the problem. Any attempt to base the initial belief on guesswork or instinct must be unscientific and unreliable. The only strict way to justify an initial degree of belief is by the equally likely method introduced in section 7.1.3. As we saw, this does not work in a continuous case.

★ 7.1.6 Conclusions on Probability

Thus probability can be considered as the limit of a frequency, as an objective number or as a subjective degree of belief. This has been a very quick look at a very deep subject, and you should be aware that there are serious differences even within these camps. Beware, too, of names: some people refer to the frequency definition of probability as 'objective', the Bayesians call the frequentists 'classical', and the frequentists call the equally likely school 'classical'.

Why have we opened this can of worms? There is no point in arguing the claims of rival schools: you can adopt whatever definition you please, and use arguments about the merits of different definitions as an amusing conversation topic. What matters is that you should be aware of what you are doing, and do not mix up thoughts, ideas, and formulae from the different definitions.

Most scientists, if challenged, would claim to belong to the frequency school. Propensities and Bayesian statistics are strictly unorthodox and heretical. However, although we claim to adopt the frequency definition, in our innermost hearts we probably think of probabilities as objective numbers, and often talk in language appropriate to Bayesian probabilities. In particular, any attempt to interpret the results of an experiment falls into the trap of repeatability.

Suppose you measure the mass of the electron as 520 ± 10 keV/c^2. This is a clear statement; you have obtained a result of 520 keV/c^2 with an apparatus of known resolution 10 keV/c^2. You may then say, on the basis of your value, that 'the mass of the electron probably lies close to 520 keV/c^2' or even make the more numerically detailed statement that 'the value lies between 510 and 530, with a 68% probability'. Either statement is, in von Mises' view, 'unscientific' and incompatible with your claimed adherence to the frequency definition. The electron has just one mass (it happens to be 511 keV/c^2) and it either lies within your error bar or outside it.

Such statements are really using subjective, Bayesian, arguments. Before the experiment you know nothing about m_e, so you consider all possibilities equally likely. Now, having made a measurement m of resolution σ, so that

$$p(m|m_e) \propto e^{-(m-m_e)^2/2\sigma^2}$$

Bayes' theorem turns this round to say

$$p(m_e|m) \propto e^{-(m-m_e)^2/2\sigma^2}.$$

If you want to do this, then that is fine, but do it with your eyes open. The conclusion rests on the initial uniform distribution which, as stressed earlier, is not automatic. You could have interpreted this as a measure of m_e^2, about which you are equally ignorant initially, and assumed that all values of m_e^2 (instead of m_e) are equally likely. Then you would get a different result.

To discuss such experimental results and confidence levels in the frequency interpretation, one is forced into a slightly contorted view-point. This is described in the next section.

> *'That's a great deal to make one word mean',*
> *Alice said in a thoughtful tone.*
> *'When I make a word do a lot of work like that',*
> *said Humpty Dumpty, 'I always pay it extra.'*
>
> *—Lewis Carroll*

> *We* warn *the reader that there is no universal*
> *convention for the term 'confidence level.'*
>
> *—The Review of Particle Properties, 1986*

★ 7.2 CONFIDENCE LEVELS

Confidence levels appear as a part of descriptive statistics, as ways of describing the spread of a distribution, especially in the tails. We will look at their definition and properties in this context first; they are very basic and simple. In the next section we go on to the more subtle business of their use in estimation, and the results of measurements.

★ 7.2.1 Confidence Levels in Descriptive Statistics

Suppose cereal packets are produced according to a Gaussian distribution of mean 520 g and standard deviation 10 g. The table of the integrated Gaussian, Table 3.2, then tells us that 68% of packets will weigh more than 510 and less than 530 g. So if we say, when challenged by a consumer group, that the weight of a packet lies in the interval 510 to 530 g, we will be correct 68% of the time. We make the statement with 68% confidence. This

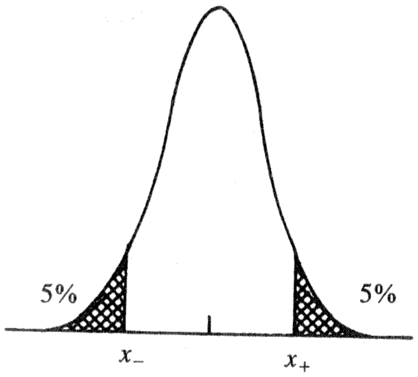

Fig. 7.1. The 90% central confidence interval for a Gaussian distribution.

is a probability according to the standard definition as a frequency limit, as the number of cereal packets coming off the production line is large.

There is a lot of choice about the confidence to quote. Common values are 68% or 1σ, 95.4% (2σ), 90% (1.64σ), 95% (1.96σ) and 99% (2.58σ). There is a trade-off between a narrow interval and low confidence. You can say with great confidence that the weight lies within very wide limits; if you want to tie it down more precisely the confidence lessens. In practice, 90 and 95% (or 2σ) limits are commonly met with; 99% limits are occasionally used by perfectionists.

For non-Gaussian distributions the correspondence listed above between confidence levels and number of σ no longer applies. If someone quotes a '2σ' result for a non-Gaussian distribution they may mean two standard deviations, or they may in fact mean, rather misleadingly, 95.4%. Care is necessary here.

Having chosen the value, there is still a choice over the range. There are three conventional ways of choosing the limits of an interval around the centre. If the limits are x_- and x_+, then, for a given confidence level C they obey the requirement

$$\text{Prob}(x_- \leqslant x \leqslant x_+) = \int_{x_-}^{x_+} P(x)\,\mathrm{d}x = C \tag{7.4}$$

and additional requirements as follows:

1. The symmetric interval: x_- and x_+ are equidistant from the mean μ, i.e. $x_+ - \mu = \mu - x_-$.
2. The shortest interval: the limits are such that the interval is as short as possible, subject to equation 7.4; i.e. $x_+ - x_-$ is a minimum.

3. The central interval: The probabilities above and below the interval are equal, i.e. $\int_{-\infty}^{x_-} P(x)\,\mathrm{d}x = \int_{x_+}^{\infty} P(x)\,\mathrm{d}x = (1-C)/2$.

The central interval is usually the most sensible and the best one to use. However, for the Gaussian distribution (and indeed for any symmetric distribution) the three definitions are all equivalent anyway, so the problem does not often appear.

Two other useful forms are the one-tailed limits, the upper and lower limits. At the stated confidence level, the packet weights lie below the upper limit, i.e.

$$\text{Prob}(x < x_+) = \int_{-\infty}^{x_+} P(x)\,\mathrm{d}x = C \tag{7.5}$$

and one does not care what their weights are at low values. Similarly, the weights lie (at the stated confidence level) above the lower limit

$$\text{Prob}(x > x_-) = \int_{x_-}^{\infty} P(x)\,\mathrm{d}x = C \tag{7.6}$$

and whether they exceed it by a little or a lot is irrelevant.

Careful! It must be emphasized that the upper limit of a 95% central confidence interval, and the 95% upper limit, are not the same thing. The former has 97.5% of the probability content below it and 2.5% above; the latter has 95% below and 5% above.

★7.2.2 Confidence Intervals in Estimation

Suppose we want to know the value of a parameter X, and have estimated it from the data, giving a result x. We know about the resolution of our measurements, and thus $V(x)$ and its square root σ. The problem is to turn our knowledge of x and σ into a statement, of the confidence level type, about the true value X.

The naive answer is to turn it round and say 'X lies within $x - \sigma$ and $x + \sigma$, with 68% confidence, and within $x - 2\sigma$ and $x + 2\sigma$, with 95% confidence'. However, as described in section 7.1.6, this apparently simple statement is dynamite, containing hidden Bayesian assumptions. Anyone still tempted to think in these terms is invited to consider the following example, which shows that applying probabilities like this is just wrong!

Example An impossible probability
The weight of an empty dish is measured as 25.30 ± 0.14 g. A sample of powder is placed on the dish, and the combined weight measured as 25.50 ± 0.14 g. By subtraction, and combination of errors, the weight of the powder is 0.20 ± 0.20 g. This is a perfectly sensible result, though the poor scientist involved should probably find a more accurate balance.

However, look what happens to the probabilities! The naive statement now says that there is a 32% chance of the weight being more than 1σ from the mean, which is evenly split, making a 16% chance that the weight is negative. This, as Euclid used to say, is absurd.

We will now approach the problem more carefully, using the frequency limit definition of probability. For a particular value of X, there is a probability distribution function for x: $P(x; X)$. For a conventional measurement of resolution σ it is a Gaussian for x with mean X and standard deviation σ; for a number of (Poisson) events it is the Poisson formula of x given a mean X. In general, it presumably peaks at or near $x = X$ and falls off to either side. From it we can construct a confidence interval—let us use a 90% central interval—so that, for a particular value of the real X, the value of the measurement x will lie (with 90% probability) within the region x_- to x_+. For a different X, there are different limits. Thus x_- and x_+ can be considered as functions of X. This can be nicely shown on a diagram (Figure 7.2). X runs vertically, and a horizontal line at a particular value of X cuts the curves shown, enabling the values of the limits x_- and x_+ for this X to be read off on the horizontal axis. The region between the two curves is called the *confidence belt*. The key to these plots is that they are *constructed horizontally* before you ever see the data, using the probability distribution $P(x; X)$, and *read vertically* when you have a measurement.

Now you make an actual measurement x. The x_- curve gives the value of X for which x is the appropriate x_-. This is the desired *upper* limit X_+. This is not saying that X has a 5% probability of exceeding X_+—a statement previously condemned as naive and even heretical. It means that if the real X is X_+ or greater, then the probability of getting a measurement smaller than this is 5% or less. Likewise the value of X for which our value of x is x_+ is the lower limit X_-. We therefore quote the 90% confidence interval for the true value X as the range X_- to X_+.

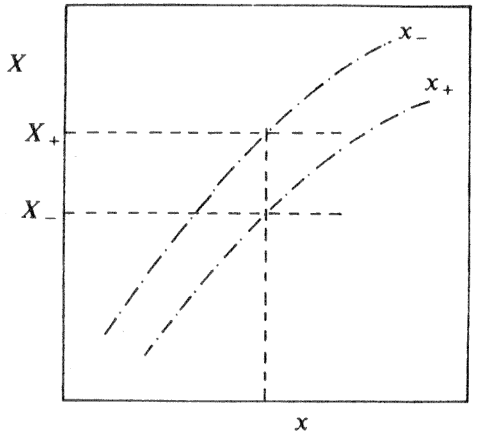

Fig. 7.2. A confidence diagram.

When X_- and X_+ are constructed in this way, we can still say the true value of X lies in the range $X_- \leqslant X \leqslant X_+$ with 90% probability. This looks like a statement about X, but in fact it is a statement about X_+ and X_-. Suppose the true value of X is X_0, and it is measured many times. The many (different) measurements will, by construction, lie within the range x_- to x_+ (inclusive) as evaluated for X_0 in 90% of all cases, while the other 10% will not. Points inside the belt are within their horizontal limits (x_- and x_+ for this X) and also their vertical limits (X_- and X_+ for this x). Points outside violate both bounds. The 90% of measurements within the x_- to x_+ range also have X_0 in the range X_- to X_+. So X_0 will lie within the limits X_- to X_+ with a probability of 90%. Although a particular statement obtained at, say, a 90% confidence level (e.g. m_e lies within 510 and 515 keV/c^2) is either right or wrong, if you take a large number of such statements then 90% of them will be true.

★7.2.3 Confidence Levels from Gaussians

For a Gaussian distribution this conversion from horizontal to vertical is very simple—indeed, deceptively so. Given a measurement x of the mean X, and knowing σ, a 90% confidence interval for X requires the values X_- and X_+ such that (looking at the confidence diagram)

$$\int_x^{\infty} \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-(x'-X_-)^2/2\sigma^2}\mathrm{d}x' = 0.05 = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-(x'-X_+)^2/2\sigma^2}\mathrm{d}x'.$$

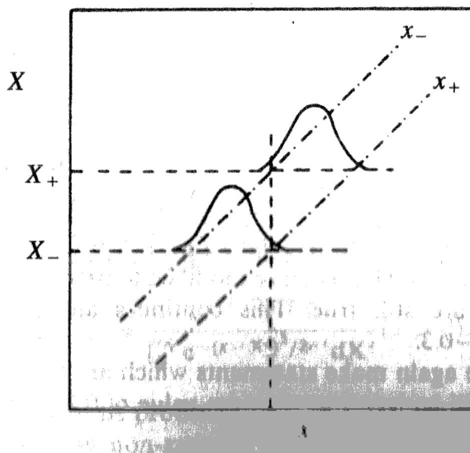

Fig. 7.3. The confidence diagram for a Gaussian

The equation for X_- requires that x lies some number of standard deviations (in this case 1.64) above X_-. This is the same as saying that X_- must lie the same number of σ below the measured x, which can be written in the form

$$\int_{-\infty}^{X_-} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x'-x)^2/2\sigma^2}\,dx' = 0.05$$

and such confidence limits can be found for Gaussian estimators using the usual table of the Gaussian integral. The curves in Figure 7.2 become, in Figure 7.3, two straight lines with unit gradient, $x_\pm = X \pm n\sigma$ when constructed horizontally, $X_\pm = x \pm n\sigma$ when read vertically, where n is 1 for 68% confidence, 1.64 for 90% confidence, etc., as given by the table of integrated Gaussians. The confidence interval for X obtained from a measurement x is merely $x \pm n\sigma$.

In fact, so it seems to us, confidence-interval theory has the defect of its principal virtue: it attains its generality at the price of being unable to incorporate prior knowledge into its statements.

—*Kendall and Stuart*

★7.2.4 Measurement of a Constrained Quantity

Now consider the case where we know that there are definite limits on X, which it would be physically impossible to exceed. Take the mass of an object as an example: irrespective of any measurement, it has to be positive.

We will use a 2σ (95.4%) central interval as an illustration. The true mass has some positive value—let suppose it is 0.1 g. This is measured with a resolution of 0.2 g, so a measurement x gives a confidence interval $x \pm 0.4$ g.

There is a 2.3% probability that the measurement will be greater than 0.5 g. From this we will quote limits which are wrong, but the 2.3% probability for this is part of the game and acceptable.

If the measurement lies in the range 0.4 to 0.5, we will quote similar limits, and this time they will be true. If it falls a bit below 0.4, say to 0.3, the limits are -0.1 to 0.7; the lower limit can be modified to 0.0 on the basis of common sense, and they are still true. This continues all the way down to a measurement of -0.3.

Below -0.3 we again make statements which are false, this time because the upper limit will be less than the true value of 0.1. Again the small 2.3% probability is acceptable.

However, if we get a measurement of -0.5, we have to quote a range of

-0.9 to -0.1. This is patently ridiculous. In using a 95.4% confidence level we know that 4.6% of our statements will be untrue, and accept the odds. We now have independent evidence that this particular statement is one of those 4.6%. In such a case, we would be pretty stupid to make it. However, our confidence level approach based on the frequency distribution can tell us nothing more.

That such measurements give nonsensical limits is obvious. More dangerous are measurements like -0.39. The 2σ limits are then -0.79 to 0.01. Changing the lower limit from -0.79 to 0.0 is permissible, in that it cannot alter the truth or falsehood of the statement. This gives a very narrow interval of 0.00 to 0.01, with 95.4% confidence. To make such a statement is strictly true, and at the same time totally dishonest. Once your quoted confidence interval covers a region of impossible values, you are in trouble.

If you get in a hole like this, Bayesian statistics provides the only means of escape. When faced with a Gaussian measurement x, of a true value X, the Bayesian does not construct any confidence diagram, but invokes equation 7.3. In this equation the conditional p (result|theory) is just the Gaussian distribution probability density for a measurement x arising from a true X with resolution σ. The p(result) in the denominator does not matter as it is taken care of in the final normalisation. p (theory) represents the intrinsic probability distribution for X. Normally this is handled by a rather disingenuous assumption of complete ignorance: nothing is known about X, so all values are equally likely, so the initial $p(X)$ is uniform and constant. Taking care of the normalisation gives

$$p(X\,|\,x) = \frac{e^{-(x-X)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

and confidence levels for X can be constructed as desired, as described in section 7.2.1. They are (because of the symmetry in the Gaussian between X and x) exactly the same as those we obtained using the frequency method.

Our extra knowledge—that X must be positive —is easily incorporated. The initial $p(X)$ is now a step function, zero for $X < 0$ and constant for $X > 0$. Equation 7.3 gives, after normalisation.

$$p(X\,|\,x) = \frac{e^{-(x-X)^2/2\sigma^2}}{\int_0^\infty e^{-(x-X')^2/2\sigma^2}dX'}\,(x > 0). \qquad (7.7)$$

Confidence levels can be produced from this as desired, using Table 3.3. The distribution is now non-symmetric so there is a choice between the symmetric, shortest, and central interval.

Example Bayesian approach to confidence
A mass is measured as -0.5 ± 0.2 g. The integral in the denominator of equation 7.7 is 0.0062, from the probability of exceeding 2.5σ in Table 3.3. This table also gives the probability of exceeding 3.24σ as 0.0006, which is 10% of the previous figure. So the 90% confidence upper limit is $-0.5 + 3.24 \times 0.2 = 0.15$ g.

Although this usage is probably the only way to make meaningful statements from such results, you do so in the knowledge that had you used another variable—X^2 or $\sqrt{X}$ or $1/X$—the resulting limits would be incompatible. Your assumption of complete ignorance means different things when applied to different forms of the same basic variable.

★7.2.5 Binomial Confidence Intervals

For the binomial distribution the observed variable (call it r) is discrete, whereas the 'true' value (call it R) is continuous. For discrete variables the integrals in equations 7.4, 7.5, and 7.6 are replaced by summations. The subtle difference in the inequality signs now matters: the two-tailed form (equation 7.4) is inclusive (r lies within the range r_- to r_+), and the terms for r_- and r_+ are included in the sum, but the one-tailed intervals (equations 7.5 and 7.6) are exclusive (r is less than $r_+, \ldots$) and the term for r_+ or r_- is excluded.

Wishing to form, say, a 95% central confidence interval for a given R it will not in general be possible to choose an r_+ such that $\sum_0^{r_+} P(r; R) = 0.975$. For safety we round r_+ up, and select it such that $\sum_0^{r_+} P(r; R) \geqslant 0.975$, and similarly round r_- down. This means that our final statement will be true at least 95% of the time, and possibly more.

The two confidence diagram curves become staircase-like, as the horizontal coordinate is discrete. Confidence limits can be constructed from the belt as before, with summations replacing integrals. Care over the detail of the definitions is required when fixing the limits of the sums. Thus if m successes are found in n binomial trials, limits on the individual probability p are given by finding p_- and p_+ such that (using the 95% central limits as an example)

$$\sum_{r=m+1}^{n} P(r; p_+, n) = 0.975 \qquad \sum_{r=0}^{m-1} P(r; p_-, n) = 0.975.$$

These are known as the *Clopper–Pearson confidence limits.*

Example A binomial confidence interval
In a sample of 20 fizzgigs, 4 are obloid. What are the 95% confidence limits on the proportion of obloid fizzgigs?

The lower limit is given by $\sum_0^3 P(r; p_-, 20) = 0.975$.

Trial and error show that for $p = 0.057$, the probabilities of 0 to 3 successes are 0.307, 0.373, 0.216, and 0.079, which sum to 0.975.

The upper limit is given by $\sum_5^{20} P(r; p_+, 20) = 0.975$, which is easier to handle as $\sum_0^4 P(r; p_+, 20) = 0.025$.

For $p = 0.437$, the probabilities of 0 to 4 successes are 0.00001, 0.0002, 0.001, and 0.005, and 0.018, which sum to 0.025.

The limits are thus, with 95% confidence, 0.057 to 0.437.

★7.2.6 Poisson Confidence Intervals

If n events are observed from a Poisson process of unknown mean N, the 90% upper limit (for example) is the value N_+ such that

$$\sum_{r=n+1}^{\infty} P(r; N_+) = 0.90 \tag{7.8a}$$

or, equivalently,

$$\sum_{r=0}^{n} P(r; N_+) = 0.10. \tag{7.8b}$$

In English, this means: if the true value of N is really N_+, the probability of getting a result n which is this small (or smaller) is only 10%, and for N larger than N_+ it is even smaller. Thus we say we are '90% confident' that N is not greater than N_+, and averaging over many such statements we will be right 9 times out of 10.

Likewise for the 90% lower limit you require N_- such that

$$\sum_{r=0}^{n-1} P(r; N_-) = 0.90. \tag{7.9}$$

These equations for N_+ and N_- (which, by the way, are real numbers, not integers) can be solved by iteration. Some are given in the following table.

TABLE 7.1.
SOME POISSON LIMITS

	Upper			Lower		
	90%	95%	99%	90%	95%	99%
$n=0$	2.30	3.00	4.61	—	—	—
1	3.89	4.74	6.64	0.11	0.05	0.01
2	5.32	6.30	8.41	0.53	0.36	0.15
3	6.68	7.75	10.05	1.10	0.82	0.44
4	7.99	9.15	11.60	1.74	1.37	0.82
5	9.27	10.51	13.11	2.43	1.97	1.28
6	10.53	11.84	14.57	3.15	2.61	1.79
7	11.77	13.15	16.00	3.89	3.29	2.33
8	12.99	14.43	17.40	4.66	3.98	2.91
9	14.21	15.71	18.78	5.43	4.70	3.51
10	15.41	16.96	20.14	6.22	5.43	4.13

Note that if no events are observed, this can give an upper limit on the ideal number, but no lower limit.

★7.2.7 Several Variables—Confidence Regions

If two (or more) variables are being estimated simultaneously then putting confidence limits on both of them is, in the words of Kendall and Stuart, 'a matter of very considerable difficulty'. One may have to be satisfied with the establishment of a *confidence region* in the parameter space, within which the true parameters lie (with a certain confidence). This is very relevant to maximum likelihood estimation, which provides a natural framework for the estimation of several variables, as discussed in section 5.3.4.

Consider first an ML estimate of a single parameter. As shown in section 5.3.3, in the large N limit the values of the parameter a at which the log likelihood function is 0.5 less than its peak value are the 'one sigma' limits of the estimate. These can be taken as Gaussian measurements and treated as in section 7.2.3, where the 'one sigma' limits were shown to give the 68% confidence interval. Thus the range of a for which the log likelihood in L is within 0.5 of its peak value constitutes the 68% confidence interval. As before, by the argument of invariance, it can also plausibly be taken as such at small N.

For more than one parameter the likelihood function is harder to plot. In the large N limit, surfaces of constant probability are ellipses for two parameters, hyperellipses for more than two. For small N the surfaces are more complicated, but still exist and can, for two parameters, be displayed. One can thus present the ellipse (or whatever) at which $\ln L$ falls off by 0.5, and say that the true parameter values lie within it, at some confidence level.

However, this level is no longer 68%. We have moved from the Gaussian to the multidimensional Gaussian; large values of the exponent are relatively more likely, and are given by the χ^2 distribution, with number of degrees of freedom equal to the number of parameters. So for two parameters the 'one sigma' confidence region gives the 39% confidence region, and for more parameters the level is even less. For a given number of parameters and desired confidence level, the value of χ^2 is found from tables (such as Table 8.1), and the boundary of the confidence region is given by the curve (or surface) at which $\ln L$ falls from its peak value by half this amount. Thus, for example, for two variables the 90% confidence region is given by the parameters for which $\ln L$ is within 2.3 of its maximum value.

★7.3 STUDENT'S *t* DISTRIBUTION

When you make a measurement of known resolution—for example, you measure the weight of a ball-bearing to be 13.5 g, using a balance which is known to have a resolution of 0.1 g—then you quote the answer with its resolution (i.e. 13.5 ± 0.1 g), and the statement is interpreted as a confidence interval for a Gaussian distribution, as discussed in section 7.2.3.

This is fine provided you know the resolution—when your balance comes with a convenient label on it telling you its accuracy. Of course it often does, and usually when making measurements you have established the performance of the apparatus. But sometimes this is not the case. (This is particularly true in the social sciences, where dispersion arises due to a spread in the basic data sample, rather than from measurement. So Student's t is a topic more familiar to doctors and economists than physicists and chemists.)

What do you do then? You have to take several measurements and look at the spread. A single measurement gives you an honest estimate, but tells you nothing about the accuracy. σ is not known *a priori*, but has to be estimated from a sample of several values: we do not have the true value σ, but only the estimate $\hat{\sigma}$. If μ is known we use (cf. equation 5.12)

$$\hat{\sigma} = \sqrt{\overline{(x-\mu)^2}}. \tag{7.10}$$

If μ is unknown we use (cf. equation 5.14)

$$\hat{\sigma} = s = \sqrt{\frac{N}{N-1}\overline{(x-\bar{x})^2}}. \tag{7.11}$$

The second case is more usual, but not universal.

Instead of the variable $(x-\mu)/\sigma$, which is distributed according to a unit Gaussian (i.e. it has mean zero and standard deviation unity), we have to deal with the variable

$$t = \frac{x-\mu}{\hat{\sigma}}. \tag{7.12}$$

t is *not* normally distributed with unit variance, as it would be if $\hat{\sigma}$ were equal to σ; the significance of a given deviation between an x and μ is less when $\hat{\sigma}$ is used in place of σ, because of the additional uncertainty in $\hat{\sigma}$. In practice, especially for small N, it is rather a poor estimate of σ.

t is described by a distribution called *Student's t distribution*, after its discoverer William Gossett, who wrote under the pen name of 'Student'. Writing

$$t = \frac{(x-\mu)/\sigma}{\hat{\sigma}/\sigma}$$

you can see that t is a unit Gaussian divided by a denominator which is (looking at equation. 7.10 or 7.11) the square root of a χ^2 sum. Our ignorance of σ in the numerator cancels our ignorance of σ in the denominator, enabling t to contain only the observed quantities x and $\hat{\sigma}$. n, the number of degrees of freedom in the χ^2, is N if equation 7.10 is used and $N-1$ for equation 7.11