Basic Statistics

There are three types of error:

1. *Gross error* – analyst makes a gross mistake (misread balance or entered wrong value into calculation).

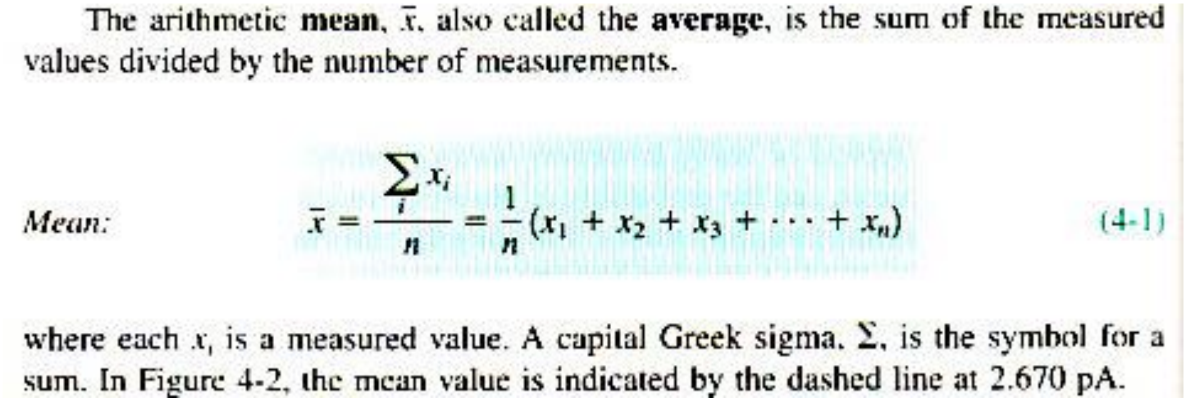
2. *Systematic error* - always too high or too low (improper shielding and grounding of an instrument or error in the preparation of standards).

3. *Random error* - unpredictably high or low (pressure changes or temperature changes).

Precision = ability to control random error.
Accuracy = ability to control systematic error.

Statistics

Experimental measurements always have some random error, so no conclusion can be drawn with complete certainty. *Statistics* give us a tool to accept conclusions that have a high probability of being correct. Deals with random error!



The arithmetic **mean**, $\bar{x}$, also called the **average**, is the sum of the measured values divided by the number of measurements.

Mean:

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) \qquad (4\text{-}1)$$

where each x_i is a measured value. A capital Greek sigma, Σ, is the symbol for a sum. In Figure 4-2, the mean value is indicated by the dashed line at 2.670 pA.

There is always some uncertainty in a measurement. Challenge is to minimize this uncertaintity!!

Standard Deviation of the Mean

Another very common way chemists represent error a measurement is to report a value called the ***standard deviation.*** ***The standard deviation of a small sampling is:***

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

Trials	Measurements
1	21.56
2	27.25
3	25.53
4	24.99
5	24.43
Mean	**24.75**
s	**2.07**

s = standard deviation of the mean

N = number of measurements or points

x_i = each individual measurement

$\bar{x}$ = sample mean

24.75 ± 2.07

RSD = (s/mean) x 100

Coefficient of variance = $(s)^2$

Standard Error of the Mean

Another very common way to represent error is to report a value called the ***standard error. The standard error is related to standard deviation:***

$$\boxed{S.E. = \frac{s}{\sqrt{N}}}$$

s = standard deviation of the mean
N = number of measurements or points

24.75 ± 0.93

Trials	Measurements
1	21.56
2	27.25
3	25.53
4	24.99
5	24.43
Mean	**24.75**
s	**2.07**
S.E.	**0.93**

Confidence Limit

Another common statistical tool for reporting the uncertainty (precision) of a measurement is the ***confidence limit (CL)***.

$$C.L. = \pm t \frac{s}{\sqrt{N}}$$

Table 22.1: Confidence Limit t-values as a function of (N-1)

N-1	**90%**	**95%**	**99%**	**99.5%**
2	2.920	4.303	9.925	14.089
3	2.353	3.182	5.841	7.453
4	2.132	2.776	4.604	5.598
5	2.015	2.571	4.032	4.773
6	1.943	2.447	3.707	4.317
7	1.895	2.365	3.500	4.029
8	1.860	2.306	3.355	3.832
9	1.833	2.262	3.205	3.690
10	1.812	2.228	3.169	3.581

Trials	Measurements
1	21.56
2	27.25
3	25.53
4	24.99
5	24.43
Mean	**24.75**
s	**2.07**
S.E.	**0.93**
C.L.	**2.56**

$$C.L. = \pm(2.776)\frac{2.07}{\sqrt{5}} = 2.56$$

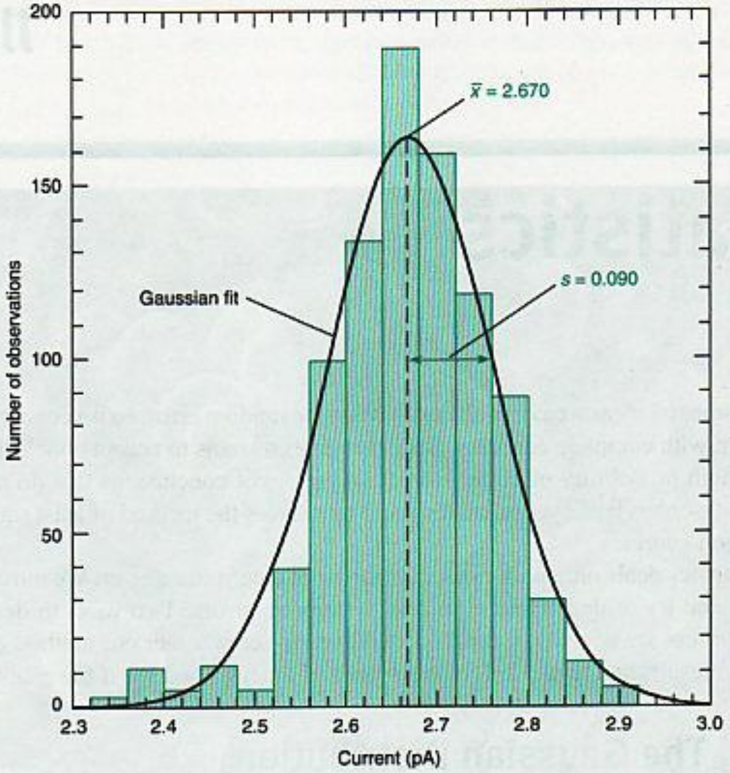
24.75 ± 2.56

Population versus Sample Size

The term *population* is used when an infinite sampling occurred or all possible subjects were analyzed. Obviously, we cannot repeat a measurement an infinite number of times so quite often the idea of a *population* is theoretical.

Sample size is selected to reflect population.

Figure 4-2 Observed cation current passing through individual channels of a frog muscle cell. The smooth line is the Gaussian curve that has the same mean and standard deviation as the measured data. The bar chart is also called a *histogram*. [Data from Nobel Lecture of B. Sakmann, *Angew. Chem. Int. Ed. Engl.* **1992**, *31*, 830.]



μ ± **1**σ	68.3%
μ ± **2**σ	95.5%
μ ± **3**σ	99.7%

Median = middle number is a series of measurements

Range = difference between the highest and lowest values

Q-Test

Q-test is a statistical tool used to identify an outlier within a data set .

$$Q_{\exp} = \frac{\left| x_q - \bar{x} \right|}{w}$$

x_q = suspected outlier
x_{n+1} = next nearest data point
w = range (largest – smallest data point in the set)

Critical Rejection Values for Identifying an Outlier: Q-test

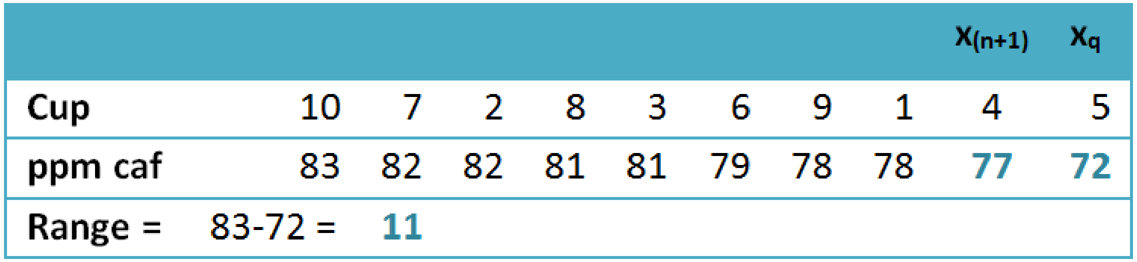
	Q_{crit}		
N	**90% CL**	**95% CL**	**99% CL**
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

Cup	ppm Caffeine
1	78
2	82
3	81
4	77
5	72
6	79
7	82
8	81
9	78
10	83
Mean	79.3
s	3.3
C.L. 95%	2.0

Calculation Q-test

Example - Perform a Q-test on the data set from Table on previous page and determine if you can statistically designate data point #5 as an outlier within a 95% CL. If so, recalculate the mean, standard deviation and the 95% CL .

Strategy – Organize the data from highest to lowest data point and use Equation to calculate Q_{exp}.



									$X_{(n+1)}$	X_q
Cup	10	7	2	8	3	6	9	1	4	5
ppm caf	83	82	82	81	81	79	78	78	77	72
Range =	83-72 =	11								

Solution – Ordering the data from Table 22.3 from highest to lowest results in

$$Q_{\exp} = \frac{|72 - 77|}{11} = 0.455$$

Using the Q_{crit} table, we see that Q_{crit}=0.466. Since Q_{exp} < Q_{crit}, you must <u>keep</u> the data point.

Grubbs Test

The recommended way of identifying outliers is to use the Grubb's Test. A Grubb's test is similar to a Q-test however G_{exp} is based upon the mean and standard deviation of the distribution instead of the next-nearest neighbor and range.

$$\boxed{G_{exp} = \frac{\left|x_q - \overline{x}\right|}{s}}$$

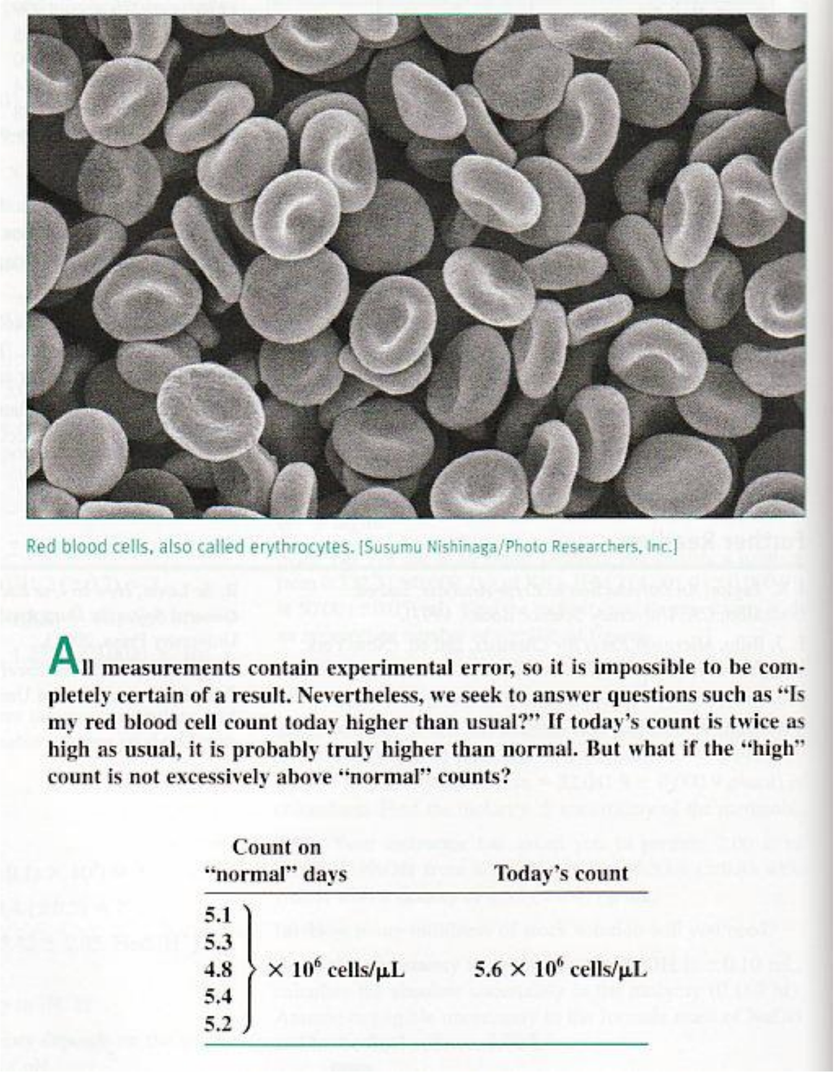
Table: Critical Rejection Values for Identifying an Outlier: G-test

		G_{crit}	
N	**90% C.L.**	**95% C.L.**	**99% C.L.**
3	1.1.53	1.154	1.155
4	1.463	1.481	1.496
5	1.671	1.715	1.764
6	1.822	1.887	1.973
7	1.938	2.020	2.139
8	2.032	2.127	2.274
9	2.110	2.215	2.387
10	2.176	2.290	2.482

If G_{exp} is greater than the critical G-value (G_{crit}) found in the Table then you are statistically justified in removing your suspected outlier .

How would you determine if the value is high or not?

Is the value high or within the confidence interval of the normal counts?



Red blood cells, also called erythrocytes. [Susumu Nishinaga/Photo Researchers, Inc.]

All measurements contain experimental error, so it is impossible to be completely certain of a result. Nevertheless, we seek to answer questions such as "Is my red blood cell count today higher than usual?" If today's count is twice as high as usual, it is probably truly higher than normal. But what if the "high" count is not excessively above "normal" counts?

Count on "normal" days		Today's count
5.1		
5.3		
4.8	$\times 10^6$ cells/μL	5.6×10^6 cells/μL
5.4		
5.2		

Mean = 5.2×10^6 counts

s = 2.3×10^5 counts

5.2 ± 0.2 (x 10^6 counts) std. dev.

$$C.L. = \pm(2.776)\frac{0.2}{\sqrt{5}} = 0.25$$

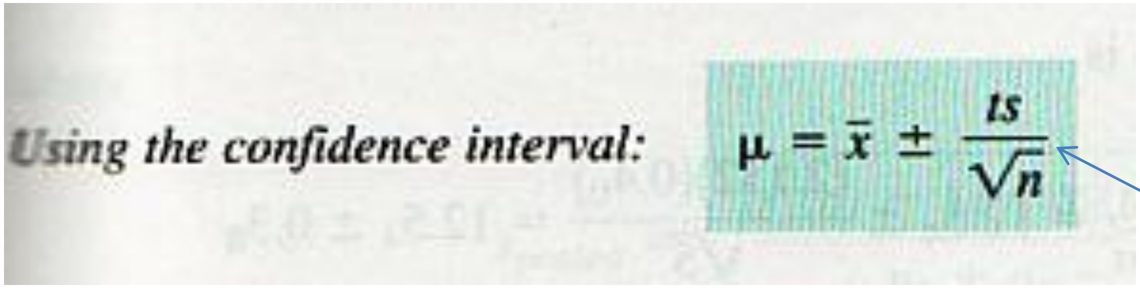
5.2 ± 0.3 (x 10^6 counts) 95% C.L.

{4.9 to 5.5 is normal range}

5.6 (x 10^6 counts) is not within the normal range at this confidence interval.

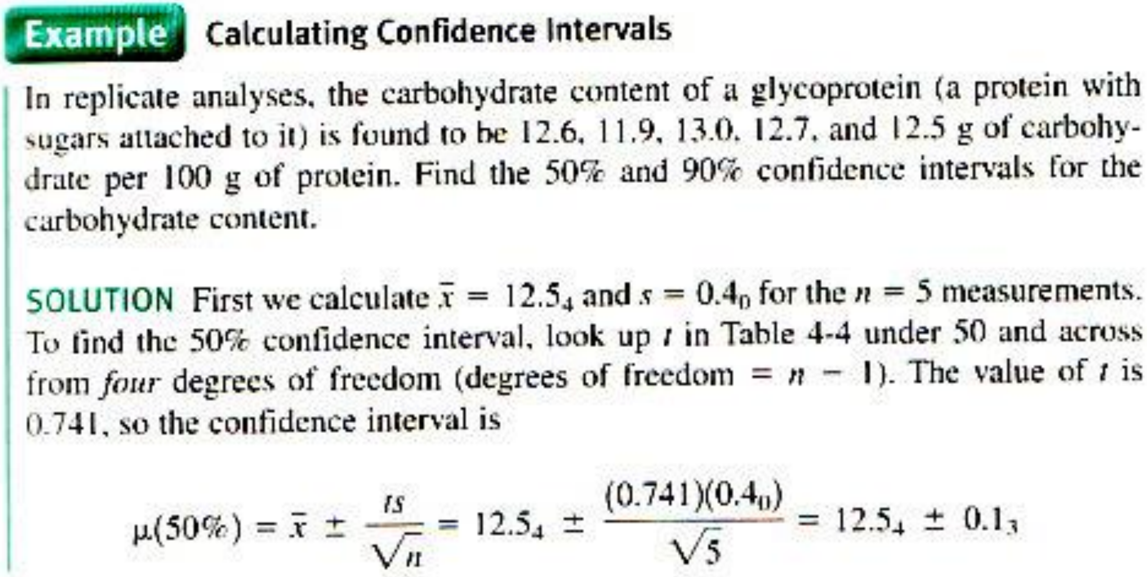
Student's t

Good to report mean values at the 95% confidence interval



Using the confidence interval: $\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$

Confidence interval



Example Calculating Confidence Intervals

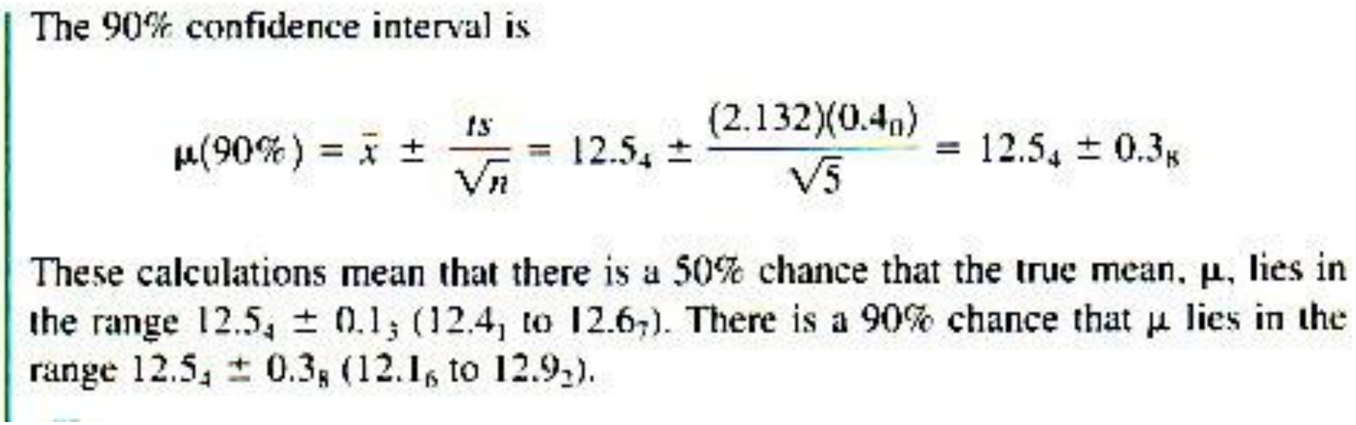
In replicate analyses, the carbohydrate content of a glycoprotein (a protein with sugars attached to it) is found to be 12.6, 11.9, 13.0, 12.7, and 12.5 g of carbohydrate per 100 g of protein. Find the 50% and 90% confidence intervals for the carbohydrate content.

SOLUTION First we calculate $\bar{x} = 12.5_4$ and $s = 0.4_0$ for the $n = 5$ measurements. To find 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 confidence interval is

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

From a limited number of measurements, it is possible to find population mean and standard deviation.

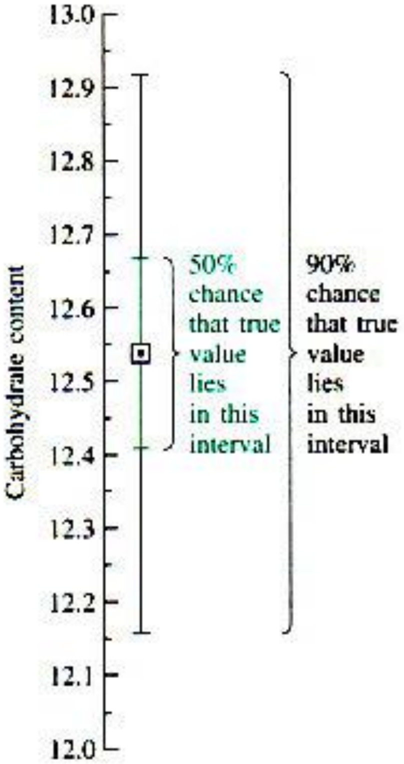
Student’s t



The 90% confidence interval is

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

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


13.0
12.9
12.8
12.7
12.6
12.5
12.4
12.3
12.2
12.1
12.0
Carbohydrate content
50% chance that true value lies in this interval
90% chance that true value lies in this interval

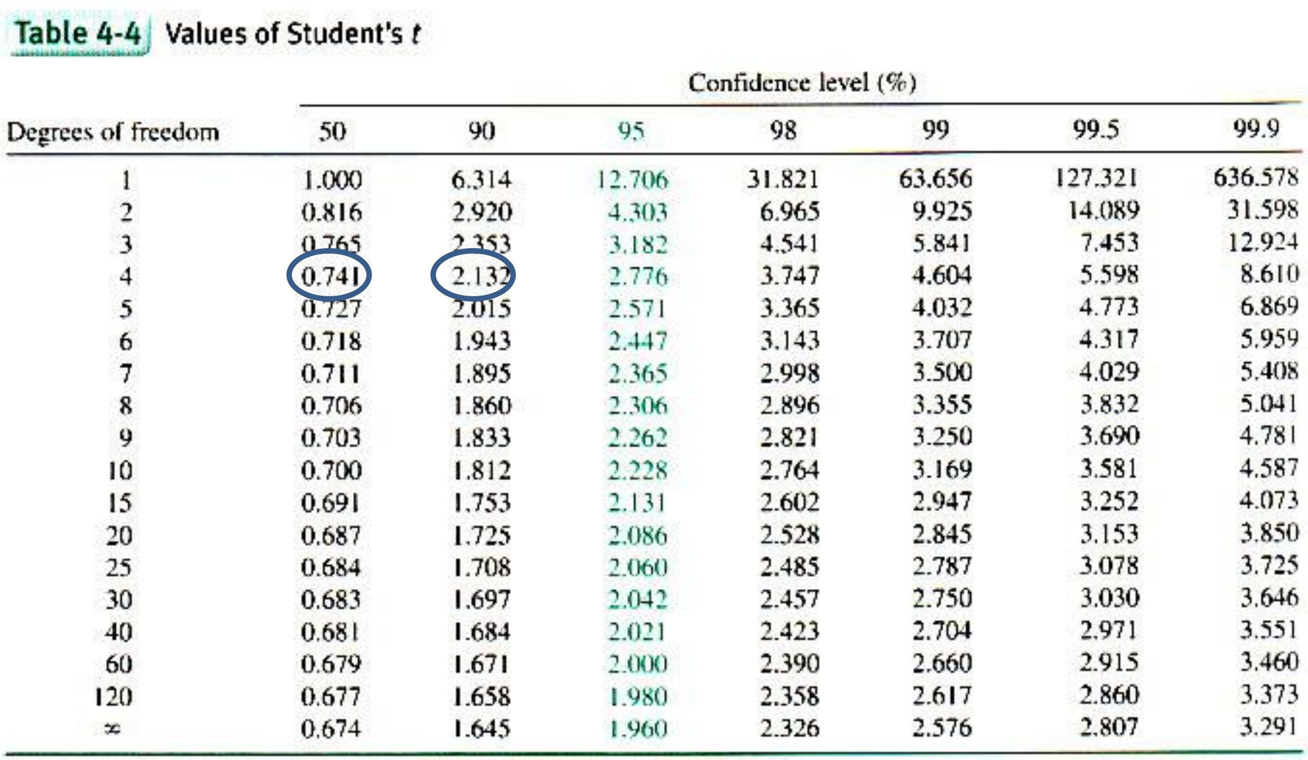


Table 4-4 Values of Student’s *t*

	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

This would be your first step, for example, when comparing data from sample measurements versus controls. One wants to know if there is any difference in the means.

Comparison of Standard Deviations

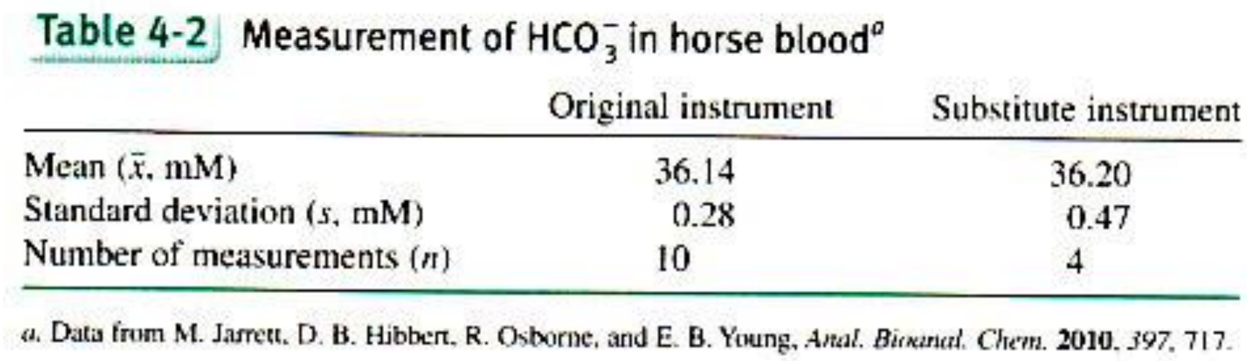


Table 4-2 Measurement of HCO_3^- in horse blood[a]

	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

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

Is *s* from the substitute instrument “significantly” greater than *s* from the original instrument?

F test (Variance test)

$$F = \frac{s_1^{\,2}}{s_2^{\,2}}$$

If $F_{calculated}$ > F_{table}, then the difference is significant.

Make $s_1 > s_2$ so that $F_{calculated} > 1$

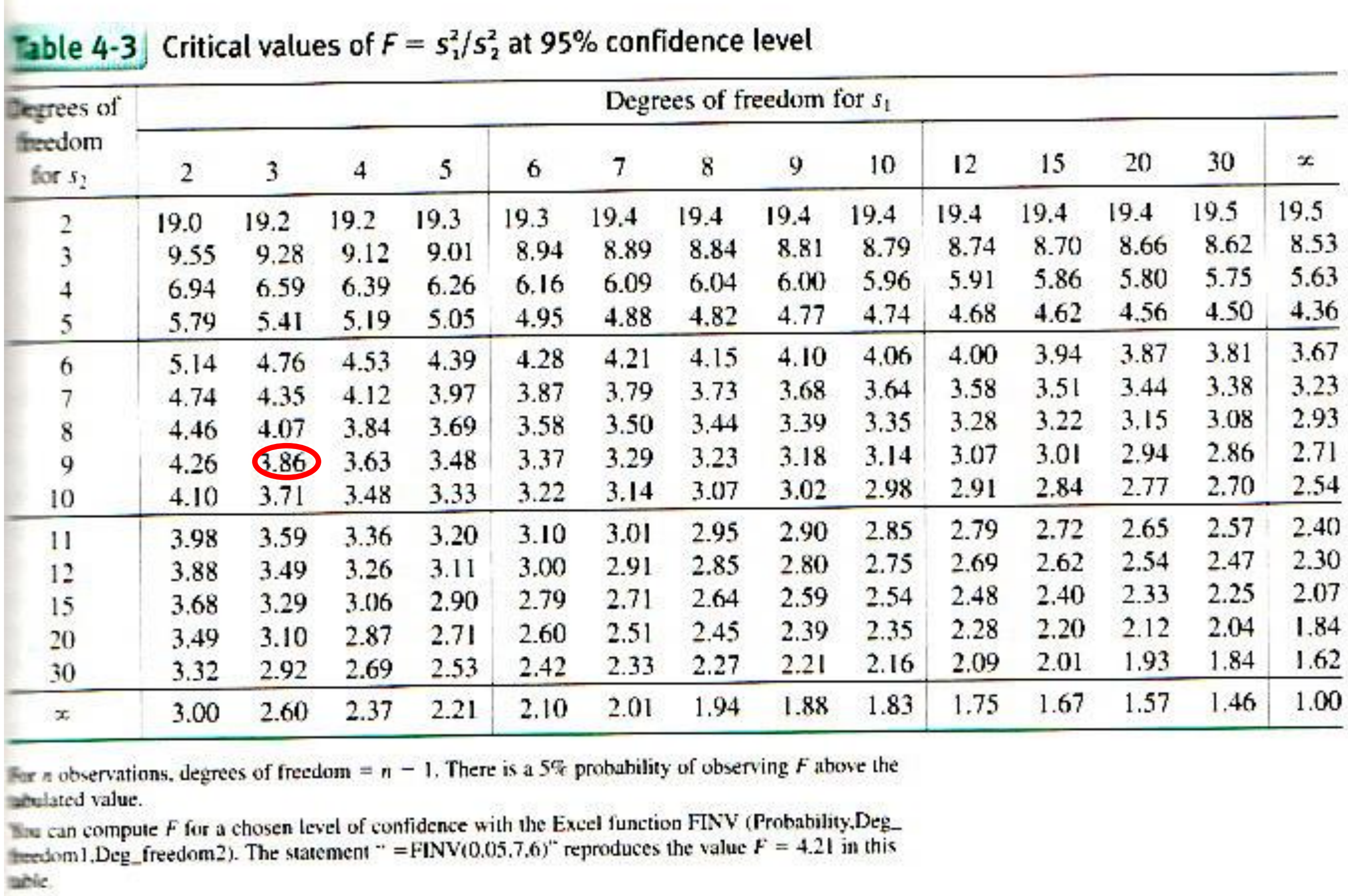


Table 4-3 Critical values of $F = s_1^2/s_2^2$ at 95% confidence level

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
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
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

For n observations, degrees of freedom = $n - 1$. There is a 5% probability of observing F above the tabulated value.

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.

$F_{calculated} = (0.47)^2/(0.28)^2 = 2.8_2$ $F_{calculated}$ $(2.8_2) < F_{table}$ (3.63)

Therefore, we reject the hypothesis that s_1 is signficantly larger than s_2. In other words, at the 95% confidence level, there is no difference between the two standard deviations.

Hypothesis Testing

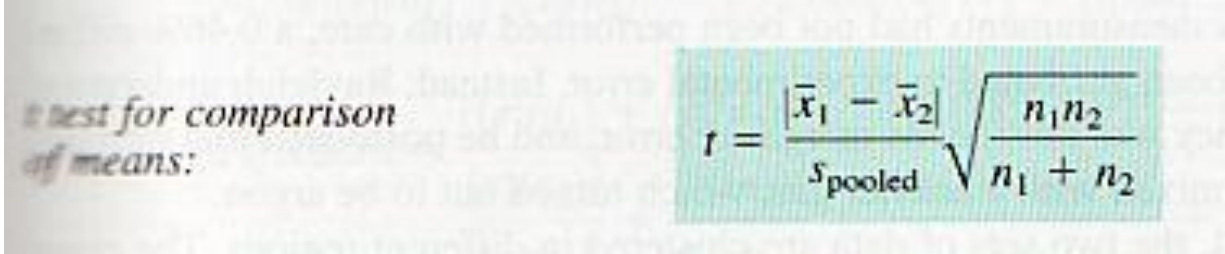
Desire to be as accurate and precise as possible. Systematic errors reduce accuracy of a measurement. Random error reduces precision.

The practice of science involves formulating and testing hypotheses, statements that are capable of being proven false using a test of observed data. The **null hypothesis** typically corresponds to a general or default position. For example, the null hypothesis might be that there is no relationship between two measured phenomena or that a potential treatment has no effect.

In statistical inference of observed data of a scientific experiment, the **null hypothesis** refers to a general or default position: that there is no relationship (no difference) between two measured phenomena, or that a potential medical treatment has no effect. Rejecting or disproving the null hypothesis – and thus concluding that there are grounds for believing that there is a relationship between two phenomena (there is a difference in values) or that a potential treatment has a measurable effect – is a central task in the modern practice of science, and gives a precise sense in which a claim is capable of being proven false.

This would be the second step in the comparison of values after a decision is made regarding the F –test.

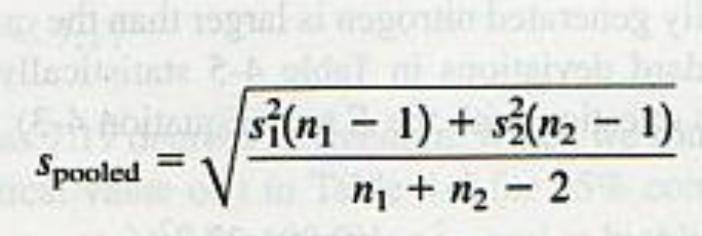
Comparison of Means



test for comparison of means:

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

This *t* test is used when standard deviations are <u>not</u> significantly different.!!!



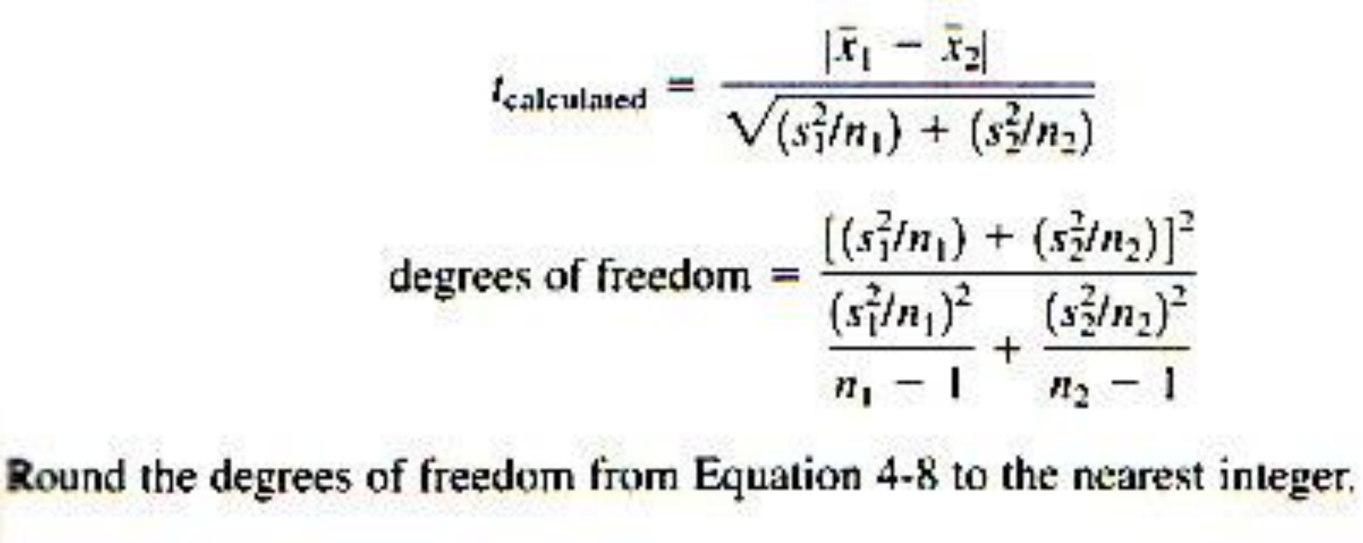
$$s_{\text{pooled}} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

s_{pooled} is a “pooled” standard deviation making use of both sets of data.

If $t_{\text{calculated}} > t_{\text{table}}$ (95%), the difference between the two means is statistically significant!

Comparison of Means

This *t* test is used when standard deviations are significantly different!!!



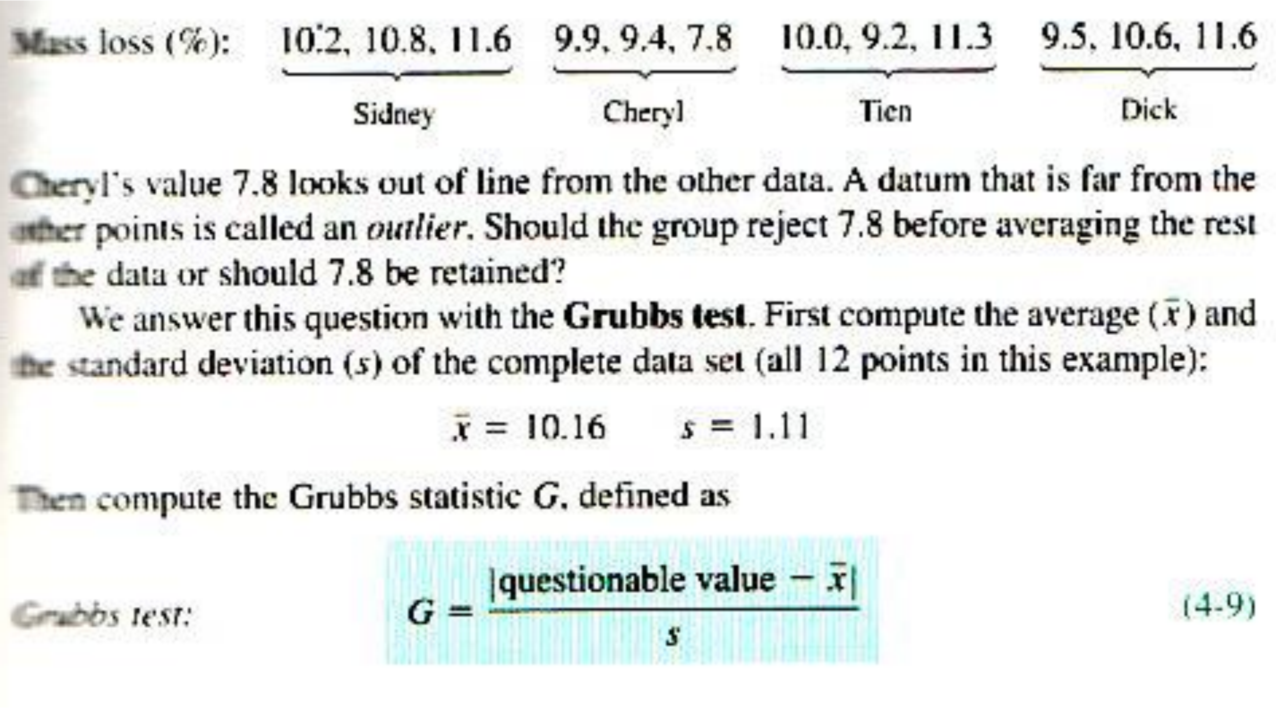
$$t_{\text{calculated}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

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

Round the degrees of freedom from Equation 4-8 to the nearest integer.

If $t_{\text{calculated}} > t_{\text{table}}$ (95%), the difference between the two means is statistically significant!

Grubbs Test for Outlier (Data Point)



Mass loss (%): $\underbrace{10.2,\ 10.8,\ 11.6}_{\text{Sidney}}$ $\underbrace{9.9,\ 9.4,\ 7.8}_{\text{Cheryl}}$ $\underbrace{10.0,\ 9.2,\ 11.3}_{\text{Tien}}$ $\underbrace{9.5,\ 10.6,\ 11.6}_{\text{Dick}}$

Cheryl's value 7.8 looks out of line from the other data. A datum that is far from the other points is called an *outlier*. Should the group reject 7.8 before averaging the rest of the data or should 7.8 be retained?

We answer this question with the **Grubbs test**. First compute the average ($\bar{x}$) and the standard deviation (s) of the complete data set (all 12 points in this example):

$$\bar{x} = 10.16 \qquad s = 1.11$$

Then compute the Grubbs statistic G, defined as

Grubbs test:

$$G = \frac{|\text{questionable value} - \bar{x}|}{s} \tag{4-9}$$

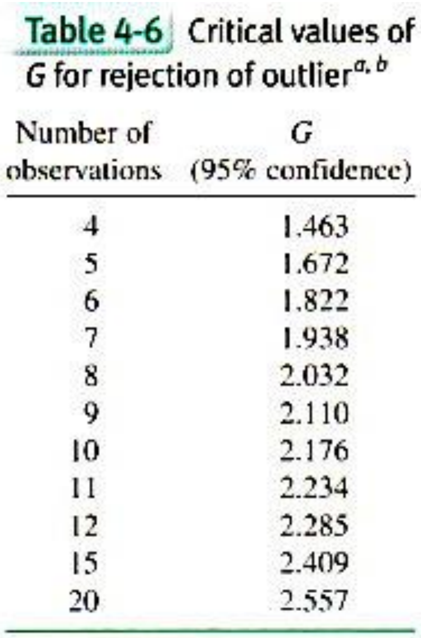


Table 4-6 Critical values of *G* for rejection of outlier[a, b]

Number of observations	*G* (95% confidence)
4	1.463
5	1.672
6	1.822
7	1.938
8	2.032
9	2.110
10	2.176
11	2.234
12	2.285
15	2.409
20	2.557

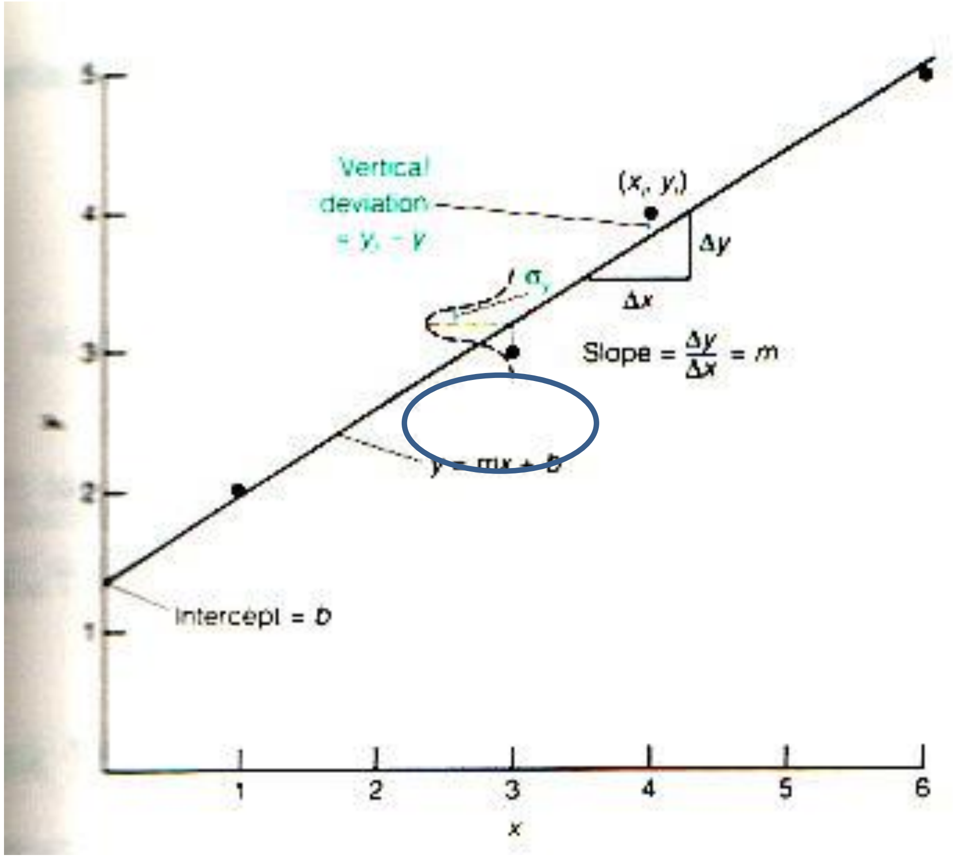
If $G_{calculated} > G_{table}$, then the questionable value should be discarded!

$G_{calculated} = 2.13$ G_{table} (12 observations) = 2.285

Value of 7.8 should be retained in the data set.

Linear Regression Analysis

The method *of least squares* finds the "best" straight line through experimental data.



Linear Regression Analysis

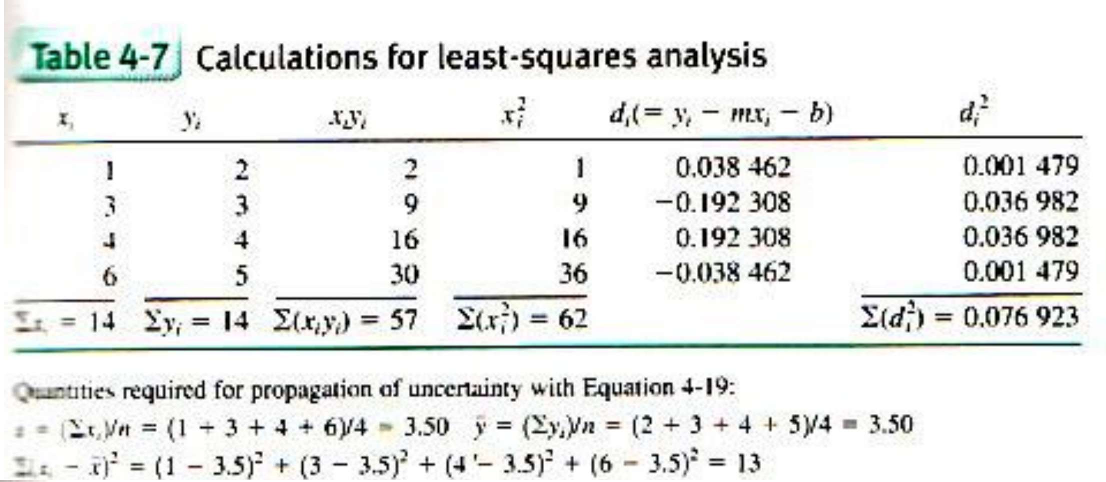


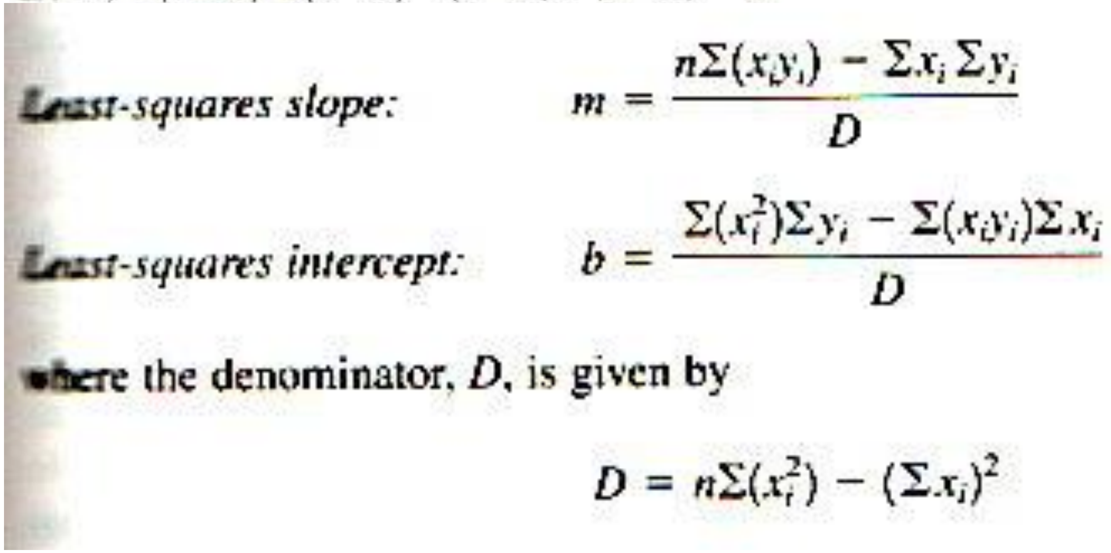
Table 4-7 Calculations for least-squares analysis

x_i	y_i	x_iy_i	x_i^2	$d_i(= y_i - mx_i - b)$	d_i^2
1	2	2	1	0.038 462	0.001 479
3	3	9	9	−0.192 308	0.036 982
4	4	16	16	0.192 308	0.036 982
6	5	30	36	−0.038 462	0.001 479
$\Sigma x_i = 14$	$\Sigma y_i = 14$	$\Sigma(x_iy_i) = 57$	$\Sigma(x_i^2) = 62$		$\Sigma(d_i^2) = 0.076\ 923$

Quantities required for propagation of uncertainty with Equation 4-19:

$\bar{x} = (\Sigma x_i)/n = (1 + 3 + 4 + 6)/4 = 3.50$ $\quad \bar{y} = (\Sigma y_i)/n = (2 + 3 + 4 + 5)/4 = 3.50$

$\Sigma(x_i - \bar{x})^2 = (1 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (6 - 3.5)^2 = 13$



Least-squares slope:

$$m = \frac{n\Sigma(x_iy_i) - \Sigma x_i \Sigma y_i}{D}$$

Least-squares intercept:

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

where the denominator, D, is given by

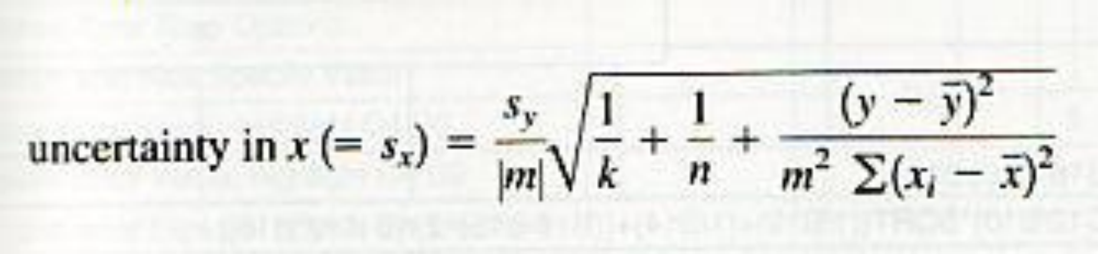
$$D = n\Sigma(x_i^2) - (\Sigma x_i)^2$$

Variability in *m* and *b* can be calculated. The first decimal place of the standard deviation in the value is the last significant digit of the slope or intercept.

Use Regression Equation to Calculate Unknown Concentration

y (background corrected signal) = m x (concentration) + b

x = (y - b)/m



$$\text{uncertainty in } x\ (= s_x) = \frac{s_y}{|m|}\sqrt{\frac{1}{k} + \frac{1}{n} + \frac{(y - \bar{y})^2}{m^2\ \Sigma(x_i - \bar{x})^2}}$$

Report x ± uncertainty in x

s_y is the standard deviation of *y*.
k is the number of replicate measurements of the unknown.
n is the number of data points in the calibration line.
y (bar) is the mean value of *y* for the points on the calibration line.
x_i are the individual values of *x* for the points on the calibration line.
x (bar) is the mean value of *x* for the points on the calibration line.