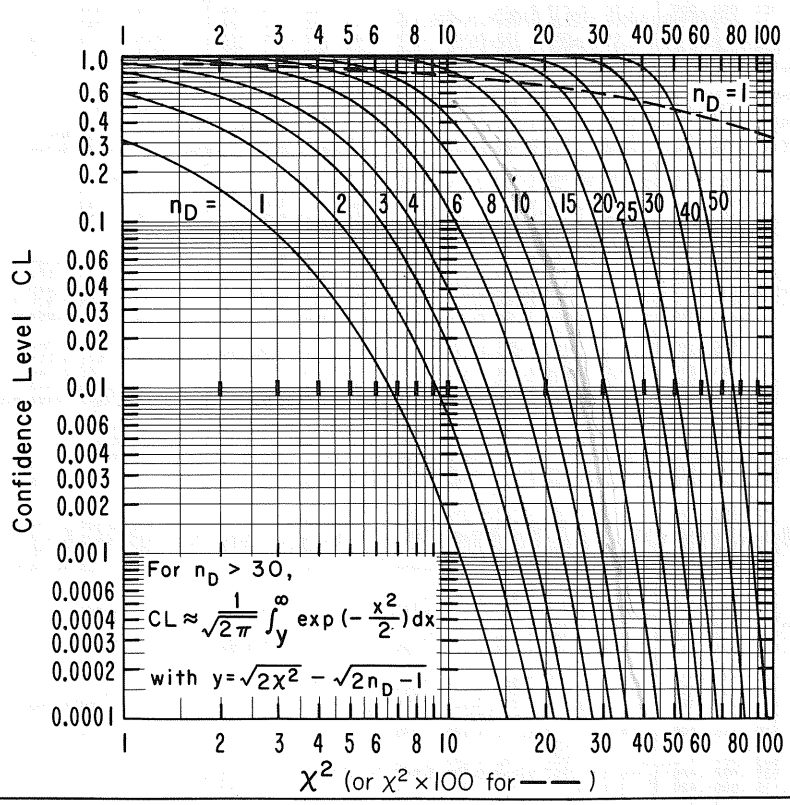
PROBABILITY AND STATISTICS

A. PROBABILITY DISTRIBUTIONS AND CONFIDENCE LEVELS

We give here properties of the three probability distributions most commonly used in high energy physics: Normal (or Gaussian), Chi-squared, and Poisson. We warn the reader that there is no universal convention for the term "confidence level" as used by physicists; thus, explicit definitions are given for each distribution, and we have attempted to choose definitions that correspond to common usage. It is explained below how confidence levels for all three distributions can be extracted from the following figure.



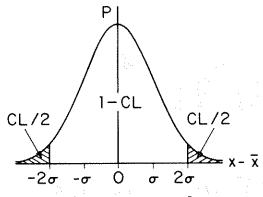
A.1. Normal Distribution

The normal distribution with mean $\bar{x}$ and standard deviation σ (variance σ^2) is:

$$P(x)dx = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\bar{x})^2/2\sigma^2} dx. \qquad (1)$$

The confidence level associated with an observed deviation from the mean, δ, is the probability that $|x-\bar{x}| > \delta$, i.e.,

$$CL = 2\int_{\bar{x}+\delta}^{\infty} dx\, P(x) \qquad (2)$$



[The small figure in Eq. (2) is drawn with $\delta = 2\sigma$.] CL is given by the ordinate of the $n_D = 1$ curve in the figure at $\chi^2 = (\delta/\sigma)^2$. The confidence level for $\delta = 1\sigma$ is 31.7%; 2σ, 4.6%; 3σ, 0.3%. The central confidence interval, 1-CL, (which is also sometimes called confidence level) for $\delta = 1\sigma$ is 68.3%; 2σ, 95.4%; 3σ, 99.7%. The odds against exceeding δ, (1-CL)/CL, for $\delta = 1\sigma$ are 2.15:1; 2σ, 21:1; 3σ, 370:1; 4σ, 16,000:1; 5σ, 1,700,000:1. Relations between σ and other measures of the width: probable error (CL = 0.5 deviation) = 0.67σ; mean absolute deviation = 0.80σ; RMS deviation = σ; half width at half maximum = 1.18σ.

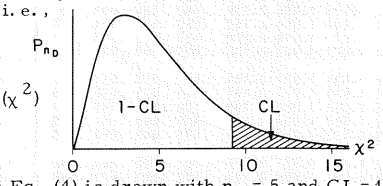
A.2. Chi-squared Distribution

The chi-squared distribution for n_D degrees of freedom is:

$$P_{n_D}(\chi^2)d\chi^2 = \frac{1}{2^h\Gamma(h)} (\chi^2)^{h-1} e^{-\chi^2/2} d\chi^2 \quad (\chi^2 \geq 0), \qquad (3)$$

where h (for "half") = $n_D/2$. The mean and variance are n_D and $2n_D$, respectively. In evaluating Eq. (3) one may use Stirling's approximation: $\Gamma(h) = (h-1)! \approx 2.507\, e^{-h} h^{(h-1/2)} \times (1 + 0.0833/h)$ which is accurate to ± 0.1% for all $h \geq 1/2$. The confidence level associated with a given value of n_D and an observed value of χ^2 is the probability of chi-squared exceeding the observed value, i.e.,

$$CL = \int_{\chi^2}^{\infty} d\chi^2\, P_{n_D}(\chi^2) \qquad (4)$$



[The small figure in Eq. (4) is drawn with $n_D = 5$ and CL = 10%.] CL is plotted as a function of χ^2 for several values of n_D in the above figure. For large n_D, χ^2 becomes normally distributed about n_D. Thus,

$$y_1 = (\chi^2 - n_D)/\sqrt{2n_D} \qquad (5)$$

becomes normally distributed with unit standard deviation. A better approximation, due to Fisher,[1] is that χ, not χ^2, becomes normally distributed, specifically

$$y_2 = \sqrt{2\chi^2} - \sqrt{2n_D - 1} \qquad (6)$$

approaches normality with unit standard deviation. For small CL's in particular, y_2 is much more accurate than y_1. Thus, for $n_D = 50$ and $\chi^2 = 80$, the true CL = 0.45%, but y_1 is 3.0 corresponding to a CL of 0.13%, while y_2 is 2.7 corresponding to a CL of 0.35%.

PROBABILITY AND STATISTICS (Cont'd)

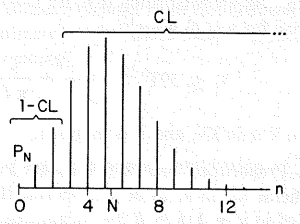
A.3. Poisson Distribution

The Poisson distribution with mean $\bar{n}$ is:

$$P_{\bar{n}}(n) = \frac{e^{-(\bar{n})}(\bar{n})^n}{n!} \quad (n = 0, 1, 2, \cdots) . \tag{7}$$

The variance is equal to the mean. Confidence levels for Poisson distributions are usually defined in terms of quantities called "upper limits" as follows: The confidence level associated with a given upper limit N and an observed value n_0 of n is the probability that $n > n_0$ if $\bar{n} = N$, i.e.,

$$CL = \sum_{n=n_0+1}^{\infty} P_N(n) = 1 - \sum_{n=0}^{n_0} P_N(n) \tag{8}$$



[The small figure in Eq. (8) is drawn with $n_0 = 2$ and CL = 90%.] A useful relation between Poisson and chi-squared confidence levels allows one to look up this quantity on the above figure. Specifically, the quantity 1-CL is given by the ordinate of the $n_D = 2(n_0+1)$ curve at $\chi^2 = 2N$. Thus, 90% confidence level upper limits for $n_0 = 0$, 1, and 2 are given by half the χ^2 value corresponding to an ordinate of 0.1 on the $n_D = 2, 4$, and 6 curves, respectively; the values are N = 2.3, 3.9, and 5.3.

- -

Tables of confidence levels for all three of these distributions, the relation between Poisson and chi-squared confidence levels, and numerous other useful tables and relations may be found in Ref. 2.

B. STATISTICS

We consider here the situation in which one is presented with N independent data, $y_n \pm \sigma_n$, and it is desired to make some inference about the "true" value of the quantity represented by these data. For this purpose we interpret each datum y_n as a single sample point drawn randomly (and independently of the other data) from a distribution having mean $\bar{y}_n$ (which we wish to estimate) and variance σ_n^2. (Identification of the true σ_n with the σ_n datum is an approximation which may become seriously inaccurate when σ_n is an appreciable fraction of y_n.) Some methods of estimation commonly used in high energy physics are given below; see Ref. 3 for numerous applications. Section B.1. deals with the case in which all $\bar{y}_n$ are the same, e.g., several different measurements of the same quantity; Sec. B.2. deals with the case in which $\bar{y}_n = \bar{y}(x_n)$, where x_n represents some set of independent variables, e.g., cross-section measurements at various values of energy and angle, $x_n = \{E_n, \theta_n\}$.

B.1. Single Mean and Variance Estimates

(1) If the y_n represent a set of values all supposedly drawn from a single distribution with mean $\bar{y}$ and variance σ^2 (i.e., the σ_n are all the same, but their common value is unknown) then

$$\bar{y}_e = \frac{1}{N}\sum_n y_n \quad \text{and} \tag{9}$$

$$\sigma_e^2 = \frac{1}{N-1}\sum_n (y_n - \bar{y}_e)^2 = \frac{N}{N-1}\left[\left(\overline{y^2}\right)_e - \bar{y}_e^2\right] \tag{10}$$

are unbiased estimates of $\bar{y}$ and σ^2. The variance of $\bar{y}_e$ is σ^2/N. If the parent distribution is normal and N is large, the variance of σ_e^2 is $2\sigma^4/N$.

(2) If the $\bar{y}_n$ all have the common value $\bar{y}$ and the σ_n are known, then the weighted average

$$\bar{y}_e = \frac{1}{w}\sum_n w_n y_n, \tag{11}$$

where $w_n = 1/\sigma_n^2$ and $w = \Sigma w_n$, is an appropriate unbiased estimate of $\bar{y}$. This choice of weighting factors in Eq. (11) minimizes the variance of the estimate; the variance is $1/w$.

B.2. Linear Least Squares Fit

A least squares fit of the function $y(x) = \Sigma_i a_i f_i(x)$ to independent data $y_n \pm \sigma_n$ at points x_n (e.g., a Legendre fit in which the f_i are Legendre polynomials and the a_i are Legendre coefficients) gives the following estimates of the parameters a_i:

$$a_{e,i} = \sum_{j,n} V_{ij} f_j(x_n) y_n/\sigma_n^2 . \tag{12}$$

Here V is the covariance matrix of the fitted parameters

$$V_{ij} = \overline{(a_{e,i} - \bar{a}_{e,i})(a_{e,j} - \bar{a}_{e,j})} , \tag{13}$$

which is given by

$$(V^{-1})_{ij} = \sum_n f_i(x_n) f_j(x_n)/\sigma_n^2 . \tag{14}$$

The variance of an interpolated or extrapolated value of y at point x, $y_e = \Sigma a_{e,i} f_i(x)$, is:

$$\overline{(y_e - \bar{y}_e)^2} = \sum_{ij} V_{ij} f_i(x) f_j(x) . \tag{15}$$

For the case of a straight line fit, $y(x) = a + bx$, one obtains the following estimates of a and b,

$$a_e = (S_y S_{xx} - S_x S_{xy})/D, \qquad b_e = (S_1 S_{xy} - S_x S_y)/D, \tag{16}$$

where

$$S_1, S_x, S_y, S_{xx}, S_{xy} = \sum (1, x_n, y_n, x_n^2, x_n y_n)/\sigma_n^2, \tag{17}$$
$$D = S_1 S_{xx} - S_x^2 .$$

The covariance matrix of the fitted parameters is:

$$\begin{pmatrix} V_{aa} & V_{ab} \\ V_{ab} & V_{bb} \end{pmatrix} = \frac{1}{D}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \tag{18}$$

The variance of an interpolated or extrapolated value of y at point x is:

$$\overline{(y_e - \bar{y}_e)^2} = \frac{1}{S_1} + \frac{S_1}{D}\left(x - \frac{S_x}{S_1}\right)^2 \tag{19}$$

C. ERROR PROPAGATION

We consider here the situation in which one wishes to calculate the value and error of a function of some other quantities with errors, e.g., in a Monte Carlo program. Let $\{y\}$ be a set of random variables with means $\{\bar{y}\}$ and covariance matrix V. Then the mean and variance of a function of these variables are approximately (to second order in $\{y-\bar{y}\}$):

$$\bar{f} \approx f(\{y\}) + \frac{1}{2}\sum_{mn} V_{mn}\left(\frac{\partial^2 f}{\partial y_m \partial y_n}\right)_{\{y\}=\{\bar{y}\}}, \tag{20}$$

$$\overline{(f - \bar{f})^2} = \sum_{mn} V_{mn}\left(\frac{\partial f}{\partial y_m}\right)_{\{y\}=\{\bar{y}\}}\left(\frac{\partial f}{\partial y_n}\right)_{\{y\}=\{\bar{y}\}} . \tag{21}$$

E.g., the mean and variance of a function of a single variable with mean $\bar{y}$ and variance σ^2 are:

$$\bar{f} \approx f(\bar{y}) + \frac{1}{2}\sigma^2 f''(\bar{y}), \tag{22}$$

$$\overline{(f - \bar{f})^2} = \sigma^2 f'(\bar{y})^2 . \tag{23}$$

Note that these equations will usually be applied by substituting some measured quantities, $\{\tilde{y}\}$ say, for the true means, $\{\bar{y}\}$. If, as is often the case, $\tilde{y}_n - \bar{y}_n$ is of order $\sqrt{V_{nn}}$, then there is no point in keeping the second order terms in Eq. (20) or (22) since the substitution itself introduces first order errors.

1. R. A. Fisher, Statistical Methods for Research Workers (Oliver and Boyd, Edinburgh and London, 1958).
2. M. Abramowitz and I. Stegun, eds., Handbook of Mathematical Functions (National Bureau of Standards, Applied Mathematics Series, Vol. 55, Washington, 1964).
3. W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet, Statistical Methods in Experimental Physics (North-Holland, Amsterdam and London, 1971).

Revised and expanded April 1974.