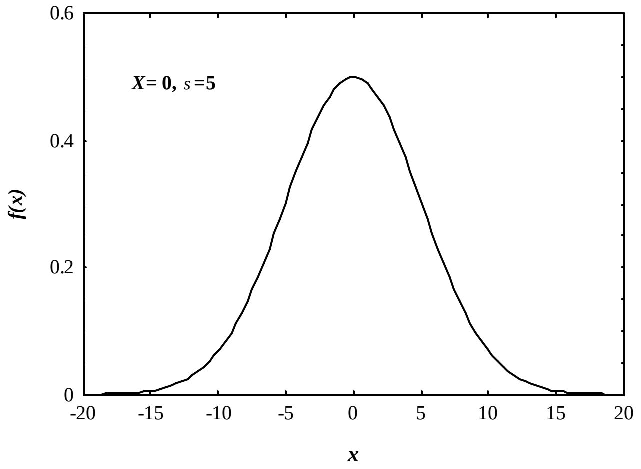
1B40 Practical Skills

Properties of the Gaussian distribution

The (Gaussian) normal distribution has the functional form

$$p(x) = \frac{1}{s\sqrt{2p}} \exp\left[\frac{-(x-m)^2}{2s^2}\right]$$

For simplicity we will set $\mu = 0$.



What fraction of the results lie between -x to +x?

The probability density function *f(x)* is the fraction of results that lie between x and $x + dx$. Define *f (x)* such that:

$$f(x) = \int_{-x}^{x} p(x')\,dx',$$

$$f(x) = \frac{1}{s\sqrt{2p}} \int_{-x}^{+x} \exp\left[-\frac{1}{2}\left(\frac{x'}{s}\right)^2\right] dx'.$$

Let $t = x'/s$ and $z = x/s$ then *f (x)* becomes

$$f(z) = \frac{1}{s\sqrt{2p}} \int_{-z}^{+z} \exp\left[-\frac{t^2}{2}\right] s\ dt.$$

This is symmetric about $t = 0$ and can be written as

$$f(z) = \frac{2}{\sqrt{2p}} \int_{0}^{+z} \exp\left[-\frac{t^2}{2}\right].$$

The function *erf(z)* is referred to as the *Error Function* and is defined by

$$erf(z) = \frac{2}{\sqrt{p}} \int_0^z \exp\left[-t^2\right] dt$$

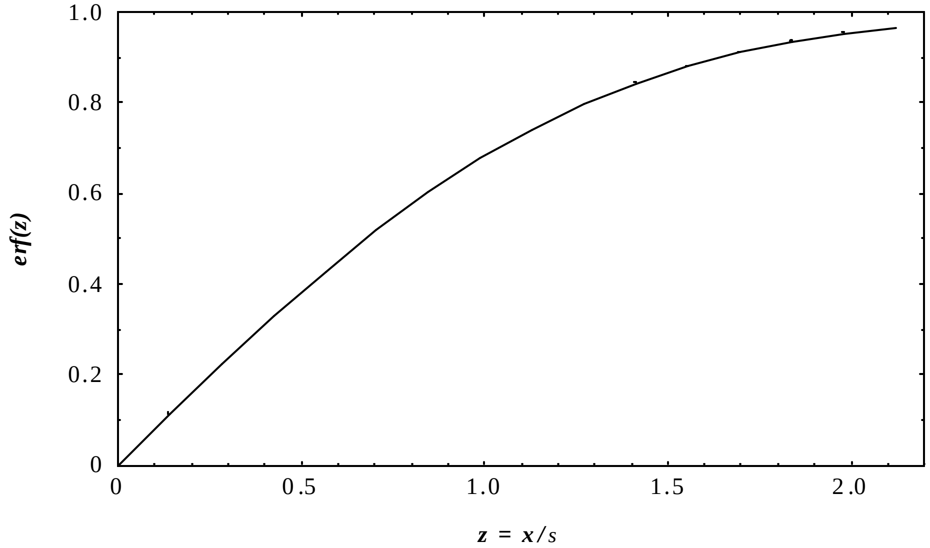
and is tabulated in mathematical tables. Thus we have

$$f(z) = erf\left(\frac{z}{\sqrt{2}}\right).$$

It represents the fraction of *p(x)* that lies within ± *z* standard deviations of the mean of the distribution.

z = x/s	*f(z)*	Approximate fraction of readings outside ±z
0	0	1
1	0.683	1/3
2	0.9540	1/20
3	0.9973	1/400
4	0.9994	1/16000

The erf function is plotted below.



Important points

- Two out of three observations lie <u>within</u> ± σ (one in three outside).
- About one in twenty observations lie outside ± 2σ
- About one in 400 observations lie outside ± 3σ
- About one in 16,000 observations lie outside ± 4σ, i.e. there is one chance in 16,000 that the true value lies outside this range.

You can use these results to check that s has been estimated correctly. Roughly two thirds of the readings should lie between $\overline{x} \pm s$. When you quote a result as $\overline{x} \pm s_m$ you imply the probability that the true value lies in the quoted range is roughly two thirds.

Treatment of suspect results

The discussion above on the properties of the normal distribution can help guide us when we have "suspect" results. Consider the following set of experimental data:

Time t (s)	Residual$\left(t_i - \overline{t}\right)^2$ (10^{-4} s^2)
5.38	25
5.42	1
5.48	25
5.30	169
5.34	81
5.29	196
5.97	2916
5.32	121
5.40	9
mean = 5.43 s	sum=39367
	$\sigma=(39367/9)^{1/2}=2\times10^{-1}$ s
	$\sigma_m=\sigma/3=7\times10^{-2}$ s

The result is $T = (5.43 \pm 0.07)$ s. The results are well clustered about the mean value except for the 5.97 value. If you recalculated the mean omitting this result you get $T = (5.36 \pm 0.02)$ s.

Are you justified in neglecting this one reading? In general you never ignore the result of an experiment in this way. However you can calculate the probability that the suspect result is valid. The standard deviation of the distribution is ±0.20 seconds. (Note in the final result the error quoted is standard error on the mean, not the standard deviation of the sample). The suspect reading is about 3 standard deviations away from the mean; the probability of its being a valid part of the distribution is about one part in 400 - i.e. not impossible, but unlikely!

You may be justified in ignoring a particularly unusual result **BUT ONLY IF YOU CAN SPOT THE SYSTEMATIC ERROR WHICH HAS LED TO THE ODD RESULT.**

You can usually do this as you are doing the experiment. If for example you are determining the value of some well known constant you may sometimes see (with hindsight) that a single unusual measurement has led you to an inaccurate value for the result. Your alternatives are:

1. Repeat the entire experiment taking care to avoid what happened before (Best).
2. Comment on what has happened, noting the likely source of error in your final result (Next best).
3. Remove the single odd measurement and recompute the answer (OK sometimes but least preferable – could be dishonest!).