PROBABILITY, STATISTICS, AND MONTE CARLO

1. PROBABILITY

1.1 General

If x is the outcome of an observation, we define the probability of x as the relative frequency with which x occurs out of a (possibly hypothetical) large set of similar observations. If x may take any value from a *continuous* range, we write $f(x;\theta)\,dx$ as the probability of observing x between x and $x+dx$. The function $f(x;\theta)$ is the *probability density function* (p.d.f.) for the *random variable* x, which may depend upon a parameter θ. If x can take on only one of a set of *discrete* values (*e.g.*, the non-negative integers), then $f(x;\theta)$ is itself a probability, but we still refer to it as a p.d.f. The p.d.f. is always normalized to unit area (unit sum, if discrete). Both x and θ may have multiple components and are then usually written as column vectors. If θ is unknown and we wish to estimate its value from a given set of data x, we may use statistics (Section 2).

The *cumulative distribution function* $F(a)$ expresses the probability that $x \le a$:

$$F(a) = \int_{-\infty}^{a} f(x)\,dx\ . \tag{1.1}$$

Here and in what follows, if x is discrete-valued, the integral is replaced by a sum. The endpoint a is expressly included in the integral or sum. Then $0 \le F(x) \le 1$, $F(x)$ is nondecreasing, and $\mathrm{Prob}(a < x \le b) = F(b) - F(a)$. If x is discrete, $F(x)$ is flat except at allowed values of x, where it has a discontinuous jump equal to $f(x)$.

Any function of random variables is itself a random variable, with (in general) a different p.d.f. The *expectation value* of any function $u(x)$ is

$$\boldsymbol{E}[u(x)] = \int_{-\infty}^{\infty} u(x)\,f(x)\,dx\ . \tag{1.2}$$

The expectation value is said to exist only if it is finite. For x and y any two random variables, $\boldsymbol{E}(x+y) = \boldsymbol{E}(x) + \boldsymbol{E}(y)$. For c and k constants, $\boldsymbol{E}(cx+k) = c\boldsymbol{E}(x) + k$.

The nth moment of a distribution is given by

$$\alpha_n = \boldsymbol{E}(x^n)\ , \tag{1.3a}$$

and the nth moment about the mean by

$$m_n = \boldsymbol{E}[(x-\alpha_1)^n]\ . \tag{1.3b}$$

The most commonly used are the mean and variance:

$$\mu \equiv \alpha_1 \tag{1.4a}$$

$$\sigma^2 \equiv \mathrm{Var}(x) \equiv m_2 = \alpha_2 - \mu^2\ . \tag{1.4b}$$

The mean is the location of the "center of mass" of the distribution of x and the variance is a measure of the square of its width. Note that $\mathrm{Var}(cx+k) = c^2\mathrm{Var}(x)$.

Any odd moment about the mean is a measure of skewness; the simplest of these is the dimensionless coefficient of skewness $\gamma_1 = m_3/\sigma^3$.

In addition to the mean, another useful indicator of the x *location* near which most of the probability is likely to concentrate is the *median* x_{med}. This is that value of x such that $F(x_{\mathrm{med}}) = 1/2$, *i.e.*, exactly half of the probability lies above and half lies below x_{med}. For a given *sample* of events, x_{med} is that observed x such that half the events have larger x and half have smaller x (as closely as possible, not counting any that have the same x as the median). If this lies between two observed x values, the sample median is set by convention to be halfway between them. If the p.d.f. for x has the form $f(x-\mu)$ and μ is both mean and median, then for a large number of events N the variance of the median approaches $1/[4Nf^2(0)]$, provided $f(0) > 0$.

Let x and y be two random variables with joint p.d.f. $f(x,y)$. The *marginal* p.d.f. of, for example, x, expressing the p.d.f. for x with y unobserved, is

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \tag{1.5}$$

and similarly for $f_2(y)$. If y is fixed, the *conditional* p.d.f. for x given the fixed y is given by

$$f(x|y) = f(x,y)/f_2(y)\ . \tag{1.6}$$

The x mean is

$$\mu_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,f(x,y)\,dx\,dy = \int_{-\infty}^{\infty} x\,f_1(x)\,dx \tag{1.7}$$

and similarly for y. The *correlation* between x and y is a measure of the dependence of one on the other:

$$\rho_{xy} = \boldsymbol{E}\left[(x-\mu_x)(y-\mu_y)\right]/\sigma_x\,\sigma_y \equiv \mathrm{Cov}[x,y]/\sigma_x\,\sigma_y\ , \tag{1.8}$$

where σ_x, σ_y are defined in analogy with Eq. (1.4b); it can be shown that $-1 \le \rho_{xy} \le 1$. The symbol "Cov" represents the covariance of x and y, a 2-variable analogue to the variance, Eq. (1.4b). Two random variables are *independent* if and only if

$$f(x,y) = f_1(x)\,f_2(y)\ . \tag{1.9}$$

If x and y are independent then $\rho_{xy} = 0$; the converse is not necessarily true except for Gaussian-distributed x and y. If x and y are independent, $\boldsymbol{E}[u(x)\,v(y)] = \boldsymbol{E}[u(x)]\,\boldsymbol{E}[v(y)]$ and $\mathrm{Var}(x+y) = \mathrm{Var}(x)+\mathrm{Var}(y)$; otherwise, $\mathrm{Var}(x+y) = \mathrm{Var}(x)+\mathrm{Var}(y)+2\mathrm{Cov}[x,y]$ and $\boldsymbol{E}[u\,v]$ does not factor.

In a *change of continuous random variables* from, *e.g.*, $\vec{x} \equiv (x_1,\ldots,x_n)$, with p.d.f. $f(x_1,\ldots,x_n)$, to $\vec{y} \equiv (y_1,\ldots,y_n)$, a one-to-one function of the x's, the p.d.f. $g(y_1,\ldots,y_n)$ is found by substitution for $(x_1,\ldots,x_n)$ in f followed by multiplication by the absolute value of the Jacobian of the transformation:

$$g(\vec{y}\,) = f\,[w_1(\vec{y}\,),\ldots,w_n(\vec{y}\,)]\,|J|\ . \tag{1.10}$$

The functions w_i express the *reverse* transformation $x_i = w_i(\vec{y}\,)$ for $i = 1,\ldots,n$, and $|J|$ is the absolute value of the determinant of the square matrix $J_{ij} = \partial x_i/\partial y_j$. Such transformations must always preserve the number of random variables, n. To transform to fewer variables, first perform (1.10) and then use Eq. (1.5) to eliminate unwanted variables. If the transformation from $\vec{x}$ to $\vec{y}$ is not one-to-one, the situation is more complex and a unique solution may not exist. To change variables for discrete random variables simply substitute; no Jacobian is necessary because in that case f is a probability rather than a probability density. If f depends upon a parameter set θ, we can change to a different parameter set $\phi = \phi(\theta)$ by simple substitution; no Jacobian is used.

1.2 Characteristic functions [1]

The characteristic function $\phi(u)$ associated with the p.d.f. $f(x)$ is essentially its Fourier transform, or the expectation value of $\exp(iux)$,

$$\phi(u) = \boldsymbol{E}(e^{iux}) = \int e^{iux} f(x)dx\ . \tag{1.11}$$

It is sufficiently useful to deserve special attention, and several of its properties follow.

We note from Eqs. (1.3a) and (1.11) that the nth moment of the distribution $f(x)$ is given by

$$i^{-n}\left.\frac{d^n\phi}{du^n}\right|_{u=0} = \int x^n f(x)dx = \alpha_n\ . \tag{1.12}$$

As a result, it is often easy to calculate all the moments of a distribution defined by $\phi(u)$ even when the inversion is not available.

If $f_1(x)$ and $f_2(y)$ have characteristic functions $\phi_1(u)$ and $\phi_2(u)$, then the characteristic function of the weighted sum $ax+by$ is $\phi_1(au)\phi_2(bu)$.

Let the (partial) characteristic function corresponding to the conditional p.d.f. $f_2(x|z)$ be $\phi_2(u|z)$, and the p.d.f. of z be $f_1(z)$. The characteristic function after integration over the conditional value is

$$\phi(u) = \int \phi_2(u|z)\,f_1(z)dz\ . \tag{1.13}$$

Suppose we can write ϕ_2 in the form

$$\phi_2(u|z) = A(u)e^{ig(u)z}\ . \tag{1.14}$$

Then

$$\phi(u) = A(u)\phi_1(g(u)) \ . \tag{1.15}$$

The semi-invariants κ_n are defined by

$$\phi(u) = \exp\left(\sum_1^\infty \frac{\kappa_n}{n!}(iu)^n\right) \ . \tag{1.16}$$

The α_n's, m_n's, and κ_n's are related algebraically, and the first few are familiar:

$$\begin{aligned} &\kappa_1 = \alpha_1 \ (= \mu, \text{ the mean}) \\ &\kappa_2 = m_2 \ = \alpha_2 - \alpha_1^2 \ (= \sigma^2, \text{ the variance}) \\ &\kappa_3 = m_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^2 \ . \end{aligned} \tag{1.17}$$

1.3 Examples of probability density functions

We describe a few p.d.f.'s commonly encountered in physics applications. Tables for most of these distributions, relations among them, and further information may be found in Refs. 1–6. Monte Carlo techniques for generating each of them may be found in Section 3.3 below.

1.3.1 Uniform distribution (continuous)

This p.d.f. assumes equal probability density for any x in an allowed range $[a, b]$:

$$\begin{aligned} f(x) &= 1/(b-a) \ , && a \le x \le b \\ &= 0, && \text{otherwise;} \end{aligned} \tag{1.18}$$

$$\boldsymbol{E}(x) = (b+a)/2 \ ; \qquad \mathrm{Var}(x) = (b-a)^2/12 \ . \tag{1.19}$$

1.3.2 Binomial distribution (discrete)

Any random process with exactly two possible outcomes is a *Bernoulli* process. If the process is repeated n times independently, and if the probability of obtaining a certain outcome (a "success") in each trial is p, then the probability of obtaining exactly r successes is given by the binomial distribution:

$$f(r; n, p) = \binom{n}{r} p^r q^{n-r} = \frac{n!}{r!(n-r)!}\, p^r q^{n-r} \ , \qquad r = 0, 1, 2, \ldots, n \ , \tag{1.20}$$

where $q = 1 - p$ and the order in which the successes and failures come is assumed irrelevant.

$$\boldsymbol{E}(r) = np \ ; \ \mathrm{Var}(r) = npq \ . \tag{1.21}$$

If r successes are observed in n_r Bernoulli trials with probability p of success, and if s successes are observed in n_s similar trials, then $t = r + s$ is also binomial with $n_t = n_r + n_s$.

1.3.3 Poisson distribution (discrete)

The Poisson distribution with mean μ is:

$$f(n; \mu) = \frac{\mu^n e^{-\mu}}{n!} \ , \ n = 0, 1, 2, \ldots \ . \tag{1.22}$$

The observed result of a Poisson process is a non-negative integer n; the parameter μ is any non-negative real number. The Poisson distribution describes the population of events in any interval of x (*e.g.*, space or time) whenever: (a) the number of events in any interval of x is independent of that in any other non-overlapping interval; (b) in any small Δx, the probability of one event is $\lambda \Delta x$ and the probability of two or more vanishes at least as fast as $(\Delta x)^2$, as $\Delta x \to 0$; and (c) λ does not depend on x. Then $\mu \equiv \lambda x$;

$$\boldsymbol{E}(n) = \mu \ ; \ \mathrm{Var}(n) = \mu \ . \tag{1.23}$$

When μ is large ($\gtrsim 7$ or 8), it is often useful to approximate the distribution of n by a Gaussian distribution of mean μ and variance $\sigma^2 = \mu$, as though n were a continuous variable. Two or more Poisson processes (*e.g.*, *signal + background*, with parameters μ_S and μ_B, respectively) which independently contribute amounts n_S and n_B to a given measurement will produce an observed number $n = n_S + n_B$, which is distributed according to a new Poisson distribution with parameter $\mu = \mu_S + \mu_B$.

1.3.4 Normal or Gaussian distribution (continuous)

The Gaussian distribution is

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \ , \ -\infty < x < \infty \ ; \tag{1.24}$$

$$\boldsymbol{E}(x) = \mu \ ; \ \mathrm{Var}(x) = \sigma^2 \ . \tag{1.25}$$

The characteristic function of a Gaussian p.d.f. with mean m and variance σ^2 is

$$\phi(u) = e^{imu - \frac{1}{2}\sigma^2 u^2} \ , \tag{1.26}$$

so the Gaussian is that unique distribution for which all semi-invariants beyond the second vanish.

For x and y independent and normally distributed, $z = x + y$ obeys $f(z; \mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

The integrated probability for x to fall in the range $\mu - \sigma$ to $\mu + \sigma$ is 0.683. Other measures of width commonly encountered are: probable error (central region containing 0.50 of the probability) $= \mu \pm 0.67\sigma$; mean absolute deviation; $\boldsymbol{E}[|x - \mu|] = 0.80\sigma$; rms deviation $= \sigma$; half-width at half-maximum $= 1.18\sigma$.

The Gaussian gets its importance in large part from the *central limit theorem*: if a continuous random variable x is distributed according to *any* p.d.f. with finite mean and variance, then the sample mean, $\overline{x}_n$ of n observations of x will have a p.d.f. that approaches a Gaussian as n increases. Therefore the end result $\sum^n x_i \equiv n\overline{x}_n$ of a large number of small fluctuations x_i will be distributed as a Gaussian, even if the x_i themselves are not.

The cumulative distribution (1.1) for a Gaussian with $\mu = 0$ and $\sigma^2 = 1$ is given by the *error function*, $\mathrm{erf}(a)$, through the following ugly relation:

$$F(a; 0, 1) = 0.5\left[1 + \mathrm{erf}(a/\sqrt{2})\right] \ . \tag{1.27}$$

The function $\mathrm{erf}(a)$ is tabulated in Ref. 2 and is available as a FORTRAN function on many computers [caution: other definitions of $\mathrm{erf}(a)$ are sometimes used]; for mean μ and variance σ^2 replace a by $[(a - \mu)/\sigma]$.

For $\vec{x}$ a set of n (not necessarily independent) Gaussian random variables x_i arranged into a column vector, their joint p.d.f. is the *multivariate Gaussian*:

$$f(\vec{x}; \vec{\mu}, V) = \frac{1}{(2\pi)^{n/2}} \ |V|^{-1/2} \tag{1.28a}$$

$$\times \ \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T \, V^{-1} \, (\vec{x} - \vec{\mu})\right] \ , \ |V| \neq 0 \ ,$$

where V is the *covariance matrix* of the x's, $V_{ii} = \mathrm{Var}(x_i)$ and $V_{ij} = \boldsymbol{E}[(x_i - \mu_i)(x_j - \mu_j)] \equiv \rho_{ij}\, \sigma_i \sigma_j$, and $|V|$ is the determinant of V. The quantity ρ_{ij} is the correlation coefficient for x_i and x_j; $|\rho_{ij}|^2 \le 1$. For $n = 2$ this becomes

$$f(x_1, x_2; \ \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\, \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \tag{1.28b}$$

$$\times \ \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}\right]\right\} \ .$$

The special case $\sigma_1 = \sigma_2$ and $\rho = 0$ is called the *Rayleigh distribution*. If V is singular, there is a linear relation among some variables; in this case one usually wants to eliminate completely dependent variables and work in a smaller number of dimensions. The marginal distribution of any x_i is a Gaussian with mean μ_i and variance V_{ii}. V is $n \times n$, symmetric, and positive definite. Therefore for any vector $\vec{X}$, the quadratic form $\vec{X}^T V^{-1} \vec{X} = c$ traces an n-dimensional ellipsoid as $\vec{X}$ varies for any given $c > 0$. If $X_i = (x_i - \mu_i)/\sigma_i$, then c is a random variable obeying the $\chi^2(n)$ distribution, which is discussed in the following section. The probability that $\vec{X}$ corresponding to a set of Gaussian random variables $\vec{x}_i$ lies *outside* the ellipsoid characterized by a given value of $c\,(= \chi^2)$ is given by Eq. (1.31) and may be read

from Fig. 1. For example, the "s-standard-deviation ellipsoid" occurs at $c = s^2$. For the two-variable case ($n = 2$) the point $\vec{X}$ lies outside the one-standard-deviation ellipsoid with 61% probability, so both X_1 and X_2 lie inside the ellipsoid with 39% probability. This assumes that μ_i and σ_i are correct. For $X_i = x_i/\sigma_i$, the ellipsoids of constant χ^2 have the same size and orientation but are centered at $\vec{\mu}$. The use of these ellipsoids as indicators of probable error is described in Sec. 2.4.1.

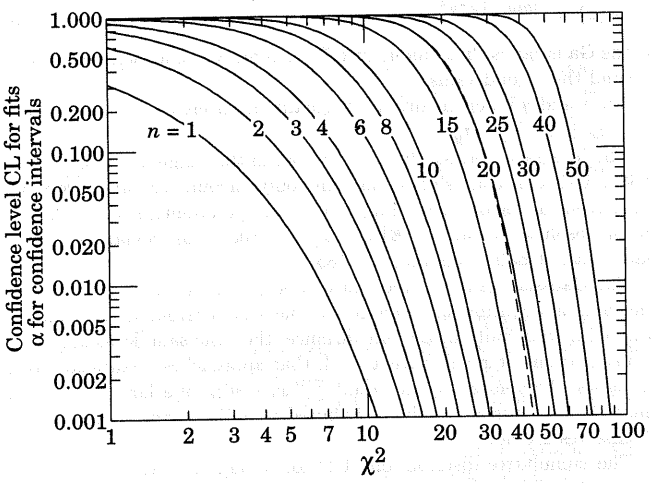

Fig. 1. χ^2 confidence level vs χ^2 for n degrees of freedom, as defined in Eq. (1.31). The curve for a given n expresses the probability that a value at least as large as χ^2 will be obtained in an experiment; *e.g.*, for $n = 10$, a value $\chi^2 \gtrsim 18$ will occur in 5% of a very large number of experiments. For a fit, CL is a measure of goodness-of-fit in that a good fit to a correct model is expected to yield a low χ^2 (Sec. 2.3.3). For a confidence interval, α measures the probability that the interval *does not* cover the true value of the quantity being estimated (Sec. 2.4). The dashed curve for $n = 20$ is calculated using the approximation of Eq. (1.32).

It is a characteristic of the multivariate Gaussian that $\rho_{ij} = 0$ is necessary and sufficient for x_i and x_j to be independent. For a given covariance matrix V, there always exist nonsingular $n \times n$ matrices H such that $HH^T = V$; H is usually upper or lower triangular in the most efficient algorithms. Then $\vec{z} = H^{-1}(\vec{x} - \vec{\mu})$ is a vector of n independent Gaussian random variables with zero mean and with covariance matrix equal to the identity.

1.3.5 The χ^2 distribution (continuous)

If $x_1, \ldots, x_n$ are independent Gaussian distributed random variables, the sum $z = \sum^n (x_i - \mu_i)^2/\sigma_i^2$ is distributed as a χ^2 with n *degrees of freedom* [$\chi^2(n)$]:

$$f(z;\, n) = \frac{1}{2^{n/2}\Gamma(n/2)}\, z^{n/2-1}\, e^{-z/2}\,, \quad z \geq 0\,; \tag{1.29}$$

$$\boldsymbol{E}(z) = n\,;\ \mathrm{Var}(z) = 2n\,. \tag{1.30}$$

Under a linear transformation to n *dependent* Gaussian variables x_i', the χ^2 at each transformed point retains its value; then $z = \vec{X}'^T\, V^{-1}\, \vec{X}'$ as in the previous section. For a set of z_i, each of which is $\chi^2(n_i)$, $\sum z_i$ is a new random variable which is $\chi^2\,(\sum n_i)$.

Fig. 1 shows the Confidence Level (CL) obtained by integrating the tail of the function given in Eq. (1.29) for n degrees of freedom:

$$\mathrm{CL}(\chi^2) = \int_{\chi^2}^{\infty} f(z;\, n)\, dz\,; \tag{1.31}$$

this area is shown schematically in Fig. 2. It is equal to 1.0 minus the cumulative distribution function $F(z = \chi^2;\, n)$. It is useful in evaluating the consistency of data with a model (see Sec. 2): The CL is the probability that a random repeat of the given experiment would observe a *worse* χ^2, assuming the correctness of the model. It is also useful for confidence intervals for statistical estimators (Sec. 2.4), when one is interested in the unshaded area of Fig. 2.

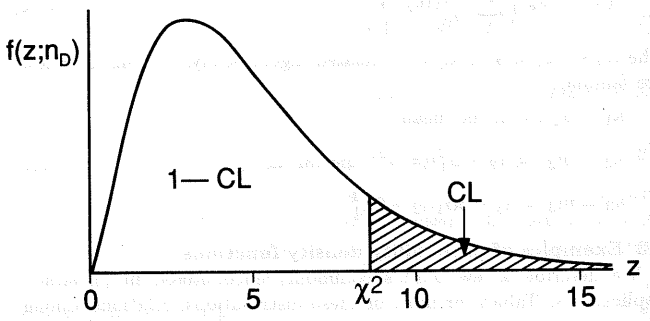

Fig. 2. Schematic illustration of the confidence level integral given in Eq. (1.31).

Since the mean of the χ^2 distribution is equal to the number of degrees of freedom, one expects to obtain $\chi^2 \approx n$ in a "reasonable" experiment. While caution is necessary because of the skewness of the distribution, the "reduced χ^{2}" $\equiv \chi^2/n$ is therefore a useful quantity. Figure 3 shows χ^2/n for useful CL's as a function of n. It contains the same information as Fig. 1, but is easier to read.

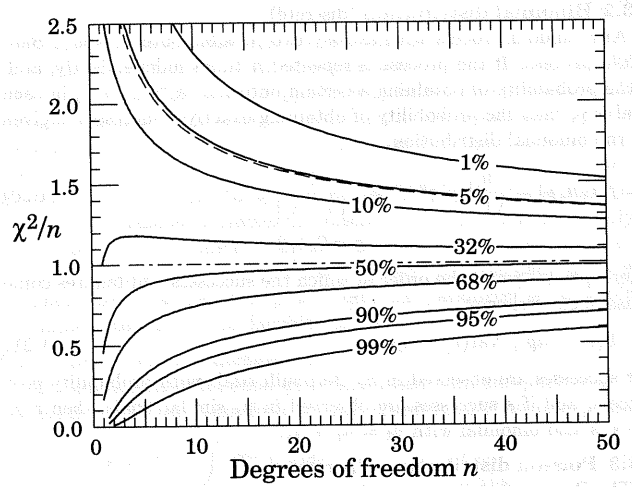

Fig. 3. Confidence limits as a function of the "reduced χ^{2}" $\equiv \chi^2/n$ and the number of degrees of freedom n. Curves are labeled by the probability of a measurement resulting in a value of χ^2/n greater than that given on the y axis; *e.g.*, for $n = 10$, a value $\chi^2/n \gtrsim 1.8$ will occur in 5% of a very large number of experiments. The dashed curve for CL = 5% is calculated using the approximation of Eq. (1.32).

It is commonly stated that for large n the CL is approximately given by [1,7]

$$\mathrm{CL} \approx \frac{1}{\sqrt{2\pi}} \int_y^{\infty} e^{-x^2/2} dx\,, \tag{1.32}$$

where $y = \sqrt{2\chi^2} - \sqrt{2n-1}$. This approximation was used to draw the dashed curves in Fig. 1 (for $n = 20$) and Fig. 3 (for CL = 5%). However, all of the functions and their inverses are now readily available in standard mathematical libraries (such as IMSL, used to generate these Figures), and so the approximation (and even such figures and tables) plays only a secondary role in practical problems.

PROBABILITY, STATISTICS, AND MONTE CARLO (Cont'd)

1.3.6 Student's t (continuous)

Suppose that x and $x_1, \ldots, x_n$ are independent and normal with mean 0 and variance 1. We then define $z = \sum_1^n x_i^2$, and

$$t = x/\sqrt{z/n}\ . \tag{1.33}$$

The variable z thus belongs to a $\chi^2(n)$ distribution. Then t is distributed according to a Student's t distribution with n degrees of freedom:

$$f(t;\, n) = \frac{1}{\sqrt{n\pi}}\,\frac{\Gamma\left[(n+1)/2\right]}{\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}\ , \tag{1.34}$$

$$-\infty < t < \infty\ ,$$

and

$$\boldsymbol{E}(t) = 0 \ \text{ for } n > 1\ ; \ \ \mathrm{Var}(t) = \frac{n}{n-2} \ \text{ for } n > 2\ . \tag{1.35}$$

Here $\Gamma(k)$ is the gamma function, equal to $(k-1)!$ if k is an integer. Student's t distribution resembles a Gaussian distribution with wide tails. As $n \to \infty$, the distribution approaches a Gaussian, and if $n = 1$, the distribution is *Cauchy*, or *Breit-Wigner*. The mean is finite for $n > 1$ and the variance is finite for $n > 2$, so for $n = 1$ or $n = 2$, t does not obey the central limit theorem.

As an example, consider the *sample mean* $\overline{x} = \sum x_i/n$ and the *sample variance* $s^2 = \sum(x_i - \overline{x})^2/(n-1)$ for normally distributed random variables x_i with unknown mean μ and variance σ^2. The sample mean has a Gaussian distribution with a variance σ^2/n, so the variable $(\overline{x} - \mu)/\sqrt{\sigma^2/n}$ is normal with mean 0 and variance 1. Similarly, $(n-1)\,s^2/\sigma^2$ is independent of this and is χ^2 distributed with $n-1$ degrees of freedom. The ratio

$$t = \frac{(\overline{x}-\mu)/\sqrt{\sigma^2/n}}{\sqrt{(n-1)\,s^2/\sigma^2\,(n-1)}} = \frac{\overline{x}-\mu}{\sqrt{s^2/n}} \tag{1.36}$$

distributes as $f(t;\, n-1)$. The unknown true variance σ^2 cancels, and t can be used to test the probability that the true mean is some particular value μ.

The distribution (1.34) is written such that n is not required to be an integer. A Student's t distribution with nonintegral $n > 0$ is useful in certain applications.

1.3.7 The gamma distribution (continuous)

If a process generating events as a function of x (*e.g.*, space or time) satisfies conditions (a)–(c) of the Poisson distribution, then the x distance from an arbitrary starting point (which may be some particular event) to the k^{th} event is belongs to a *gamma* distribution:

$$f(x;\, \lambda, k) = \frac{x^{k-1}\,\lambda^k\,e^{-\lambda x}}{\Gamma(k)}\ ,\ 0 < x < \infty\ . \tag{1.37}$$

$\Gamma(k)$ is the gamma function, equal to $(k-1)!$ if k is an integer. The Poisson parameter μ is λ per unit x;

$$\boldsymbol{E}(x) = k/\lambda\ ;\ \mathrm{Var}(x) = k/\lambda^2\ . \tag{1.38}$$

The special case $k = 1$ is called the *exponential* distribution. A sum of k' exponential random variables x_i is distributed as $f(\sum x_i;\, \lambda, k')$. Eq. (1.37) allows $k > 0$ to be nonintegral. If $\lambda = 1/2$ and $k = n/2$, the gamma and $\chi^2(n)$ distributions are identical.

2. STATISTICS

2.1 General

A probability density function with known parameters enables us to predict the frequency with which a random variable will take on a particular value (if discrete) or lie in a given range (if continuous). In *parametric* statistics we have the opposite problem of estimating the parameters of the p.d.f. from a set of actual observations.

We refer to the true p.d.f. as the *population*; the data form a *sample* from this population. A *statistic* is any function of the data, plus known constants, which does not depend upon any of the unknown parameters. A statistic is a random variable if the data have random errors. An *estimator* is any statistic whose value is intended as a meaningful guess for the value of an unknown parameter; we denote estimators with hats, *e.g.*, $\widehat{\theta}$.

Often it is possible to construct more than one reasonable estimator. Let θ represent the true value of a parameter to be estimated; θ is a vector if there is more than one parameter. Then if $\widehat{\theta}$ is an estimator for θ, desirable properties for $\widehat{\theta}$ are: (a) *Unbiased*; bias $b = \boldsymbol{E}(\widehat{\theta}) - \theta$, where the expectation value is taken over a hypothetical set of similar experiments in which $\widehat{\theta}$ is constructed the same way. The bias may be due to statistical properties of the estimator or to *systematic* errors in the experiment. If we can estimate the average bias b we usually subtract it from $\widehat{\theta}$ to obtain a new $\widehat{\theta}' \equiv \widehat{\theta} - b$. However, b may depend upon θ or other unknowns, in which case we usually try to choose an estimator which minimizes its average size. (b) *Minimum variance*; the minimum possible value of $\mathrm{Var}(\widehat{\theta})$ is given by the Rao-Cramér-Frechet bound:

$$\mathrm{Var}_{\min} = \left[1 + \partial b/\partial\theta\right]^2/I(\theta)\ ; \tag{2.1}$$

$$I(\theta) = \boldsymbol{E}\left\{\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln f(x_i;\, \theta)\right]^2\right\}\ .$$

The sum is over all data and b is the bias, if any; the x_i are assumed independent and distributed as $f(x_i;\theta)$, and the allowed range of x must not depend upon θ. The ratio $\epsilon = \mathrm{Var}_{\min}/\mathrm{Var}(\widehat{\theta})$ is the *efficiency*. An *efficient* estimator (with $\epsilon = 1$) exists only for certain cases. The square root of the variance expresses the expected spread of $\widehat{\theta}$ about its average value, as would be observed in a large number of repeats of the same measurement. (c) *Minimum mean-squared error* (mse); $\mathrm{mse} = \boldsymbol{E}[(\widehat{\theta} - \theta)^2] = V(\widehat{\theta}) + b^2$. The mse combines the error due to any bias quadratically with the variance, which expresses only the spread about $\boldsymbol{E}(\widehat{\theta})$, as distinct from θ, the true value. (d) *Robust*; a robust estimator is not sensitive to errors in our assumptions, *e.g.*, to departures from the assumed p.d.f. due to such factors as noise.

These criteria (and others) allow us to evaluate any procedure for obtaining $\widehat{\theta}$. In many cases these criteria conflict. The bias, variance, and mse may depend on the unknown θ. In this case the optimum prescription for $\widehat{\theta}$ may depend on the range in which we assume θ to lie.

Following are techniques in common use for obtaining estimators and their standard errors $\sigma(\widehat{\theta}) = \sqrt{\mathrm{Var}(\widehat{\theta})}$. When the conditions of the central limit theorem are satisfied, the interval $\widehat{\theta} \pm \sigma(\widehat{\theta})$ forms a 68.3% *confidence interval*. This is a random interval in that its endpoints depend upon the randomly sampled data; its meaning here will be taken to be that in 68.3% of all similar experiments the interval will include the true value θ. One should be aware that in most practical cases the central limit theorem is only approximately satisfied and accordingly confidence intervals which depend on that are only approximate. Confidence intervals are discussed in Section 2.4 below.

2.2 Data with a common mean

(1) Suppose we have a set of N independent measurements y_i assumed to be unbiased measurements of the same unknown quantity μ with a common, but unknown, variance σ^2 resulting from measurement error. Then

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2.2}$$

$$\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \widehat{\mu})^2 = \frac{N}{N-1}\left(\boldsymbol{E}(y^2) - \widehat{\mu}^2\right) \tag{2.3}$$

are unbiased estimators of μ and σ^2. The variance of $\widehat{\mu}$ is σ^2/N. If the common p.d.f. of the y_i is Gaussian, these statistics are independent. Then, for large N, the variance of $\widehat{\sigma}^2$ is $2\sigma^4/N$. If the y_i are Gaussian or N is large enough that the central limit theorem applies, then $\widehat{\mu}$ is an efficient estimator for μ. Otherwise $\widehat{\mu}$ is sometimes subject to large fluctuations, *e.g.*, if the p.d.f. for y_i has long tails. In this case the median of the y_i may be a more *robust* estimator for μ, provided the median and mean are expected to lie at the same point in the p.d.f. for y. For Gaussian y, the median has asymptotic (large-N) efficiency

$2/\pi \approx 0.64$. The Student's t distribution provides an example in which there are large tails. In this case, for large N the efficiency of the sample median relative to the sample mean is (∞, ∞, 1.62, 1.12, 0.96, 0.80, 0.64) for (1, 2, 3, 4, 5, 8, ∞) degrees of freedom.

If σ^2 is known, $\widehat{\mu}$ as given in Eq. (2.2) is still the best estimator for μ; if μ is known, substitute it for $\widehat{\mu}$ in Eq. (2.3) and replace $N-1$ by N, to obtain a somewhat better estimator $\widehat{\sigma}^2$.

(2) If the y_i have different, known, variances σ_i^2, then

$$\widehat{\mu} = \frac{1}{w}\sum^{N} w_i\, y_i \ , \tag{2.4}$$

is an unbiased estimator for μ with smaller variance than Eq. (2.2), where $w_i = 1/\sigma_i^2$ and $w = \sum w_i$. The variance of $\widehat{\mu}$ is $1/w$.

2.3 The method of maximum likelihood

2.3.1 General

"From a theoretical point of view, the most important general method of estimation so far known is the *method of maximum likelihood.*" [1]. We suppose that a set of independently measured quantities $\vec{x}$ came from a p.d.f. $f(\vec{x};\ \vec{\theta})$, where $\vec{\theta}$ is an unknown set of parameters. The method of maximum likelihood consist of finding the set of values of $\vec{\theta}$, $\widehat{\theta}$, which maximizes the joint probability density for all the data, given by

$$\mathcal{L}(\vec{\theta}) = \prod_i f(\vec{x}_i;\ \vec{\theta}) \ , \tag{2.5}$$

where $\mathcal{L}$ is called the likelihood. It is usually easier to work with $\ln \mathcal{L}$, and since both are maximized for the same set of $\vec{\theta}$, it is sufficient to solve the *likelihood equation*

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_n} = 0 \ . \tag{2.6}$$

The solution is called the *maximum likelihood estimate* of $\vec{\theta}$. The importance of the approach is shown by the following proposition, proved in Ref. 1:

If an efficient estimate $\widehat{\theta}$ of $\vec{\theta}$ exists, the likelihood equation will have a unique solution equal to $\widehat{\theta}$.

In evaluating $\mathcal{L}$, it is important that any normalization factors in the f's which involve $\vec{\theta}$ be included. However, we will only be interested in the maximum of $\mathcal{L}$ and in ratios of $\mathcal{L}$ at different $\vec{\theta}$'s; hence any multiplicative factors which do not involve the parameters we want to estimate may be dropped; this includes factors which depend on the data but not on $\vec{\theta}$.

If the solution to Eq. (2.6) is at a maximum, $\partial \ln \mathcal{L}/\partial\theta_n$ will have negative slope in its vicinity. In many practical problems, one often uses nonlinear algorithms for finding the maximum, and must be alert to various possibilities for error: (a) Eq. (2.6) may yield a minimum, therefore one must check the second derivative; (b) there may be more than one maximum—one must try to find the global maximum; (c) the global maximum may lie at a boundary of the physical region, in which case Eq. (2.6) will not find it.

If an unbiased, efficient estimator exists, this method will find it. If $\partial \ln \mathcal{L}/\partial\theta_n$ is linear in the vicinity of the root, an efficient estimator is guaranteed; other efficient cases are discussed in the literature. For large data samples, the central limit theorem will usually assure this condition in some significant neighborhood of zero; hence the estimator is usually efficient in that case, provided certain conditions are met (*e.g.*, that the solution does not lie on a boundary). In this case, in the neighborhood of the maximum $\ln \mathcal{L}$ is a downward-curving parabola and $\mathcal{L}$ is proportional to a Gaussian.

The results of two or more experiments may be combined by forming the product of the $\mathcal{L}$'s, or the sum of the $\ln \mathcal{L}$'s.

Under a one-to-one change of parameters from $\vec{\theta}$ to $\vec{\phi} = \vec{\phi}(\vec{\theta})$, the maximum likelihood estimate is simply $\widehat{\phi} = \vec{\phi}(\widehat{\theta})$, given the solution for $\widehat{\theta}$ for $\vec{\theta}$. That is, the maximum likelihood solution for $\vec{\phi}$ is found by simple substitution of $\widehat{\theta}$ into the transformation equation. It is possible that the new solution $\widehat{\phi}$ will be a biased solution for the true value of $\vec{\phi}$ even if $\widehat{\theta}$ is not biased, and vice-versa. In the asymptotic limit (of large amounts of data) both $\widehat{\theta}$ and $\widehat{\phi}$ will (usually) converge to unbiased solutions, but at different rates.

Except in special cases like the least-squares method, the value of the likelihood function at the solution does not necessarily tell us whether the final fit was a sensible description of the data or not. To evaluate this, one may: (a) prepare histograms of the data projected on various axes and make χ^2 (or other) comparisons with the fitted model projected upon the same axes; and/or (b) do numerous Monte Carlo simulations of the experiment under the hypothesis that the fitted parameters are correct, fit each of these, and compare the experimental likelihood (or $\ln \mathcal{L}$) with those obtained from these simulations. If the experimental likelihood is lower than that of some agreed-upon fraction of these results, one should question the appropriateness of the p.d.f. f. At the same time one can check for bias in the solution.

2.3.2 Error estimates

The covariance matrix V may be estimated from

$$V_{nm} = \left(\boldsymbol{E}\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_n\, \partial\theta_m}\bigg|_{\widehat{\theta}} \right] \right)^{-1} \ . \tag{2.7}$$

If $\partial \ln \mathcal{L}/\partial\theta_n$ is linear, the "expectation" operation in Eq. (2.7) has no effect because the second derivative of $\ln \mathcal{L}$ is constant. Otherwise, it may be approximated by taking the average of the quantity in square brackets over a range of θ_n and θ_m near the solution. For complex cases it may be more practical to evaluate s-standard-deviation errors from the contour

$$\ln \mathcal{L}(\vec{\theta}) = \ln \mathcal{L}_{\max} - s^2/2 \ , \tag{2.8}$$

where $\ln \mathcal{L}_{\max}$ is the value of $\ln \mathcal{L}$ at the solution point (compare with $\chi^2(\vec{a}') = \chi^2_{\min} + 1$ and the discussion in the least-squares case, below). The extreme limits of this contour parallel to the θ_n axis give an approximate s-standard-deviation confidence interval in θ_n. These intervals may not be symmetric and they may even consist of two or more disjoint intervals. This procedure gives one-standard-deviation errors in θ_n equal to $\sqrt{V_{nn}}$ of Eq. (2.7) if the estimator is efficient. If it is not efficient, the level of confidence implied by the value of s is only approximate.

2.3.3 Method of least squares

By far the most common case of the maximum likelihood approach is the *method of least squares.* We suppose a set of N measurements at points x_i. The ith measurement y_i is assumed to be chosen from a Gaussian distribution with mean $F(x_i;\ \vec{a})$ and variance σ_i^2. Then

$$-\tfrac{1}{2}\ln \mathcal{L} \equiv \chi^2 = \sum_1^N \frac{(y_i - F(x_i;\ \vec{a}))^2}{\sigma_i^2} \ . \tag{2.9}$$

Finding the set of parameters $\vec{\alpha}$ which maximizes $\mathcal{L}$ is equivalent to finding the set which minimizes χ^2.

At the outset it should be said that the method of least squares is sometimes applied in cases where the distribution is not Gaussian or not known to be Gaussian. In such cases it can still be used, but it is then not a special case of the maximum likelihood method, and the theorems having to do with that approach no longer apply.

In many practical cases one further restricts the problem to the situation in which $F(x_i;\ \vec{a})$ is a linear function of the a_m's,

$$F(x_i;\ \vec{a}) = \sum_n a_n\, f_n(x) \ , \tag{2.10}$$

where the f_n are k linearly independent functions (*e.g.*, 1, x, x^2, ..., or Legendre polynomials) which are single-valued over the allowed range of x. We require $k \le N$, and at least k of the x_i must be distinct. We wish to estimate the linear coefficients a_n. Later we will discuss the nonlinear case.

If the point errors $\epsilon_i = y_i - F(x_i;\ \vec{a})$ are Gaussian, then the minimum χ^2 will be distributed as a χ^2 random variable with $n = N - k$ degrees of freedom. We can then evaluate the goodness-of-fit (confidence level) from Figs. 1 or 3, as per the earlier discussion. The confidence level expresses the probability that a *worse* fit would be obtained in a large number of similar experiments under the assumptions that: (a) the model $y = \sum a_n\, f_n$ is correct and (b) the errors ϵ_i are Gaussian and unbiased with variance

PROBABILITY, STATISTICS, AND MONTE CARLO (Cont'd)

σ_i^2. If this probability is larger than an agreed-upon value (0.001, 0.01, or 0.05 are common choices), the data are *consistent* with the assumptions; otherwise we may want to find improved assumptions. As for the converse, most people do not regard a model as being truly *inconsistent* unless the probability is as low as that corresponding to four or five standard deviations for a Gaussian (6×10^{-3} or 6×10^{-5}; see Sec. 2.4.1). If the ϵ_i are not Gaussian, the method of least squares still gives an answer, but the goodness-of-fit test would have to be done using the correct distribution of the random variable which is still called "χ^2."

Finding the minimum of χ^2 in the linear case is straightforward:

$$-\frac{1}{2}\frac{\partial\chi^2}{\partial a_m} = \sum_i f_m(x_i)\left(\frac{y_i - \sum_n a_n\, f_n(x_i)}{\sigma_i^2}\right)$$

$$= \sum_i \frac{y_i\, f_m(x_i)}{\sigma_i^2} - \sum_n a_n \sum_i \frac{f_n(x_i)\, f_m(x_i)}{\sigma_i^2} \,. \qquad (2.11)$$

With the definitions

$$g_m = \sum_i y_i\, f_m(x_i)/\sigma_i^2 \qquad (2.12)$$

and

$$\left(V_{\widehat{a}}^{-1}\right)_{mn} = \sum_i f_n(x_i)\, f_m(x_i)/\sigma_i^2 \,, \qquad (2.13)$$

the k-element column vector of solutions $\widehat{a}$, for which $\partial\chi^2/\partial a_m = 0$ for all m, is given by

$$\widehat{a} = V_{\widehat{a}}\ \vec{g} \,. \qquad (2.14)$$

More generally, the measured y_i's are not independent. Then the set of σ_i^2's must be replaced by the $N \times N$ covariance matrix V_y. Then, if H is the $N \times k$ matrix with element $H_{in} = f_n(x_i)$, the solution $\widehat{a}$ is given by the solution to the *normal equation*

$$(H^T\, V_y^{-1}\, H)\, \widehat{a} = H^T\, V_y^{-1}\, \vec{y} \,, \qquad (2.15a)$$

or, formally,

$$\widehat{a} = (H^T\, V_y^{-1}\, H)^{-1}\, H^T\, V_y^{-1}\, \vec{y} \equiv D\vec{y} \,, \qquad (2.15b)$$

where $\vec{y}$ is the N-element vector of measured y_i's. The normal equations may be solved by numerical methods much more computationally efficient than brute application of Eq. (2.15b). In particular, $H^T\, V_y^{-1}\, H$ is sometimes singular or nearly singular. In such cases there is at least one f_n which may be expressed as a linear combination of others (or nearly so) when evaluated at the data points. The best procedure is usually to drop such functions from the expansion (or set $\widehat{a}_n = 0$). See Press [8], Maindonald [9], or Basilevsky [10] for discussions.

In terms of the $k \times N$ matrix D, the standard covariance matrix for the $\widehat{a}$ is estimated by

$$V_{\widehat{a}} = D\, V_y\, D^T \,. \qquad (2.16)$$

If the measured y_i's are independent, V_y is diagonal with ii^{th} element σ_i^2 and $V_{\widehat{a}}$ is obtained from Eq. (2.13) above.

The expected covariance [see Eq. (1.8)] of $\widehat{a}_n$ and $\widehat{a}_m$ is estimated by

$$\boldsymbol{E}\left[(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)\right] = (V_{\widehat{a}})_{nm} \,. \qquad (2.17)$$

Even when the y_i's are independent (diagonal V_y), $\widehat{a}_n$ and $\widehat{a}_m$ may not be (nondiagonal $V_{\widehat{a}}$). For the model function $y = \sum a_n\, f_n(x)$, the estimated variance of an interpolated or extrapolated value of y at a point x is

$$\boldsymbol{E}\left[(y - \widehat{y})^2\right] = \sigma^2(y)$$

$$= \sum_{n,m} (V_{\widehat{a}})_{nm}\, f_n(x)\, f_m(x) \,. \qquad (2.18)$$

If y is not linear in the fitting parameters a_n, or if the errors σ_i depend upon y and therefore on a_n, the solution vector may have to be found by iteration of Eqs. (2.12)–(2.14) or Eq. (2.15b). The same results may be obtained by numerical techniques from the sum of squares, χ^2, directly, if we have a reasonable first guess $\vec{a}_0$ for the solution vector:

$$\widehat{a} = \vec{a}_0 - \left(\frac{\partial^2\chi^2}{\partial a^2}\right)_{\vec{a}_0}^{-1} \cdot \left.\frac{\partial\chi^2}{\partial a}\right|_{\vec{a}_0} \qquad (2.19a)$$

and

$$V_{\widehat{a}} = 2\left(\frac{\partial^2\chi^2}{\partial a^2}\right)_{\widehat{a}}^{-1} \,, \qquad (2.19b)$$

where $\partial\chi^2/\partial a$ is a k-element vector whose n^{th} element is $\partial\chi^2/\partial a_n$, $\partial^2\chi^2/\partial a^2$ is a $k \times k$ matrix with mn^{th} element $\partial^2\chi^2/(\partial a_m \cdot \partial a_n)$, and all derivatives are to be evaluated at the points indicated. If "χ^2" is a true χ^2, the second-derivative matrix is independent of $\vec{a}$; therefore the shape of the χ^2 as a function of $\vec{a}$ is a paraboloid and Eq. (2.19a) will give the solution immediately. Otherwise one may need to iterate Eq. (2.19a) to arrive at a solution (Newton-Raphson method).

Note that in Eq. (2.15b), one needs only a matrix proportional to V_y to find $\widehat{a}$. Hence, for example, if the variances σ_i^2 of the errors are unknown but assumed equal and independent, and $\boldsymbol{E}(\epsilon_i) = 0$, one can still solve for $\widehat{a}$. One cannot, however, solve for $V_{\widehat{a}}$ or evaluate goodness-of-fit. These can be estimated from the *residuals*, $r_i = \widehat{y}(x_i) - y_i$, where $\widehat{y}(x_i)$ is the fitted curve at x_i, because study of the r_i enables one to estimate V_y. In addition, the residuals can be used to look for evidence of bias such as trends in the data not incorporated in the model [3].

Note that the errors on the solution $\widehat{a}$ are independent of the value of χ^2 at minimum—they depend only upon the shape about the minimum. Eq. (2.19b) implies that one-standard-deviation limits on the elements of $\widehat{a}$ are given by the set of $\vec{a}'$ such that

$$\chi^2(\vec{a}') = \chi^2_{\min} + 1 \,; \qquad (2.20)$$

compare with Eq. (2.8) for the general maximum-likelihood case. This equation, which defines a contour in $\vec{a}$-space, is often convenient for estimating errors in applications of least-squares techniques to *nonlinear* cases, where the second derivative [Eq. (2.19b)] may be a rapidly varying function of $\vec{a}$. In general, contours at s standard deviations may be found by replacing the 1 in Eq. (2.20) by s^2. If the problem is highly nonlinear, all such contours are at best only approximations to desired exact confidence regions which would have some given probability of covering the true value of $\vec{a}$. It may be that Eq. (2.20) will define a set of disjoint regions. In addition, iteration of Eq. (2.19a) may require sophisticated techniques [8] to reach convergence in a practical amount of computation. For example, in cases involving many variables in $\vec{a}$, especially if the correlations are not small, simplex or other techniques which do not involve explicit calculation of derivatives are often to be preferred. Such techniques are designed to find their way through complicated nonlinear problems without diverging to infinite $\vec{a}$ (unless the minimum is actually at infinity).

Least-squares estimation requires that an error matrix V_y be known (a matrix proportional to V_y will suffice to find an estimator). For counting experiments it is therefore necessary to group the data in bins in order to associate a Poisson error with each bin. In this case y_i is the bin height and the error depends on the expectation value of the theory in each bin, N_i^{th}, as estimated by the best fit of the model. Thus the requirements of the Gauss-Markov theorem are not satisfied, since the errors are not fixed. Many experimenters arrange the bins to contain enough expected events (say $\gtrsim 7$ or 8) that the Gaussian approximation to the Poisson (Sec. 1.3.3) is accurate, in which case the expected error is the square root of the theoretical height and "χ^2" is approximately a true χ^2. If an approximate error is used, based on the actual observed height $N_i^{\rm obs}$ rather than the theoretical height N_i^{th}, the Gauss-Markov conditions would be satisfied except that a bias favoring downward fluctuations will occur.

This is because a fluctuation in the data which goes down from the true expectation value will be assigned a smaller error and therefore a greater weight than an equal fluctuation upward. For bins with few events, a procedure that converges to the above when N_i^{th} is large and yields correct error estimates for all N_i^{th} is to define

$$\chi^2 = \sum_i \left[2(N_i^{th} - N_i^{\rm obs}) + 2N_i^{\rm obs} \ln(N_i^{\rm obs}/N_i^{th}) \right] . \tag{2.21}$$

This assumes that $N_i^{\rm obs}$ is the outcome of a Poisson process, with Poisson parameter $\mu = N_i^{th}$, in the i^{th} bin. In bins where $N_i^{\rm obs} = 0$, the second term is zero. For any N_i^{th}, s-standard-deviation error estimates are constructed as in Eq. (2.20) and subsequent discussion. If we drop the requirement that χ^2 converge to a true χ^2 for large numbers of events in each bin, then minimizing "χ^2" $= 2\sum_i [N_i^{th} - N_i^{\rm obs} \ln(N_i^{th})]$ will give the same answer and errors, with slightly faster execution, as the above.

In the more general maximum likelihood case, the small-number distributions are well known and there are no corresponding requirements concerning large numbers or even of binning.

Example: straight-line fit

For the case of a *straight-line fit*, $y(x) = a_1 + a_2\, x$, one obtains, for independent measurements y_i, the following estimates of a_1 and a_2,

$$\begin{aligned} \widehat{a}_1 &= (S_y\, S_{xx} - S_x\, S_{xy})/D \ , \\ \widehat{a}_2 &= (S_1\, S_{xy} - S_x\, S_y)/D \ , \end{aligned} \tag{2.22}$$

where

$$S_1,\, S_x,\, S_y,\, S_{xx},\, S_{xy} = \sum (1,\, x_i,\, y_i,\, x_i^2,\, x_i\, y_i)/\sigma_i^2 \ , \tag{2.23}$$

respectively, and

$$D = S_1\, S_{xx} - S_x^2 \ .$$

The covariance matrix of the fitted parameters is:

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \ . \tag{2.24}$$

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

$$(\widehat{y} - y_{\rm true})^2 \Big|_{\rm est} = \frac{1}{S_1} + \frac{S_1}{D} \left(x - \frac{S_x}{S_1} \right)^2 \ . \tag{2.25}$$

2.4 Errors and confidence intervals

2.4.1 Gaussian errors

If the data are such that the distribution of the estimator(s) satisfies the central limit theorem discussed in Sec. 1.3.4, the Gaussian distribution is the basis of the error analysis. If there is more than one parameter being estimated, the multivariate Gaussian is used. We define a *confidence interval* as being an interval constructed from the data to have probability *at least* $1 - \alpha$ (α is called the *confidence coefficient*) of covering the true value of θ. For the univariate case with known σ,

$$1 - \alpha = \int_{\widehat{\mu}-\delta}^{\widehat{\mu}+\delta} f(x;\, \widehat{\mu},\, \sigma^2)\, dx \tag{2.26}$$

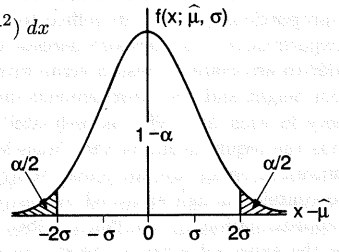

Fig. 4. Illustration of a two standard-deviation confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by α, are as shown.

is the probability that the true value of μ will fall within $\pm\delta$ ($\delta > 0$) of the measured $\widehat{\mu}$. This interval will cover μ in a fraction $1 - \alpha$ of all similar measurements. Fig. 4 shows a $\delta = 2\sigma$ confidence interval unshaded. The choice $\delta = \sqrt{{\rm Var}(\widehat{\mu})} \equiv \sigma$ gives an interval called the *standard error* which has $1 - \alpha = 68.33\%$ if σ is known. Other frequently used choices for δ, in terms of α are:

α (%)	δ	α (%)	δ
31.73	1σ	20	1.28σ
4.55	2σ	10	1.64σ
0.27	3σ	5	1.96σ
6.3×10^{-3}	4σ	1	2.58σ
5.7×10^{-5}	5σ	0.1	3.29σ
2.0×10^{-7}	6σ	0.01	3.89σ

For other δ, find α as the ordinate of Fig. 1 on the $n = 1$ curve at $\chi^2 = (\delta/\sigma)^2$. We can set a one-sided (upper or lower) limit by excluding above $\widehat{\mu} + \delta$ (or below $\widehat{\mu} - \delta$); α's for such limits are 1/2 the values in the table above.

Note that we have increased confidence that the interval covers the true value as $1 - \alpha$ increases, or χ^2 increases. We must be careful to distinguish this case from the other major use of Fig. 1, evaluation of goodness-of-fit (Sec. 2.3.3). In that case we have increased confidence in the fit as χ^2 decreases. In an attempt to reduce possible confusion in this discussion, we will use the α notation (which corresponds to notation used in hypothesis testing [3]) when discussing confidence intervals and CL notation when discussing goodness-of-fit. Elsewhere in this Review, where the confusion between fit confidence level and interval (usually an upper or lower limit) confidence level does not arise, we follow the common practice of using "CL" to refer to the confidence level of the interval. This CL is understood to represent $1 - \alpha$.

If the variance σ^2 of the estimator is not known, but must be estimated from the data, then we need to incorporate the error in $\widehat{\sigma}$ into our confidence interval using Student's t distribution. If we have N data points with which we estimate k parameters, the Gaussian approximation is adequate for $N - k \gg 1$. Otherwise replace δ by a factor $T\widehat{\sigma}$, T being defined by

$$1 - \alpha = \int_{-T}^{T} f(x; N - k)\, dx \ , \tag{2.27}$$

where f is defined in Eq. (1.34). T is tabulated in Ref. 2 and here:

	α (%)					
$N - k$	31.67	10.00	5.00	4.55	1.00	0.27
1	1.84	6.31	12.71	13.97	63.66	235.78
2	1.32	2.92	4.30	4.53	9.92	19.21
3	1.20	2.35	3.18	3.31	5.84	9.22
4	1.14	2.13	2.78	2.87	4.60	6.62
5	1.11	2.01	2.57	2.65	4.03	5.51
10	1.05	1.81	2.23	2.28	3.17	3.96
20	1.03	1.72	2.09	2.13	2.85	3.42
∞	1.00	1.64	1.96	2.00	2.58	3.00

For multivariate θ we must consider pairwise correlations. Assuming a multivariate Gaussian, Eq. (1.28a), and subsequent discussion the standard error ellipse for the pair $(\widehat{\theta}_m, \widehat{\theta}_n)$ may be drawn as in Fig. 5.

The minimum χ^2 or maximum likelihood solution is at $(\widehat{\theta}_m, \widehat{\theta}_n)$. The standard errors σ_m and σ_n are defined as shown, where the ellipse is at a constant value of $\chi^2 = \chi^2_{\rm min} + 1$ or $\ln \mathcal{L} = \ln \mathcal{L}_{\rm max} - 1/2$. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{mn}\, \sigma_m\, \sigma_n}{\sigma_m^2 - \sigma_n^2} \ . \tag{2.28}$$

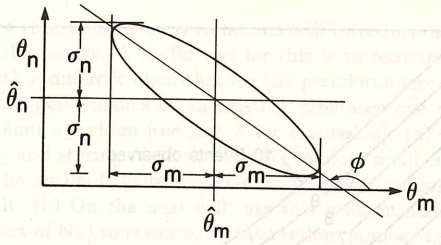

Fig. 5. Standard error ellipse for the estimators $\widehat{\theta}_m$ and $\widehat{\theta}_n$. In this case the correlation is negative.

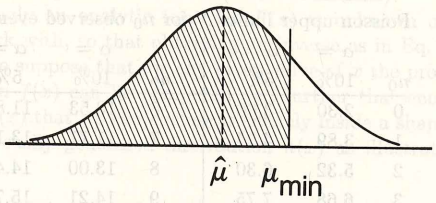

Fig. 6. An example of a bounded physical region with Gaussian errors. In this case the estimator $\widehat{\mu}$ has fallen within the unphysical region due to random error.

For non-Gaussian or nonlinear cases, one may construct an analogous contour from the same χ^2 or $\ln\mathcal{L}$ relations. Any other parameters $\widehat{\theta}_\ell, \ell \neq m, n$, must be allowed freely to find their optimum values for every trial point.

For any unbiased procedure (*e.g.*, least squares or maximum likelihood) being used to estimate k parameters θ_i, $i = 1, \ldots, k$, the probability $1 - \alpha$ that the true values of all k lie within the s-standard deviation ellipsoid may be found from Fig. 1. Read the ordinate as α; the correct value of α occurs on the $n = k$ curve at $\chi^2 = s^2$. For example, for $k = 2$, the probability that the true values of θ_1 and θ_2 simultaneously lie within the one-standard-deviation error ellipse ($s = 1$), centered on $\widehat{\theta}_1$ and $\widehat{\theta}_2$, is 39%. This probability only assumes Gaussian errors, unbiased estimators, and that the model describing the data in terms of the θ_i is correct.

2.4.2 Gaussian errors—bounded physical region

In certain statistical problems the true value of the parameter to be estimated, μ, is constrained to lie within a bounded *physical region* (*e.g.*, the mass of a neutrino is bounded from below by 0). However, due to random measurement error, real measured values may or may not occur inside the physical region. For this case no completely satisfactory approach exists, but here we suggest a technique for obtaining limits within the physical region approximately at specified confidence levels. The "classical" statistical techniques of the previous section can still be used for confidence intervals at some exact α. However, such limits are useful mainly in the statistical sense where it is assumed that no bound exists. In bad cases, the limit may exclude the physical region entirely, or extend into it a small distance and create the false impression of a powerful limit close to the edge of the physical region.

We assume a measurement x, which represents one observation (or the result of combining multiple measurements as in Sec. 2.2) from a Gaussian of true (but unknown) mean μ and known, fixed, variance σ^2. We *estimate* μ by $\widehat{\mu} = x$ and attempt to construct a confidence interval for μ from the resultant Gaussian, as above. If $\widehat{\mu}$ or a significant portion of the probability lies in the unphysical region (Fig. 6), the result, while statistically perfectly correct as stated, is physically unsatisfactory.

If we assume μ is bounded from below by $\mu_{\min}$ (the argument for μ bounded from above is similar), we may estimate a reasonable upper limit for μ at the $1 - \alpha$ (*e.g.*, 90% or 95%) level by the following procedure: (1) *renormalize* the Gaussian probability distribution for x such that the integral of Eq. (1.24) with $\mu = \widehat{\mu}$ over x from $\mu_{\min}$ to infinity (i.e., over the physical region), unshaded in the figure below, is equal to 1.0; (2) find the value μ_1 such that the integral over x of the renormalized distribution from $\mu_{\min}$ to μ_1 is equal to the desired value of $1 - \alpha$; (3) set μ_1 to be the desired upper limit with confidence $1 - \alpha$. In fact, it can be shown that this is *conservative*, in the sense that the probability that this interval actually covers the true value of μ is $\geq 1 - \alpha$

The "classical" approach as described above can be derived formally by the application of Bayes' theorem with the explicit assumption that all values of the parameter are equally probable. This means, for example, that limits on m^2 are different than limits on m. A recent treatment is given by James and Roos [11].

For $\mu - \mu_{\min} \gg \sigma$, this technique, which may be applied for any measured x (physical or unphysical), converges smoothly to that of the previous section since x is then effectively confined to the physical region.

One should exercise caution for values of x which lie many standard deviations outside the physical region. It may be that the particular probability model (Gaussian with variance σ^2) may not be a correct description of the measurement process (*e.g.*, the true variance may have unanticipated components and be $> \sigma^2$, or there may be a bias), in which case confidence levels of this sort will not be correct.

If $\widehat{\mu} < \mu_{\min}$, some authors prefer to use a fixed upper limit calculated for $\widehat{\mu} = \mu_{\min}$ or $\widehat{\mu} = \mu_{\min} + \sigma$, rather than allow the upper limit to decrease as $\widehat{\mu}$ decreases. In any case, averaging of experiments requires that $\widehat{\mu}$ and its variance be quoted, in addition to any upper limits, even if $\widehat{\mu}$ is unphysical.

2.4.3 Poisson processes—upper limits

Because the outcome of a Poisson process is an integral number of events, n_0, it is usually not possible to set confidence intervals for the true Poisson parameter μ at a certain exact α. For large n_0 an approximate interval can be set using the Gaussian approximation, Sec. 1.3.3, and the techniques of Sec. 2.4.1.

For small n_0 we can define an upper limit N for μ as being that value of μ such that it would be at least $1 - \alpha$ (*e.g.*, 90% or 95%) probable that a random observation of n would then lie above the observed n_0. Thus

$$1 - \alpha = \sum_{n=n_0+1}^{\infty} f(n; N) \; ; \qquad \alpha = \sum_{n=0}^{n_0} f(n; N) \; . \tag{2.29}$$

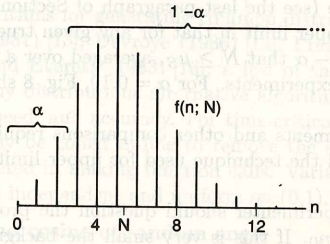

Fig. 7. Illustration of Eq. (2.29) Poisson probabilities for an assumed mean of N. With an observed count $n_0 = 2$, $N = 5.3$ as shown gives summed probability $1 - \alpha = 90\%$.

Fig. 7 illustrates the case with $n_0 = 2$ and $1 - \alpha = 90\%$, for which it may be shown that $N = 5.3$. For any given n_0 and desired α we can obtain N from the χ^2 Confidence Level figure because of a relation between the Poisson and the χ^2: read the ordinate as α, find χ^2 on the curve for $n = 2(n_0 + 1)$; then $N = \chi^2/2$. Some useful values are:

Poisson upper limits N for n_0 observed events

n_0	$\alpha =$ 10%	$\alpha =$ 5%	n_0	$\alpha =$ 10%	$\alpha =$ 5%
0	2.30	3.00	6	10.53	11.84
1	3.89	4.74	7	11.77	13.15
2	5.32	6.30	8	13.00	14.44
3	6.68	7.75	9	14.21	15.71
4	7.99	9.15	10	15.41	16.96
5	9.27	10.51			

The meaning of these upper limits is that, for a given true μ, the probability is *at least* $1-\alpha$ that one will observe n_0 which will result in N which is $\geq \mu$. The probability for that to occur may be higher than $1-\alpha$; for example, if $\mu \leq 2.30$ a "90%" upper limit will actually exceed μ 100% of the time. Note from Eq. (2.29) that for $n_0 = 0$, $N = \ln[1/(1-\alpha)]$.

2.4.4 Poisson processes with background [12]

If we observe n_0 events in a Poisson process which has two components, signal and background, estimating a limit on the signal is more complicated. Let μ_S be the unknown mean (the Poisson parameter) for the signal and μ_B be the mean for the sum of all backgrounds. Assume μ_B is known with negligible error; however we don't know n_B, the actual number of events resulting from the background. We do know that $n_B \leq n_0$. If $\mu_B + \mu_S$ is large, the Gaussian approximation to the Poisson distribution (see Sec. 1.3.3) is usually adequate, and one can define confidence intervals or limits as above, assuming $\widehat{n}_B \approx \mu_B$ and therefore $\widehat{\mu}_S = n_0 - \mu_B$ with variance equal to n_0 (larger than $\widehat{\mu}_S$ to allow for the error in $\widehat{n}_B$).

Otherwise an upper limit can be defined by extension of the argument of the preceding section. Let N be the desired upper limit on μ_S with confidence coefficient α. Set N to be that value of μ_S such that any random repeat of the current experiment with $\mu_S = N$ and the same μ_B would observe *more* than n_0 events in total *and* would have $n_B \leq n_0$, all with probability $1-\alpha$. For any assumed N and μ_B we can calculate this probability:

$$1-\alpha = 1 - \frac{e^{-(\mu_B+N)} \sum_{n=0}^{n_0} \frac{(\mu_B+N)^n}{n!}}{e^{-\mu_B} \sum_{n=0}^{n_0} \frac{\mu_B^n}{n!}} \,. \tag{2.30}$$

We adjust N to obtain a desired α. For $\mu_B = 0$ this converges to (2.29). As in that case (see the last paragraph of Section 2.4.3) this gives a *conservative* upper limit in that for any given true μ_S we get a true probability $\geq 1-\alpha$ that $N \geq \mu_S$, averaged over a large set of identically performed experiments. For $\alpha = 0.10$, Fig. 8 shows N as a function of n_0 and μ_B.

Averaging of experiments and other comparisons require that n_0 and μ_B be quoted and the technique used for upper limit extraction be given.

If $\mu_B \gg n_0$ the experimenter should question the probability of observing n_B as that n_0. If this is very small the background, μ_B, may not have been calculated properly and the upper limit for μ_S obtained under those assumptions may be too low. For example, in Fig. 8, the dashed portions of the curves lie in the region where n_0 is expected to exceed the observed value 99% of the time (or more), even in the complete absence of signal. In these regions one should be cautious about accepting the results of the measurement.

As in the Gaussian case (2.4.2), whenever $n_0 < \mu_B$ some experimenters may prefer to use N calculated as if $n_0 \approx \mu_B$ rather than the smaller value obtained from the observed n_0.

2.5 Propagation of errors

Suppose we have a set of N random variables y_i which may be direct measurements or derived estimators $\widehat{\theta}$, and we have a covariance matrix $V(y)$ for these. We can make a transformation to a different

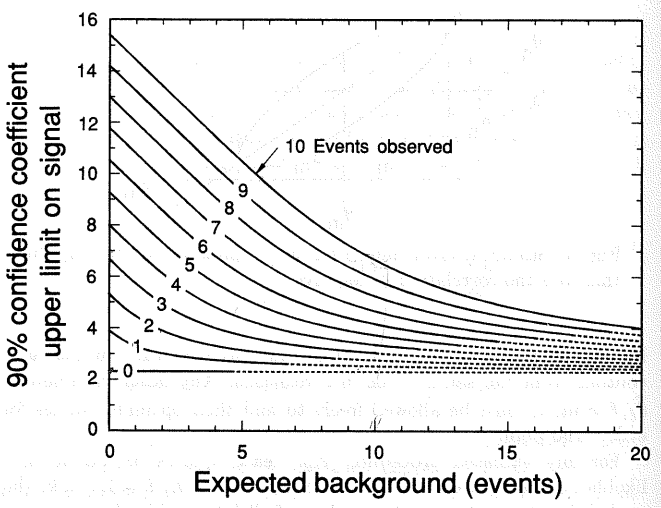

Fig. 8. 90% confidence coefficient upper limit on the number of signal events as a function of the expected number of background events. For example, if the expected background is 8 events and 5 events are observed, then the signal is 4.0 (approximately) or less with 90% confidence. Dashed portions indicate regions where it is to be expected that the number observed would exceed the number actually observed $\geq 99\%$ of the time, even in the complete absence of signal.

set of variables $f_j \equiv f_j(y)$, $j = 1, \ldots, M$ $(M \leq N)$ and obtain best estimates for the f_j from

$$\widehat{f}_j \approx f_j(\widehat{y}) + \frac{1}{2} \sum_{k,n}^{N} V_{kn}(\widehat{y}) \left[\frac{\partial^2 f_j}{\partial y_k \, \partial y_n} \right]_{\widehat{y}} \tag{2.31}$$

with covariance matrix

$$V_{ij}(\widehat{f}) \approx \sum_{n,m} \left. \frac{\partial f_i}{\partial y_n} \right|_{\widehat{y}} \left. \frac{\partial f_j}{\partial y_m} \right|_{\widehat{y}} V_{nm}(\widehat{y}) \,. \tag{2.32}$$

For a single-valued function f of a single measurement y with variance σ^2 (i.e., $M = 1, N = 1$), this becomes

$$\begin{aligned} \widehat{f} &\approx f(\widehat{y}) + \frac{1}{2}\sigma^2 f''(\widehat{y}) \\ V(\widehat{f}) &\approx \sigma^2 [f'(\widehat{y})]^2 \,, \end{aligned} \tag{2.33}$$

where the primes denote differentiation with respect to y, evaluated at $\widehat{y}$.

These approximations are based on a Taylor expansion of f about the true value of y. If f is approximately linear in y over a range of roughly $\widehat{y}_i \pm \sigma(y_i)$, the approximation is good and the second-order terms in (2.31) and (2.33) can be neglected. This is what is usually done. However, if linearity is badly violated (*e.g.*, $f \propto 1/y$ and $\widehat{y}$ is no more than a few σ from zero), it should be recognized that propagation of errors will give very approximate results. In such cases $\widehat{f} \approx f(\widehat{y})$ may be a biased estimator for f even if $\widehat{y}$ is unbiased for y, and the second-order terms in (2.31) and (2.33) will help to reduce that bias.

3. MONTE CARLO TECHNIQUES

Monte Carlo techniques are used to simulate on a computer random behavior which is too complex to be derived analytically. Most calculations are based upon *pseudorandom* numbers, a reproducible sequence of numbers generated on the open interval (0,1) in such a way that they satisfy various statistical tests for a uniform distribution, with independent numbers. (Caution: some commercial random number generators fill the *closed* interval [0,1]. The occurrence of 0 or 1 can sometimes cause problems for the algorithms below). No such numbers are truly uniform and independent. Many commercial random number generators sacrifice randomness in favor of speed. It

is not rare that unforeseen correlations will introduce non-negligible errors in the results. A useful test for this is to recompute the same results with a different algorithm for the pseudorandom numbers. To improve the performance of an existing generator one may use the *Bays-Durham algorithm* [see Ref. 8 for discussion]: (a) Initialize by generating and storing N (*e.g.*, $N = 97$) random numbers in an array v, using the available generator. Generate a new random number u and save it. (b) On the next call, use this u as an address $j = 1+$ (integer part of Nu) to select v_j as the random number to be returned. Also save this v_j as u for the next call. Replace v_j in the array with a new random number using the available generator. On the next call, go to (b).

A second problem sometimes encountered in computations requiring long sequences of random numbers is that all pseudorandom number generators will eventually begin over and repeat the same sequence. One may choose algorithms which minimize the number used. One may also use two or three different generators in different parts of the program.

Monte Carlo simulations of complex processes break them down into a sequence of steps. At each step a particular outcome is chosen from a set of possibilities according to a certain p.d.f. To do this we must transform our uniform random numbers into random numbers sampled from different distributions on different ranges.

Two techniques are in wide use to do this. We will discuss only single variable cases; multiple variable cases use straightforward extensions of these techniques. We assume we are in possession of a random number u chosen from a uniform distribution on (0,1).

3.1 Inverse transform method

If the desired probability density function is $f(x)$ on the range $-\infty < x < \infty$, its cumulative distribution function (expressing the probability that $x \le a$) is given by Eq. (1.1). If a is chosen with probability density $f(a)$, then the integrated probability up to point a, $F(a)$, is itself a random variable which will occur with uniform probability density on $[0, 1]$. Ignoring the endpoints, we can then find a unique x distributed as $f(x)$ for $f(x)$ *continuous*, for a given u if we set

$$u = F(x) \ , \tag{3.1}$$

provided we can find an inverse of F, defined by

$$x = F^{-1}(u) \ , \tag{3.2}$$

as is illustrated in Fig. 9

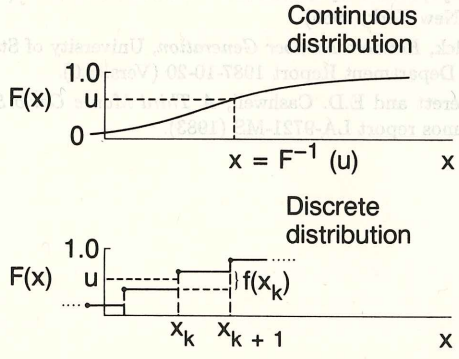

Fig. 9. Use of a random number u chosen from a uniform distribution (0,1) to find a random number x from a distribution with cumulative distribution function $F(x)$.

For a *discrete* distribution, $F(x)$ will have a discontinuous jump of size $f(x_k)$ at each allowed $x_k, k = 1, 2, \cdots$. Choose u from a uniform distribution on (0,1) as before. Find x_k such that

$$F(x_{k-1}) < u \le F(x_k) \equiv \text{Prob}\ (x \le x_k) = \sum_{i=1}^{k} f(x_i) \ ; \tag{3.3}$$

then x_k is the value we seek (note: $F(x_0) \equiv 0$).

3.2 Acceptance-rejection method (Von Neumann)

Very commonly an analytic form for $F(x)$ is unknown or too complex to work with, so that obtaining an inverse as in Eq. (3.2) is impractical. We suppose that for any given value of x the probability density function $f(x)$ can be computed and further that enough is known about $f(x)$ that we can enclose it entirely inside a shape which is C times an easily generated distribution $h(x)$ as illustrated in Fig. 10.

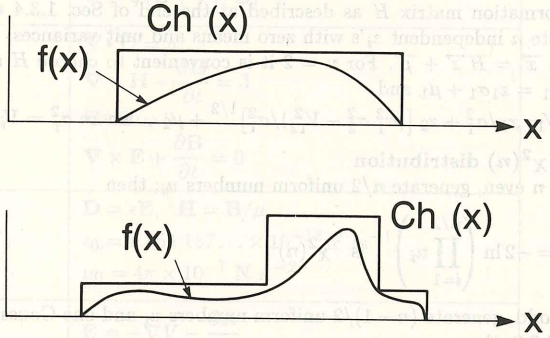

Fig. 10. Illustration of the acceptance-rejection method. Random points are chosen inside the upper bounding figure, and rejected if the ordinate exceeds $f(x)$. Lower figure illustrates importance sampling.

Frequently $h(x)$ is uniform or is a normalized sum of uniform distributions. Note that both $f(x)$ and $h(x)$ must be normalized to unit area and therefore the proportionality constant $C > 1$. To generate $f(x)$, first generate a candidate x according to $h(x)$. Calculate $f(x)$ and the height of the envelope $C\,h(x)$; generate u and *test* if $uC\,h(x) \le f(x)$. If so, accept x; if not reject x and try again. If we regard x and $uC\,h(x)$ as the abscissa and ordinate of a point in a two-dimensional plot, these points will populate the entire area $C\,h(x)$ in a smooth manner; then we accept those which fall under $f(x)$. The efficiency is the ratio of areas, which must equal $1/C$; therefore we must keep C as close as possible to 1.0. Therefore we try to choose $C\,h(x)$ to be as close to $f(x)$ as convenience dictates, as in the lower part of Fig. 10. This practice is called *importance sampling*, because we generate more trial values of x in the region where $f(x)$ is most important.

3.3 Algorithms

Many algorithms for generating common distributions are given by Rubinstein (1981) [13], Devroye (1986) [14], Press (1986) [8], Walck (1987) [15], and Everett (1983) [16]; a few of these are reproduced here. For many distributions alternative algorithms exist, varying in complexity, speed, and accuracy. For time-critical applications, these algorithms may be coded in-line to remove the significant overhead often encountered in making function calls. Variables named "u" are assumed to be independent and uniform on (0,1).

3.3.1 Sine and cosine of random angle

Generate u_1 and u_2. Then $v_1 = 2u_1 - 1$ is uniform on $(-1,1)$, and $v_2 = u_2$ is uniform on (0,1). Calculate $r^2 = v_1^2 + v_2^2$. If $r^2 > 1$, start over. Otherwise, the sine (S) and cosine (C) of a random angle are given by

$$S = 2v_1v_2/r^2 \quad \text{and} \quad C = (v_1^2 - v_2^2)/r^2 \ .$$

3.3.2 Gaussian distribution

If u_1 and u_2 are uniform on (0,1), then

$$z_1 = \sin 2\pi u_1 \sqrt{-2 \ln u_2} \quad \text{and} \quad z_2 = \cos 2\pi u_1 \sqrt{-2 \ln u_2}$$

are independent and Gaussian distributed with mean 0 and $\sigma = 1$.

There are many faster variants of this basic algorithm. For example, construct $v_1 = 2u_1 - 1$ and $v_2 = 2u_2 - 1$, which are uniform on $(-1,1)$.

PROBABILITY, STATISTICS, AND MONTE CARLO (Cont'd)

Calculate $r^2 = v_1^2 + v_2^2$, and if $r^2 > 1$ start over. If $r^2 < 1$, it is uniform on (0,1). Then

$$z_1 = v_1\sqrt{\frac{-2\ln r^2}{r^2}} \quad \text{and} \quad z_2 = v_2\sqrt{\frac{-2\ln r^2}{r^2}}$$

are independent numbers chosen from a normal distribution with mean 0 and variance 1. $z_i' = \mu + \sigma z_i$ distributes with mean μ and variance σ^2.

For a multivariate Gaussian it often is simplest to find a transformation matrix H as described at the end of Sec. 1.3.4 and generate n independent z_i's with zero means and unit variances; then return $\vec{x} = H\vec{z} + \vec{\mu}$. For $n = 2$ it is convenient to choose H such that $x_1 = z_1\sigma_1 + \mu_1$ and

$x_2 = V_{12}\, x_1/\sigma_1^2 + z_2\left[(\sigma_1^2\,\sigma_2^2 - V_{12}^2)/\sigma_1^2\right]^{1/2} + \mu_2$, where $\sigma_i^2 = V_{ii}$.

3.3.3 $\chi^2(n)$ distribution

For n even, generate $n/2$ uniform numbers u_i; then

$$y = -2\ln\left(\prod_{i=1}^{n/2} u_i\right) \quad \text{is} \quad \chi^2(n)\ .$$

For n odd, generate $(n-1)/2$ uniform numbers u_i and one Gaussian z as in 3.3.2; then

$$y = -2\ln\left(\prod_{i=1}^{(n-1)/2} u_i\right) + z^2 \quad \text{is} \quad \chi^2(n)\ .$$

For $n \gtrsim 30$ the much faster Gaussian approximation for the χ^2 may be preferable: generate z as in 3.3.2 and use $y = \left[z + \sqrt{2n-1}\,\right]^2/2$; if $z < -\sqrt{2n-1}$ reject and start over.

3.3.4 Binomial distribution

If $p \le 1/2$ in Eq. (1.20), iterate until a successful choice is made: begin with $k = 1$; compute $P_k = q^n$ [for $k \ne 1$ use $P_k \equiv f(r_k; n, p)$, Eq. (1.20)] and store P_k into B; generate u. If $u \le B$ accept $r_k = k - 1$ and stop; otherwise increment k by 1 and compute next P_k and add to B; generate a new u and repeat. If we arrive at $k = n+1$, stop and accept $r_{n+1} = n$. If $p > 1/2$ it will be more efficient to generate r from $f(r; n, q)$, i.e., with p and q interchanged, and then set $r_k = n - r$.

3.3.5 Poisson distribution

Iterate until a successful choice is made: Begin with $k = 1$ and set $A = 1$ to start. Generate u. Replace A with uA; if now $A < \exp(-\mu)$, where μ is the Poisson parameter, accept $n_k = k - 1$ and stop. Otherwise increment k by 1, generate a new u and repeat, always starting with the value of A left from the previous try. For large $\mu(\gtrsim 10)$ it may be satisfactory (and much faster) to approximate the Poisson distribution by a Gaussian distribution [Sec. 1.3.4] and generate z from $f(z;0,1)$; then accept $x = \max(0, [\mu + z\sqrt{\mu} - 0.5])$ where [ ] signifies the greatest integer $\le$ the expression.

3.3.6 Student's t distribution

For $n > 0$ degrees of freedom (n not necessarily integer), generate x from a Gaussian with mean 0 and $\sigma^2 = 1$ according to the method of 3.3.2. Next generate y, an independent gamma random variate with $k = n/2$ degrees of freedom. Then $z = x\sqrt{2n}/\sqrt{y}$ is distributed as a t with n degrees of freedom.

For the special case $n = 1$, the *Breit-Wigner* distribution, generate u_1 and u_2; set $v_1 = 2u_1 - 1$ and $v_2 = 2u_2 - 1$. If $v_1^2 + v_2^2 \le 1$ accept $z = v_1/v_2$ as a Breit-Wigner distribution with unit area, center at 0.0, and FWHM 2.0. Otherwise start over. For center M_0 and FWHM Γ, use $W = z\Gamma/2 + M_0$.

Revised April 1992.

1. H. Cramér, *Mathematical Methods of Statistics,* Princeton Univ. Press, New Jersey (1958).
2. M. Abramowitz and I. Stegun, eds., *Handbook of Mathematical Functions* (Dover, New York, 1972).
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).
4. L. Lyons, *Statistics for Nuclear and Particle Physicists* (Cambridge University Press, New York, 1986).
5. S.L. Meyer, *Data Analysis for Scientists and Engineers* (John Wiley and Sons, Inc., New York, 1975).
6. A.G. Frodesen, O. Skjeggestad, and H. Tøfte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Oslo, Norway, 1979).
7. R.A. Fischer, *Statistical Methods for Research Workers,* 8th edition, Edinburgh and London (1941).
8. W.H. Press *et al.*, *Numerical Recipes* (Cambridge University Press, New York, 1986).
9. W.H. Maindonald *et al.*, *Statistical Computation* (John Wiley and Sons, Inc., New York, 1984).
10. A. Basilevsky *et al.*, *Applied Matrix Algebra in the Statistical Sciences* (North Holland, New York, 1983).
11. F. James and M. Roos, Phys. Rev. **D44**, 299 (1991).
12. O. Helene, Nucl. Instr. and Meth. **212**, 319 (1983).
13. R.Y. Rubinstein, *Simulation and the Monte Carlo Method* (John Wiley and Sons, Inc., New York, 1981).
14. L. Devroye, *Non-Uniform Random Variate Generation* (Springer-Verlag, New York, 1986).
15. Ch. Walck, *Random Number Generation,* University of Stockholm Physics Department Report 1987-10-20 (Vers. 3.0).
16. C.J. Everett and E.D. Cashwell, *A Third Monte Carlo Sampler,* Los Alamos report LA-9721-MS (1983).