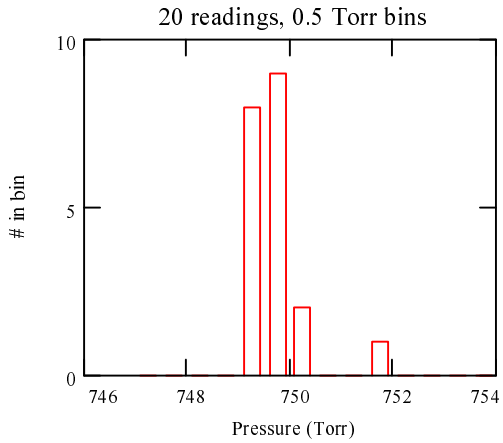
Part 2. Error Distributions and Error Functions

1. Getting the Error Distribution from Experiment. If you measure some quantity (e.g., the barometric pressure) several times (say, twenty), you will probably get several different answers. A histogram of such readings is shown below

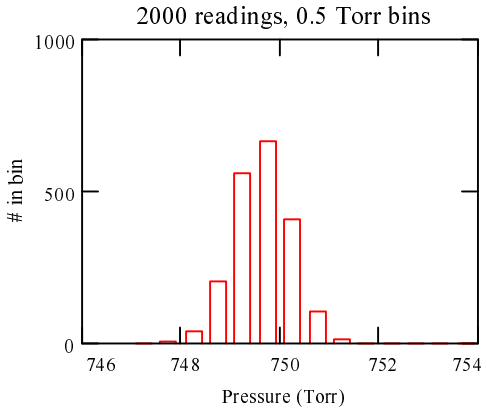


There are several parameters you can calculate from your observed distribution of values. You can calculate estimates of the average and standard deviation of the distribution

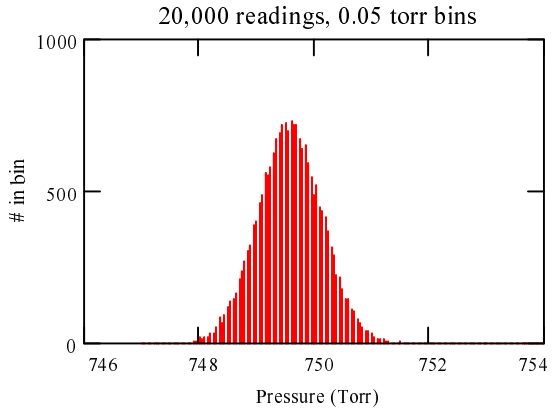
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (20), \text{ and}$$

$$S = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \qquad (21).$$

Those are parameters that apply to your particular sample. If you repeat the experiment, the new values of $\bar{x}$ and S will (probably) not be the same, in fact the distribution would probably not be exactly the same, even if the conditions (temperature in room, atmospheric conditions, etc.) were the same. If you took 2,000 readings, the histogram might look like this



If you took 20,000 readings and reduced the bin size by a factor of ten you begin to get a good picture of the underlying distribution function. The histogram might look like the following



What relation do the $\bar{x}$ and S have to the actual barometric pressure? If each measurement is just like any other so far as we know, then some unknown mechanism is making changes in our values before we get them. We assume that the probability of obtaining a certain value b_i in any one trial is given by a probability distribution, $P(b_i)$. This distribution is not known to us, but might be expected to look something like the figure above. If you could make an infinite number of measurements with infinitesimally small bin sizes a continuous distribution would result. The continuous distribution P(b) has the definition that P(b)db gives the probability that a measurement will give a result in the range [b, b+db], i.e a differential form of the bins above. So, the probability that a measured value will be between b_1 and b_2 is

$$P(b \in [b_1, b_2]) = \int_{b_1}^{b_2} P(b)db \tag{22}$$

If we do the experiment, we get *some* answer; so the underlying continuous distribution must be normalized:

$$\int_{-\infty}^{\infty} P(b)db = 1 \tag{23}$$

The distribution P(b) controls the probability of getting a particular answer on any one experimental trial, and is called the *parent distribution*. The distribution actually obtained by the experimenter is called the *sample distribution*. The set of values actually obtained in an experiment is simply one of very many possible sets. That idea underlies all statistical analysis of data. The "likelihood" of each possible set is controlled by the parent distribution. The sample distribution will become more and more similar to the parent distribution as the number of samples become larger, approaching equality as $N \to \infty$. If we knew the parent distribution P(b), we would know the true answer and there would be no reason to do the experiments. So the real trick in experiments is to try to

guess the parent distribution from the available samples. Often, Greek letters are used to represent parameters of the parent distribution (mean μ, standard deviation σ, etc.) and Roman letters are used for the sample distribution ($\overline{x}$, S, etc.). Unlike real experiments, the histograms above were generated from a normal or Gaussian distribution, i.e. a Bell curve. Before examining the properties of the normal distribution, let's first spend some time discussing the characteristics of probability distributions in general.

2 Things to Do with Probability Distributions. If we have a quantity x described by a probability distribution P(x) or $P(x_i)$, then the following can be calculated:

2.1 Normalization

$$1 = \begin{cases} \int_{-\infty}^{\infty} P(x)dx & \text{continuous} \\ \sum_{i=1}^{N} P(x_i) & \text{discreet} \end{cases} \tag{24}.$$

2.2 Average value of a variable

$$< x >= \begin{cases} \int_{-\infty}^{\infty} xP(x)dx & \text{continuous} \\ \sum_{i=1}^{N} x_i P(x_i) & \text{discreet} \end{cases} \tag{25}.$$

2.3 Average value of a function of a variable

$$< f(x) >= \begin{cases} \int_{-\infty}^{\infty} f(x)P(x)dx & \text{continuous} \\ \sum_{i=1}^{N} f(x_i)P(x_i) & \text{discreet} \end{cases} \tag{26}.$$

3. A Discreet Distribution: Possible Results in Throws of Two Dice. All the possible outcomes of a toss of two dice are enumerated in the following Table. Note that the distribution is normalized; the sum of the fractions in the right column is 1. The average throw is $2\frac{1}{36}+3\frac{1}{18}+4\frac{1}{12}+5\frac{1}{9}+6\frac{5}{36}+7\frac{1}{6}+8\frac{5}{36}+9\frac{1}{9}+10\frac{1}{12}+11\frac{1}{18}+12\frac{1}{36}=7$. The average *squared* throw is $4\frac{1}{36}+9\frac{1}{18}+16\frac{1}{12}+25\frac{1}{9}+36\frac{5}{36}+49\frac{1}{6}+64\frac{5}{36}+81\frac{1}{9}+100\frac{1}{12}+$

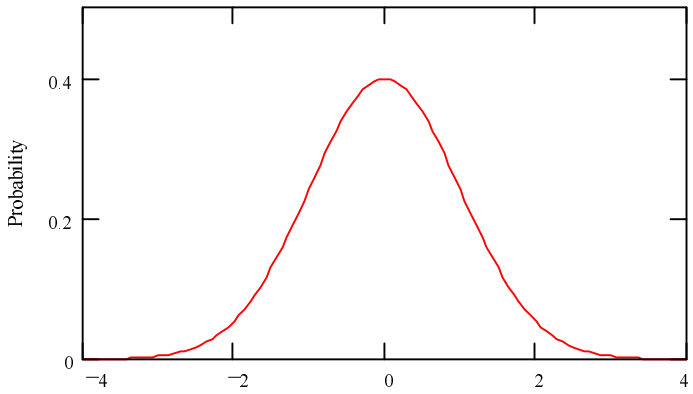
Result	Combinations	# Comb.	Prob.
2	1+1	1	1/36
3	1+2,2+1	2	1/18
4	1+3,2+2,3+1	3	1/12
5	1+4,2+3,3+2,4+1	4	1/9
6	1+5,2+4,3+3,4+2,5+1	5	5/36
7	1+6,2+5,3+4,4+3,5+2,6+1	6	1/6
8	2+6,3+5,4+4,5+3,6+2	5	5/36
9	3+6,4+5,5+4,6+3	4	1/9
10	4+6,5+5,6+4	3	1/12
11	5+6,6+5	2	1/18
12	6+6	1	1/36

$121\frac{1}{18}+144\frac{1}{36}=54.8\overline{3}$. Note that the average squared throw is not 49.

4 The Normal or Gaussian Distribution. A particularly important continuous distribution is the normal, or Gaussian, distribution, given by

$$P(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]} \qquad (27),$$

where x is the independent variable and μ and σ are parameters describing the distribution. The normal distribution is shown in the figure below for σ=1, i.e. for deviations in units of standard deviations.



The height at the maximum is $1/(\sigma\sqrt{2\pi})$. The ratio of the height at positions ±σ away from the center to the height at the center is

$$\frac{P(\mu\pm\sigma)}{P(\mu)}=e^{-1/2}=0.6065 \qquad (28).$$

So the range of ±1 standard deviation about the mean is defined by the points at which the distribution of deviations dies away to ~61% of its maximum height. Another perhaps more useful observation concerns the fraction of measurements expected to fall within ±1 standard deviation about the mean, which we get by integrating P(x) from μ-σ to μ+σ

$$P(x\in[\mu-\sigma,\mu+\sigma])=\int_{\mu-\sigma}^{\mu+\sigma}P(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{\mu-\sigma}^{\mu+\sigma}e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}dx=\frac{1}{\sqrt{2\pi}}\int_{-1}^{1}e^{-\frac{1}{2}u^2}du\,(\,) \qquad (29),$$

where the change of variable u = (x-μ)/σ, du = dx/σ has been made. This integral cannot be done analytically for limits other than [0,±∞] and [-∞,∞]. A Table of its values for different limits is given below as generated by the Mathcad equations.

$$i := 0..30 \quad j := 0..9 \qquad z_{i,j} := \frac{i}{10} + \frac{j}{100} \qquad P_{i,j} := \frac{1}{\sqrt{2\cdot\pi}} \cdot \int_{-z_{i,j}}^{z_{i,j}} \exp\left(-\frac{1}{2}\cdot u^2\right) du \qquad (30).$$

The value of z is the sum of values in the first column and first row. Integration over

Integral of the Gaussian Distribution P vs. z where $z=|x-\mu|/\sigma$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0	0.0080	0.0160	0.0239	0.0319	0.0399	0.0478	0.0558	0.0638	0.0717
0.1	0.0797	0.0876	0.0955	0.1034	0.1113	0.1192	0.1271	0.135	0.1428	0.1507
0.2	0.1585	0.1663	0.1741	0.1819	0.1897	0.1974	0.2051	0.2128	0.2205	0.2282
0.3	0.2358	0.2434	0.251	0.2586	0.2661	0.2737	0.2812	0.2886	0.2961	0.3035
0.4	0.3108	0.3182	0.3255	0.3328	0.3401	0.3473	0.3545	0.3616	0.3688	0.3759
0.5	0.3829	0.3899	0.3969	0.4039	0.4108	0.4177	0.4245	0.4313	0.4381	0.4448
0.6	0.4515	0.4581	0.4647	0.4713	0.4778	0.4843	0.4907	0.4971	0.5035	0.5098
0.7	0.5161	0.5223	0.5285	0.5346	0.5407	0.5467	0.5527	0.5587	0.5646	0.5705
0.8	0.5763	0.5821	0.5878	0.5935	0.5991	0.6047	0.6102	0.6157	0.6211	0.6265
0.9	0.6319	0.6372	0.6424	0.6476	0.6528	0.6579	0.6629	0.668	0.6729	0.6778
1.0	0.6827	0.6875	0.6923	0.697	0.7017	0.7063	0.7109	0.7154	0.7199	0.7243
1.1	0.7287	0.733	0.7373	0.7415	0.7457	0.7499	0.754	0.758	0.762	0.766
1.2	0.7699	0.7737	0.7775	0.7813	0.785	0.7887	0.7923	0.7959	0.7995	0.8029
1.3	0.8064	0.8098	0.8132	0.8165	0.8198	0.823	0.8262	0.8293	0.8324	0.8355
1.4	0.8385	0.8415	0.8444	0.8473	0.8501	0.8529	0.8557	0.8584	0.8611	0.8638
1.5	0.8664	0.869	0.8715	0.874	0.8764	0.8789	0.8812	0.8836	0.8859	0.8882
1.6	0.8904	0.8926	0.8948	0.8969	0.899	0.9011	0.9031	0.9051	0.907	0.909
1.7	0.9109	0.9127	0.9146	0.9164	0.9181	0.9199	0.9216	0.9233	0.9249	0.9265
1.8	0.9281	0.9297	0.9312	0.9328	0.9342	0.9357	0.9371	0.9385	0.9399	0.9412
1.9	0.9426	0.9439	0.9451	0.9464	0.9476	0.9488	0.95	0.9512	0.9523	0.9534
2.0	0.9545	0.9556	0.9566	0.9576	0.9586	0.9596	0.9606	0.9615	0.9625	0.9634
2.1	0.9643	0.9651	0.966	0.9668	0.9676	0.9684	0.9692	0.97	0.9707	0.9715
2.2	0.9722	0.9729	0.9736	0.9743	0.9749	0.9756	0.9762	0.9768	0.9774	0.978
2.3	0.9786	0.9791	0.9797	0.9802	0.9807	0.9812	0.9817	0.9822	0.9827	0.9832
2.4	0.9836	0.984	0.9845	0.9849	0.9853	0.9857	0.9861	0.9865	0.9869	0.9872
2.5	0.9876	0.9879	0.9883	0.9886	0.9889	0.9892	0.9895	0.9898	0.9901	0.9904
2.6	0.9907	0.9909	0.9912	0.9915	0.9917	0.992	0.9922	0.9924	0.9926	0.9929
2.7	0.9931	0.9933	0.9935	0.9937	0.9939	0.994	0.9942	0.9944	0.9946	0.9947
2.8	0.9949	0.995	0.9952	0.9953	0.9955	0.9956	0.9958	0.9959	0.996	0.9961
2.9	0.9963	0.9964	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972
	0.9973	0.9974	0.9975	0.9976	0.9976	0.9977	0.9978	0.9979	0.9979	0.998

the range of ± 1 standard deviation about the mean gives a value of 0.68269 (underlined in table), so ~68% of the samples taken in a normally distributed experiment should lie within $\pm\sigma$, of the mean. The corresponding percentage for ± 2 standard deviations is 95.45%.

We now have a simple example of a *confidence interval*: if samples are taken from a probability distribution which is normal with a mean μ and standard deviation σ, within what interval will 90% of the samples fall? Look in the table for a value of z which gives an integrated probability of 0.9. We find that $\int_{-z}^{z} P(u)du \geq 0.90$ for $z \geq 1.65$, so we can say that 90% of the values will fall between $\mu-1.65\sigma$ and $\mu+1.65\sigma$. This gives the 90% confidence interval for a *single measurement*, and assumes that we know both μ and σ ahead of time.

Finally, we are back to the point of trying to estimate the characteristics of the parent distribution from the observed data. Our sample distribution is simply one of many possible ones, though we happened not to find the others. But it's all we have. At this point we have several options. Sometimes the best is to use computer-based

techniques known collectively as Monte Carlo methods. Classical statistical analysis is based on a different procedure: we *assume* a particular form of parent distribution, and then use the "maximum likelihood principle" to find the parameters of the parent distribution which give the largest probability that the observed data set would occur.

Sometimes the form of parent distribution one should assume is well known from the type of experiment. For example, in experiments which involve counting the number of events (photons arriving at a detector, decays observed from a radioactive sample, etc.) occurring in different time intervals, the Poisson distribution is usually appropriate. More often, the sources of random error in the experiment are not well known, an implicit appeal to the central limit theorem is made, and the normal distribution is assumed to apply. If we assume that the parent distribution is normal, then it is easy to show (B&R pp. 53-55) that the best estimate of the mean μ of the parent distribution is just $\bar{x}$, the mean of the sample distribution

4.1 Error in the Mean Value for Normal Distributions. Now we want to ask the question: How certain are we of the true value? If the sample distribution looks like the third Fig. in section **3**, then it seems like we know the value of the mean to considerably better precision than at 68% confidence (±S). The sample standard deviation S gives the spread of *individual measurements*, so that approximately 68% of them will be within S of the mean. That is true no matter how many measurements we take. However, the value of the mean becomes better and better determined as more measurements are made.

Imagine that we make M sets of measurements, with N measurements in each set. Index the different sets with u, and measurements inside a set with i. (We have MN measurements in all.) Define some notation:

x_{ui}	i^{th} measurement in set u
$\bar{x}_u$	mean of set u
$\bar{X}$	mean of all measurements
$d_{ui} = x_{ui} - \bar{X}$	deviation of individual measurement from total mean
$D_u = \bar{x}_u - \bar{X}$	deviation of mean of set from total mean (31).

The variance of the individual measurements is begin

$$\sigma^2 = \frac{1}{MN}\sum_{u=1}^{M}\sum_{i=1}^{N} d_{ui}^2 \qquad (32).$$

The variance of the means is $$\sigma_m^2 = \frac{1}{M}\sum_{u=1}^{M} D_u^2 \qquad (33).$$

Look at D_u: $$D_u = \bar{x}_u - \bar{X} = \left[\frac{1}{N}\sum_{i=1}^{N} x_{ui}\right] - \bar{X} = \frac{1}{N}\sum_{i=1}^{N}(x_{ui} - \bar{X}) = \frac{1}{N}\sum_{i=1}^{N} d_{ui} \qquad (34),$$

that is, the deviation of mean $\bar{X}_u$ from $\bar{X}$ is the mean of the individual deviations in set u. Inserting that into the expression for variance of the means (two back) gives,

$$\sigma_m^2 = \frac{1}{M}\sum_{u=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N} d[ui]\right)^2 = \frac{1}{MN^2}\sum_{u=1}^{M}\left(\sum_{i=1}^{N} d_{ui}\right)^2 \qquad (35).$$

To get any farther, we must assume that the individual deviations d_{ui} are symmetric about zero; that is, deviations are equally probable above and below the true value. In that case, the squared sum simplifies. Think about the different terms:

$$\left(\sum_{i=1}^{N} d_{ui}\right)^2 = d_{u1}^2 + d_{u2}^2 + d_{u3}^2 + \ldots + d_{uN}^2 \quad + \quad 2d_{u1}d_{u2} + 2d_{u1}d_{u3} + \ldots + 2d_{u2}d_{u3} + \ldots \qquad (36).$$

The d_{ui}^2 terms will always be positive. The "cross terms", though, will be positive sometimes and negative sometimes; in the limit of many measurements, they will add to 0. So

$$\left(\sum_{i=1}^{N} d_{ui}\right)^2 \approx \sum_{i=1}^{N} d_{ui}^2 \,. \qquad (37),$$

if N is large and the deviations are symmetrically distributed. With that simplification, we have

$$\sigma_m^2 \approx \frac{1}{MN^2}\sum_{u=1}^{M}\sum_{i=1}^{N} d_{ui}^2 \qquad (38).$$

Comparing this to the formula for σ^2 to the first equation in this section, we see

$$\sigma_m^2 = \frac{\sigma^2}{N} \qquad \sigma_m = \frac{\sigma}{\sqrt{N}} \qquad (39).$$

(This derivation comes from Chap. 12 of Young.) SGN gives (p. 43) a shorter but harder to understand derivation in the case that N is not so large, getting the expected result that

$$S_m = \frac{S}{\sqrt{N}} \qquad (40).$$

5. Student-t Distribution (Small N). Now, in our hypothetical collection of M sets of measurements, how are the observed $\overline{X_u}$ distributed? In the case where N is large, so that S gives a good approximation to σ (the std. deviation of the parent distribution), the means $\overline{X_u}$ are distributed normally, with std. deviation $\sigma_m = \sigma/\sqrt{N}$ as just shown. If N is rather small, say $N \leq 30$, then we cannot really make the approximation $N \to \infty$, so that S is not as good approximation of σ. In that case (see SGN p. 45), the means have a different distribution called the Student-t distribution. If N is small, this distribution is

rather wide; our single sample might in fact have a mean $\bar{x}$ rather far from the true mean μ. As N becomes large, the Student-t, distribution looks more and more like the normal distribution. The Student-t distribution is given by

$$P(\tau) = k_{norm}\left(1+\frac{\tau^2}{N-1}\right)^{-N/2} \qquad (41),$$

where $\tau = (\bar{x} - \mu) / S_m$, and k_{norm} is a normalization constant. (The expression for k_{norm} is given in equation (30) on p. 45 of SGN 5th edition, and equation (31) on p. 47 of the 6th edition.) Confidence limits are found with this distribution the same way they are found for the normal distribution. With a table of integrals of the distribution, integration limits are chosen which make the integral equal to the desired fraction (.95 for 95%, and so on.) Table 2 on p. 46 of SGN 5th ed. or Table 3 on p. 49 of SGN 6th ed. gives values of t which give various confidence intervals:

$$0.95 = \int_{\bar{x}-tS_m}^{\bar{x}+tS_m} P(\tau)d\tau \qquad (42),$$

for a 95% confidence limit, and so on. Keep in mind what this confidence limit actually means: if you were to repeat the experiment many times, and each time claim that the mean μ was within the limits $\bar{x} \pm tS_m$ you obtained on that particular repetition, you expect to be right 95% (or whatever) of the time.

6. Rejection of Data. What do you do if one measurement in a set of supposedly identical ones seems quite different from the others, so that you suspect that you made a mistake somehow? First, you look to see if there is evidence of a problem - an error in arithmetic done in the notebook, for example. Barring that, you have two choices to make: 1) In the absence of any real evidence to indicate a problem with the "outlier", you might decide that it simply should be kept. You may want to report the median instead of, or in addition to, the mean as your measurement of a central value. The median is much less susceptible to the influence of outlier points. 2) You may apply a statistical test, discarding the outlier value if there is less than some critical probability that it came from the same parent distribution as the others.

The second option - applying a statistical test - is much more palatable if you actually know something about the parent distribution for your experiment. If you have only a few experimental points, and you don't have some good theoretical reason to postulate a particular form of parent distribution, it is almost certainly safer to retain the outlier point and report both the mean and median.

6.1 Never Reject Data without a Good, Defendable Reason. A data point that can not be fit by a certain theory may not fit because a mistake of some sort was made. On the other hand, the data may not fit because of inadequacies in the theory. This is one way that new phenomena are discovered. Do not be the one who Qtests a new discovery away. If you must reject a point and still don't know why, it should still be reported in

the data set, even if not included in the analysis. If you must reject some data point, the following methods are recommended; but they may not be applied iteratively. Only one point may be rejected by these methods. If more than one point is outlier, find out why.

6.2 Q Test. This is extremely easy to perform, and works well for samples with small numbers of points. It *assumes* that the experiment is governed by the normal distribution, but it does not assume that a good estimate of σ is available. It is not applicable to experiments with nonnormal error distributions, such as counting experiments with small numbers of counts per sample. To perform a Q test, calculate the value of Q for your sample:

$$Q = \frac{|x_{suspect} - x_{closest}|}{|x_{highest} - x_{lowest}|} \qquad (43).$$

(The suspect value will, of course, be either the largest or smallest of the set.) If the value of Q is larger than the critical value given in Table 1 of SGN, you may discard the suspect value. You should then recalculate $\overline{x}$, S_m, and the confidence interval based on the new (smaller) number of observations.

Note that the Q-test can be applied *only once* to a data set. If you have more than one screwball value, you must either live with them or redo the experiment. In routine practice, the Q test at 90% confidence is acceptable. For particularly important observations, it is important to decide in advance what criteria will be used for data rejection. Deciding in advance is crucial to eliminate bias. Background information on the Q test can be found in R. B. Dean and W. J. Dixon, *Analytical Chemistry* **23**, 636 (1951), and Dixon and Massey, *Introduction to Statistical Analysis*, 3rd ed.(McGraw-Hill, 1969).

6.3 Chauvenet's Criterion. A second statistical test is *Chauvenet's criterion*. It has the advantage that it can be used for any form of parent distribution, but the disadvantage that the parent distribution *and its parameters* must be known. (Note that the Q-test did not require a good value of σ.) The idea is simple: a data point should be rejected if the parent distribution predicts that fewer than half an event should appear which deviates as much from the mean as the questionable experimental point. Fractions smaller than 1/2 may also be used, but should be agreed upon beforehand as discussed above.

For example, suppose that you have performed a particular measurement in an automated way, and made 120 measurements on a sample. The histogram of your measurements indicates that the experiment does have a normal parent distribution. The mean of the measured values is 12.637 and the standard deviation S, which should be a good estimate of σ since you have 120 samples, is 0.058. You have one point at 12.872 which you suspect to be erroneous. Should you reject that point?

You must calculate the number of measurements which are expected to lie at least as far afield as 12.872. To do that, first calculate the fraction of measurements expected to lie closer than that to the mean using the gaussian distribution. The limit z is $(x-\mu)/\sigma$ = (12.872-12.637)/0.058 = 4.05. Using the integral of the normal distribution, we find that

the integral of the normal distribution from -4.05 to 4.05 is approximately 0.99995 (interpolating between 4.0 and 4.1). The fraction expected to be outside that range is therefore 1-0.99995 = 0.00005 (per event basis), so we should expect (120)(0.00005) = .006 of an event to be as far from the mean as 12.872. Since 0.006<1/2, we reject that point.

The criterion is used in exactly the same way with other distribution functions. You must know enough about your parent distribution to calculate (or look up) the relevant integral. For experiments where you have only a few measurements and no *a priori* information about the parent distribution, the Q test is better. If you have at least 10 points, or you do know something about the parent distribution, Chauvenet's criterion is a good choice.