PROBABILITY AND STATISTICS
APPLIED TO HIGH-ENERGY PHYSICS

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

Lectures given in the
Academic Training Programme of CERN

1968 - 1969
II ELEMENTS OF PROBABILITY (continued)

2. Special distributions (continued)
   2.4 Bayesian inference from normal distributions
      2.4.1 Inference about the mean
      2.4.2 Inference about the variance
      2.4.3 Inference about the mean and the variance
   2.5 Real life
      2.5.1 Truncation
      2.5.2 Detection efficiency
      2.5.3 Experimental resolution

3. Convergence and the law of large numbers
   3.1 Tchebycheff theorem
      3.1.1 Bienaymé - Tchebycheff Inequality
   3.2 Convergence
      3.2.1 Convergence in distribution
      3.2.2 Paul Levy's theorem
      3.2.3 Convergence in probability
      3.2.4 Stronger types of convergence
   3.3 Law of large numbers
      3.3.1 Monte Carlo integration
      3.3.2 Central Limit theorem
      3.3.3 Example: Gaussian Random Number Generator
   3.4 Asymptotic properties of special distribution
2.4 Bayesian interference from normal distributions

In this Section we deal with the Bayesian approach to the distributions of mean and variance of data with a normal distribution. Where previously we could discuss the distributions of $\bar{X}$ and $S^2$ without reference to how to use these, when we take the Bayesian approach we are forced much closer to the problem of inference, and of actually saying something about the unknown parameters involved.

We consider here a set of observations $X_1, \ldots, X_n$ of normal distribution, $\mathcal{N}(\mu, \sigma^2)$.

2.4.1 Inference about the mean

Here $\mu$ is unknown, $\sigma^2$ is known. The observations $X$ come from a distribution $\mathcal{N}(\theta, \sigma^2)$ where $\theta$ is unknown. Taking the prior density of $\theta$, summarising prior knowledge, as $\mathcal{N}(\mu_0, \sigma_0^2)$, we obtain the posterior distribution of $\theta$ as:

$$p(\theta | X) = \frac{p(X | \theta) p(\theta)}{\int p(X | \theta) p(\theta) d\theta}$$

Then, ignoring terms not involving $\theta$, we have

$$p(\theta | X) \propto \exp \left[ -\frac{1}{2} \left( \frac{\bar{X} - \theta}{\sigma} \right)^2 n + \frac{1}{2} \left( \frac{\theta - \mu_0}{\sigma_0} \right)^2 \right]$$

therefore $p(\theta | X)$ is normal, with mean

$$\frac{n \bar{X}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}$$
and variance

\[ \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \]

The mean of the posterior distribution for \( \theta \) is thus a weighted average of the prior mean \( \mu_0 \) and the observed mean \( \bar{x} \), weighted by their respective precisions \( 1/\sigma_0^2 \) and \( n/\sigma^2 \) whilst the precision (1/variance) of the posterior distribution is the sum of the prior and observational precisions.

Using the posterior density for \( \theta \), summarising all our knowledge of \( \theta \), inferences, confidence limits, etc. can be made.

Since the posterior distribution is formed by a product of the likelihood and prior functions, the form of the prior outside some region close to \( \bar{x} \) is of little importance. It can be shown that reasonable priors lead to posterior densities close to the above results.
2.4.2 Inference about the variance

Here $\mu$ is known, $\sigma^2$ is unknown. The observations $X$ come from a distribution $N(\mu, \theta)$. We now take the prior density of $\theta$ such that $(\nu_0 \sigma_0^2) / \theta$ is distributed as a $\chi^2$ variable with $\nu_0$ degrees of freedom,

$$p(\theta) \propto \theta^{-\frac{1}{2}\nu_0 - 1} e^{-\frac{\nu_0 \sigma_0^2}{\theta}}$$

Thus a prior estimate (with $\nu_0$ degrees of freedom) of $\theta$ is $\sigma_0^2$.

Let

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

The likelihood of the observations is then:

$$p(X | \theta) = (2\pi \theta)^{-\frac{n}{2}} e^{-\frac{nS^2}{2\theta}}$$

The posterior density is then:

$$p(\theta | X) \propto \theta^{-\frac{1}{2}(\nu_0 + n) - 1} e^{-\frac{\nu_0 \sigma_0^2 + nS^2}{2\theta}}$$

Thus the posterior distribution of $\theta$ is such that

$$\frac{\nu_0 \sigma_0^2 + nS^2}{\theta}$$

is a $\chi^2$ variable with $n + \nu_0$ degrees of freedom.

In choosing this distribution for our prior knowledge of $\theta$, we are saying that small and large values, relative to $\sigma_0^2$, are unlikely, that the

![Fig. II.12](image-url)
most probable value is \( \sigma_0^2 \nu_0 / \nu_0 + 2 \), and that the mean value is \( \sigma_0^2 \nu_0 / \nu_0 - 2 \).

The spread of the distribution, as measured by the variance, is given by

\[
\sqrt{\nu} = \frac{2 \sigma_0^4 \nu_0}{(\nu_0 - 2)(\nu_0 - 4)} \sim \frac{2 \sigma_0^4}{\nu_0}
\]

for large \( \nu_0 \). Therefore the precision of our knowledge is measured by the parameter \( \nu_0 \), while the value \( \sigma_0^2 \), between the most likely and the mean values, gives a prior estimate of the true value, \( \sigma^2 \).

The posterior density has the same form, with \( \sigma_0^2 \) replaced by

\[
\frac{2 \nu_0 \sigma^2}{\nu_0 + n} + n \sigma^2
\]

and \( \nu_0 \) by \( (\nu_0 + n) \). A posterior estimate is thus given by \( \sigma_1^2 \), the precision of this estimate being measured by \( (\nu_0 + n) \).

Vague prior knowledge can be represented by a uniform distribution of \( (\ln \theta) \), corresponding to the above prior distribution when \( \nu_0 \to 0 \). In this case, the posterior distribution for \( \theta \) is such that \( n s^2 / \theta \) is distributed as a \( \chi^2 \) variable with \( n \) degrees of freedom.

2.4.3 Inference about the mean and the variance

Here \( \mu \) is unknown, \( \sigma^2 \) is unknown. The observations \( X \) now come from a normal distribution \( N(\theta_1, \theta_2) \). It is then necessary to define a joint prior distribution for \( \theta_1 \) and \( \theta_2 \). Suppose \( \theta_1 \) and \( \theta_2 \) to be independent, that is, if we were given some value for \( \theta_1 \), our beliefs about \( \theta_2 \) would not change. Taking prior distribution for \( \theta_1 \) as uniform from \( (-\infty, \infty) \) and for \( \ln \theta_2 \) as uniform from \( (-\infty, \infty) \), we have the joint posterior distribution of \( \theta_1 \) and \( \theta_2 \) as:

\[
p(\theta_1, \theta_2) \propto \theta_2^{\frac{n+2}{2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta_1)^2 / 2 \theta_2 \right].
\]

\[
\sum_{i=1}^{n} (x_i - \theta_1)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \theta_1)^2
\]

so that

\[
p(\theta_1, \theta_2) \propto \theta_2^{\frac{n+2}{2}} \exp \left[ \frac{(n-1)\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \theta_1)^2}{2 \theta_2} \right]
\]
Integrating this density with respect to $\theta_2$, we obtain the posterior distribution of $\theta_1$, namely

$$p(\theta_1 | x) \propto \left\{ 1 + \frac{n(\bar{x} - \theta)^2}{(n-1)s^2} \right\}^{-\frac{1}{2}n}.$$

Therefore $t = \sqrt{n}(\bar{x} - \theta_1)/s$ has a $t$ distribution with $(n - 1)$ degrees of freedom. Similarly, we obtain the posterior distribution of $\theta_2$ as

$$p(\theta_2 | x) \propto \theta_2^{-\frac{n-1}{2}-1} \exp \left[ -\frac{(n-1)s^2}{2\theta_2} \right].$$

that is, $\theta_2$ has a posterior distribution such that $[(n-1)s^2]/\theta_2$ has a $\chi^2$ distribution with $(n - 1)$ degrees of freedom.

Thus if the variance of the distribution is known, then:

$$\sqrt{n} \left( \theta_1 - \bar{x} \right)/\sigma \sim N(0,1).$$

If the variance is not known: $\sqrt{n}(\theta_1 - \bar{x})/s$ has a $t$ distribution with $(n - 1)$ degrees of freedom. The greater width of the $t$ distribution relative to the normal reflects the difference in knowledge prior to obtaining the observations. Similarly, if the mean is known, the posterior distribution of the variance, $\theta_2$, is such that

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\theta_2}$$

is a $\chi^2$ with $n$ degrees of freedom. If the mean is not known $\sum(X_i - \mu)^2/\theta_2$ is a $\chi^2$ with $(n - 1)$ d.f. Here the reduction in prior knowledge is reflected in the reduction in the degrees of freedom from $n$ to $(n - 1)$.

Consider again the joint posterior distribution of $\theta_1$ and $\theta_2$. The distribution of $\theta_1$, conditional on $\theta_2$, has a density:

$$p(\theta_1 | \theta_2, x) \propto \theta_2^{-\frac{1}{2}} \exp \left[ -\frac{n(\bar{x} - \theta_1)^2}{2\theta_2} \right].$$

that is $N[\bar{x}, (\theta_2/n)]$. The parameters $\theta_1, \theta_2$ are therefore not independent, contrary to the prior assumption. The larger is $\theta_2$, the greater is the spread of $\theta_1$, and the less precise our knowledge of $\theta_1$. Each observation $X_i$ has variance $\theta_2$, and has more scatter about $\mu$ the larger is $\theta_2$, and thus provides less information about $\mu$. 
2.5 Real life

In the preceding paragraphs, the most important kinds of distributions have been presented with some useful properties. It is highly dangerous to assume simply that all useful distributions have some idealized shape. Although ideal distributions are physically realisable under certain ideal conditions, real life is unfortunately not so simple, and we have to do more often with combinations or distortions of these ideal distributions. In this section we show how some typical "real distributions" can be handled using the techniques developed above for ideal distributions.

2.5.1 Truncation

The most obvious modification of ideal distributions comes from the fact that the range of our measurements is not from $-\infty$ to $+\infty$. Usually we measure variable $x$ between finite limits $A$ and $B$, so that our p.d.f. is modified as follows.

$$f(x)dx \rightarrow \frac{f(x)dx}{\int_{A}^{B}f(x)dx} = \frac{f(x)dx}{F(B) - F(A)}$$

(108)

Although this is usually a complicating effect, it is sometimes very useful. Such a case is the Cauchy or Breit-Wigner distribution. We have seen above that this distribution has no moments if considered over the full infinite range, but if it is chopped off at $+A$ and $-A$, we get for its new normalization, mean, and variance:

$$q(x) = \frac{1}{\pi} \int_{-A}^{A} \frac{1}{1 + x^2} dx$$

(109)

$$= \frac{1}{2 \arctan A} \left( 1 + \frac{1}{1 + x^2} \right)$$

$$E(X) = \frac{1}{2 \arctan A} \int_{-A}^{A} \frac{x}{1 + x^2} dx = 0$$

(110)

$$V(X) = \frac{1}{2 \arctan A} \int_{-A}^{A} \frac{x^2}{1 + x^2} dx$$

$$= \frac{A}{\arctan A} - 1$$

(111)
Taking the limit of this variance as $A \to \infty$, we see why the expectation does not exist:

$$
\lim_{A \to \infty} \nu(x) = \infty
$$

The tails of the Cauchy distribution fall off so slowly that an arbitrarily large variance can be obtained by considering the function far away from the origin. (This is not the case for the Gaussian distribution.)

### 2.5.2 Detection efficiency

Truncation is actually only a special case of a more general class of distortions which we may call detection efficiency. Since all events do not have the same probability of being observed by our apparatus, the ideal p.d.f. $f(x)$ is distorted, and becomes

$$
g(x) = \frac{\int_{y} f(x) \, \mathcal{P}(y|x) \, e(x,y) \, dy}{\int_{x,y} f(x) \, \mathcal{P}(y|x) \, e(x,y) \, dy \, dx}.
$$

where $y$ denotes a set of ancillary variables and $e(x,y)$ is the probability density for observing an event at $x$ under conditions $y$. If $e(x,y)$ can be expressed only as a function of $x$, the problem is greatly simplified.

$$
g(x) = \frac{\int f(x) \, e(x) \, dx}{\int f(x) \, e(x) \, dx}.
$$

Apart from normalization, we have simply the product of two functions, and although the detection probability $e(x)$ is often known only numerically from a Monte Carlo calculation, this case presents no special difficulties. The general case, Eq. (114) will be treated further in Part IV.

### 2.5.3 Experimental resolution

Another frequent source of distortion of ideal distributions is that of experimental uncertainty in measurements. That is, an event having a true value $x$ will yield a measured value $x'$ with a probability density, sometimes called the resolution function:

$$
r(x, x')
$$
Then if the true density of $X$ is $f(X)$, the measured density will be

$$g(X') = \int r(X, X') f(X) dX$$  \hspace{1cm} (117)

This is fundamentally different from the case of detection efficiency, for now the "true" variable $X$ has been integrated out and been replaced by a measured variable $X'$. This can give rise, for example, to a measured value at a point where the ideal density is zero. In such cases, it is clearly essential to take account of this distortion.

As described above in Section 2.3.2, it is often assumed that the resolution function $r(X, X')$ is a normal distribution

$$r(X, X') = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{(X'-X)^2}{2\sigma^2} \right]$$  \hspace{1cm} (118)

Some important properties of this ideal resolution function may be demonstrated easily:

i) If the distribution being measured is $N(\mu, \tau^2)$ and the resolution function is (118), then the resulting measured distribution is $N[\mu, (\sigma^2 + \tau^2)]$:

$$f(X) \sim \frac{1}{\sqrt{2\pi} \sigma} \exp \left[ -\frac{(X-\mu)^2}{2\sigma^2} \right]$$  \hspace{1cm} (119)

$$g(X') \sim \frac{1}{\sqrt{2\pi} \sqrt{\sigma^2 + \tau^2}} \exp \left[ -\frac{(X'-\mu)^2}{2(\sigma^2 + \tau^2)} \right]$$  \hspace{1cm} (120)

That is, the variance of the final Gaussian distribution is the sum of the variances of the original Gaussian distribution and the Gaussian resolution function.

ii) When the original distribution $f(x)$ is a Cauchy or Breit-Wigner distribution, the integral (117) cannot be reduced to any simple form, but two limiting cases can be cited. When the resolution width is small compared to the Breit-Wigner width, the resulting shape is approximately given by the Breit-Wigner alone. When the opposite is true, the Breit-Wigner width being small compared to the resolution width, the resulting distribution is nearly $N(\mu, \sigma^2)$, where $\mu$ is the expectation of the truncated Breit-Wigner, and $\sigma^2$ is the resolution function variance.
iii) If the resolution function for a given measurement is Gaussian [Eq. (118)], but the variance $\sigma^2$ is not the same for all measurements in an experiment, then the total resolution function for the experiment is not Gaussian, as was pointed out at the end of Section 2.3.2. Suppose, for example, that the measurement variance $\sigma^2$ depends on some ancillary variables (such as the position of an event in a track chamber) and its distribution can be described by a density function $q(\sigma^2)$. Then our resolution function will be

$$r(x, x') = \int_0^\infty q(\sigma^2) \cdot \frac{1}{\sqrt{2\pi} \sigma^2} \exp\left(\frac{(x-x')^2}{2\sigma^2}\right) \, d\sigma$$  \hspace{1cm} (121)$$

In this case, $r$ will be Gaussian only if $q(\sigma^2)$ is a delta-function (i.e. $\sigma^2$ is constant). More usually, our apparatus will produce events with different precision, depending on quantities like track length, which may be nearly independent of the quantity we want to measure. Let us consider two different error distributions $g(\sigma^2)$ and see what kind of resolution functions they give:

**Example 1**

$$q_1(\sigma) = \frac{2}{\sqrt{2\pi} \sigma^2} e^{-\frac{1}{2\sigma^2}}$$

For large values of $\sigma$, the above distribution goes to zero as $1/\sigma^2$, and for small values of $\sigma$, there is a reasonably sharp cut-off $[q_1(\sigma)$ and all its derivatives are zero at $\sigma = 0]$ as is shown below

The resolution function corresponding to the above error distribution is found as follows

$$r_1(x, x') = \int_0^\infty \frac{2}{\sqrt{2\pi} \sigma^2} e^{-\frac{1}{2\sigma^2}} \exp\left(\frac{(x-x')^2}{2\sigma^2}\right) \, d\sigma$$

$$= \frac{1}{\pi} \frac{\sigma}{1 + \sigma^2 (x-x')^2}$$
This is just the Cauchy or Breit-Wigner distribution, which we have seen above to have no mean and infinite variance if it is not truncated. Since, in practice, this distribution is always truncated, we do not care about the long tails which give rise to the infinite variance, and we need only consider whether the behaviour of \( q_1(\sigma) \) is reasonable at small values of \( \sigma \). If this behaviour is approximately valid, then the resolution function takes on a Breit-Wigner shape, and this allows us to fit both wide and narrow resonances using the same general shape. Conversely, if a Breit-Wigner shape is used to fit a resonance that is narrow compared with experimental resolution, then the error distribution \( q_1(\sigma) \) is implicitly assumed to be valid for small \( \sigma \).

Example 2

\( q_2(\sigma) \) = uniform distribution in \( 1/\sigma \) between \( 1/\sigma = A \) and \( 1/\sigma = B \).

\[
q_2(\sigma) = \begin{cases} \text{uniform distribution in } 1/\sigma \text{ between } 1/\sigma = A \text{ and } 1/\sigma = B, \end{cases}
\]

The uniform distribution in \( 1/\sigma \) is a good approximation to reality in certain cases, for example if the precision depends on the length of track in the bubble chamber, and all track lengths are equally probable between two cut-off values

Let \( t = \frac{1}{\sigma} \), then \( q(t) = \frac{1}{B - A} \) \( \text{if } t \in [A, B] \).

and \( r_2(x, x') = \frac{1}{(B - A)\sqrt{2\pi}} \int_A^B t e^{-\frac{(x' - x)^2}{2t}} dt \).

If we set \( x' = 0 \), we can find the expectation and variance of \( r_2(x, 0) \):

\[
E(x) = 0, \quad \text{and} \quad V(x) = \frac{1}{AB}.
\]
This says that the variance of this resolution function can be arbitrarily large if the lower limit on precision becomes arbitrarily small. Suppose now we wish to estimate $E(X)$ by taking an average of several observations from this distribution, with a fixed maximum precision $B$. We clearly get a smaller variance by throwing away events with the worst precision, but this also reduces the size of the sample ($N$). If we throw away all events with variance between the minimum $A$ and some value $A'$, then the resulting variance in the mean will be

$$\sqrt{\overline{X}} = \frac{1}{N} \sqrt{\langle X \rangle} \leq \frac{1}{A'B(\overline{B} - A')}.$$ 

The minimum variance then results when $A' = B/2$. Therefore, with the error distribution $q_2$, if it is desired to estimate the mean by an unweighted average, the minimum variance is obtained by not using any events with precision $(1/\sigma) < \frac{1}{2}$ of the maximum precision. The seeming paradox here is resolved by noting that a simple unweighted average is not the best estimator of the mean, but this will be discussed later under estimation theory.

### 3. CONVERGENCE AND THE LAW OF LARGE NUMBERS

In this chapter the most fundamental theorems are presented, which lay the groundwork for the more practical results of the following chapters. The reader interested only in practical results may skip quickly through this chapter, except for 3.3.2, the Central limit theorem, which is of both theoretical and practical interest.

#### 3.1 Tchebycheff theorem

This theorem, with its corollary, the Bienaymé-Tchebycheff Inequality, is one of the basic tools of convergence theorems.

Let $h(X)$ be a non-negative function of the random variable $X$. The Tchebycheff Theorem then says that for every $k > 0$,

$$\text{Prob}(h(X) \geq k) \leq \frac{E(h(X))}{k} \quad (122)$$

This says that independently of the shape of $h(X)$, an upper limit can be set for the probability that the function $h(X)$ will exceed some value $k$,.
if the expectation of \( h(X) \) is known. The proof is quite simple:

\[
E(h(X)) = \int h(x) f(x) \, dx
\]

\[
\Rightarrow k \int_R f(x) \, dx = k P(h(X) \geq k)
\]  

(123)

where \( R \) is the region where \( h(X) \geq k \).

Unfortunately, the theorem is too general to be useful in practical calculations since the upper limit it sets can usually be improved by using more knowledge about the function \( h(X) \). It is, however, useful in demonstrating convergence theorems.

3.2 Bienaymé Tchebycheff Inequality

In equation (122) above, if we make the substitutions

\[
h(X) \rightarrow (X - E(X))^2
\]

and

\[
k \rightarrow (k \cdot t)^2
\]

we get

\[
P \left( |X - E(X)| \geq k \cdot t \right) \leq \frac{1}{k^2}
\]

(124)

This is the Bienaymé Tchebycheff Inequality and it gives an upper limit on the probability of exceeding any given number of standard deviations, independent of the shape of the function provided its standard deviation is known. Like the Tchebycheff Theorem, it is too general to be really useful outside the purely theoretical domain, but it can have some practical applications when very little is known about the function under consideration. If it is known that the function is symmetric, an inequality stronger than (124) can be used.

3.2 Convergence

Physicists usually feel that they know what they mean by convergence and limiting processes from their familiarity with ordinary calculus where
this concept is important. In statistics, however, convergence is a more complicated idea since we have in addition to cope with random fluctuations and probabilities. Because of this, it is necessary to define different kinds of convergence.

3.2.1 Convergence in Distribution

This is the weakest kind of convergence. Consider a sequence \( \{X_1, \ldots, X_n\} \) of random variables with cumulative distribution functions \( F_1(X) \ldots F_n(X) \) then it is said that the sequence \( X_n \) converges in distribution (when \( n \to \infty \)) to \( X \), of cumulative distribution \( F \), if for every point where \( F \) is continuous we have

\[
\lim_{n \to \infty} F_n(x) = F(x)
\]

(125)

Example: Let \( X_n \) be distributed as \( N(0, 1/n) \); then \( X_n \) converges in distribution to a delta-function p.d.f. at zero. It is easy to show that for \( X \neq 0 \), \( F_n(X) \to F(X) \), where

\[
\begin{align*}
F(X) &= 0, \quad X < 0, \\
F(X) &= 1, \quad X > 0
\end{align*}
\]

(126)

At the point \( X = 0 \), equation (125) does not hold, since \( F_n(0) = \frac{1}{2} \) and \( F(0) = 0 \). But \( F(X) \) is not continuous at this point, and our definition does therefore indeed hold.

3.2.2 Paul Levy's Theorem

Convergence properties are often demonstrated most easily using the characteristic function \( \phi \). The theorem which allows us to do this is due to Paul Levy: if \( \phi_n \) converges to a certain function \( \phi \) and if the real part of \( \phi(t) \) is continuous at \( t=0 \), then

i) \( \phi \) is a characteristic function

ii) \( F_n \) converges to \( F \) of which \( \phi \) is the characteristic function.

3.2.3 Convergence in Probability

The sequence \( \{X_1, \ldots, X_n\} \) is said to converge in probability to \( X \) if for any \( \varepsilon > 0 \) and any \( \eta > 0 \), a value of \( N \) can be found such that

\[
\operatorname{Prob} \left( \left| X_n - X \right| > \varepsilon \right) < \eta
\]

(127)
for all \( n \geq N \). The convergence in distribution is weaker than this kind of convergence since the former says nothing about the "distance" of \( X_n \) from \( X \). In fact it can be shown that convergence in probability implies convergence in distribution whereas the reverse is not generally true.

### 3.2.4 Stronger Types of Convergence

Other kinds of convergence encountered in mathematical statistics are known as almost certain convergence and convergence in quadratic mean. They both imply convergence in probability. Although they are important theoretically, they are of no interest practically and will not be discussed here.

### 3.3 The Law of Large Numbers

This is the most important application of the convergence theorems. There are in fact two laws of large numbers, known as the weak law and the strong law, corresponding to weaker and stronger convergence. The distinction between the two does not interest us here, and we simply state the general result.

Let \( \{X_n\} \) be a sequence of independent random variables, each having mean \( \mu \) and variance \( \sigma_i^2 \). Then if \( \mu \) exists and all the \( \sigma_i^2 \) are finite, the quantity

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

converges in probability to \( \mu \).

### 3.3.1 Monte Carlo Integration

Perhaps the most common example of the use of the Law of Large Numbers is the calculation of an integral by the Monte Carlo Method. By the definition of expectation we have

\[
E(\bar{g}) = \int_a^b g(u) f(u) \, du
\]

(129)

We take for \( f \) the normalized uniform distribution

\[
f(u) = \frac{1}{b - a} \quad a \leq u \leq b
\]
Now from (128) we know that if \( u_i \) are chosen uniformly and independently at random over the interval \((a, b)\), the corresponding function values \( g_i \) will satisfy
\[
\frac{1}{n} \sum_{i=1}^{n} g_i \xrightarrow{\text{a.s.}} \mathbb{E}(g) \quad a.s. \quad n \to \infty \tag{130}
\]
Combining (129) and (130) gives us the usual formula for Monte Carlo integration
\[
\int_a^b f(x) dx \approx \frac{b-a}{n} \sum_{i=1}^{n} g(u_i) \xrightarrow{\text{a.s.}} \int_a^b g(u) du \tag{131}
\]
with \( u_i \) chosen uniformly from \( a \) to \( b \).

Note that the above method is quite different from usual numerical methods such as the trapezoidal rule. Under certain general conditions, all these methods converge to the correct value of the integral, but the error on the integral for finite values of \( n \) depends on the method and on the function \( g \). For example, the variance of the Monte Carlo estimate \( I \) is simply
\[
\text{var}(I) = \frac{\text{var}(g)}{n}
\]
A complete discussion of Monte Carlo Methods, although closely related to statistical problems in physics, is beyond the scope of this course, and the reader is referred to the bibliography for further reference.

3.3.2 The Central Limit Theorem

This powerful theorem is of central importance in both theoretical and practical problems in statistics. In fact it has already been used in these lectures but we take the time here to state it fully and to discuss it in some detail.

We just recall the result from 1.4.3 that if we have a sequence of independent random variables \( X_i \) each from a distribution with mean \( \mu_i \) and variance \( \sigma_i^2 \), then the distribution of the sum \( S = \sum X_i \) will have a mean \( \bar{\mu} = \sum \mu_i \) and a variance \( \bar{\sigma}^2 = \sum \sigma_i^2 \). This holds for any distributions provided that the \( X_i \) are independent and the individual means and variances exist. Nothing is said about the distribution of the sum except for its mean and variance.
so that the characteristic function of \( \mathbf{Z}_n = \frac{\sum_{i=1}^{n} X_i}{\sigma \cdot \sqrt{n}} \) is

\[
\phi(t) = \left[ \phi\left( \frac{t}{\sigma \cdot \sqrt{n}} \right) \right]^{n}
\]

It follows that

\[
\ln \phi(t) = n \ln \phi\left( \frac{t}{\sigma \cdot \sqrt{n}} \right)
\]

\[
= n \ln \left[ 1 + \frac{t^2}{2} \frac{\phi(t^2)}{n \sigma^2} + o\left( \frac{t^2}{n} \right) \right]
\]

\[
= \frac{(it)^2}{2} + o(1) \text{ for any fixed } t
\]

Thus

\[
\phi(t) \rightarrow e^{-\frac{1}{2}t^2}
\]

= characteristic function of standard normal distribution

One obtains the same result if \( X_i \) have different variances \( \sigma_i^2 \), provided the variances \( \sigma_i^2 \) are all finite, or at least do not approach infinity as fast as \( i \).

One measure of when we reach normality is given by the measures of \textbf{skewness} and \textbf{kurtosis}, namely

\[
\gamma_1 = \frac{\mu_3}{\sqrt{3} \sigma^3} \text{ and } \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3
\]

For the normal distribution these are zero. From the characteristic function we can obtain the moments \( \mu_i \), and hence measure how far the distribution is from normality.

Thus

\[
\ln \phi(t) = \mu_1 i t + \frac{\sigma^2 i^2 t^2}{2!} + \frac{\mu_3 i^3 t^3}{3!} + \frac{(\mu_4 - 3 \mu_2^2) i^4 t^4}{4!} + \ldots
\]

or

\[
\phi(t) = \sum_{j=0}^{\infty} \frac{\mu_j}{j!} (i t)^j
\]

(135)
The coefficients $Y_1$, $Y_2$ are at least some measure of how far from normality we are.

For example for the $\chi^2$ distribution with $n$ degrees of freedom, we have

$$\phi(t) = \left(1 - 2it\right)^{\frac{-n}{2}}.$$ 

and

$$\ln \phi(t) = -\frac{n}{2} \ln (1 - 2it)$$

$$\quad = -\frac{n}{2} \left[2it + \frac{4(2it)^2}{2} + \frac{2^3(2it)^3}{3} + \frac{2^4(2it)^4}{4} + \ldots \right]$$

Hence

$$\mu = n,$$

$$\sigma^2 = \frac{2n}{3},$$

$$\mu_3 = 8n,$$

$$\mu_4 - 3\mu^2 = 48n,$$

and so

$$Y_1 = \frac{8n}{(2n)^{3/2}} = 2 \sqrt{\frac{2}{n}}$$

and

$$Y_2 = \frac{48n}{4n^2} = \frac{12}{n}$$

It is generally accepted that $\chi^2$ with $n > 30$ is reasonably normal. Then

$$Y_1 = \frac{2\sqrt{2}}{\sqrt{30}} \sim 0.5,$$

$$Y_2 = \frac{12}{30} = 0.4$$

We may therefore say that for practical purposes, skewness and kurtosis less than 0.5 are negligible.
3.3.3 Example: Gaussian Random Number Generator

Most Monte Carlo calculations require a set of uniformly distributed random numbers, so high-speed computers usually have such a random number generator available as a library subroutine. However, it is often useful to have a generator producing random numbers which are distributed not uniformly but normally, say $N(0, 1)$. Using the Central Limit Theorem we can easily make such a Gaussian generator if we already have a uniform generator. If $u_i$ is the $i^{\text{th}}$ number from a uniform generator, take

$$g = \sum_{i=1}^{n} u_i - \frac{n}{2} \sqrt{\frac{n}{12}}.
$$

From equation (132) and the fact that $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$ for the uniform distribution, it is immediately seen that (137) becomes $N(0, 1)$ as $n \to \infty$. In practice, $g$ is already very close to normal when $n = 10$ for instance. The largest discrepancies occur in the tails of the distribution, which are truncated:

$$-\sqrt{3n} \leq g \leq \sqrt{3n}.
$$

A safe and convenient choice is $n = 12$, which produces values as far out as six standard deviations, and yields the simple formula

$$g = \sum_{i=1}^{12} u_i - 6.
$$

Note however, that exact methods, based on change of variables, exist and may sometimes be more efficient.

3.4 Asymptotic Properties of Special Distributions

Some of the special distributions discussed earlier converge asymptotically to certain other types of distributions in an appropriate limit. Rather than discuss each case in detail, we summarize these asymptotic properties in the figure below. For details see the section on distribution.
Fig. II.15