Statistical Treatment of Experimental Data

Every experimental physicist requires at least an elementary knowledge of statistical methods for interpreting results in a meaningful way. Summarized here are a few of the fundamental formulas for treating data. This is only an introduction, and you are urged to acquire a more detailed understanding of error analysis. Some useful references are:

1. *Data Reduction and Error Analysis for the Physical Sciences*, P. R. Bevington (McGraw-Hill, New York, 1969).

2. *Notes on Statistics for Physicists,* J. Orear (U.S. Atomic Energy Commision, 1958).

3. *Statistical Treatment of Experimental Data*, H. D. Young (McGraw-Hill, 1972).

4. *Errors of Observation and Their Treatment,* J. Topping (Institute of Physics, London, 1955).

These references can be found in the Science Library.

Suppose you make N independent observations of a physical quantity generating a set of N data values, $x_1, x_2, \ldots x_i, \ldots x_N$. The following statistical quantities can be calculated:

The mean value is $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

The deviation of each datum from the mean is

$$d_i = x_i - \bar{x} \quad .$$

The sample mean absolute deviation is

$$\bar{d} = \frac{1}{N-1}\sum_{i=1}^{N} |d_i| \quad .$$

The sample standard deviation is

$$\sigma = \left[\frac{1}{N-1}\sum_{i=1}^{N} d_i^{\,2}\right]^{\frac{1}{2}}$$

while σ^2 is the variance. Note that $\bar{d}$,σ, and σ^2 are always positive.

The standard error of the mean (*i.e.*, the uncertainty of the mean) from a set of N readings is given by the formula

$$\sigma_m = \frac{\sigma}{\sqrt{N}} = \left[\frac{\sum d_i^2}{N(N-1)}\right]^{\frac{1}{2}} \quad .$$

Propagation of errors in a compound quantity $z = f(x, y, \ldots)$ is given by

$$\sigma_z = \left[\sigma_x^2\left(\frac{\partial z}{\partial x}\right)^2 + \sigma_y^2\left(\frac{\partial z}{\partial y}\right)^2 + \ldots\right]^{\frac{1}{2}}$$

assuming the variables are uncorrelated. For example, if $a = b^2 c$, then

$$\sigma_a = \left[(2bc)^2\sigma_b^2 + b^4\sigma_c^2\right]^{\frac{1}{2}}$$

and

$$\frac{\sigma_a}{a} = \left[\left(\frac{2\sigma_b}{b}\right)^2 + \left(\frac{\sigma_c}{c}\right)^2\right]^{\frac{1}{2}} \quad .$$

The Gaussian distribution or normal error function may be thought of as a distribution about a mean of a large number of observations, subject to the random error of measurement. The equation for a Gaussian distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right]$$

where x is the value obtained from a single observation, $\bar{x}$ represents the mean of the distribution, and σ is the standard deviation. Letting $Z = (x-\bar{x})/\sigma$, the fraction of readings between Z and $Z+dZ$ is $f(Z)\,dZ$ where

$$f(Z) = \frac{1}{\sqrt{2\pi}}e^{-Z^2/2} \quad .$$

The fraction of readings between $-Z$ and $+Z$ is

$$O(Z) = \sqrt{2/\pi}\int_0^z e^{-t^2/2}\,dt \quad .$$

Values of Z, f(Z) and $O(Z)$ are given in Table I.

W. P. Beyermann, 1/17/03

Table I The Gaussian Distribution

Z	$f(Z)$	$O(Z)$
0.0	0.399	0.000
0.1	0.397	0.080
0.2	0.391	0.159
0.3	0.381	0.236
0.4	0.368	0.311
0.5	0.352	0.383
0.6	0.333	0.451
0.7	0.312	0.516
0.8	0.290	0.576
0.9	0.266	0.632
1.0	0.242	0.693
1.1	0.218	0.729
1.2	0.194	0.770
1.3	0.171	0.806
1.4	0.150	0.838
1.5	0.130	0.866
1.6	0.111	0.890
1.7	0.094	0.911
1.8	0.079	0.928
1.9	0.066	0.943
2.0	0.054	0.954
2.2	0.035	0.972
2.4	0.022	0.984
2.6	0.014	0.991
2.8	0.008	0.995
3.0	0.0044	0.9973
3.6	0.0009	0.9995
4.0	0.0001	0.99994

If N is large ($N > \sim 100$), σ can be estimated from the full-width at half maximum, FWHM, of the histogram. For a Gaussian distribution, $\sigma = 0.4248$ FWHM.

When N is small ($N \le 4$) proceed with caution. Any formula with such a small number of readings can only give a rough estimate, and common sense may be a better guide. Remember that the error in σ_m is surprisingly large (about 25% for $n = 9$ and about 10% for $n = 50$). Beware of quoting too many significant figures and be cautious with elaborate calculations.

Example

In some of the experiments the detected photons or particles will be counted and stored as a distribution (*e.g.*, the number of photons versus their energy). A raw typical spectrum for the 1.17 MeV emission from ^{60}Co is displayed in Fig. 1. After subtracting off the background, which was determined by a linear fit to a few points at each end, the spectrum is plotted in Fig. 2 as a function of energy.

The total number of counts in the peak is N = 123,227. The following quantities can be calculated from the spectrum.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 1169.952 \text{ keV}$$

$$\sigma = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{\frac{1}{2}} = 1.226 \text{ keV}$$

$$\sigma_m = \frac{\sigma}{\sqrt{N}} = 3.493\times10^{-3} \text{ keV}$$

$$\bar{x} \pm \sigma_m = 1169.952 \pm 0.003 \text{ keV}$$

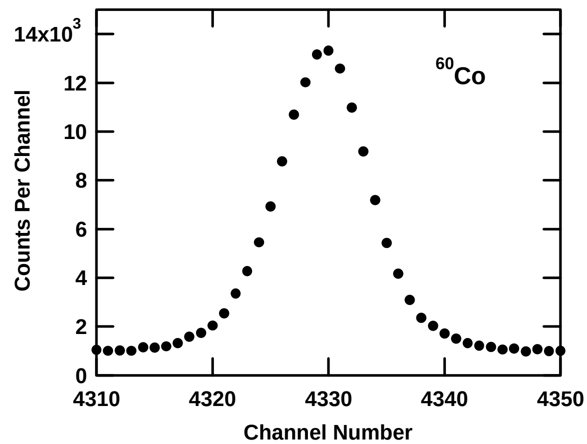

Fig. 1 Counts per channel versus channel number for the 1.17 MeV γ-ray emission from ^{60}Co.

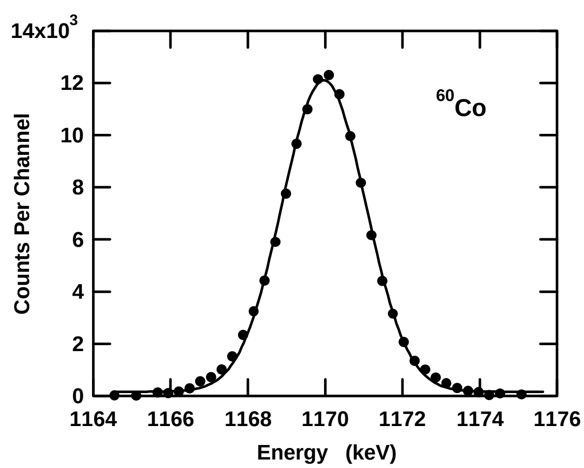

Fig. 2 Counts per channel with the background subtracted as a function of energy for the 1.17 MeV γ-ray emission from ^{60}Co.

This peak should be a Gaussian distribution for very large *N*. The solid line in Fig. 2 is a fit to the data with the Gaussian function

$$f(E) = C_0 + C_1 \exp\left[-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2\right]$$

where C_0 = 166, C_1 = 11940, $\bar{x}$ = 1169.97 keV, σ = 1.08 keV.

The discrepancy in σ is because the distribution is not an exact Gaussian. Finally, σ could be determined from the FWHM.

$$\sigma = 0.425 \times FWHM = 0.425 \times 2.50 \text{ keV} = 1.06 \text{ keV} \quad .$$