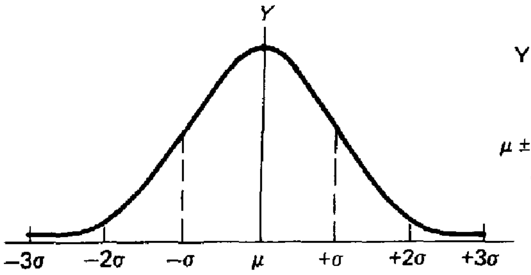
BASIC STATISTICS

(maybe even simplified)

Make a zillion measurements, each with its own indeterminate error (noise), and the measurements might distribute normally (gaussian) around the average value, $\bar{x}$.



Normal (Gaussian) distribution.

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

$\mu \pm \sigma$ = 68% of area under curve
$\mu \pm 1.96\sigma$ = 95% of area under curve
$\mu \pm 3\sigma$ = 99.7% of area under curve

In the absence of determinate error, $\bar{x}$ approaches the true value for the population mean, μ. In a population, the measurements will distribute about μ ± 1, 2, and 3σ. σ is called the standard deviation and points equidistant apart which contain a certain portion of the area under the curve as shown in the diagram. If the number of measurement is small such that the standard deviation of the sample is not necessarily the same as the population from which the sample is taken, the symbol *s* is used instead if only to differentiate it from the population standard deviation. When N (the number of measurements) is large (>20) $\bar{x} \rightarrow \mu$ and $s \rightarrow \sigma$. In small data sets, *s* may or may not approximate σ.

But what is s (or σ)? Is variance (where s and σ come from) an important word? How are they calculated? How can these be used for anything?

Read on and we'll analyze a real (almost) set of data. Then we will report the result properly.

The experiment...

Determine the density of soft glass by measuring the volume of water displaced by a mass of the glass.

The Results...

… of 4 experiments

d (g/mL)
2.21
1.98
2.01
2.36

Now what?

Get the average, of course!

$$\boxed{\bar{x} = \frac{\sum x}{N}}$$

d (g/mL)
2.21
1.98
2.01
2.36

$\bar{d}$ = 2.14

This is good. But, you say, this is an average of values. Assuming it is a good average, how could you show the individual error in each value?

d	$\bar{d}$
2.21	2.14
1.98	2.14
2.01	2.14
2.36	2.14

How about calculating <u>absolute variation from the mean</u> for each value...

d	$\bar{d}$	$d - \bar{d}$
2.21	2.14	0.07
1.98	2.14	-0.16
2.01	2.14	-0.13
2.36	2.14	0.22

The sum of all the variations from the mean, then, should give a good indication of the variability in the data, right?

d	$\bar{d}$	$d - \bar{d}$
2.21	2.14	0.07
1.98	2.14	-0.16
2.01	2.14	-0.13
2.36	2.14	0.22
	Σ =	0.00

Whoops. What happened?

Since we assumed only *indeterminate error* in each density value then the sum of the absolute errors <u>should</u> add to zero.

So how can we get an idea of the magnitude of the variation in the data? One possibility is the average of the absolute deviations (variations). Simply ignore the signs of the variations and take their average...

d	$\bar{d}$	$\left\|d - \bar{d}\right\|$
2.21	2.14	0.07
1.98	2.14	0.16
2.01	2.14	0.13
2.36	2.14	0.22
	Σ =	0.58

average deviation = 0.19 (@3 df)

Why 3 degrees of freedom (df)? In statistics we nearly always subtract 1 dof each time we use the same data for a new statistical value. Rule-of-thumb: dof = N - # of times same data is used. Some examples of this will follow.

Problems with average deviation:

1) Average deviation overestimates distribution of large dataset taken from the same sample, and
2) taking absolute value (a step function of sorts) rarely distributes normally.

So...how can we get rid of signs (without taking the absolute value) that will be mathematically sound *and* distribute normally?

While you're thinking on that, we'll do a short song and dance...

You've probably got it by now... square the deviations:

d	$\bar{d}$	$d - \bar{d}$	$\left(d - \bar{d}\right)^2$
2.21	2.14	0.07	0.0049
1.98	2.14	-0.16	0.0256
2.01	2.14	-0.13	0.0169
2.36	2.14	0.22	0.0484

The signs are effectively neutralized and we could always get back by taking the square root (in fact we will in just a little bit).

Oh yeah... Since these are individual variations we need the average:

d	$\bar{d}$	$d - \bar{d}$	$\left(d - \bar{d}\right)^2$
2.21	2.14	0.07	0.0049
1.98	2.14	-0.16	0.0256
2.01	2.14	-0.13	0.0169
2.36	2.14	0.22	0.0484
		Σ =	0.0958
		average deviation =	0.0319 $^{g^2}/_{mL^2}$

The average of the squares of the variations is call the <u>variance</u>. It's a sign-corrected indication of the distribution of the data. Its symbol is s^2 (sample variance) or σ^2 (population variance). Notice that we divided by 3 (not 4).

This because the df of the data set is *N*-1 in calculating variance. Now look at the units - ${}^{g^2}\!/_{mL^2}$. They doesn't match the original measurements. So what can we do?

Take the square root. The square root of the average of the squared variations is called the <u>standard deviation</u> and given the symbol s (sample standard deviation) or σ (population standard deviation).

$$\text{standard deviation} = \sqrt{0.0319\ {}^{g^2}\!/_{mL^2}} = 0.179\ {}^{g}\!/_{mL}$$

Taking what we've done and combining it, then, into one mathematical package...

$$\boxed{\text{Sample standard deviation} = \sqrt{\frac{\sum_{1}^{N}(x_i - \bar{x})^2}{N - 1}}}$$

When the sample size get sufficiently large, then s → σ and the equation becomes

$$\boxed{\text{Population standard deviation} = \sqrt{\frac{\sum_{1}^{N}(x_i - \bar{x})^2}{N}}}$$

Now our answer has the correct units and is directly comparable to the average...

d = 2.14 g/mL

s = 0.18 g/mL   (No more significant figures than SF's to the right of the decimal. In fact, since the standard deviation is in the tenth's place, the value deviates in the tenths place and the hundredths place is not really significant. The value should be reported as *d* = 2.1 g/mL; s = 0.2 g/mL.

N = 4   It is important when using standard deviation to report N for comparison or for use in further statistical calculations

One way of reporting the data then is to say the actual density, μ, is somewhere in the range of 2.14 ± 3(0.18) g/mL with 99% confidence (i.e. 3 s). This is not the best way of reporting the data though.

Another way to look at it is with <u>relative standard deviation</u>

density = $2.14 \pm \frac{0.18}{2.14}$, N=4 or density = 2.14 g/mL ± 0.08 CV , N=4 (CV = coefficient of variance)

or

density = $2.14 \pm \frac{0.18}{2.14}$x100, N=4 or density = 2.14 g/mL (± 8%RSD)

or

density = $2.14 \pm \frac{0.18}{2.14}$x1000, N=4 or density = 2.14 g/mL (± 80 ppt)   (ppt RSD)

Finally, an acceptable (albeit suspect) method to report data (especially for comparison) is standard error.

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

$$\text{density} = 2.14 \pm \frac{0.18}{\sqrt{4}} = 2.14 \pm 0.09 \text{ g/mL}$$

or with confidence limits we can say that μ is inside of

$$2.14 \pm 3(0.09) \text{ g/mL} \quad \text{or} \quad 2.14 \pm 0.27 \text{ g/mL (99\%)}$$

Note that the above discussion is about precision. If we are interested in the statistical confidence we may place on the data - that is, the probability that a particular result is in error due to noise or that the value is indeed not representative of the population - then more sophisticated statistics are in order.

We can get a realistic value (statistically speaking) of the confidence limits by using Student's (Gosset's)-*t* if the data set is small or σ for a population is unknown. *z*-values may be used when σ for a population is known. Values of *t* and *z* are tabulated in many reference books.

Report results using Student's-*t* or *z*-values as follows:

$$\boxed{\bar{x} \pm \frac{ts}{\sqrt{N}}}$$

calculate *s*, get *t* from table at desired CL

$$\boxed{\bar{x} \pm \frac{z\sigma}{\sqrt{N}}}$$

calculate or know σ, get *z* from table at desired CL

This most certainly not a complete description of statistics you must know as a chemist. For additional reading to help you understand the use of statistics in chemistry see *Practical Statistics for Analytical Chemists,* Robert A. Anderson, Von Nostrand Reinhold, 1987.