

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

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 uncertainty!!

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:

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

s = standard deviation of the mean
N = number of measurements or points
 x_i = each individual measurement

\bar{x} = sample mean

RSD = (s/mean) x 100

Coefficient of variance = (s)²

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

24.75 ± 2.07

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:*

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

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

$$C.L. = \pm(2.776) \frac{1.9}{\sqrt{5}} = 2.4$$

$$24.75 \pm 2.56$$

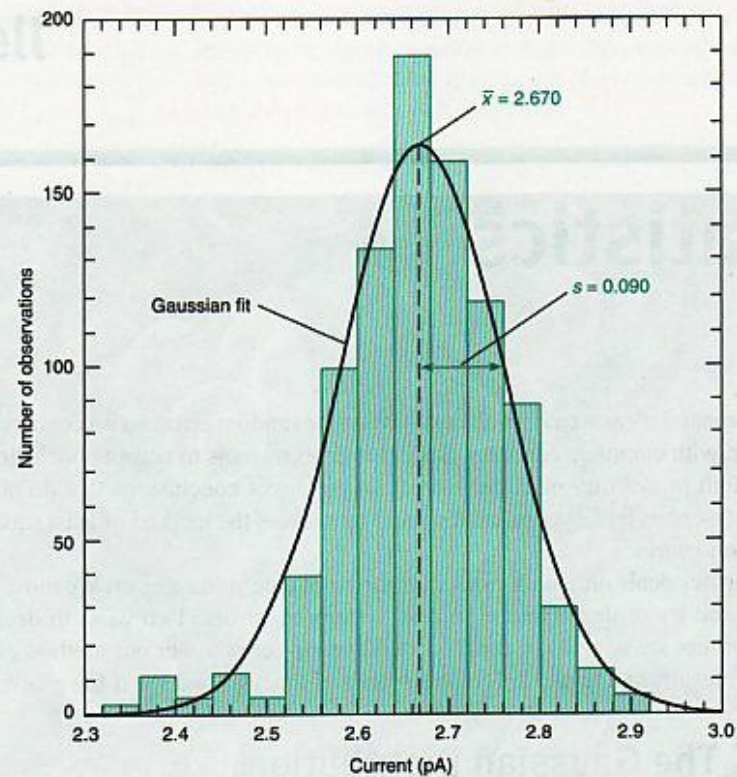
Population versus Sample Size

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.

$\mu \pm 1\sigma$	68.3%
$\mu \pm 2\sigma$	95.5%
$\mu \pm 3\sigma$	99.7%

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.]



Median = middle number in 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 .

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

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

<u>N</u>	Q_{crit}		
	<u>90% CL</u>	<u>95% CL</u>	<u>99% CL</u>
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

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

Using the Q_{crit} table, we see that $Q_{\text{crit}} = 0.466$. Since $Q_{\text{exp}} < Q_{\text{crit}}$, you must keep 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.

$$G_{\text{exp}} = \frac{|x_q - \bar{x}|}{s}$$

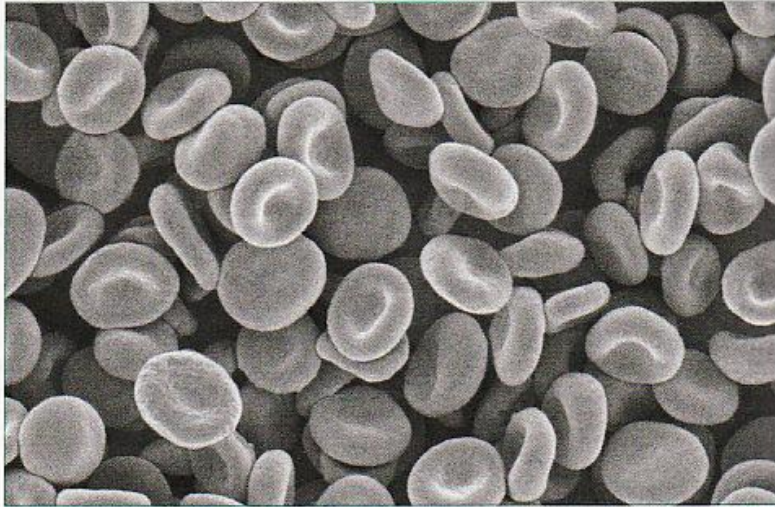
Table: Critical Rejection Values for Identifying an Outlier: G-test

<u>N</u>	G_{crit}		
	<u>90% C.L.</u>	<u>95% C.L.</u>	<u>99% C.L.</u>
3	1.153	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	5.6 × 10 ⁶ cells/μL
5.3	
4.8	
5.4	
5.2	

$$\text{Mean} = 5.1 \times 10^6 \text{ counts}$$

$$s = 1.9 \times 10^6 \text{ counts}$$

$$5.1 \pm 1.9 (\times 10^6 \text{ counts}) \text{ std. dev.}$$

$$C.L. = \pm(2.776) \frac{1.9}{\sqrt{5}} = 2.4$$

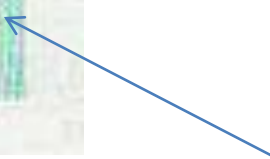
$$5.1 \pm 2.4 (\times 10^6 \text{ counts}) \text{ 95\% C.L.}$$

5.6 (× 10⁶ counts) is 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:

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


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

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

Confidence interval

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 $12.5_4 \pm 0.1_3$ (12.4₁ to 12.6₇). There is a 90% chance that μ lies in the range $12.5_4 \pm 0.3_8$ (12.1₆ to 12.9₂).

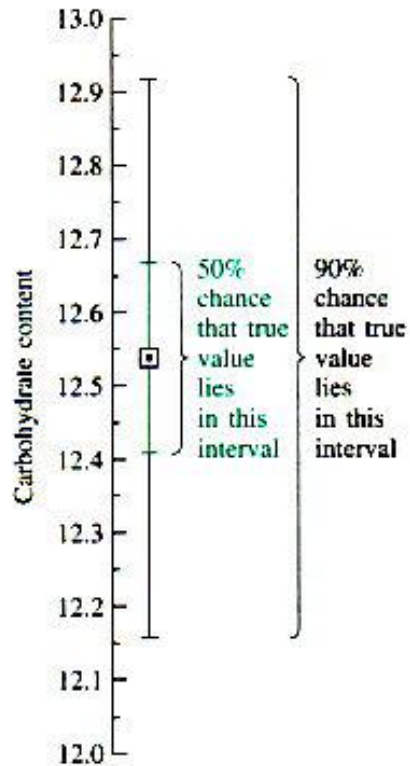


Table 4-4 Values of Student's *t*

Degrees of freedom	Confidence level (%)						
	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