PROBABILITY AND STATISTICS
APPLIED TO HIGH-ENERGY PHYSICS

D. Drijard, W.T. Fadde,
F. James, M. Roos, B. Sadoulet

Lectures given in the
Academic Training Programme of CERN
1958 - 1959
Contents

II ELEMENTS OF PROBABILITY (continued)

2. Special distributions
   2.1 A useful tool. The characteristic function
      2.1.1 Definition and properties
      2.1.2 Probability generating function
   2.2 Discrete distributions
      2.2.1 Binomial
      2.2.2 Multinomial
      2.2.3 Poisson
      2.2.4 Geometric and negative binomial
   2.3 Continuous distributions
      2.3.1 Uniform
      2.3.2 Normal one-dimensional
      2.3.3 Normal many-dimensional
      2.3.4 Exponential
      2.3.5 F and $\chi^2$
      2.3.6 Student's t
      2.3.7 F
      2.3.8 Cauchy or Breit-Wigner
2. **SPECIAL DISTRIBUTIONS**

The probability densities and cumulative distributions encountered in real life are often well approximated by a few mathematical functions. In this section we shall therefore present several such "ideal" distributions, their properties, and some techniques for handling them.

2.1 **A useful tool -- the characteristic function**

The Fourier transform of the probability density function is called the characteristic function of the random variable. It has many useful and important properties. In particular, the distribution of sums of random variables is most easily handled by means of characteristic functions. For many distributions, the moments are also best obtained from the properties of these functions.

2.1.1 **Definition and properties**

Given a random variable $X$ with density $f(X)$, the characteristic function is defined as:

\[
\phi_X(t) = \mathbb{E}(e^{itX}) = \begin{cases} 
\int_{-\infty}^{\infty} f(x) e^{itx} dx & (t \text{ real}) \\
\sum_{k} p_k e^{ikt} & (X \text{ discrete})
\end{cases}
\]  

The characteristic function $\phi_X(t)$ determines completely the probability distribution of the random variable. In particular, if $F(X)$ is continuous everywhere and $dF(X) = f(X)dx$, then

\[
\phi_X(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(t) e^{-itx} dt
\]

The function $\phi(t)$ has the properties

\[
\phi(0) = 1, \quad |\phi(t)| \leq 1.
\]
If a and b are constants, 
\[ \phi_{aX+b}(t) = e^{ibt} \phi_X(at) \]  
(40)

since 
\[ E(e^{it(aX+b)}) = E(e^{ibt} e^{iatX}) = e^{ibt} \phi_X(at) \]

If X and Y are independent random variables, with characteristic functions \( \phi_X(t) \), \( \phi_Y(t) \), then the characteristic function of the sum \( X + Y \) is 
\[ \phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t). \]

Proof:
\[ \phi_{X+Y}(t) = E(e^{it(X+Y)}) = E(e^{itX} e^{itY}) = E(e^{itX}) E(e^{itY}) = \phi_X(t) \cdot \phi_Y(t). \]

This property is much simpler than the corresponding property for the p.d.f. of the sum of independent variables:
\[ f(z = X + Y) = \int_{-\infty}^{\infty} f_X(z-t) f_Y(t) \, dt \]  
(41)

In general, the characteristic function of the sum of n independent random variables is 
\[ \phi_{X_1 + X_2 + \ldots + X_n}(t) = \prod_{i=1}^{n} \phi_{X_i}(t), \]  
(42)

The relation between the characteristic function and the moments of a distribution can be found from the formal expansion 
\[ \phi_X(t) = E(e^{itX}) = E\left(\sum_{r=0}^{\infty} \frac{(itX)^r}{r!}\right) \]  
\[ = \sum_{r=0}^{\infty} \frac{(it)^r}{r!} \mu_r, \]  
(43)
Thus $\mu_1$ appears as the coefficient of $(it)^r/r!$ in the expansion of $\phi(t)$.

Note:

i) If a distribution has a variance $\sigma^2$ and mean $\mu$, then

$$\phi(t) = 1 + it\mu + \frac{1}{2!}(\sigma^2 + \mu^2)(it)^2 + \cdots$$

ii) If the mean of a distribution is $\mu$, then the transformation

$$e^{-i\mu t} \phi(t) = \mathbb{E}(e^{it(x-\mu)}) = \phi(x-\mu)$$

generates the central moments about the mean.

It now follows that

$$\mu_r = \left. \frac{1}{i^r} \left( \frac{d^r}{dt^r} \phi(t) \right) \right|_{t=0}$$

Examples

i) Normal distribution

Suppose that $X$ is distributed as $N(\mu, \sigma^2)$.

Then

$$\phi(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{i\mu t} e^{-\frac{1}{2}(x-\mu)^2 \sigma^2} \, dx$$

$$= e^{i\mu t} - \frac{1}{2} t^2 \sigma^{-2}$$

Thus

$$e^{-iut} \phi(t) = \frac{e^{i\mu t}}{\sigma^{-2}} \left( \frac{\sigma^2}{2} \right)^r$$

Therefore

$$\mu_{2r-1} = 0$$

and

$$\mu_{2r} = \left( \frac{2r}{e^2} \right) \left( \frac{\sigma^2}{2} \right)^r$$

(47)

ii) Cauchy distribution

$$\phi(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{i\mu t}}{1 + x^2} \, dx = \frac{1}{e}$$
\( \phi(t) \) has no derivatives at \( t = 0 \), and so this distribution has no moments. In particular, the mean of the distribution does not exist.

2.1.2 Probability generating function

One can define a characteristic function for a discrete random variable. If \( X \) is an integer random variable, and

\[
P(X = r) = \rho_r
\]

then

\[
\phi_X(t) = \sum_{r=0}^{\infty} \rho_r e^{ir} \tag{48}
\]

However, it is more convenient to write \( z = e^{it} \). Then we call

\[
G(z) = E(z^X) = \sum_{r=0}^{\infty} \rho_r z^r \tag{49}
\]

the probability generating function of \( X \). The properties of \( G(z) \) are as follows:

- \( G(z) \) is regular for, at least, \( |z| < 1 \), since
  \[
  \rho_r \geq 0 \quad \text{and} \quad \sum_{r=0}^{\infty} \rho_r = 1
  \]

Thus

\[
G'(z) = \sum_{r=0}^{\infty} r \rho_r z^{r-1}
\]

Thus

\[
G'(1) = \sum_{r=0}^{\infty} r \rho_r \approx E(X) = \mu \tag{50}
\]

Thus

\[
G''(z) = \sum_{r=0}^{\infty} r (r-1) z^{r-2} \rho_r
\]

Thus

\[
G''(1) = E\{X(X - 1)\} = E(X^2) - E(X) \tag{51}
\]

and

\[
V(X) = G''(1) + G'(1) - [G'(1)]^2 \tag{52}
\]

This procedure is frequently the most direct method of obtaining the mean and variance of a discrete distribution.
2.2 Discrete distributions

2.2.1 Binomial distribution

Suppose that an experiment is repeated \( n \) times, and there are only two possible outcomes: \( A \) with probability \( p \), and \( \bar{A} \) with probability \( 1 - p \). Call \( A \) a success and \( \bar{A} \) a failure. The number of successes \( X \) in \( n \) trials is a random number (\( n \) is not a random number). The probability of getting \( X = r \) successes in \( n \) trials is given by the binomial distribution

\[
\rho_r = \binom{n}{r} p^r (1 - p)^{n-r}
\]  
\( r = 0, 1, 2, \ldots, n \)  
\( \sum_{r=0}^{n} \rho_r = 1 \)

Note that this probability density is properly normalized

\[
\sum_{r=0}^{n} \binom{n}{r} p^r (1 - p)^{n-r} = 1
\]

The expectation is

\[
E(X) = \sum_{r=0}^{n} r \rho_r
\]

Inserting Eq. (53) into Eq. (55) and using Eq. (54) we get

\[
E(X) = np \sum_{r=0}^{n-1} \binom{n-1}{r} p^r (1 - p)^{n-r} = np
\]

The variance is obtained from the probability generating function

\[
G(z) = E(z^X) = \sum_{r=0}^{n} \binom{n}{r} p^r (1 - p)^{n-r} z^r = \sum_{r=0}^{n} \binom{n}{r} (pz)^r q^{n-r}
\]

where \( q = (1 - p) \)

\[
G'(1) = np \left[ (q + p)z \right]_{z=1} = np
\]

\[
G''(1) = np(p(z) + (npz)^{n-1})_{z=1} = np
\]

\[
\text{Var}(X) = np - np^2 = np(1 - p)
\]

Note that a comparison of Eqs. (56) and (57) offers an example of the general rule (50).
Examples

i) Suppose that the outcome $A$ corresponds to getting an event in a histogram bin $j$, and $\bar{A}$ corresponds to getting an event in any other histogram bin. The probability $p$ of getting a success $A$ is actually the integral of the p.d.f. over the bin (sometimes it is well approximated by the product of the bin width and the value of the p.d.f. at the middle of the bin). The probability of getting $r$ events in the bin $j$ and $n-r$ events outside $j$ is given by Eq. (53). The expected number of events in $j$ is $np$ from Eq. (56) and the variance is $np(1-p)$ from Eq. (58).

ii) Consider a study of forward-backward asymmetry. Suppose that we end the experiment when $n$ events have been collected ($n$ is not a random variable). Let $F$ and $B$ denote the number of events in the forward and backward hemisphere, respectively:

$$n = F + B$$

Then $F$ behaves according to the binomial law

$$\binom{n}{r} p^r (1-p)^{n-r}$$

with mean $np$ and variance $np(1-p)$.

For large $n$ the variance is well approximated by $F(1-p)$ and the standard deviation by $\sqrt{F(1-p)}$ (not by $\sqrt{F}$).

Usually the asymmetry is defined as

$$r = \frac{F - B}{F + B} = \frac{2F}{n} - 1$$

It follows that $p_r$ is given by the law

$$p_r = \left( \frac{n}{\frac{1}{2} n (r+1)} \right) p^r (1-p)^{n-r}$$

with the variance $4p(1-p)/n$. For large $n$ the variance is well approximated by

$$\frac{4F B}{n^3} = \frac{4F B}{(F + B)^3}$$

and thus

$$\sqrt{\frac{2}{(F + B)}} \sqrt{\frac{F B}{(F + B)^3}}$$
iii) A desired event has a probability \( p \) of occurring in an experimental trial. We want to know how many trials must be carried out in order to have a probability \( \alpha \) of at least one event occurring. Now \( X \) is the number of events in \( n \) trials, but \( n \) is unknown. We want to find \( n \) such that

\[
\Pr(X \geq 1) \geq \alpha.
\]

\[
1 - \Pr(X = 0) \geq \alpha.
\]

\[
\Pr(X = 0) \leq 1 - \alpha
\]

\[
(1 - p)^n \leq 1 - \alpha.
\]

Since \( (1 - p) < 1 \), \( k \log(1 - p) > \log(1 - \alpha) \).

\[
k \geq \frac{\log(1 - \alpha)}{\log(1 - p)}.
\]

2.2.2 Multinomial distribution

The generalization of the binomial distribution to the case of more than two possible outcomes of an experiment is called the multinomial distribution. An example is a histogram containing \( n \) events (a fixed number) distributed in \( k \) bins. There are now \( k \) random variables: the number \( X_j \) of events in bin \( j \), with \( j = 1, k \). The numbers \( X_j \) form a \((k - 1)\)-dimensional vector since for all results

\[
\sum_{j=1}^{k} X_j = n
\]

Let the probability of observing one event in bin \( j \) be \( p_j \), and assume that all \( p_j \), \( j = 1, k \), are independent.

The probability of obtaining a histogram with \( r_1 \) events in bin 1, \( r_2 \) events in bin 2, etc., is then

\[
\Pr(X = r) = \frac{n!}{r_1! \cdots r_k!} \prod_{i=1}^{k} p_i^{r_i}
\]

\[
= n! \frac{1}{r_1! \cdots r_k!} \prod_{i=1}^{k} p_i^{r_i}.
\]

Let us derive a few results from the density (60).

The expected number of events in one bin, \( j \), is

\[
E(X_j) = n p_j
\]
with variance
\[ \sqrt{(x_i')} = \eta p_1 (1 - p_1) \] (62)
as derived in the one-dimensional case in the previous section.

Consider two bins \((i, j)\) together. The p.d.f. is then
\[ p(x_i = r_i, x_j = r_j) = \frac{n!}{c_i! c_j! (n - c_i - c_j)!} p_i^{c_i} p_j^{c_j} (1 - p_i)(1 - p_j) \] (63)

The probability generating function is now multidimensional, and has the form
\[ G(z_1, z_2, \ldots, z_k) = (p_1 + p_2 z_1 + \ldots + p_k z_k)^n \]
\[ p(x_i = r_i, x_j = r_j) \]

The expectation of the product \(x_i x_j\) is then the coefficient of \(z_i^{r_i} z_j^{r_j}\)
\[ E(x_i; x_j) = \sum_{c_i, c_j, c_3 = n} \frac{n!}{c_i! c_j! c_3!} p_i^{c_i} p_j^{c_j} p_k^{c_3} \]
\[ = \frac{n (n-1)}{c_i! c_j! c_3!} \frac{n!}{c_i! c_j! (n - c_i - c_j)!} p_i^{c_i - 1} p_j^{c_j - 1} p_k^{c_3} \]
\[ = n (n-1) p_i p_j \] (64)

From Eqs. (61-64) and (27) we find
\[ \text{cov}(x_i, x_j) = n (n-1) p_i p_j - \eta p_i \eta p_j \]
\[ = -\eta p_i \eta p_j \] (65)

Thus the number of events in any two bins \(i\) and \(j\) of a histogram are negatively correlated, with correlation coefficient
\[ \text{corr}(x_i, x_j) = \frac{-\eta p_i \eta p_j}{\sqrt{(1 - p_i)(1 - p_j)}} \] (66)
Note that if $p_i << 1$ (many bins) Eq. (62) can be approximated by

$$V(X_j) \sim n p_j \sim X_j$$

and the standard deviation of the number of events in a bin becomes

$$\sigma_j \sim \sqrt{X_j} \quad (67)$$

2.2.3 Poisson

This is a limiting case of the binomial distribution for $p \to 0$, as $k \to \infty$, with $kp$ finite and fixed. Then

$$p_r = P(X = r) = \frac{\mu^r e^{-\mu}}{r!} \quad (68)$$

where

$$\mu = kp$$

The probability-generating function is

$$G(z) = \sum_{r=0}^{\infty} \frac{\mu^r e^{-\mu} z^r}{r!} = e^{-\mu} (1 - \mu(z-1))$$

$$G'(z) = \mu e^{\mu(z-1)} \quad , \quad G'(1) = \mu$$

$$G''(z) = \mu^2 e^{\mu(z-1)} \quad , \quad G''(1) = \mu^2$$

$$\mu(X) = \mu$$

$$\nu(X) = \mu$$

Mean = Variance for the Poisson distribution.

**Examples:**

i) Suppose particles are emitted from a radioactive substance at an average rate of $\gamma$ particles/unit time, in such a way that probability of emission in $\delta t$ is $\gamma \delta t$, and probability of more than one emission in $\delta t$ is $o(\delta t)$. Then the distribution of the number of particles emitted $X$, in a fixed time interval $t$, is Poisson, with mean $\gamma t$:

$$P(X = r) = \frac{(\gamma t)^r e^{-\gamma t}}{r!}$$

ii) The distribution of bubbles along a track of constant momentum in a bubble chamber also is of the Poisson form.

iii) Consider the ratio of particles emitted in the forward (F) and backward (B) direction. Suppose that the numbers F and B are independent random variables distributed according to the Poisson law.
F has mean $\mu_F$ and variance $\mu_F'$, which can be approximated by $F$. The asymmetry is defined by

$$r = \frac{F - B}{F + B}$$

Assuming that the error on $F$ and $B$ is small (large number of events), we may use a differential formula to obtain an approximation to the variance of $r$:

$$\frac{dr}{r} = \frac{d(F - B)}{(F - B)} - \frac{d(F + B)}{(F + B)} + 2 \frac{dF - FdB}{(F - B)(F + B)}.$$  

$$E(\frac{dr^2}{r^2}) = 4 \frac{B^2 E(F^2) + F^2 E(B^2)}{(F - B)^2 (F + B)^2} \approx \frac{ABF}{(F - B)^2 (F + B)^2}.$$

$$V(r) \approx \frac{ABF}{(F + B)^3}.$$

This is the same result as in example (ii) of the binomial distribution (2.2.1). Note, however, the following differences:

a) The last formula is an approximation (we use the differentiation).

b) In the binomial case the number of events was fixed a priori.

If $n$ is not fixed, then the correct distribution is

- a Poisson law for the total number of events $n$

$$\frac{e^{-\mu} \mu^n}{n!}$$

- a binomial law for $F$ and $B$ conditional on $n$.

Therefore

$$f(n, F, B) = e^{-\mu} \frac{\mu^n}{n!} \binom{n}{F}^F (1 - \mu)^B \approx e^{-\mu} \frac{(\mu F)^F (\mu (1 - \mu))^B}{F! B!}$$

which is the product of the two Poisson laws for $F$ and $B$.

c) The differential method, being an approximation used in the Poisson calculation, did indeed lead to a result different from the rigorous binomial calculation, but the difference vanished when we replaced $\mu_F$ and $\mu_B$ by $F$ and $B$, respectively.
2.2.4 Geometric and negative binomial

Consider again the case when the probability of success is \( p \) and the probability of failure \( q = 1 - p \). The probability of waiting \( r \) trials until the first success occurs is given by the geometric distribution

\[
P(X = r) = pq^{r-1}
\]

with expectation and variance

\[
E(X) = \frac{1}{p}, \quad V(X) = \frac{(1-p)}{p^2}
\]

In particular

\[
P(X = 1) = p, \quad P(X = 2) = pq
\]

Let \( Y \) denote the number of trials up to and including the \( m \)th success. Obviously the probability that all the \( Y \) trials are successes is

\[
P(Y = m) = p^m
\]

The probability of having at least one failure in \( Y = m + 1 \) trials \((m \) successes, the \( m \)th trial being a success\) is then

\[
P(Y = m+1) = \binom{m}{m-1} p^{m-1} q p
\]

In general the probability of waiting \( r \) trials for \( m \) successes is given by the negative binomial distribution

\[
P(Y = r, r \geq m) = \binom{r-1}{m-1} p^m (1-p)^{r-m}
\]

(73)

with parameter \( p \) and index \( m \). This density has the expectation and variance

\[
E(Y) = \frac{m}{p}, \quad V(Y) = \frac{(1-p)m}{p^2}
\]

and the generating function

\[
G(z) = \left(\frac{p z}{1 - q z}\right)^m
\]

(74)
2.3 Continuous distributions

2.3.1 Uniform

The uniform or rectangular density is given by

\[ f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases} \]

and its cumulative distribution is

\[ F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } b < x \end{cases} \]

Its expectation and variance is

\[ E(X) = \frac{a+b}{2} \]
\[ \text{Var}(X) = \frac{(b-a)^2}{12} \]

The rounding-off errors in arithmetical calculations are uniformly distributed.

2.3.2 Normal one-dimensional

The most important theoretical distribution in statistics is the normal probability density function, or Gaussian, usually abbreviated \( N(\mu, \sigma^2) \). It depends on two parameters and has the form

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

(77)

\[ \text{Point of inflexion} \]

Fig. II.7
The cumulative distribution is

\[ F(x) = \int_{-\infty}^{x} f(x') \, dx' \]

\[ = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \, dx' \]

\[ = \Phi \left( \frac{x - \mu}{\sigma} \right) \]

(78)

where \( \Phi(x) \) is called the standard normal distribution function.

The function \( N(0,1) \) is called the standard normal density.

The expectation \( E(X) = \mu \) and variance \( V(X) = \sigma^2 \) can be obtained by simple integration.

Note that the standard deviation \( \sigma \) is not the width of the p.d.f. at half the height. The width at half-height is 1.176 \( \sigma \). The probability content of various intervals is shown below:

\[ P\left( -1.64 \leq \frac{X - \mu}{\sigma} \leq 1.64 \right) = 0.95 \]

\[ P\left( -1.96 \leq \frac{X - \mu}{\sigma} \leq 1.96 \right) = 0.95\% \]

\[ P\left( -2.58 \leq \frac{X - \mu}{\sigma} \leq 2.58 \right) = 0.99 \%

\[ P\left( -3.29 \leq \frac{X - \mu}{\sigma} \leq 3.29 \right) = 0.999\%

The characteristic function of the distribution is

\[ \phi(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}} \]

(79)

\[ \Phi \left( \frac{X - \mu}{\sigma} \right) = \frac{(2\pi)^{\frac{1}{2}}}{r^{\frac{1}{2}}} \]

\[ \mu^{n+1} = \left( \frac{2\pi}{\sigma} \right)^{\frac{1}{2}} \]

A linear combination of independent normal variables is also normal.

Suppose \( X, Y \) are independent normal variables, with means \( \mu_X, \mu_Y \) and variances \( \sigma_X^2 \) and \( \sigma_Y^2 \), respectively. Then \( aX \) and \( bY \) are independent
normal variables, with characteristic functions
\[
\phi_x(t) = \exp \left[ it \mu_x - \frac{1}{2} t^2 \sigma_x^2 \right]
\]
\[
\phi_y(t) = \exp \left[ it \mu_y - \frac{1}{2} t^2 \sigma_y^2 \right]
\]
To find the characteristic function of \( Z = aX + bY \), we have
\[
\phi_Z(t) = \phi_x(t) \phi_y(t) = \exp \left[ it(a \mu_x + b \mu_y) - \frac{1}{2} t^2 (a^2 \sigma_x^2 + b^2 \sigma_y^2) \right]
\]
Thus \( Z \) is also normal, with mean \( a \mu_x + b \mu_y \) and variance \( (a^2 \sigma_x^2 + b^2 \sigma_y^2) \).

It follows that, if
\[
Z = \sum_{i=1}^{n} X_i
\]
where \( X_i \) are independent normal, then \( Z \) is also normal.

Suppose that the random variables \( X_1, \ldots, X_n \) are independent normal \( N(\mu, \sigma^2) \), with identical mean \( \mu \) and variance \( \sigma^2 \). Then the average
\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]
and the estimate of the variance
\[
S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
are also independently distributed.

This property is unique to the normal distribution. Given that the \( X_i \) are independent random variables, then \( \bar{X} \), and \( S^2 \) are independent if, and only if, the \( X_i \) have a normal distribution.

As an example of the above, suppose we have a peaked mass distribution and would like to find the centre and the width of this peak. Suppose further that the estimates we use are the average, Eq. (80), and estimate of variance, Eq. (81). Now considering each of the two quantities above as random variables, we may also estimate the covariance matrix for these as in Section 1.4.3 above. Then the correlations (off-diagonal elements) will be zero if and only if the original mass peak was Gaussian. Any other distribution (say a Breit-Wigner) will yield
estimates for the central mass and width which are correlated. Since
the error on the central mass depends on the width, the situation
becomes complicated when the two estimates are correlated. (Correlations
of this type are discussed further in Part IV.) In fact, the average
$\bar{X}$ being a linear combination of normal variables, is normal, $N(\mu, \sigma^2/n)$,
with mean $\mu$ and variance $\sigma^2/n$. The quantity $(n-1) s^2/\sigma^2$ can be shown
to be distributed as a $\chi^2$ variable with $n-1$ degrees of freedom
(see further Sec. 2.3.5). In general if a quantity $U$ is such that
$$\frac{U}{\sigma^2}$$
is distributed as a $\chi^2$ with $\nu$ degrees of freedom, we call $U$ a normal
theory estimate of $\sigma^2$ with $\nu$ degrees of freedom. Let us define the
coefficient of kurtosis as
$$y_2 = \frac{M_4}{(M'_2)^2} - 3 = \frac{4^{th \text{ Central moment}}}{\text{variance squared}} - 3$$
(82)
If $y_2 < 0$, then $(\bar{X} - \mu)^2$ positively/negatively correlated with $s^2$.
[$y_2$ measures the "peakiness" ($y_2 > 0$) or "flatness" ($y_2 \leq 0$) relative
to the normal.] If the measurements $X_i$ are not normal, the distribution of $s^2$ is roughly proportioned to $\chi^2$ with density function
$$p.d.f. = \frac{\text{normal theory} p.d.f.}{(1 + \frac{1}{2} y_2)}$$
(83)
Provided that $\sigma^2$ is finite, the non-normal distribution of $\bar{X}$ becomes
asymptotically normal, $N(\mu, \sigma^2/n)$. (This is a consequence of the central
limit theorem, which we shall discuss in Section 3 of this chapter.)

Another interesting fact about the normal distribution is that the
probability density of a sample of independent observations from a
standardized normal distribution is constant on the sphere
$$\sum X_i^2 = \text{constant}$$
since the density is a function of $\Sigma X^2$ only. This property of radial
symmetry is not possessed by any other distribution. Bartlett (1934)
used the following proof.
If the density can be written

$$L = \prod_{i=1}^{n} f(x_i) = f\left(\frac{x^2}{\sigma_i^2}\right)$$

we must have

$$\frac{\partial L}{\partial x_i} = \frac{\partial \log L}{\partial x_i} = 0$$

for $\Sigma x^2 = $ constant. Using Lagrange multiplier $\lambda$, we therefore have

$$\frac{\partial \log f(x_i)}{\partial x_i} + \lambda x_i = \lambda$$

for all $i$.

$$\log f(x_i) = -\frac{1}{2} \lambda x_i^2 + k_i$$

$$f(x_i) \propto e^{-\frac{1}{2} \lambda x_i^2}$$

A final interesting characteristic is that if $X$ is a vector of independent standardised variables, not necessarily identically distributed, and if a non-trivial transformation $X = CY$ gives a vector $Y$ of independent standardised variables, then each $X_i$ is normal, the transformation is orthogonal, and each $Y_i$ is normal.

The question naturally arises as to what extent a series of measurements can be considered a priori to be normally distributed. It is often assumed, for example, that a person making repeated measurements of the distance between two fixed points will obtain a set of measurements that is normally distributed with mean equal to the "true" distance and width given by the precision of the method used. Mathematicians usually assume this to be an experimental fact, and experimentalists often assume it to be a mathematical theorem. In fact it is neither, but there are suggestions from both sides (mathematical and experimental) that it is at least a good approximation to the actual distribution. Therefore, normality is assumed because it is simple to use (many mathematically simple properties can be demonstrated only for normal distributions) and because it is, to some extent, empirically supported. Two important cases are cited here, one of which is certainly normal and the other certainly not normal.

i) The sum or the mean value of many independent observations. In this case the central limit theorem can be invoked (see Section 3.3.2 in this portion) and we obtain the result that the sum or mean of several
measurements will be (asymptotically) normally distributed, even if the original measurements are not normally distributed (they need not even be distributed similarly). For example, suppose that we are studying the $\pi^-$ meson emerging from a target placed in a high-energy beam, and we wish to measure the mean value of the $\pi^-$ momentum. We know that in general a $\pi^-$ coming from a given reaction will not have a Gaussian momentum spectrum, and furthermore the total momentum spectrum of $\pi^-$ from all reactions also will not be a normal distribution. However, if we take as our random variable the mean value of momentum for $n$ consecutive $\pi^-$, then repeated observations of this kind will yield a distribution of mean $\pi^-$ momentum that becomes normal as $n$ becomes large.

ii) A set of normally-distributed measurements, each having the same mean but different variances. In this case, even though each individual measurement is a normally distributed random variable, the over-all distribution is the sum of normal distributions of different widths, which is never normal. An example is the measurement in a bubble chamber experiment of the mass of a narrow resonance or stable particle. If the mass determination from one event is assumed to be normally distributed with standard error $\Delta M$, then the over-all experimental mass spectrum obtained will be normal only if events are selected so that they all have the same $\Delta M$. In the more usual case where there is a spread of $\Delta M$ values (often a large spread) the over-all distribution may be quite far from normal. This important case is described in more detail in Section 2.4.3 below.

To summarize the two examples above, we can say that under certain general conditions the sum of independent random variables is normal, but the sum of normal distributions (of different variances) is not normal.

2.3.3 Normal many-dimensional

In generalising the normal distribution to many dimensions, it is natural to look for a density function which has, as its exponent, a quadratic form in the component variables:

$$
\rho(X) \propto e^{\sum_{j,k} \frac{1}{2} a_{jk} (X_j - \mu_j)^2} \prod_{i=1}^n \left( \prod_{j=1}^p \frac{1}{\sqrt{2\pi} \sigma_j} e^{-\frac{(X_j - \mu_j)^2}{2\sigma_j^2}} \right)
$$
As in the one-dimensional case, $\mu$ and $\sigma$ are location and scale parameters which may be removed by a standardising transformation. Let us use matrix notation, and write this density function as

$$f(x) \propto e^{x^T \rho \left[-\frac{1}{2} (x - \mu)^T A (x - \mu)\right]}$$

The condition that this function $f(x)$ is a density function imposes the restraint that $A$ is positive semi-definite, in order that $\int f(x) dx$ converges. The constant of proportionality is obtained from the condition

$$\int f(x) dx = 1$$

from which one obtains

$$f(x) = \left(\frac{\pi}{\lambda}\right)^{-\frac{n}{2}} |A|^{\frac{1}{2}} e^{x^T \rho \left[-\frac{1}{2} (x - \mu)^T A (x - \mu)\right]}$$

The characteristic function is

$$\phi(t) = e^{x^T \rho \left(i t - \frac{1}{2} t^T A^{-1} t\right)}$$

The mean of the $j^{th}$ variable $X_j$ is

$$E(X_j) = \left[\delta \phi \frac{\partial}{\partial t_j}\right]_{t=0} = \mu_j$$

and its variance is

$$V(X_j) = E[(X_j - \mu_j)^2]$$

$$= \left[\delta^2 \phi \frac{\partial}{\partial t_j^2}\right]_{t=0}$$

$$= \alpha_{jj}$$

the $j^{th}$ diagonal element of $A$. The covariance of the $j^{th}$ and $k^{th}$ components is

$$E[(X_j - \mu_j)(X_k - \mu_k)] = \left[\delta^2 \phi \frac{\partial}{\partial t_j \partial t_k}\right]_{t=0}$$

$$= \alpha_{jk}.$$
the j-k:th element of $A^{-1}$. For convenience, we may define the covariance matrix of $X$ as

$$V = E\left[(X - \mu)(X - \mu)^T\right] = A^{-1}$$

where

$$\rho_{j,k} = \frac{E[(X_j - \mu_j)(X_k - \mu_k)]}{\sqrt{\sigma_{X_j} \sigma_{X_k}}}$$

is defined as the correlation between components $X_j$, $X_k$.

The many-dimensional normal density is then

$$f(X) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} e^{-\frac{1}{2} (X - \mu)^T V^{-1} (X - \mu)}$$

(84)

For $n=2$, $X_1=X$ and $X_2=Y$ this density becomes

$$f(X, Y) = \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} e^{-\frac{1}{2} \left( \frac{(X - \mu_X)^2}{\sigma_X^2} + \frac{(Y - \mu_Y)^2}{\sigma_Y^2} - 2\rho \frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y} \right)}$$

(85)

Here we have explicitly parametrized the p.d.f. in terms of its expectations $\mu_X$, $\mu_Y$, standard deviations $\sigma_X$, $\sigma_Y$ and correlation $\rho$.

The marginal distribution of $X$ is also normal, $N(\mu_X, \sigma_X^2)$, with mean $\mu_X$ and variance $\sigma_X^2$. The conditional density

$$f(X|Y) = \frac{1}{\sqrt{2\pi \sigma_X(1-\rho^2)}} e^{-\frac{1}{2} \left( \frac{(X - \mu_X - \rho \sigma_X Y)}{\sigma_X(1-\rho^2)} \right)^2}$$

(86)

is normal, with mean $\mu_X + \rho (\sigma_X/\sigma_Y) (Y - \mu_Y)$ and variance $\sigma_X^2 (1-\rho^2)$.

If $X$ is normal, $V$ is non-singular and $\mu = 0$, the quantity $X^T V^{-1} X$ is called the covariance form of $X$. This has a $\chi^2$-distribution with $n$ degrees of freedom.
The many-dimensional (or multi-variate) distribution is interesting for many reasons:

i) it is a function only of the means, variances, and two-variable correlations;

ii) the p.d.f. has a many-dimensional "bell-shaped" form;

iii) contours of constant probability density are given by the quadratic form;

iv) any section through the distribution, say at $X_n = \text{constant}$, gives again a normal distribution in $n - 1$ dimensions of the form (84), with covariance matrix $V_{n-1}$ obtained by removing $n^{th}$ row and column of $V$ and inverting the resultant submatrix.

v) any projection onto a lower space gives a marginal distribution which is again normal, of the form (84), with covariance matrix obtained by deleting appropriate rows and columns of $V$.

vi) a set of variables, each of which is a linear function of a set of normal variables, has itself a many-dimensional normal distribution.

vii) The average vector $\bar{X}$, and covariance matrix, $S$, of a set of independent observations, $X_i$, $i=1, \ldots, n$, are independently distributed if, and only if, the parent distribution of the $X_i$ is many-dimensional normal.

2.3.4 Exponential

Consider events occurring randomly in time, with average $\lambda$ events per unit time. Then the probability $P_n(t)$ of $n$ events occurring in a time interval $t$ has the Poisson distribution:

$$P_n(t) = \frac{1}{n!} \cdot e^{-\lambda t} \cdot (\lambda t)^n$$

The probability of no event in time $t$ is then $\exp(-\lambda t)$. Consider the time interval $X$ between two successive events. In this interval, no events occur. Therefore, for fixed $t$:

$$P(X > t) = e^{-\lambda t}$$

We define the exponential (cumulative) distribution as

$$F(X) = P(X \leq t) = 1 - e^{-\lambda t}, \lambda > 0, t > 0$$

(87)
with probability density function

$$f(t) = \lambda e^{-\lambda t}$$  \hspace{1cm} (88)

expectation

$$E(t) = \frac{1}{\lambda}$$

and variance

$$V(t) = \frac{1}{\lambda^2}$$

Thus the standard deviation equals the mean: \(1/\lambda\).

The exponential distribution has "no memory": if no event has occurred up to time \(y\), the probability of no event in a subsequent time \(x\) is independent of \(y\). For fixed \(y\)

$$P(X > x + y \mid X > y) = e^{-\lambda(x+y)}$$

and

$$e^{-\lambda x} = P(X > x)$$

2.3.5 \(\Gamma\) and \(\chi^2\)

The family of distributions called \(\Gamma\), and defined by

$$f(x) = \frac{a (a x)^{b-1} e^{-ax}}{\Gamma(b)}, \ x > 0, \ a > 0, \ b > 0$$  \hspace{1cm} (89)

is sometimes useful, being capable of representing a variety of slopes (cf. Fig. II. 8).

Fig. II.8
Here a is a scale parameter.

Note that \( b = 1 \) gives the exponential distribution. This family includes the \( \chi^2 \) distribution as a sub-set, namely Eq. (39), with \( b = n/2 \), is \( \chi^2 \) with \( n \) degrees of freedom.

**Example:** The distribution of the time interval between every second event in a series of events occurring at random is given by the gamma distribution, with \( b = 2 \). In general, the sum of \( k \) intervals is a gamma distribution with \( b = k \) (a \( \chi^2 \) with \( n = 2k \)).

Suppose that \( X_1, \ldots, X_n \) are independent, standard normal variables, \( N(0,1) \). Then the sum of squares

\[
S_n = \sum_{i=1}^{n} X_i^2
\]

is said to have a chi-squared distribution \( \chi^2(n) \), with \( n \) degrees of freedom.

The probability density function for \( S_\cdot \) is

\[
P(S) = \frac{1}{\Gamma(n/2)} \left( \frac{S}{2} \right)^{n/2 - 1} e^{-S/2}
\]

and the characteristic function is

\[
\phi(t) = \left( 1 - 2it \right)^{-n/2}
\]

The expectation and variance are

\[
\mu = n, \quad \sigma^2 = 2n
\]

![Fig. II.9](image)
If \( S_n, S_m \) have independent \( \chi^2 \) distributions with \( n \) and \( m \) degrees of freedom, respectively, then the sum

\[
S_k = S_n + S_m
\]

has a \( \chi^2 \) distribution with \( k = n + m \) degrees of freedom. This is obvious from the definition (90).

The function

\[
R_n = \sqrt{S_n}
\]

has the \( \chi(n) \) distribution. It has the density function

\[
\mathcal{P}(R) = \frac{R^{\frac{n}{2}-1} e^{-\frac{R}{2}}}{\Gamma\left(\frac{n}{2}\right)}
\]

Asymptotically, the \( \chi^2(n) \) and \( \chi(n) \) distributions are both normal.

More specifically it can be shown that for large \( n \) the quantities

\[
Z_n = \frac{\chi(n) - n}{\sqrt{2n}}
\]

and

\[
Z'_n = \sqrt{2} \chi(n) - \sqrt{2n - 1}
\]

are standard normal, \( N(0,1) \).

When the independent variables \( \chi_i \) are normal, \( N(\mu_i, 1) \) (not standard normal, as in the previous case), then \( S_n \) of Eq. (90) has a non-central \( \chi^2 \) distribution with \( n \) degrees of freedom.

The characteristic function of the distribution of \( S_n \) is

\[
E(\mathcal{C}^{i'tS_n}) = \mathcal{C}^{\frac{2}{1 - \frac{t^2}{2}}} \left[ 1 - \frac{t^2 \frac{n}{2}}{2} \right]^{\frac{n}{2}}
\]

where

\[
\Delta = \sum_{j=1}^{n} \mu^2
\]
is called the non-centrality parameter. Note that $\mu_1, \ldots, \mu_n$ enter the distribution only through $\Delta$. If $\Delta = 0$, (i.e., $\mu_i = 0$, $i = 1, \ldots, n$), we recover the central $\chi^2$ distribution.

We can approximate the non-central $\chi^2(n, \Delta)$ distribution by a central $\chi^2(m)$ distribution having the same mean and variance:

$$\chi^2(n, \Delta) \sim \beta \chi^2(m) \quad (98)$$

where

$$\beta = 1 + \frac{\Delta}{n + \Delta}$$

$$m = n + \frac{\Delta^2}{n + 2\Delta} = \frac{(n + \Delta)^2}{n + 2\Delta}$$

2.3.6 Student's $t$

In Section 2.3.2 we have seen that if the measurements $X_1, \ldots, X_n$ are normal, $N(\mu, \sigma^2)$, then $\sqrt{n} (\bar{X} - \mu)$ is also normal, $N(0, \sigma^2)$, and $S^2$ is a normal theory estimate of $\sigma^2$ with $(n - 1)$ degrees of freedom. The quantities $\bar{X}$ (the average) and $S$ (the estimated standard deviation) are defined by Eqs. (80) and (81).

It follows from the normality of the $X_i$ that $\bar{X}$ and $S$ are independent, and vice versa. $\bar{X}$ and $S^2$ are estimates of the parameters based on a finite number of measurements and, consequently, as a measure of the spread of the distribution, $S$ is too small, being measured relative to $\bar{X}$. The correct distribution to use, taking account of the random variation in both $\bar{X}$ and $S$, is that of a ratio between them. A suitable ratio is

$$t = \frac{\sqrt{n} (\bar{X} - \mu)}{S} \quad (99)$$

which has a distribution called Student's $t$ distribution with $n$ degrees of freedom.

In general, $t = X/S$ has a Student's $t$ distribution if $X$ is $N(0, \sigma^2)$, $S$ is a normal theory estimate of $\sigma$ with $n$ degrees of freedom $\sqrt{nS^2/\sigma^2}$ is
distributed as $\chi^2(n)$, and $S$ and $X$ are independent.

The p.d.f. of $t$ is

$$p(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{(1 + \frac{L^2}{n})^{\frac{n+1}{2}}}$$  \hspace{1cm} (100)$$

with expectation

$$\mu = 0,$$

variance

$$\sigma^2(t) = \frac{n}{n-2}, \quad (n > 2),$$

and 4th moment

$$\mu_4 = \frac{3n^2}{(n-4)(n-2)}, \quad (n > 4).$$

These results can be obtained directly or from

$$E(t) = E(X) \cdot E\left(\frac{1}{S}\right) = 0$$

$$E(t^4) = E(X^4) E\left(\frac{1}{S^4}\right)$$

since $X$ and $S$ are independent.

The Student's distribution is symmetrical around $t = 0$. For $n = 1$ it specializes to the Cauchy distribution and for $n \to \infty$ it approaches the standard normal distribution $N(0, 1).

From the p.d.f. one can construct confidence limits for $\mu$, or tests of the hypothesis that $\mu = \mu_0$ (the "null hypothesis", $H_0$).

If $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are distributed as $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively, all observations being independent, then

$$\left( \overline{X} - \overline{Y} \right) - \left( \mu_X - \mu_Y \right)$$

has the distribution

$$N\left( c, \sigma^2 \left( \frac{1}{m} + \frac{1}{n} \right) \right).$$
The pooled estimate of \( \sigma^2 \),

\[
S^2 = \frac{\sum_{i=1}^{m} (X_i - \bar{X})^2 + \sum_{j=1}^{n} (Y_j - \bar{Y})^2}{m+n-2}
\]  

(101)

is a normal theory estimate of \( \sigma^2 \) with \( m + n - 2 \) degrees of freedom.

Now \( S^2 \) and \( \bar{X} - \bar{Y} \) are independent [since neither sum in Eq. (102) depends on both \( \bar{X} \) and \( \bar{Y} \)].

Therefore

\[
t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sum_{i=1}^{m} (X_i - \bar{X})^2 / m + \sum_{j=1}^{n} (Y_j - \bar{Y})^2 / n}}
\]  

(102)

has a \( t_{m+n-2} \) distribution. This provides a test of the difference in means of the two sets of data, \( X \) and \( Y \).

If \( t' = X/Y \), where \( X \) is \( N(\Delta \sigma, \sigma^2) \), and \( \sqrt{Y}/\sigma^2 \) is \( \chi^2(\nu) \), (i.e., \( Y \) is a n.t. estimate of \( \sigma^2 \)), then \( t' \) has a non-central t distribution with \( \nu \) degrees of freedom and non-centrality parameter \( \Delta \). \( X, Y \) are assumed to be independent.

Approximately,

\[
\Phi(t' = \frac{X}{Y} < k) = \Phi\left( X - kY < 0 \right)
\]

Now \( X \) is normal, and \( Y \) is fairly normal.

Also,

\[
E \left( X - kY \right) = \sigma \left( \Delta - k \right)
\]

and

\[
\nu \left( X - kY \right) \approx \sigma^2 \left( 1 + \frac{k^2}{\nu} \right)
\]

Hence

\[
\Phi \left( \frac{X}{Y} < k \right) \approx \Phi \left( \frac{k - \Delta}{\sqrt{1 + k^2 / \nu}} \right)
\]

(103)

where \( \Phi \) denotes the normal probability integral.

One application of the non-central t distribution is in determining the power of the normal t test. A second application is concerned with inferences about tolerances.

We wish to make inferences about

\[
\mathcal{Z} = \frac{L - k}{\sigma}
\]

for instance, confidence intervals on the proportion of a population lying beyond a given point \( L \).
2.3.7 F

Let \( S_1^2, S_2^2 \) be independent and normal theory estimates of \( \sigma_1^2 \) and \( \sigma_2^2 \), with \( \nu_1 \) and \( \nu_2 \) degrees of freedom, respectively, i.e.

\[
\frac{S_1^2}{\sigma_1^2} \text{ distributed as } \chi^2(\nu_1),
\]

and

\[
\frac{S_2^2}{\sigma_2^2} \text{ distributed as } \chi^2(\nu_2).
\]

Then

\[
F = \frac{S_1^2}{S_2^2} \frac{\sigma_2^2}{\sigma_1^2}
\]

is said to have the F distribution with \((\nu_1, \nu_2)\) degrees of freedom (or variance-ratio distribution). The probability density function is

\[
\rho(F) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\sqrt{\nu_1 \nu_2} \Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right)} \frac{F^{\nu_1/2 - 1}}{(\nu_1 + \nu_2 F)^{\nu_2/2}}
\]

This distribution is positively skew, and tends to normality as \( \nu_1, \nu_2 \to \infty \), but only slowly \((\nu_1, \nu_2 > 50)\). The expectation is

\[
E(F) = E\left(\frac{S_1^2}{\sigma_1^2}\right), E\left(\frac{S_2^2}{\sigma_2^2}\right)
\]

\[
= \frac{\nu_2}{\nu_2 - 2}, \quad (\nu_2 > 2)
\]

and the variance

\[
V(F) = \frac{2 \nu_2^2 (\nu_1^2 + \nu_2^2 - 2 \nu_2)}{\nu_1 (\nu_1 - 2)^2 (\nu_2 - 4)}, \quad (\nu_2 > 4)
\]

The quantity \( Z = \frac{1}{2} \ln F \) has a distribution much closer to normal, and with mean and variance approximately give by

\[
E(Z) = -\frac{1}{2} \left(\frac{1}{\nu_1} - \frac{1}{\nu_2}\right)
\]

\[
V(Z) = \frac{1}{2} \left(\frac{1}{\nu_1^2} + \frac{1}{\nu_2^2}\right)
\]
With $v_1 = 1$, the $F$ distribution specializes to $F = t^2_{v_2}$. As $v_2 \to \infty$, $F$ approaches 
\[
\frac{1}{F^{v_1}} \chi^2(v_1)
\]

A function equivalent to $F$ for tests is the function
\[
\mathcal{U} = \frac{\sqrt{v_1} \sqrt{F}}{v_2 + v_1 F}
\]
which is a monotonic function of $F$.

### 2.3.8 Cauchy or Breit-Wigner

The Cauchy distribution
\[
f(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \quad (-\infty < x < \infty)
\]
represents a pathological case since, as we have noted in Section 1.4.1, the expectation is undefined. In fact, all moments diverge. The Cauchy distribution is identical to the physically important Breit-Wigner distribution.

The characteristic function of the Cauchy distribution is
\[
\phi(t) = e^{-|t|}
\]
which has no Taylor expansion around the origin. Thus the moments are not defined.

We will see in Section 2.4.1 how the Cauchy distribution becomes manageable by truncation.