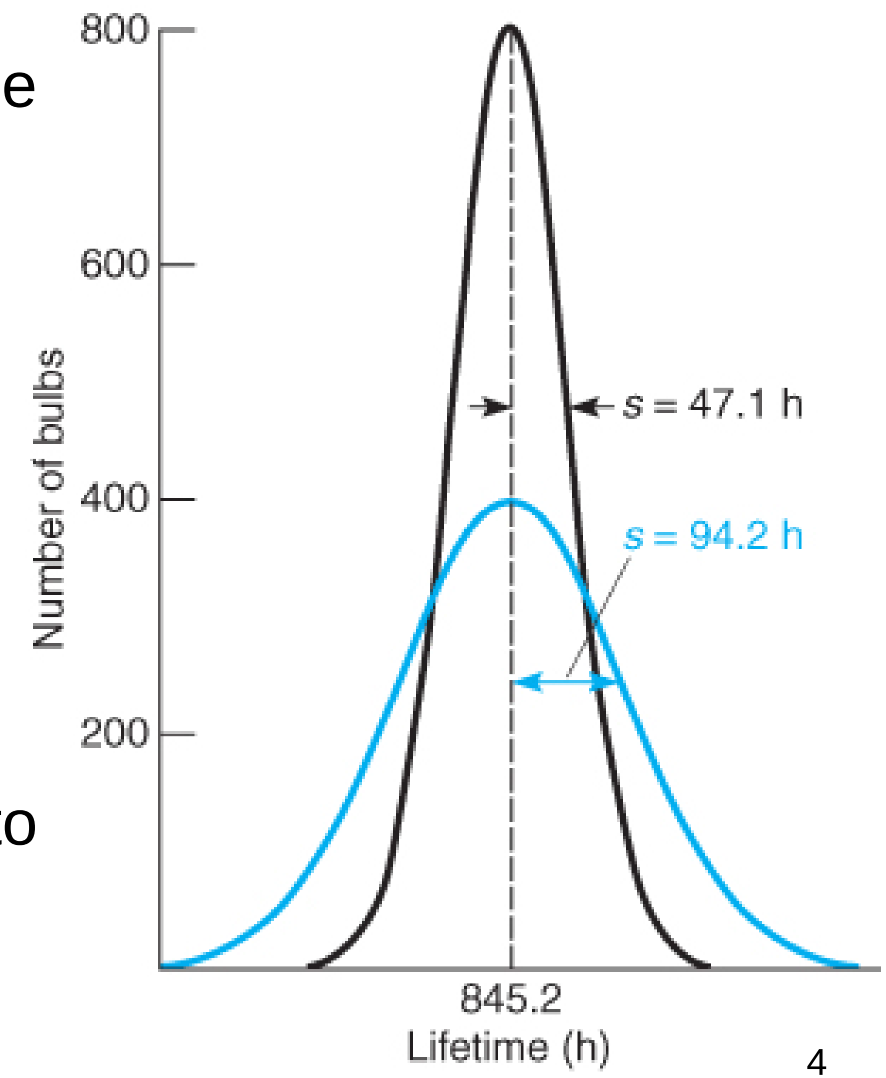
- The results of many measurements of an experimental quantity follow a **Gaussian distribution**.
- The measured **mean**, ***x***, approaches the true mean, ***μ***, as the number of measurements becomes very large.



4-1: Standard Deviation

- The broader the distribution, the greater is ***s***, the **standard deviation**.
- About two-thirds of all measurements lies within 1S from the mean (i.e., μ ±s accounts for ~67% of measurements),

- certain interval is proportional to the area of that interval.



4-1: Standard Deviation

- For *n* measurements, an estimate of the standard deviation is

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

- The relative standard deviation expressed as a % is known as the **coefficient of variation**.

$$\textbf{coefficient of variation} = 100 \times \frac{s}{\bar{x}}$$

4-1: Standard Deviation

EXAMPLE **Mean and Standard Deviation**

Find the average, standard deviation, and coefficient of variation for 821, 783, 834, and 855.

Solution The average is

$$\bar{x} = \frac{821 + 783 + 834 + 855}{4} = 823._2$$

To avoid accumulating round-off errors, retain one more digit in the mean than was present in the original data. The standard deviation is

$$s = \sqrt{\frac{(821 - 823.2)^2 + (783 - 823.2)^2 + (834 - 823.2)^2 + (855 - 823.2)^2}{(4 - 1)}} = 30._3$$

The average and the standard deviation should both end at the *same decimal place*. For $\bar{x} = 823._2$, we will write $s = 30._3$. The coefficient of variation is the percent relative uncertainty:

$$\text{Coefficient of variation} = 100 \times \frac{s}{\bar{x}} = 100 \times \frac{30._3}{823._2} = 3.7\%$$

TEST YOURSELF If each of the four numbers 821, 783, 834, and 855 in the example is divided by 2, how will the mean, standard deviation, and coefficient of variation be affected? (***Answer:*** $\bar{x}$ and s will be divided by 2, but the coefficient of variation is unchanged)

4-1: Standard Deviation and Probability

- The mathematical equation for the Gaussian curve is given below. [Equation 4-3]

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

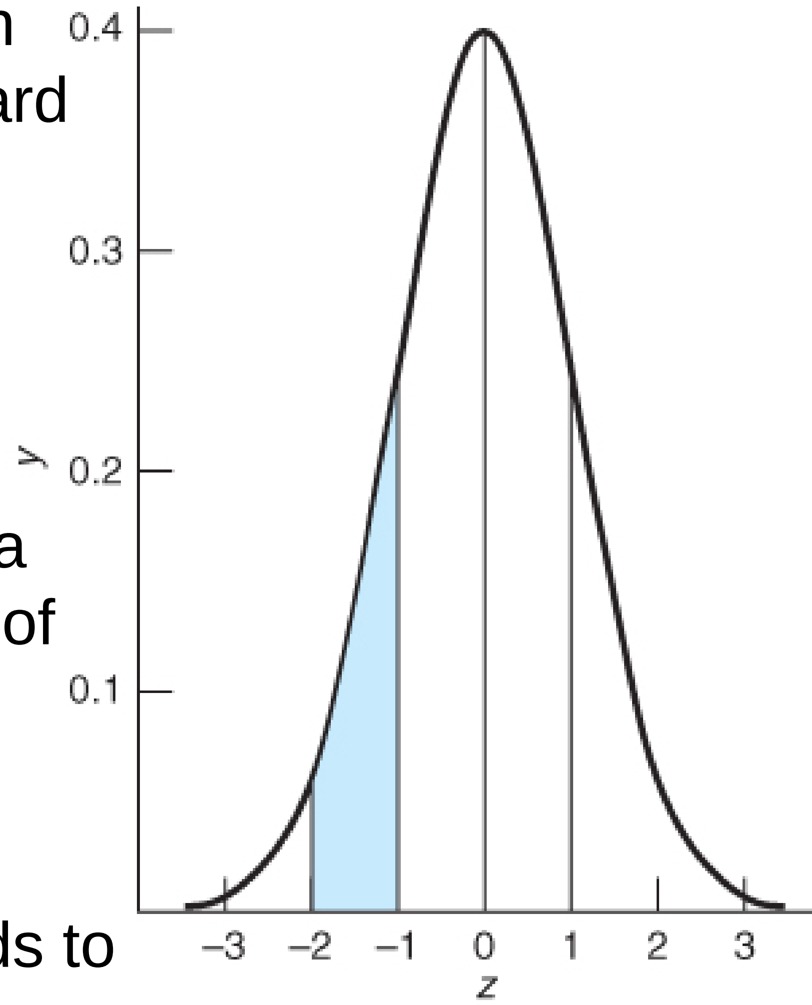
- The maximum occurs for $x = \mu$.
- The probability of observing a value within a certain interval is proportional to the area of that interval.
- About two-thirds of all measurements lie within $\pm 1s$ and 95% lie within $\pm 2s$.

4-1: Area Under the Gaussian Distribution

- Express deviations from the mean value in multiples, *z*, of the standard deviation.
- Transform *x* into *z*, given by:

$$z \approx \frac{x - \bar{x}}{s}$$

- The probability of measuring *z* in a certain *range* is equal to the *area* of that range.
- For example, the probability of observing *z* between -2 and -1 is 0.136. This probability corresponds to the shaded area in Figure 4-3.



Standard Normal Distribution (Probability Distribution)

Now if we plot the frequency of an occurrence of a given value of Z as a function of Z...

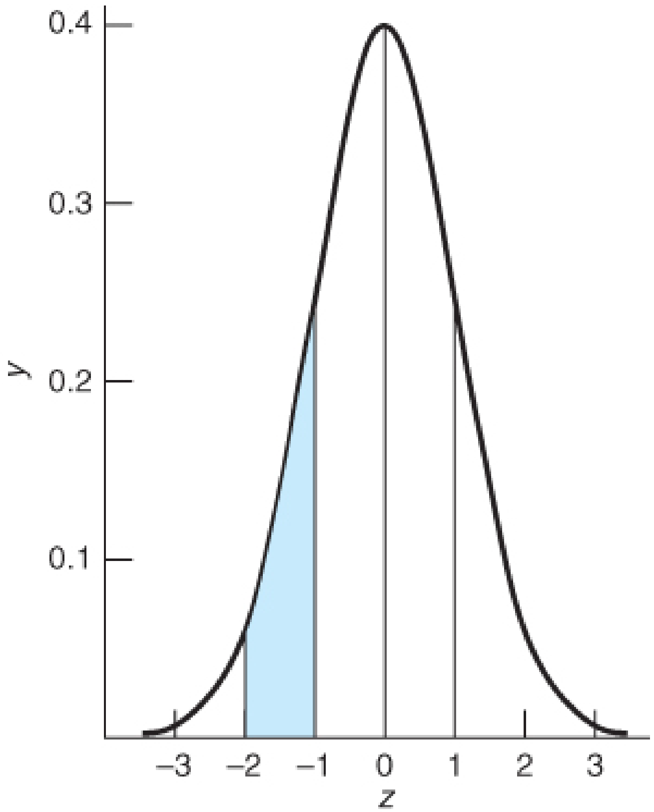


FIGURE 4-3 A Gaussian curve in which $\mu_z = 0$ and $\sigma_z = 1$. A Gaussian curve whose area is unity is **called a normal error curve**. The abscissa $z = (x - \mu)/\sigma$ is the distance away from the mean, measured in units of the standard deviation. When $z = 2$, we are two standard deviations away from the mean.

$y = \exp[-Z^2/2] / (2\pi\sigma^2)^{1/2}$

with $\sigma_z = 1$ and $\mu_z = 0$:

$$y = \exp[-Z^2/2] / (2\pi)^{1/2}$$

4-1: Area Under the Gaussian Distribution

TABLE 4-1 Ordinate and area for the normal (Gaussian) error curve, $y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

$\|z\|$[a]	y	Area[b]	$\|z\|$	y	Area	$\|z\|$	y	Area
0.0	0.398 9	0.000 0	1.4	0.149 7	0.419 2	2.8	0.007 9	0.497 4
0.1	0.397 0	0.039 8	1.5	0.129 5	0.433 2	2.9	0.006 0	0.498 1
0.2	0.391 0	0.079 3	1.6	0.110 9	0.445 2	3.0	0.004 4	0.498 650
0.3	0.381 4	0.117 9	1.7	0.094 1	0.455 4	3.1	0.003 3	0.499 032
0.4	0.368 3	0.155 4	1.8	0.079 0	0.464 1	3.2	0.002 4	0.499 313
0.5	0.352 1	0.191 5	1.9	0.065 6	0.471 3	3.3	0.001 7	0.499 517
0.6	0.333 2	0.225 8	2.0	0.054 0	0.477 3	3.4	0.001 2	0.499 663
0.7	0.312 3	0.258 0	2.1	0.044 0	0.482 1	3.5	0.000 9	0.499 767
0.8	0.289 7	0.288 1	2.2	0.035 5	0.486 1	3.6	0.000 6	0.499 841
0.9	0.266 1	0.315 9	2.3	0.028 3	0.489 3	3.7	0.000 4	0.499 904
1.0	0.242 0	0.341 3	2.4	0.022 4	0.491 8	3.8	0.000 3	0.499 928
1.1	0.217 9	0.364 3	2.5	0.017 5	0.493 8	3.9	0.000 2	0.499 952
1.2	0.194 2	0.384 9	2.6	0.013 6	0.495 3	4.0	0.000 1	0.499 968
1.3	0.171 4	0.403 2	2.7	0.010 4	0.496 5	∞	0	0.5

a. $z = (x - \mu)/\sigma$

b. The area refers to the area between $z = 0$ *and* $z =$ *the value in the table. Thus the area from* $z = 0$ *to* $z = 1.4$ *is 0.419 2. The area from* $z = -0.7$ *to* $z = 0$ *is the same as from* $z = 0$ *to* $z = 0.7$*. The area from* $z = -0.5$ *to* $z = +0.3$ *is (0.191 5 + 0.117 9) = 0.309 4. The total area between* $z = -\infty$ *and* $z = +\infty$ *is unity.*

4-1: Area Under the Gaussian Distribution

EXAMPLE **Area Under a Gaussian Curve**

Suppose the manufacturer of the bulbs used for Figure 4-1 offers to replace free of charge any bulb that burns out in less than 600 hours. If she plans to sell a million bulbs, how many extra bulbs should she keep available as replacements?

Solution We need to express the desired interval in multiples of the standard deviation and then find the area of the interval in Table 4-1. Because $\bar{x} = 845.2$ and $s = 94.2$, $z = (600 - 845.2)/94.2 = -2.60$. The area under the curve between the mean value and $z = -2.60$ is 0.495 3 in Table 4-1. The entire area from $-\infty$ to the mean value is 0.500 0, so the area from $-\infty$ to -2.60 must be $0.500\,0 - 0.495\,3 = 0.004\,7$. The area to the left of 600 hours in Figure 4-1 is only 0.47% of the entire area under the curve. Only 0.47% of the bulbs are expected to fail in fewer than 600 h. If the manufacturer sells 1 million bulbs a year, she should make 4 700 extra bulbs to meet the replacement demand.

TEST YOURSELF If the manufacturer will replace bulbs that burn out in less than 620 hours, how many extras should she make? (***Answer:*** $z \approx -2.4$, area $\approx 0.008\,2 = 8\,200$ bulbs)

4-1: Standard Deviation of the Mean

- The standard deviation s is a measure of the uncertainty of individual measurements.
- The **standard deviation of the mean**, σ_n is a measure of the uncertainty of the mean of n measurements.

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

- Uncertainty decreases by a factor of 2 by making four times as many measurements and by a factor of 10 by making 100 times as many measurements.
- Instruments with rapid data acquisition allow us to average many experiments in a short time in order to improve precision.

4-2: Comparing Standard Deviations

- The ***F* test** is used to decide whether two standard deviations are significantly different from each other.

$$F_{\text{calculated}} = \frac{s_1^2}{s_2^2}$$

- If $F_{\text{calculated}}$ (s_1^2/s_2^2) > F_{table}, then the two data sets have less than a 5% chance of coming from distributions with the same population standard deviation.

TABLE 4-2 **Measurement of HCO_3^- in horse blood**

	Original instrument	Substitute instrument
Mean ($\bar{x}$, mM)	36.14	36.20
Standard deviation (s, mM)	0.28	0.47
Number of measurements (n)	10	4

Data from M. Jarrett, D. B. Hibbert, R. Osborne, and E. B. Young, Anal. Bioanal. Chem. *2010,* 397, *717.*

4-2: Comparing Standard Deviations

TABLE 4-3 **Critical values of $F = s_1^2/s_2^2$ at 95% confidence level**

Degrees of freedom for s_2	Degrees of freedom for s_1													
	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
2	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5
3	9.55	9.28	9.12	9.01	8.94	8.89	8.84	8.81	8.79	8.74	8.70	8.66	8.62	8.53
4	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.63
5	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.36
6	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.67
7	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.58	3.51	3.44	3.38	3.23
8	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	2.93
9	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.71
10	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.70	2.54
11	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.40
12	3.88	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.30
13	3.81	3.41	3.18	3.02	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.21
14	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.13
15	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.07
16	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.19	2.01
17	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.15	1.96
18	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.11	1.92
19	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.07	1.88
20	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.84
30	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.62
∞	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.00

Critical values of F for a one-tailed test of the hypothesis that $s_1 > s_2$. There is a 5% probability of observing F above the tabulated value if the two sets of data come from populations with the same population standard deviation.

You can compute F for a chosen level of confidence with the Excel function FINV(Probability, deg_freedom1, deg_freedom2). The statement "= FINV(0.05,7,6)" reproduces the value F = 4.21 in this table. The statement "= FINV(0.1,7,6)" gives F = 3.01 for 90% confidence.

4-2: Comparing Standard Deviations

TABLE 4-2 **Measurement of HCO_3^- in horse blood**

	Original instrument	Substitute instrument
Mean ($\bar{x}$, mM)	36.14	36.20
Standard deviation (s, mM)	0.28	0.47
Number of measurements (n)	10	4

Data from M. Jarrett, D. B. Hibbert, R. Osborne, and E. B. Young, Anal. Bioanal. Chem. ***2010**, 397, 717.*

EXAMPLE **Is the Standard Deviation from the Substitute Instrument Significantly Greater Than That of the Original Instrument?**

In Table 4-2, the standard deviation from the substitute instrument is $s_1 = 0.47$ ($n_1 = 4$ measurements) and the standard deviation from the original instrument is $s_2 = 0.28$ ($n_2 = 10$).

Solution To answer the question, find F with Equation 4-6:

$$F_{\text{calculated}} = \frac{s_1^2}{s_2^2} = \frac{(0.47)^2}{(0.28)^2} = 2.8_2$$

In Table 4-3, we find $F_{\text{table}} = 3.86$ in the column with 3 degrees of freedom for s_1 (degrees of freedom $= n - 1$) and the row with 9 degrees of freedom for s_2. *Because $F_{calculated}$ ($= 2.8_2$) $<$ F_{table} ($= 3.86$), we reject the hypothesis that s_1 is significantly larger than s_2.* You will see in the next section on hypothesis testing that there is more than a 5% chance that the two sets of data are drawn from populations with the same population standard deviation.

TEST YOURSELF If there had been $n = 13$ replications in both data sets, would the difference in standard deviations be significant? (***Answer:*** Yes. $F_{\text{calculated}} = 2.8_2 > F_{\text{table}} = 2.69$)

4-3: Student's *t*

- **Student's *t*** is a statistical tool.
- It is used to find **confidence intervals** and it is also used 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.

4-3: Student's *t*

TABLE 4-4 **Values of Student's *t***

	Confidence level (%)						
Degrees of freedom	50	90	95	98	99	99.5	99.9
1	1.000	6.314	12.706	31.821	63.656	127.321	636.578
2	0.816	2.920	4.303	6.965	9.925	14.089	31.598
3	0.765	2.353	3.182	4.541	5.841	7.453	12.924
4	0.741	2.132	2.776	3.747	4.604	5.598	8.610
5	0.727	2.015	2.571	3.365	4.032	4.773	6.869
6	0.718	1.943	2.447	3.143	3.707	4.317	5.959
7	0.711	1.895	2.365	2.998	3.500	4.029	5.408
8	0.706	1.860	2.306	2.896	3.355	3.832	5.041
9	0.703	1.833	2.262	2.821	3.250	3.690	4.781
10	0.700	1.812	2.228	2.764	3.169	3.581	4.587
15	0.691	1.753	2.131	2.602	2.947	3.252	4.073
20	0.687	1.725	2.086	2.528	2.845	3.153	3.850
25	0.684	1.708	2.060	2.485	2.787	3.078	3.725
30	0.683	1.697	2.042	2.457	2.750	3.030	3.646
40	0.681	1.684	2.021	2.423	2.704	2.971	3.551
60	0.679	1.671	2.000	2.390	2.660	2.915	3.460
120	0.677	1.658	1.980	2.358	2.617	2.860	3.373
∞	0.674	1.645	1.960	2.326	2.576	2.807	3.291

In calculating confidence intervals, σ may be substituted for s in Equation 4-7 if you have a great deal of experience with a particular method and have therefore determined its "true" population standard deviation. If σ is used instead of s, the value of t to use in Equation 4-7 comes from the bottom row of this table.

Values of t in this table apply to two-tailed tests illustrated in Figure 4-9a. The 95% confidence level specifies the regions containing 2.5% of the area in each wing of the curve. For a one-tailed test, we use values of t listed for 90% confidence. Each wing outside of t for 90% confidence contains 5% of the area of the curve.

4-3: Confidence Intervals

- From a limited number of measurements (*n*), we **cannot** find the true population mean, ***μ,*** or the true standard deviation, ***σ.***
- What we can determine are *x* and ***s***, the **sample mean** and the **sample standard deviation**.
- The confidence interval allows us to state, at some level of confidence, a range of values that include the true population mean.

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

t **is the student t-value (@ specific df & C.L.**
s **is the standard deviation**
$\bar{x}$ **is the mean**
n **is the number of trials**

4-3: Confidence Intervals

EXAMPLE **Calculating Confidence Intervals**

The carbohydrate content of a glycoprotein (a protein with sugars attached to it) is found to be 12.6, 11.9, 13.0, 12.7, and 12.5 wt% (g carbohydrate/100 g glycoprotein) 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 4-4 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

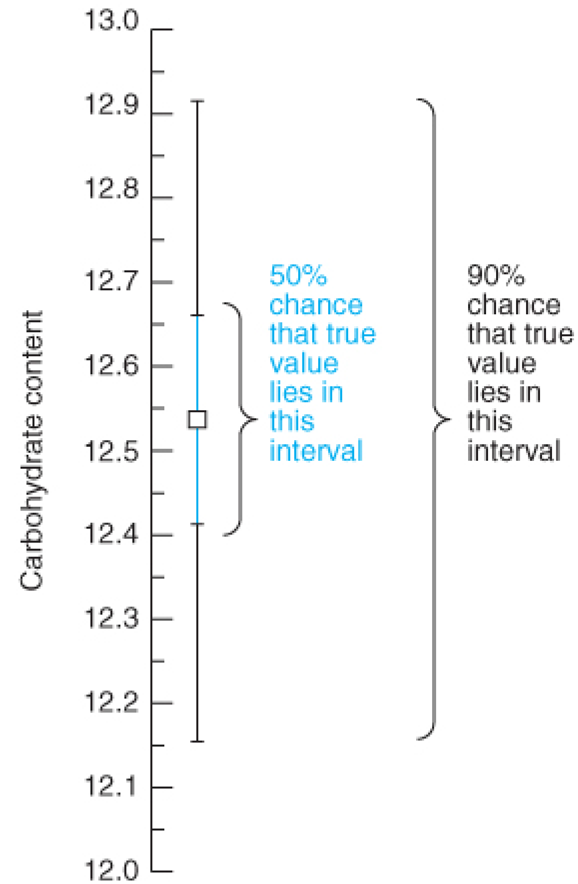
$$50\% \text{ confidence interval} = \bar{x} \pm \frac{ts}{\sqrt{n}} = 12.5_4 \pm \frac{(0.741)(0.4_0)}{\sqrt{5}} = 12.5_4 \pm 0.1_3 \text{ wt\%}$$

The 90% confidence interval is

$$90\% \text{ confidence interval} = \bar{x} \pm \frac{ts}{\sqrt{n}} = 12.5_4 \pm \frac{(2.132)(0.4_0)}{\sqrt{5}} = 12.5_4 \pm 0.3_8 \text{ wt\%}$$

If we repeat sets of five measurements many times, half of the 50% confidence intervals are expected to include the true mean, μ. Nine-tenths of the 90% confidence intervals are expected to include the true mean, μ.

TEST YOURSELF Carbohydrate measured on one more sample was 12.3 wt%. Using six results, find the 90% confidence interval. (***Answer:*** $12.5_0 \pm (2.015)(0.3_7)/\sqrt{6} = 12.5_0 \pm 0.3_1$ wt%)



4-3: Student's *t*

- **Student's *t*** test is also used to **compare mean values** measured by different methods.

4-3: Hypothesis *t* Tests

Student's *t* test is also used to compare experimental results. There are three different cases to consider.

1. Compare mean $\bar{x}$ with μ.
2. Compare two means $\bar{x}_1$ and $\bar{x}_2$.
3. Paired data.
    - Compare same group of samples using two methods.
    - Different samples!
    - Samples are not duplicated.

4-3: Hypothesis *t* Tests

Case 1 We measure a quantity several times, obtaining an average value and standard deviation. Does our measured value agree with the accepted value?

Case 2 We measure a quantity multiple times by two different methods that give two different answers, each with its own standard deviation. Do the two results agree with each other?

Case 3 Sample A is measured once by method 1 and once by method 2; the two measurements do not give exactly the same result. Then a different sample, designated B, is measured once by method 1 and once by method 2. The procedure is repeated for *n* different samples. Do the two methods agree with each other?

4-4: Using the Student's *t* Test

- Make a **null hypothesis**, $\boldsymbol{H_0}$**,** and an **alternate hypothesis**, $\boldsymbol{H_A}$.

  - Null Hypothesis $\boldsymbol{H_0}$
    - Difference explained by random error
  - Alternate Hypothesis $\boldsymbol{H_A}$
    - Difference cannot be explained by random error

1) Calculate $\boldsymbol{t}$
2) Compare $\boldsymbol{t}$ with $\boldsymbol{t_{table}}$ at a given confidence level and degrees of freedom.
3) If $\boldsymbol{t} < \boldsymbol{t_{table}}$, accept the null hypotheses H_0.
4) If $\boldsymbol{t} > \boldsymbol{t_{table}}$, reject the null hypotheses H_0 and accept H_A.

The Equations

(1)

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

(2) 4-9a

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s_{\text{pooled}}} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

or

4-9b

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}}$$

(3) 4-12

$$t = \frac{\left|\bar{d}\right|}{s_d} \cdot \sqrt{n}$$

4-4: Case 2: Additional Equations

Pooled standard deviation
***s*pooled (equation 4-10a)**
If s_1 and s_2 are **not significantly** different (according to **F-test**), then Spooled can be calculated as follows:

$$s_{\text{pooled}} = \sqrt{\frac{s_1^{\;2}\cdot(n_1-1)+s_2^{\;2}\cdot(n_2-1)}{n_1+n_2-2}}$$

And
df **=** $\mathbf{n_1 + n_2 - 2}$

If s_1 and s_2 are **significantly** different (according to **F-test**), then Degrees of freedom (equation 4-10b)

$$df = \left(\frac{\left[\frac{s_1^{\;2}}{n_1}+\frac{s_2^{\;2}}{n_2}\right]^2}{\left[\frac{\left(\frac{s_1^{\;2}}{n_1}\right)^2}{(n_1+1)}+\frac{\left(\frac{s_2^{\;2}}{n_2}\right)^2}{(n_2+1)}\right]} \right) - 2$$

4-4: Example: Case 1

- A coal sample is certified to contain 3.19 wt% sulfur. A new analytical method measures values of 3.29, 3.22, 3.30, and 3.23 wt% sulfur, giving a mean of $\bar{x} = 3.26_0$ and a standard deviation $s = 0.041$.
- Does the answer using the new method agree with the known answer?
- Calculate the confidence interval for the new method.
- If the known answer is not within the calculated 95% confidence interval, then the results **do not** agree.

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

4-4: Example: Case1

Solution:

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

$$95\%\ \text{confidence interval} = 3.26 \pm \frac{(3.182)(0.041)}{\sqrt{3}}$$

Typo, should be 4

$$\text{C.I.} = 3.26 \pm 0.06_5$$

The 95% confidence interval ranges from 3.195 to 3.325 wt%.

- The known answer (3.19 wt%) is just outside the 95% confidence interval.
- There is less than a 5% chance that new method agrees with the known answer.

4-4: Student's *t*

- Student's *t* is used to compare mean values measured by different methods.
- If the standard deviations are not significantly different (as determined with the *F*-test), find the pooled standard deviation with Equation 4-10a and compute *t* with Equation 4-9a.
- If *t* is greater than the tabulated value for $n_1 + n_2 - 2$ degrees of freedom, then the two data sets have less than a 5% chance ($p < 0.05$) of coming from distributions with the same population mean.
- If the standard deviations are significantly different, compute the degrees of freedom with Equation 4-10b and compute *t* with Equation 4-9b.

4-4: Example: Case 2 (standard deviations similar)

- Are 36.14 and 36.20 mM significantly different from each other? Do a *t* test to find out.

TABLE 4-2 Measurement of HCO_3^- in horse blood

	Original instrument	Substitute instrument
Mean ($\bar{x}$, mM)	36.14	36.20
Standard deviation (s, mM)	0.28	0.47
Number of measurements (n)	10	4

Data from M. Jarrett, D. B. Hibbert, R. Osborne, and E. B. Young, Anal. Bioanal. Chem. *2010, 397, 717.*

The *F* test indicates that the two standard deviations are not significantly different. Therefore calculate ***t*** using ***s*-pooled**.

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s_{\text{pooled}}} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

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

4-4: comparison of standard deviations with F test

TABLE 4-2 **Measurement of HCO_3^- in horse blood**

	Original instrument	Substitute instrument
Mean ($\bar{x}$, mM)	36.14	36.20
Standard deviation (s, mM)	0.28	0.47
Number of measurements (n)	10	4

Data from M. Jarrett, D. B. Hibbert, R. Osborne, and E. B. Young, Anal. Bioanal. Chem. ***2010**, 397, 717.*

EXAMPLE **Is the Standard Deviation from the Substitute Instrument Significantly Greater Than That of the Original Instrument?**

In Table 4-2, the standard deviation from the substitute instrument is $s_1 = 0.47$ ($n_1 = 4$ measurements) and the standard deviation from the original instrument is $s_2 = 0.28$ ($n_2 = 10$).

Solution To answer the question, find F with Equation 4-6:

$$F_{\text{calculated}} = \frac{s_1^2}{s_2^2} = \frac{(0.47)^2}{(0.28)^2} = 2.8_2$$

In Table 4-3, we find $F_{\text{table}} = 3.86$ in the column with 3 degrees of freedom for s_1 (degrees of freedom $= n - 1$) and the row with 9 degrees of freedom for s_2. *Because $F_{calculated}$ ($= 2.8_2$) $<$ F_{table} ($= 3.86$), we reject the hypothesis that s_1 is significantly larger than s_2.* You will see in the next section on hypothesis testing that there is more than a 5% chance that the two sets of data are drawn from populations with the same population standard deviation.

TEST YOURSELF If there had been $n = 13$ replications in both data sets, would the difference in standard deviations be significant? (***Answer:*** Yes. $F_{\text{calculated}} = 2.8_2 > F_{\text{table}} = 2.69$)

4-4: Example: Case 3 (paired data)

Nitrate concentrations in eight different plant extracts were measured using two different methods (shown in columns A and B) below in Figure 4-8.

Is there a significant difference between the methods?

	A	B	C	D
1	Comparison of methods for measuring nitrate			
2				
3		Nitrate (ppm) in plant extract		
4		Spectrophotometry		
5	Sample	with Cd reduction	Experimental biosensor	Difference (d_i)
6	1	1.22	1.23	0.01
7	2	1.21	1.58	0.37
8	3	4.18	4.04	–0.14
9	4	3.96	4.92	0.96
10	5	1.18	0.96	–0.22
11	6	3.65	3.37	–0.28
12	7	4.36	4.48	0.12
13	8	1.61	1.70	0.09
14			Mean difference =	0.114
15			Standard deviation of differences =	0.401
16			$t_{calculated}$ =	0.803
17			t_{table} =	2.365
18	D6 = C6–B6			
19	D14 = AVERAGE(D6:D13)			
20	D15 = STDEV(D6:D13)			
21	D16 = ABS(D14)*SQRT(A13)/D15 (ABS = absolute value)			
22	D17 = TINV(0.05,A13-1)			

4-4: Example: Case 3 (paired data)

- For a given sample, calculate the differences between the methods (column D). Average the differences in order to find $\bar{d}$ and s_d (standard deviation).

$$t = \frac{|\bar{d}|}{s_d} \cdot \sqrt{n} \qquad t = \frac{|0.114|}{0.401} \cdot \sqrt{8} = 0.803$$

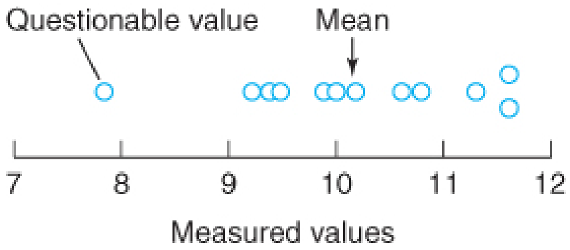
$t_{\mathbf{table}} = 2.365$

$t < t_{table}$, accept the null hypothesis

4-6: Removing outliers

The **Grubbs test** helps you to decide whether or not a questionable datum (**outlier**) should be discarded.

The mass loss from 12 galvanized nails was measured.



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; ($\bar{x}$= 10.16, s = 1.11).

Should the value 7.8 be discarded or retained?

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

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

Because $G_{calculated} < G_{table}$*, the questionable point should be retained.*

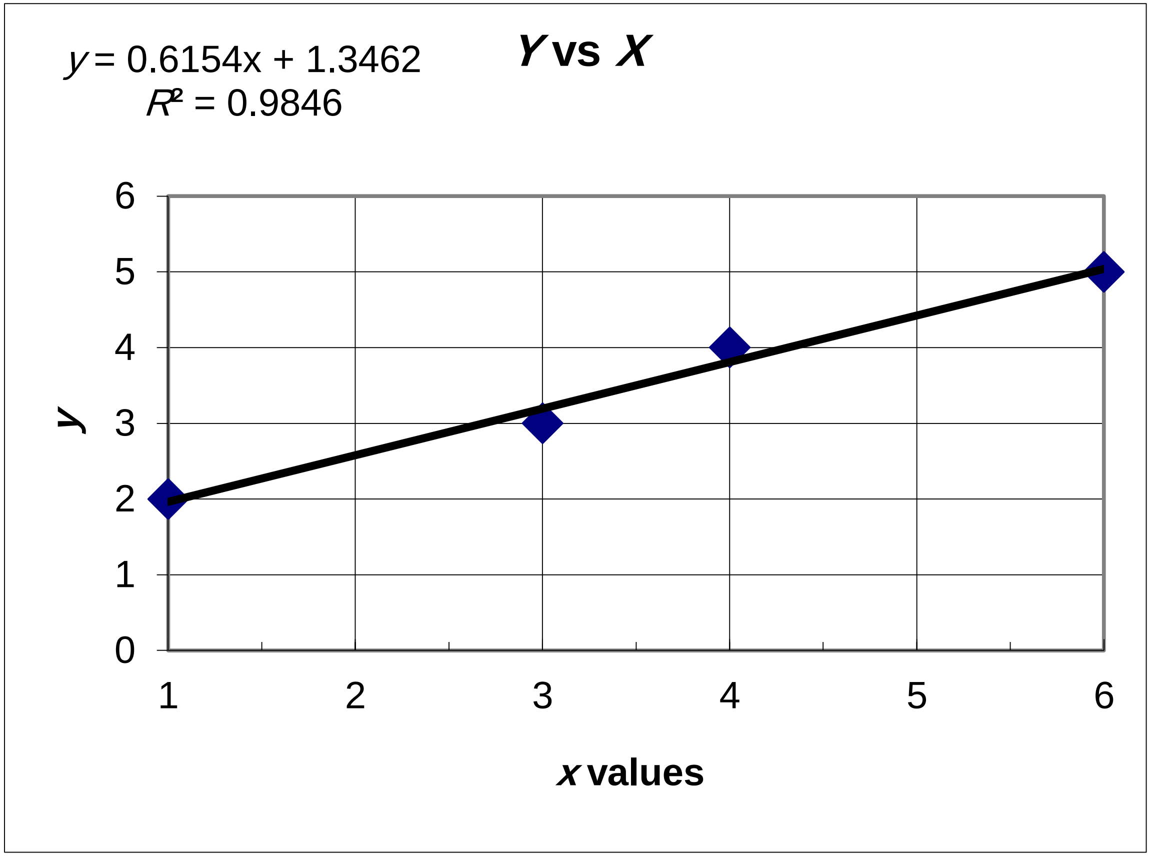
$$G_{\text{table}} = 2.285 \text{ (from Table 4 - 6)}$$

4-7: Method of Least Squares

- The method of **least squares** is used to determine the equation of the "best" straight line through experimental data points, $y_i = mx_i + b$. We need to find m and b.
- Equations 4-16 to 4-18 and 4-20 to 4-22 provide the least-squares **slope** and **intercept** and their standard uncertainties.

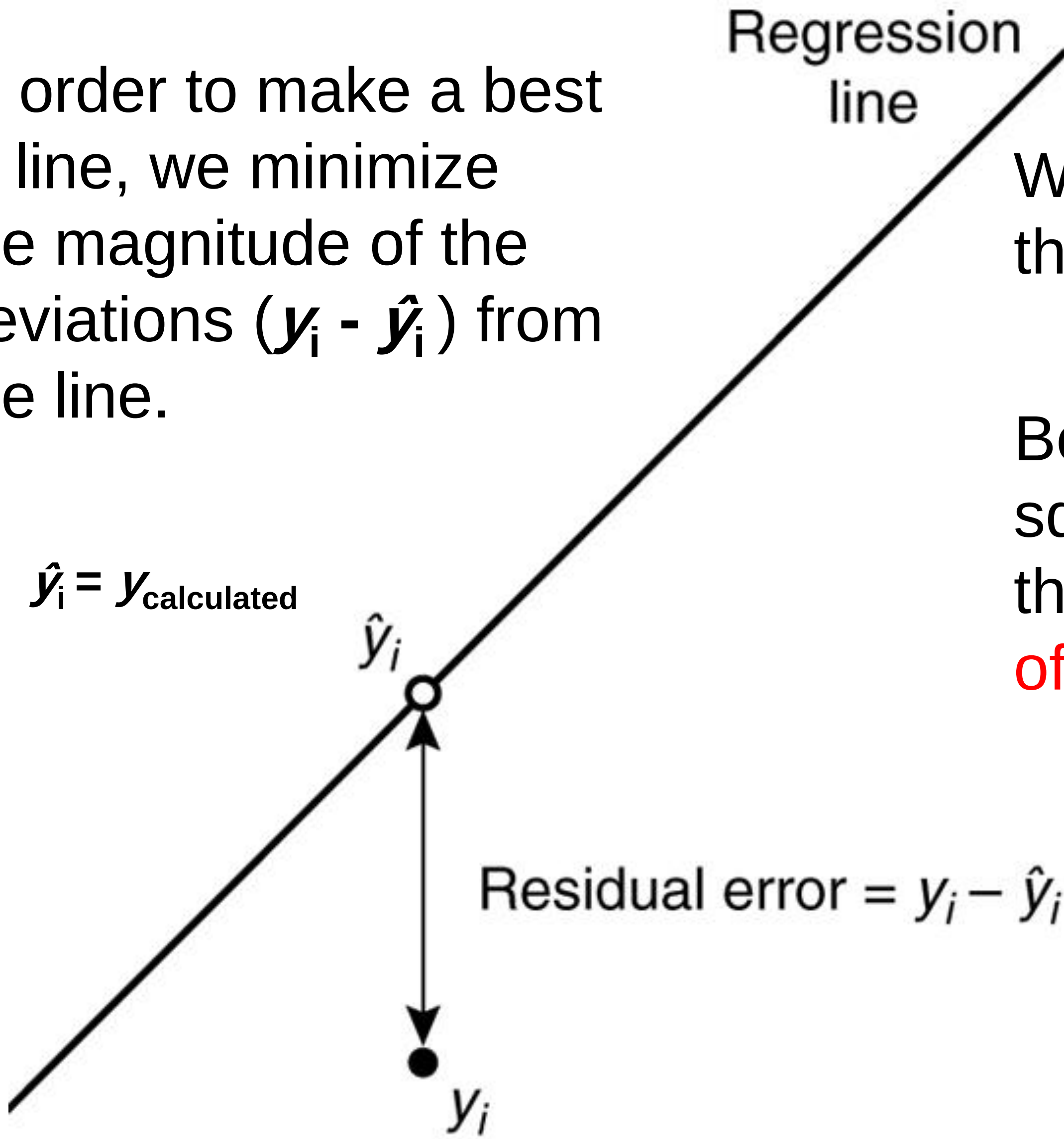
Find the Best-Fit Line through the Data

x	*y*
1	2
3	3
4	4
6	5



In order to make a best fit line, we minimize the magnitude of the deviations ($\boldsymbol{y}_i$ - $\boldsymbol{\hat{y}}_i$) from the line.

$\boldsymbol{\hat{y}}_i = \boldsymbol{y}_{\textbf{calculated}}$



We want to minimize the total residual error!

Because we minimize the squares of the deviations, this is called the method of least squares.

$$\text{Total residual error} = \sum (y_i - \hat{y}_i)^2$$

4-7: Equations for Least-Square Parameters

$$\mathbf{m} = \frac{\begin{vmatrix} \sum_{i=1}^{i} (x_i y_i) & \sum_{i=1}^{i} x_i \\ \sum_{i=1}^{i} y_i & n \end{vmatrix}}{D}$$

Equation 4-16

$$\mathbf{b} = \frac{\begin{vmatrix} \sum (x_i^2) & \sum x_i y_i \\ \sum x_i & \sum y_i \end{vmatrix}}{D}$$

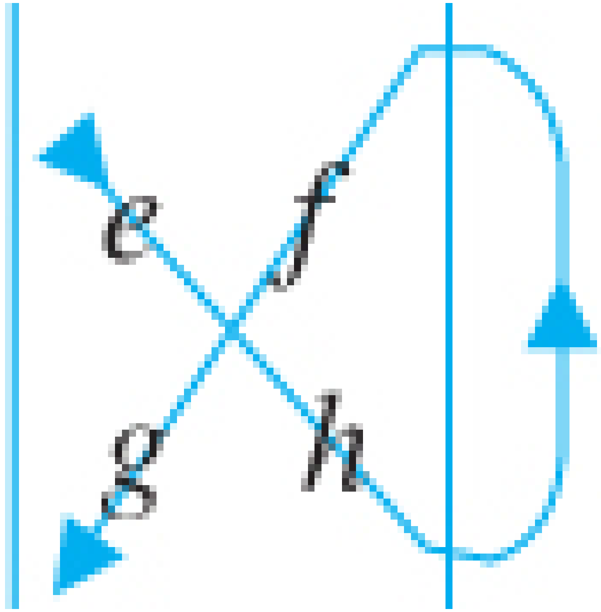
Equation 4-17

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

Equation 4-18

Operation of Determinates

D = eh - fg



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

Translation of least-squares equations:

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

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

4-7: Equations for Least-Square Parameters

Least-squares "best" line

Slope: $$m = \begin{vmatrix} \Sigma(x_i y_i) & \Sigma x_i \\ \Sigma y_i & n \end{vmatrix} \div D$$

Intercept: $$b = \begin{vmatrix} \Sigma(x_i^2) & \Sigma(x_i y_i) \\ \Sigma x_i & \Sigma y_i \end{vmatrix} \div D$$

where D is

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

Translation of least-squares equations:

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

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

4-7: Calculating the Uncertainty

Equations for estimating the **standard uncertainties** in ***y***, the slope ***m***, and the intercept ***b*** are given below.

$$s_y = \sqrt{\frac{\sum \left(d_i{}^2\right)}{n-2}}$$

Equation 4-20

$$u_m{}^2 = \frac{s_y{}^2 n}{D}$$

$$u_b{}^2 = \frac{s_y{}^2 \sum \left(x_i{}^2\right)}{D}$$

Equation 4-21

Equation 4-22

4-7: Results for the Least-Square Analysis

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

TABLE 4-7 Calculations for least-squares analysis

x_i	y_i	$x_i y_i$	x_i^2	$d_i(= y_i - mx_i - b)$	d_i^2
1	2	2	1	0.038 46	0.001 479 3
3	3	9	9	−0.192 31	0.036 982
4	4	16	16	0.192 31	0.036 982
6	5	30	36	−0.038 46	0.001 479 3
$\Sigma x_i = 14$	$\Sigma y_i = 14$	$\Sigma(x_i y_i) = 57$	$\Sigma(x_i^2) = 62$		$\Sigma(d_i^2) = 0.076\ 923$

4-8: Calculating the Uncertainty

- Equation 4-27 estimates the standard uncertainty in *x* from a measured value of *y* with a calibration curve.

$$s_c = \frac{s_y}{|m|}\sqrt{\frac{1}{k} + \frac{1}{n} + \frac{(y_c - \bar{y})^2}{m^2 \sum (x_i - \bar{x})^2}}$$

m = slope

k = number of replicate measurements for unknown

n = number of data points for calibration line

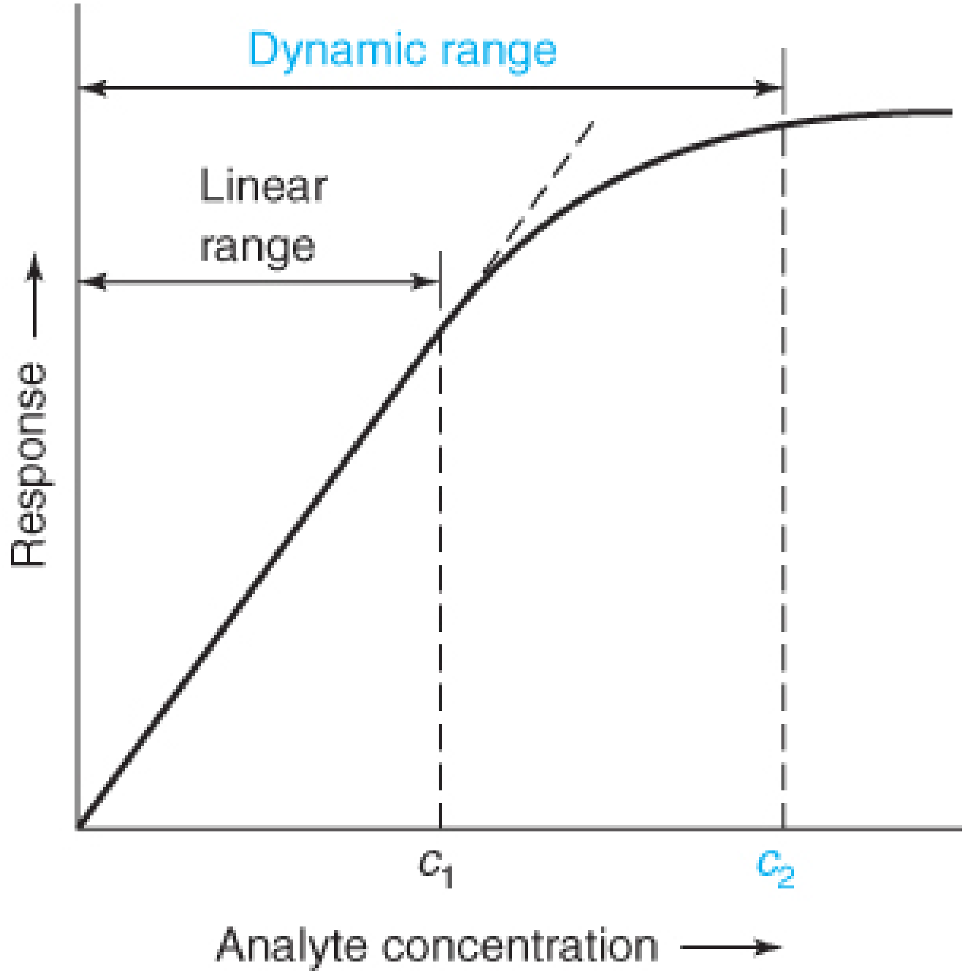
$\bar{y}$ = mean of *y* values in calibration line

y_c = mean value of measured *y* for unknown *x*

- A spreadsheet simplifies least-squares calculations and graphical display of the results.

4-8: Calibration Curve

- A calibration curve shows the response of a chemical analysis to known quantities (standard solutions) of analyte.



4-8: Calibration Curve

- When there is a linear response, the corrected analytical signal (signal from sample - signal from **blank**) is proportional to the quantity of analyte.
- The **linear range** of an analytical method is the range over which response is proportional to concentration.
- The **dynamic range** is the range over which there is a measurable response to analyte, even if the response is not linear.

4-8: Calibration Curve

EXAMPLE **Using a Linear Calibration Curve**

An unknown protein sample gave an absorbance of 0.406 and a blank had an absorbance of 0.104. How many micrograms of protein are in the unknown?

Solution The corrected absorbance is 0.406 − 0.104 = 0.302, which lies on the linear portion of the calibration curve in Figure 4-13. Rearranging Equation 4-25 gives

$$\mu\text{g of protein} = \frac{\text{absorbance} - 0.004_7}{0.016\,3_0} = \frac{0.302 - 0.004_7}{0.016\,3_0} = 18.2_4\ \mu\text{g} \qquad \textbf{(4-26)}$$

TEST YOURSELF What mass of protein gives a corrected absorbance of 0.250? (*Answer:* 15.0_5 μg)

4-8: Blank Solutions

- **Blank solutions** are prepared from the same reagents and solvents used to prepare standards and unknowns, but blanks have no intentionally added analyte.
- The blank tells us the response of the procedure to impurities or to interfering species in the reagents.
- The blank value is subtracted from measured values of standards prior to constructing the calibration curve.
- The blank value is subtracted from the response of an unknown prior to computing the quantity of analyte in the unknown.