DETERMINING THE STATISTICAL SIGNIFICANCE OF EXPERIMENTAL RESULTS

Lectures presented at the 1980 CERN School of Computing, Vraona, Attiki, Greece 14 September-27 September 1980
CONTENTS

1. ESTIMATION OF PARAMETERS. .................................................. 1
   Statistics as the inverse of Probability. ................................. 2
   Normal Theory (Gaussian measurements). ................................. 2
   The universality of the Gaussian distribution. .......................... 3
   Real-life resolution functions. ............................................. 4
   Confidence Intervals (classical theory). .................................. 4
     Example: confidence intervals for the binomial parameter. ........... 6
   Confidence Intervals (Bayesian theory). ................................... 8
     Bayes' Theorem ............................................................. 8
     The Bayesian use of Bayes' Theorem .................................... 8
   Use of the likelihood function. ............................................. 9
     The second derivative of the log-likelihood function. ............... 11
     The method of MINOS ...................................................... 11
   Combination and propagation of Uncertainties. ......................... 12
     The sum or difference of two variables. ............................... 12
     Local theory, or the propagation of small errors. ................... 12
     Error on the ratio of two continuous variables. ..................... 13
     Error on the ratio of two discrete variables. ....................... 13
     The propagation of large errors. ..................................... 14
   Multiparameter estimation. ............................................... 17
     The Error Matrix ....................................................... 17
     MINOS with several free parameters ................................ 18
     Probability content of confidence regions .......................... 19

2. TESTING OF HYPOTHESES. .................................................... 23
   The test statistic and other definitions. ................................ 23
     Example: Separation of two types of events. .......................... 24
   Choosing a test. ......................................................... 25
   The Neyman-Pearson test. ............................................... 26
   Composite hypotheses. .................................................... 27
     The maximum likelihood ratio. ....................................... 28
   Small-sample problems. .................................................. 28

3. GOODNESS-OF-FIT. ............................................................ 29
   Choosing a goodness-of-fit test ......................................... 29
   Distribution-free tests. ................................................. 30
   Comparing two one-dimensional distributions. .......................... 31
     The Chi-square test. .................................................. 31
     The Kolmogorov test. .................................................. 32
     The Smirnov-Cramer-Von Mises test. ................................ 33
   Comparing multidimensional distributions. ............................ 33
DETERMINING THE STATISTICAL SIGNIFICANCE OF EXPERIMENTAL RESULTS

F. James

CERN, Geneva, Switzerland

The results of an experiment are usually expressed as one of the following:

1. The measurement of one or more numerical values, parameters of a theory, or

2. Evidence that one theoretical hypothesis is more likely to be true than another hypothesis, or

3. Evidence for or against one hypothesis being true (without mention of other possible hypotheses).

In each of these cases, it is of the greatest importance to be able to understand and present the statistical significance of the results, namely the accuracy of the measurements or the strength of the evidence for or against hypotheses.

In these lectures we present the theoretical basis for assigning numerical values to the significance of such results, and discuss the practical application of that theory.

In order to squeeze a reasonable amount of significant material into the time allotted, it is necessary to assume some prior knowledge on the part of the reader of these notes, namely a good basic understanding of matrix algebra and differential and integral calculus, and enough background in probability and statistics to understand such terms as expectation, variance, Gaussian distribution, Poisson distribution, and binomial distribution. The reader who feels unsure about such elementary statistical concepts is invited to consult a standard textbook. One such text which is intended for physicists and which will be referred to in these lectures is Eadie et al. (1971).

1. ESTIMATION OF PARAMETERS.

In this chapter we address ourselves to parameter measurement (for which we use the statistical term estimation), and in particular to the determination of measurement uncertainties, known to physicists as error calculation and known to statisticians as interval estimation.
1.1 STATISTICS AS THE INVERSE OF PROBABILITY.

Probability theory is a branch of pure mathematics, and although physicists often use it just as they use algebra, it has nothing to say directly about the physical world. The classical problem in probability is: 'Given the probabilities of certain elementary events, find the probability of a more complicated combination of such events.' For example, given that events occur randomly in time, with equal probability per unit time, and given that this probability per unit time is known to be, for example, one per second, what is the probability of observing exactly three events in an interval of one second? As is well-known, this problem was solved centuries ago by Poisson, and the answer is found by evaluating the Poisson function:

\[ P(n) = \frac{k^n e^{-k}}{n!} \]

where \( k = 1 \), the expected number of events in the time interval, and \( n = 3 \), so the probability is 0.0613.

Now let us consider the inverse problem: 'Given that a certain number of events have been observed to occur in a given time interval, and given that events are known to occur randomly with constant, but unknown probability \( k \) per unit time, what can we say about the value of the constant \( k \)?' Here we leave the domain of pure probability and enter the realm of statistics, also a branch of mathematics, but considerably more recent and much closer to experimental science. Already the question posed is more vague than we would expect for a proper mathematical problem, and as we shall see below, the answer furnished by statistics is open to at least two different interpretations according to whether one adopts the Bayesian or non-Bayesian point of view. All this is not really surprising since the question we are asking (and therefore the mathematics we are going to use to solve it) is basically inductive whereas we are used to mathematics being deductive.

Before answering our statistical question and attempting to interpret the meaning of the answer, let us consider a somewhat simpler case where the interpretation is independent of our philosophical viewpoint.

1.2 NORMAL THEORY (GAUSSIAN MEASUREMENTS).

A large class of phenomena yield variables which are distributed according to a normal or Gaussian distribution.¹ A discussion of where and why we should expect to meet such a distribution is given in the subsection below; let us assume for the moment that we are dealing with a random variable such as the measurement of the length of a table, where the variable is distributed according to:

\[ P(x|x' \leq x + dx) = c \exp[-(x-x_e)^2/2s^2] \, dx \]

¹ As used in these lectures, the words normal and Gaussian are completely equivalent.
where e is the true length of the table and s is the standard deviation of the measurement.

The above formula should be read: "The probability of observing (measuring) a value between x and x + dx is proportional to dx and to the exponential of -(x - e)^2 / 2s^2." [The normalisation constant c does not interest us here, but it is important as the total probability for observing any x must be one.] In order to calculate this probability one must know the true value e, which in practice we can only determine by using measurements x, so the whole process appears to be circular. In fact, although we don't know e, we do know a value of x, our measurement, and interestingly enough we notice that the formula is symmetric upon interchange of x and e, so that we can consider that it also gives the distribution of e for a given x.

As we will see in more detail below, the precise interpretation of the Gaussian formula will depend on which of the two schools of statistics we choose to follow:

1. The classical approach is that e is unknown but fixed, and we can talk only about probability distributions for observing x.

2. The Bayesian approach allows for distributions of degrees of belief in the value of e.

In any case the symmetry of the Gaussian distribution in e and x causes the numerical determination of the statistical significance of the measurement to be exactly the same in the two cases, namely the area under the relevant part of the Gaussian probability density curve. For example, the probability that e and x are less than s apart (one standard deviation) is about 68%; less than 2s apart is about 95%, etc.

1.2.1 The universality of the Gaussian distribution.

Physicists are so used to the usual rules-of-thumb connecting standard deviations and significance levels (for example, "two standard deviations is a 5% effect") that they often do not realize that the usual correspondence is only true for variables (measurements) which are Gaussian distributed. The assumption is that everything is Gaussian distributed, and indeed simple measurements of continuous physical phenomena usually do appear to assume a Gaussian shape, when a series of independent identical measurements are made and the results plotted as a histogram. On the other hand it is easy to see that this cannot be an exact and universal effect, since if it were true for some variable y, it would in general not be true for any function of y such as y^2.

There is in fact some mathematical grounds for "universal normality", as expressed by the Central Limit Theorem. This theorem states that the sum of n random variables approaches a Gaussian distribution for large n, no matter how the individual variables are distributed. This amazing but well-known theorem, which apparently is contradicted by the reason-
ing given directly above,\textsuperscript{2} would explain universal normality if a complex measurement could be considered as the sum of a large number of component measurements. For example, if the inaccuracy in measuring some parameter results from the additive combination of a large number of smaller elementary inaccuracies, the total inaccuracy should be approximately Gaussian, for identical but independent measurements.

1.2.2 \textbf{Real-life resolution functions.}

If the sum of \textit{random variables} is asymptotically Gaussian distributed, even when the individual random variables are not, the sum of \textit{probability distributions} is in general not at all Gaussian, even if all the component distributions \textit{are} Gaussian, as long as they have different widths.

Let us assume, for example, that the process of measuring the length of a table with some measuring instrument provides Gaussian distributed values of the length, with a given standard deviation \(s\). We make a series of measurements with this instrument, but then continue with another instrument having a different accuracy and therefore a different value of \(s\). Now if we look at the combined histogram of all measurements together, we will see a superposition of two Gaussians, with an overall shape which is not at all Gaussian, even though the basic process is Gaussian.

This situation is familiar in high energy physics where the \textit{resolution function} is in general a superposition of Gaussians of different widths lying in a continuous range corresponding to the fact that the accuracy with which one can measure the momentum and angles of a particle track are continuous functions of the momentum and angles themselves. The resulting resolution function is always more sharply peaked than a Gaussian (due to the measurements with exceptionally small errors) and also has longer tails than a Gaussian (due to measurements with the largest errors).

1.3 \textbf{CONFIDENCE INTERVALS (CLASSICAL THEORY).}

The classical theory of confidence intervals allows us to find an interval (a range of parameter values) which, with a given probability, will contain the true value of the parameter being measured. We show here how such intervals can be constructed, at least in theory, for the most general case.

\textsuperscript{2} The resolution of this paradox resides in the nature of convergence to limits for distributions of random variables. The convergence will be different for different functions of the random variable, even if the limit is the same.
Let $t$ be an estimate (measurement) of a parameter whose true value (unknown) is $\theta$. The value of $t$ for a given experiment will be a function of the measured data $x$:

$$ t = t(x) $$

For any given value of $\theta$, we can in principle always find the expected distribution (probability density) of $t$, which we denote by $f(t|\theta)$. This function describes the behaviour of the measurement apparatus. For the simplest case, it would simply be a Gaussian centered at $t=\theta$, with width given by the measurement accuracy of the apparatus. For the most complicated experiments it is necessary to find $f$ numerically by Monte Carlo simulation of the measuring system.

Now consider a range of values of $t$ from $t_1$ to $t_2$, denoted by $\langle t_1, t_2 \rangle$. Since $f(t|\theta)$ is known, we can calculate, for any given value of $\theta$, the probability of obtaining a value of $t$ in the range $\langle t_1, t_2 \rangle$:

$$ P(t_1 \leq t \leq t_2 | \theta) = \int_{t_1}^{t_2} f(t'|\theta) dt' $$

Similarly, for a given value of $\theta$, $0 < \theta < 1$, it is possible to find values of $t_1$ and $t_2$ satisfying the above relation, namely that the probability of $t$ lying in the range $\langle t_1, t_2 \rangle$ is $\beta$. In fact, there are in general many different ranges which will satisfy this relation; it can be made unique by requiring in addition that the range be central:

$$ \int_{-\infty}^{+\infty} f(t'|\theta) dt' = \int_{t_2}^{+\infty} f(t'|\theta) dt' = \alpha $$

and therefore $\alpha = (1-\beta)/2$

The above relations, for a given confidence level $\beta$, uniquely define two functions $t_1(\theta)$ and $t_2(\theta)$. Since the functions are only defined implicitly, the actual calculation of function values would in general have to be performed numerically by iteration, but it is in principle always possible, and may even be quite easy in certain simple but important cases. Thus we can draw two curves in $(t, \theta)$-space, representing $t_1(\theta)$ and $t_2(\theta)$. Let us take the $\theta$-axis is horizontal, and the $t$-axis vertical. Then along any vertical line (constant $\theta$), the distance between the two curves represents by construction a probability content $\beta$ (that is, the probability of obtaining an estimate $t$ in that range is $\beta$ for that value of $\theta$.)

The usefulness of the diagram comes from the fact that this is also true along a horizontal line: namely, for any given value of $t$, the probability content of the range of $\theta$ lying between the two curves is also $\beta$. The diagram is constructed vertically, but can also be read horizontally. To convince oneself that this is true requires some mental gymnastics and a good understanding of the meaning of the probability content of a range of parameter values in the classical sense, but
it is rigorously true. Since the classical interpretation of confidence intervals can best be understood by opposition to the Bayesian interpretation, this should become clearer after the discussion of Bayesian theory below.

1.3.1 Example: confidence intervals for the binomial parameter.

The technique outlined above was used, for example, by Clopper and Pearson to find the confidence limits for the parameter of a binomial distribution. This is the distribution which describes processes where only two kinds of events are possible, and a given event has a constant probability of being of one kind, independent of the other events. [Examples: measuring the branching ratio for a particle with only two decay modes; measuring a forward-backward or left-right asymmetry] Then, if the probability of one event being of one type is \( p \), the probability of \( n \) events out of a total sample of \( N \) being of that type is given by the well-known binomial formula:
\[ P(n) = \binom{N}{n} \frac{n}{p} (1-p)^{N-n} \]

where \( \binom{N}{n} \) is the binomial coefficient \( N!/(N-n)!n! \)

Figure 2: Clopper-Pearson confidence intervals for the binomial distribution

This was used to construct (vertically) the diagram of Figure 2, which was in turn used horizontally to solve the inverse problem: Given observed values of \( N \) and \( n \), what confidence limits can be placed on the value of \( p \)?

Clopper and Pearson give elaborate diagrams which allow the reader to construct confidence intervals for essentially any \( N, n, \) and \( B \). Later it was recognized that these values are in fact just those of Fisher's F-distribution with appropriately transformed arguments, so that existing tables or subroutines of F can be used to determine exact confidence intervals for \( p \). Additional details concerning this problem and the way in which it arises in the context of physics experiments, may be found in James and Roos (1980).
1.4 CONFIDENCE INTERVALS (BAYESIAN THEORY).

The Bayesian theory of parameter interval estimation, which is not accepted by all statisticians, is based on an extension of the domain of validity of Bayes' theorem beyond what a "classical" statistician would deem proper. We therefore begin this section by recalling Bayes' theorem, a fundamental theorem of probability theory.

1.4.1 Bayes' Theorem

Consider a set of elements, each of which may belong to set A, or to set B, or to neither or to both A and B. Then the probability of a random element belonging to both A and B is the probability of its belonging to A given that it already belongs to B, multiplied by the probability of its belonging to B. The same is clearly true with A and B inverted, so that we can write:

\[ P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A) \]

This is Bayes' Theorem. It gives a relationship between a conditional probability \( P(A|B) \), and the inverse conditional probability \( P(B|A) \).

\[ P(B|A) = \frac{P(A|B)P(B)}{P(A)} \]

1.4.2 The Bayesian use of Bayes' Theorem

We have assumed known the conditional probability density \( f(t|\theta) \), namely the distribution of estimates \( t \) we would obtain if the true value of the parameter were known. Our experiment however gives us just the inverse: it provides a value of \( t \) and we wish to make a probability statement about \( \theta \). Apparently Bayes' Theorem is just what we need since it allows us to express a conditional probability in terms of its inverse. Straightforward application of the theorem to the case of interest gives:

\[ P(\theta|t) = \frac{P(t|\theta)P(\theta)}{P(t)} \]

(Each of the above "probabilities" is in fact a probability density and should be followed by a d\( \theta \) or a dx, but these differentials clearly cancel, so that we can work directly with the probability densities.)

Now let us examine each of the factors in the above expression:

1. \( P(\theta|t) \) is just what we want (says the Bayesian), the probability density for the true value of the parameter \( \theta \), given the measured value of \( t \). The "classical" statistician says that although \( \theta \) is unknown, it has only one true fixed value, and it does not make sense to talk about the probability of its taking on any value. The Bayesian counters by calling \( P(\theta|t) \) the "degree of belief" in
the value $\theta$, and says that as long as the true value is not
known, this quantity behaves like a probability.

2. $P(t|\theta)$ is nothing but our old friend $f(t|\theta)$.

3. $P(\theta)$ presents a serious problem both in interpretation and in
practical calculation. This quantity is known as the prior know-
ledge of the parameter $\theta$. The Bayesian claims that we nearly
always have prior knowledge about the phenomena we study, and
that the experimental results will always be interpreted in terms
of this knowledge anyway, so why not build it into the results
from the beginning? Even the Bayesian admits however to practi-
cal difficulties in expressing vague knowledge about $\theta$, and espe-
cially in expressing complete ignorance, a subject to which sta-
tisticians have devoted considerable effort. Moreover, the
non-Bayesian insists that it must be possible to express the
results of an experiment independently of other outside know-
ledge, but this is apparently impossible in the Bayesian for-
mat ion.

4. $P(t)$, the a priori probability of observing $t$, is apparently even
more intractable than $P(\theta)$, but in fact can be reduced to the
same problem since we can express it as:

$$
\int_{-\infty}^{+\infty} P(t|\theta) P(\theta) d\theta
$$

1.5 USE OF THE LIKELIHOOD FUNCTION.

In practice, Bayesian theory of confidence intervals is not used by phys-
sicists, probably because of the problem of subjectivity involved in the
prior knowledge, and also because of practical difficulties arising in
computation. Similarly, the classical technique of construction of
exact confidence intervals is applied primarily to the solution of
rather general problems (such as the example of 1.3.1) and is rarely
used to estimate errors for particular experiments. The reason why
these fundamental exact techniques are not used is that there exists a
much simpler approximate method for obtaining interval estimates by
using the likelihood function directly.

Consider the log-likelihood function for the simplest possible exper-
iment: the direct measurement of a quantity ($\theta$) using a measuring engine
which produces Gaussian-distributed errors of zero bias and known vari-
ance $\sigma^2$:

$$
\ln L = \ln f(t|\theta) = \ln \left| \frac{1}{\sigma\sqrt{2\pi}} \exp\left[ -\frac{(t-\theta)^2}{2\sigma^2} \right] \right| 
= -\frac{(t-\theta)^2}{2\sigma^2} + \text{const.}
$$
Figure 3. Parabolic log-likelihood for Gaussian measurement.

This is the parabola shown in Figure 3. We know by construction that this represents a measurement with standard deviation = σ. That is, if θ₀ is the value of θ for which ln L has its maximum, then:

\[ P(θ₀⁻σ ≤ θ ≤ θ₀+σ) = 0.68 \]
\[ P(θ₀⁻2σ ≤ θ ≤ θ₀+2σ) = 0.95 \]

etc.

Now if the value of σ is not known, it can be measured from the shape of the parabola in either of two ways:

1. By measuring the second derivative of the log-likelihood function at its maximum (or in fact anywhere, since it is constant). Then:

\[ σ^2 = \left| \frac{∂^2 \ln L}{∂θ^2} \right|^{-1} \]

2. Or by taking σ = |θ₁ - θ₀|, where θ₁ is either of the two points where
\[ \ln L(\theta_1) = \ln L(\theta_0) - 1/2. \]

We will call this second way the method of MINOS.\(^3\)

For the case at hand, both methods give the same value of \( \sigma \). We consider below application of these methods to the more general case of a likelihood function arising from any experiment.

1.5.1 The second derivative of the log-likelihood function.

Largely for reasons of computational simplicity, this is the most common method of calculating parameter uncertainties directly from the likelihood function. In the case where several parameters are estimated simultaneously the full error matrix of the parameters is obtained simply by inverting the second derivative matrix of the log-likelihood function (usually at its maximum). This method is considered an excellent approximation whenever the log-likelihood function is indeed parabolic in shape, (in which case the second derivative matrix is constant) which in turn will happen whenever there is a large amount of data, small measurement errors, or a model which is a linear function of the parameters to be estimated.

1.5.2 The method of MINOS

This method is much more recent than the second-derivative method; it was probably originated by Shepple (CERN, unpublished) around 1965 and first implemented in a general program MINROS (now obsolete, superseded by MINUIT) shortly afterward. It is believed to be valid (at least to a good approximation) also in the general non-linear case where the log-likelihood function is not parabolic.

The justification for its validity for non-parabolic cases lies in the fact that it is invariant to any (even non-linear) transformation of the parameter \( \theta \). That is, if the 68% confidence interval as determined by MINOS for parameter \( \theta \) is \( \langle \theta_1, \theta_2 \rangle \), and we make a transformation of variables to \( \Phi = \Phi(\theta) \), then the 68% confidence interval in \( \Phi \) will be \( \langle \Phi_1, \Phi_2 \rangle \), where \( \Phi_1 = \Phi(\theta_1) \) and \( \Phi_2 = \Phi(\theta_2) \). Now there must be some (in general non-linear) transformation of \( \theta \) which would transform the log-likelihood function to an exact parabola, for which the method is believed to give a very good approximation to the exact interval. And since the method is invariant under such a transformation, it must also give a good approximate answer in the general case, unlike the second-derivative method.

\(^3\) MINOS is the name of the subroutine in MINUIT which implements this method. Other references to MINUIT will be made in these lectures for the benefit of MINUIT users. MINUIT is available from the CERN Program Library and is described in James and Roos (1975).
Note that this method will in general yield asymmetric intervals, with different parameter uncertainties in the positive and negative directions. Intervals may also be non-linear in the sense that two-standard-deviation errors [given by $\ln L(\theta_1) = \ln L(\theta_0) - 2.0$] are not necessarily twice as big as one-standard-deviation errors [given by $\ln L(\theta_1) = \ln L(\theta_0) - 0.5$].

1.6 COMBINATION AND PROPAGATION OF UNCERTAINTIES.

It often happens that one wants to know the uncertainty of a certain value which is not measured directly but is a known function of other quantities whose uncertainties are known. For example, we may estimate the mass of the $\mu$ lepton using the mass of the $\pi$ meson (and its uncertainty) and the mass difference $\pi-\mu$, with its uncertainty. Physicists call this propagation of error, since the error in the final result arises from combining the errors of the two component values. It can be considered simply as a transformation of variable rather than a true statistical problem, but it so often arises in a statistical context that it is appropriate to discuss it here.

1.6.1 The sum or difference of two variables.

It follows directly from the definition of expectation and variance that the expectation of the sum of two independent random variables is the sum of the expectations of the variables, and the variance of the sum is the sum of the variances. (It is important to note that this is true only when the two component variables (measurements) are independent.) This means that the standard deviation of the sum (or difference) is the square root of the sum of the squares of the individual standard deviations. As long as the individual variables are independent and their variances are finite, this is an exact result for any number of measurements and for any distribution of deviations. If, in addition, the individual measurements are Gaussian, the sum or difference will also be Gaussian, and in fact -- by the Central Limit Theorem -- the sum or difference will always be "more Gaussian" than the individual distributions.

1.6.2 Local theory, or the propagation of small errors.

Apart from the simplest case of sums or differences, the exact calculation of uncertainties of transformed variables can be extremely complicated and highly dependent on the exact distributions involved. For this reason one usually uses only the local properties of the distributions around the estimated parameter values, which is usually an excellent approximation, especially when errors are indeed small. Thus one linearizes the transformations by using only the first derivatives of the new variables with respect to the old variables:
(\Delta R)^2 = \sum \frac{\partial R}{\partial \theta_i} V_{ij} \frac{\partial R}{\partial \theta_j} + ... \\

where R=R(\theta_1, \theta_2, ...) is the new variable and V_{ij} is the variance-covariance matrix of the component variables \theta. The three dots indicate that there are higher order terms containing higher derivatives of R with respect to \theta, but the usual linear approximation is to neglect the higher terms and keep only the sum given here.

1.6.3 Error on the ratio of two continuous variables.

If x and y are two independent variables with variances \sigma_x^2 and \sigma_y^2, then a straightforward application of the linear approximation above gives for the variance of the ratio R=x/y:

(\Delta R/R)^2 = \sigma_x^2/x^2 + \sigma_y^2/y^2

This is a well-known rule-of-thumb, often used by physicists. It is interesting to see how close it is to the exact answer, assuming x and y to be Gaussian-distributed variables. The shocking answer is that the exact variance of R in this case is infinite! At first glance, hardly a good approximation. In fact, as long as \sigma_x \ll |x| and \sigma_y \ll |y| the approximation is good locally, and gives about the right width to the distribution of R, but it underestimates the extreme tails of the distribution which cause the integrated square deviation of R to diverge. The approximation is in some sense closer to what the physicist really wants to know than the exact answer, and indeed the exact 68% confidence interval for R (which is of course finite) is close to that given by the linear approximation. The physicist should however be aware of the fact that the distribution of R deviates strongly from a Gaussian in the tails, especially when the errors in x and y are not small compared with the values of x and y (especially y).

1.6.4 Error on the ratio of two discrete variables.

An important case where the local approximation does not work is where small samples of discrete data are involved. Fortunately, exact methods are often relatively easy to apply for these cases, so good solutions can be found, but the usual approximations must be avoided.

Consider the measurement of a branching ratio for the decay of an unstable particle with only two decay modes. We observe, in a given time interval, n decays of one mode and N-n decays of the other. The branching ratio may be estimated as R=n/N, and it is tempting to estimate the uncertainty on R by combining uncertainties of \sqrt{n} and \sqrt{N} using the formula given above for the error of a ratio. Two difficulties arise: (1) n and N are not independent, and (2) the linear approximation is very poor when n and N are small (say less than 10). The exact solution of this problem follows from the fact that the distribution
involved is really binomial (since there are only two outcomes for a decay). A complete treatment of this problem including examples of how it arises in physics experiments (and how some experimenters have published incorrect error analyses by using the approximation when it was not justified) is given in James and Roos (1980).

1.6.5 The propagation of large errors.

The question now arises of what to do in the general case for continuous variables when the linear approximation for error propagation is suspected of being poor. Straightforward calculation of the distribution of the new variable R involves complicated integrals over the component distributions which, even if they are independent Gaussians, quickly become intractable, and one must resort to numerical calculations even in relatively simple cases.

One such case came up recently in the analysis of an experiment by Reines, Sobel, and Pasierb which gives evidence for the instability of the neutrino. This result is of the greatest importance in high energy physics since it has generally been believed that all neutrinos were massless and could not decay. In view of the consequences of neutrino decay, it is necessary to determine the significance of these results accurately. The final result of the experiment is the measurement of the ratio of two cross sections, let us call this R. Expressed in terms of the elementary quantities measured in the experiment, it can be written as:

\[
R = \frac{d}{(b-c) - 2 \left(1 - \frac{k^2d}{ke}\right) a}
\]

where

- \(a = 3.84 \pm 1.33\)
- \(b = 74 \pm 4\)
- \(c = 9.5 \pm 3\)
- \(d = 0.112 \pm 0.009\)
- \(e = 0.32 \pm 0.002\)
- \(k = 0.89\)

Straightforward application of the linear approximation gives:

\[R \approx 0.191 \pm 0.073\]

But theoretical calculations show that the neutrino is unstable if R is less than about 0.42. Therefore, based on approximate error analysis, the result appears to be very significant: 3.2 standard deviations or about one chance in a thousand that the neutrino is stable.
However, two of the elementary quantities have large errors, and two quantities enter into the formula twice, producing correlations. In addition, there are several fractions, which we have seen cause non-Gaussian distributions, so let us try to calculate the exact confidence intervals for R. The easiest (and perhaps the only) way to do this is by Monte Carlo. Choose values of a,b,c,d,e randomly according to the appropriate Gaussian distributions (we will be optimistic and assume that at least the elementary measurements are Gaussian with known variances), and plot the resulting values of R. The FORTRAN program to do this is so simple that I include it here (Calls to subroutines beginning with H are for the HBOOK histogramming package; NORRAN is a Gaussian random number generator; all subroutines called here are from the CERN Program Library):

```
PROGRAM REINES(INPUT,OUTPUT)
C  CALCULATION OF ERROR ON NEUTRAL TO CHARGED CURRENT
C  NEUTRINO INTERACTIONS, D'APRES REINES AND ROOS.
C
C  SET UP HISTOGRAM OF R
CALL HBOOK1(1,10H N OVER D , 50, 0., 0.5, 0.)
C
C  FILL HISTOGRAM BY LOOPTING OVER RANDOM SAMPLES OF R
  DO 100 I= 1, 10000
  CALL NORRAN(XN)
  XN = XN*1.33 + 3.84
  CALL NORRAN(X112)
  X112 = X112 * .009 + 0.112
  CALL NORRAN(X74)
  X74 = X74 * 4. + 74.
  CALL NORRAN(X95)
  X95 = X95 * 3. + 9.5
  CALL NORRAN(X32)
  X32 = X32 * 0.02 + 0.32
  X89 = 0.89
  D1 = X112*(X74-X95)/(X89*X32)
  D2 = 2.0 * XN * (1.0 - (X89*X112/X32))
  XXX = XN/(D1-D2)
  CALL HFILL (1,XXX)
  100 CONTINUE
C
C  ASK FOR PRINTING OF HISTOGRAM, WITH INTEGRATED CONTENTS
  CALL HINTEG(1, 3YES)
  CALL HISTOO
C
C  STOP
END
```

The histogram showing the distribution of the 10000 Monte Carlo values of R is shown in Figure 4. Those of you familiar with the reading of HBOOK output will quickly find the significant number, namely the number of entries falling above 0.42. This is almost 4%, so that the true significance of the result is only 4% instead of the apparent 0.1%. Notice also the skew, non-Gaussian distribution of R.
Figure 4. Distribution of R for experiment of Reines et al.
1.7 MULTIPARAMETER ESTIMATION.

In addition to the difficulties described above, a special class of problems arises in interpreting errors when there are more than one free parameters. These problems are quite separate from those described above and are really much simpler in principle, although in practice confusion often arises.

1.7.1 The Error Matrix

The error matrix, also called the covariance matrix, is the inverse of the second derivative matrix of the (log-likelihood or chi-square) function with respect to its free parameters, usually assumed to be evaluated at the best parameter values (the function minimum). The diagonal elements of the error matrix are the squares of the individual parameter errors, including the effects of correlations with the other parameters.

The inverse of the error matrix, the second derivative matrix, has as diagonal elements the second partial derivatives with respect to one parameter at a time. These diagonal elements are not therefore coupled to any other parameters, but when the matrix is inverted, the diagonal elements of the inverse contain contributions from all the elements of the second derivative matrix, which is 'where the correlations come from'.

Although a parameter may be either positively or negatively correlated with another, the effect of correlations is always to increase the errors on the other parameters in the sense that if a given free parameter suddenly became exactly known (fixed), that would always decrease (or at least not change) the errors on the other parameters. In order to see this effect quantitatively, the following procedure can be used to 'delete' one parameter from the error matrix, including its effects on the other parameters:

1. Invert the error matrix, to yield the second-derivative matrix.

2. Remove the row and column of the inverse corresponding to the given parameter, and

3. Re-invert the resulting (smaller) matrix.

This reduced error matrix will have its diagonal elements smaller or equal to the corresponding elements in the original error matrix, the difference representing the effect of knowing or not knowing the true value of the parameter that was removed at step two. This procedure is exactly that performed by MINUIT when a FIX command is executed. Note that it is not reversible, since information has been lost in the deletion. The MINUIT commands RESTORE and RELEASE therefore cause the error matrix to be considered lost and it must be recalculated entirely.
1.7.2 **MINOS with several free parameters**

The MINOS algorithm is described in some detail in the MINUIT long-write-up and will not be repeated here, but we will add some supplementary 'geometrical interpretation' for the multidimensional case (which is the usual case -- in fact, early versions of MINOS had a bug which prevented them from working in the one-parameter case because it had not occurred to the authors that anybody would use it for only one parameter!).

Let us consider that there are just two free parameters, and draw the contour line connecting all points where the function takes on the value $F_{\text{min}} + \text{UP}$. (The CONTOUR command will do this for you from MINUIT). For a linear problem, this contour line would be an exact ellipse, the shape and orientation of which are described in Eadie et al, p.196 (fig. 9.4). For our problem let the contour be as in Figure 5 below. If MINOS is requested to find the errors in parameter one (the x-axis), it will find the extreme contour points A and B, whose x-coordinates, relative to the x-coordinate at the minimum (X), will be respectively the negative and positive MINOS errors of parameter one.

![Figure 5 - MINOS errors for parameter 1](image-url)
1.7.3 Probability content of confidence regions

For an n-parameter problem MINOS performs minimizations in \((n-1)\) dimensions in order to find the extreme points of the hypercontour of which a two-dimensional example is given in Figure 5, and in this way takes account of all the correlations with the other \(n-1\) parameters. However, the errors which it calculates are still only single-parameter errors, in the sense that each parameter error is a statement only about the value of that parameter. This is represented geometrically by saying that the confidence region expressed by the MINOS error in parameter one is the cross-hatched area of Figure 6, extending to infinity at both the top and bottom of the figure.

![Figure 6 - MINOS error confidence region for parameter 1](image)

If UP is set to the appropriate one-std.-dev. value, then the precise meaning of the confidence region of Figure 6 is: "The probability that the true value of parameter one lies between A and B is 68.3%" (the probability of a Normally-distributed parameter lying within one std.-dev. of its mean). That is, the probability content of the cross-hatched area in Figure 6 is 68.3%. No statement is made about the simultaneous values of the other parameter(s), since the cross-hatched area covers all values of the other parameter(s).

If it is desired to make simultaneously statements about the values of two or more parameters, the situation becomes considerably more com-
plicated and the probabilities get much smaller. The first problem is that of choosing the shape of the confidence region, since it is no longer simply an interval on an axis, but a hypervolume. The easiest shape to express is the hyperrectangle given by:

\[ A < \text{param } 1 < B \\
C < \text{param } 2 < D \\
E < \text{param } 3 < F, \quad \text{etc.} \]

\[ \text{param 2} \]

\[ \text{param 1} \]

---

Figure 7 - Rectangular confidence region for parameters 1 and 2

This confidence region for our two-parameter example is the cross-hatched area in Figure 7. However, there are two good reasons not to use such a shape:

1. Some regions inside the hyperrectangle (namely the corners) have low likelihoods, lower than some regions just outside the rectangle, so the hyperrectangle is not the optimal shape (does not contain the most likely points).

2. One does not know an easy way to calculate the probability content of these hyperrectangles (see Eadie et al, p.196-197, esp. fig. 9.5a).

For these reasons one usually chooses regions delimited by contours of equal likelihood (hyperellipsoids in the linear case). For our two-parameter example, such a confidence region would be the cross-hatched region in Figure 8, and the corresponding probability statement is: "The probability that parameter one and parameter two simultaneously take on values within the one-std.-dev. likelihood contour is 39.3%. 

- 20 -
The probability content of confidence regions like those shaded in Figure 8 becomes very small as the number of parameters NPAR increases, for a given value of UP. Such probability contents are in fact the probabilities of exceeding the value UP for a chisquare function of NPAR degrees of freedom, and can therefore be read off from tables of chisquare. Table 1 below gives the values of UP which yield hypercontours enclosing given probability contents for given number of parameters.
<table>
<thead>
<tr>
<th>Number of Parameters</th>
<th>Confidence level (probability contents desired inside hypercontour of $\chi^2 = \chi^2_{\text{min}} + \text{UP}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50%</td>
</tr>
<tr>
<td>1</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>1.39</td>
</tr>
<tr>
<td>3</td>
<td>2.37</td>
</tr>
<tr>
<td>4</td>
<td>3.36</td>
</tr>
<tr>
<td>5</td>
<td>4.35</td>
</tr>
<tr>
<td>6</td>
<td>5.35</td>
</tr>
<tr>
<td>7</td>
<td>6.35</td>
</tr>
<tr>
<td>8</td>
<td>7.34</td>
</tr>
<tr>
<td>9</td>
<td>8.34</td>
</tr>
<tr>
<td>10</td>
<td>9.34</td>
</tr>
<tr>
<td>11</td>
<td>10.34</td>
</tr>
</tbody>
</table>

[If FCN is $-\log(\text{likelihood})$ instead of chisquare, all values of UP should be divided by 2.]
2. TESTING OF HYPOTHESES.

In this chapter, we consider choosing between two or more well-defined hypotheses, for example measuring the parity of an elementary particle, which can only be positive or negative.

2.1 THE TEST STATISTIC AND OTHER DEFINITIONS.

In order to discuss the theory of hypothesis testing, a few basic concepts need to be defined, with the corresponding notation. The notation, although rather arbitrary, is fairly well standardized throughout the statistical literature, and we adopt here the most commonly used symbols.

The two hypotheses under consideration may have free parameters whose values must also be estimated (composite hypotheses), but we assume for the moment that they are completely defined (simple hypotheses). The simplest or most important of the two hypotheses is denoted $H_0$ and called the null hypothesis, probably because it often corresponds to "no effect" or "zero dependence". The other hypothesis, $H_1$, is called the alternative hypothesis. Our aim is to choose between the two hypotheses, based on some observations, and especially to be able to express the significance of those observations in distinguishing between the two hypotheses.

We denote the space of all observations by $W$. This includes all possible outcomes of our experiment. We will want to divide this space into two regions: the critical region, $\omega$ is the set of all observations for which we would reject $H_0$, and the acceptance region, $W-\omega$, for which we would accept $H_0$. This will be done with the help of a function $X(W)$ called the test statistic. The study of hypothesis-testing is thus reduced to the study of the properties of different test statistics and the resulting regions of rejection and acceptance.

We can now define the important quantities in terms of probabilities of obtaining different values of the test statistic:

1. The level of significance of the test, $\alpha$, is the probability of $X$ falling in the critical region $\omega$ when $H_0$ is in fact true:

$$ P(X \in \omega | H_0) = \alpha $$

2. The power of the test, $1-\beta$, is the probability of $X$ falling in the critical region when $H_1$ is true:

$$ P(X \in \omega | H_1) = 1-\beta $$

or, alternatively:

$$ P(X \in (W-\omega) | H_1) = \beta $$
3. The error of the first kind, or loss, occurs when $H_0$ is rejected even though it is true. The probability for this happening is clearly $\alpha$.

4. The error of the second kind, or contamination, occurs when $H_0$ is accepted even though the alternative $H_1$ is true. The probability for this is from the above definitions $\beta$.

A good statistical procedure for testing hypotheses will therefore be one which chooses the function $X$ and the region $\omega$ so as to minimize both $\alpha$ and $\beta$.

2.1.1 Example: Separation of two types of events.

As an example let us suppose we are studying elastic proton-proton scattering, and we have a sample of events which contains both true elastic events and those with additional $\pi^0$ production. On the basis of the measured momenta of the protons alone, we have to decide, for each event, whether an unseen $\pi^0$ was also produced, in order to obtain finally the biggest and purest possible sample of elastic events. For each event, the two hypotheses are as follows:

$$H_0: \quad p+p \rightarrow p+p$$
$$H_1: \quad p+p \rightarrow p+p+\pi^0$$

We could choose as test statistic the missing mass for the event, a function of the measured momenta which, if the momenta were measured exactly, would be equal to the $\pi^0$ mass for unwanted events and zero for elastic events. Because of measurement errors, the expected distributions of missing mass under the two hypotheses will have some width, and in practice may appear as in Figure 9. It is clearly not possible to choose $\omega$ so as to make both $\alpha$ and $\beta$ zero at the same time, although either one could be made arbitrarily small at the expense of increasing the other. Physically, this means that if we set the acceptance level so as to lose very few true elastic events, we will also have to accept a large number of background events; if, on the other hand, we require a very pure sample of elastic events, we will have to settle for a big loss of number of events. Some compromise is necessary, based on the physics to be done. In order to reduce both $\alpha$ and $\beta$ simultaneously, we would have to find a better test statistic. For example, if it is known that the $\pi^0$ are usually very fast, it may be better to use missing energy or missing momentum to discriminate between the two types of events. The choice of test will be discussed in the next subsection.
2.2 CHOOSING A TEST.

The properties of a test between two simple hypotheses can be seen from the diagram of Figure 10, which shows $\beta$ as a function of $\alpha$ for four different tests. Such a curve for any admissible test must lie below the diagonal indicated, since for any point above the diagonal the probability of accepting the wrong hypothesis would be greater than the probability of accepting the right one, and surely it must be possible to do better than that.
Figure 10. Basic properties of four different tests.

The best test will correspond to the lowest (α, β) curve, since that will have the lowest value of β for a given value of α. Thus test one is always worse than the other three, test two is sometimes better than test three (namely if we are interested in small values of β rather than small values of α), and test four is always better than the other three.

If the test is being applied in order to select data for later analysis, additional criteria may be important, namely that the selection procedure not introduce a bias in the distributions of interest for the data selected, or at least that the bias so introduced be calculable.

2.3 THE NEYMAN-PEARSON TEST.

In the case of completely defined simple hypotheses, there is one test which can be shown to be always the best, in the sense of giving always the smallest β for a given α. This test, the Neyman-Pearson test, may be very complex computationally, especially when the space of observables is many-dimensional, but can always be defined in principle.
For a given value of the significance level $\alpha$, the most powerful test will be that one which has the best critical region in the space of the observations. That is, among all those regions $\omega(\alpha)$ which satisfy

$$\int_{\omega(\alpha)} f(X|H_0) \, dX = \alpha$$

we wish to find that region which minimizes $\beta$, or maximizes the power

$$1 - \beta = \int_{\omega(\alpha)} f(X|H_1) \, dX$$

$$= \int_{\omega} \frac{f(X|H_1)}{f(X|H_0)} f(X|H_0) \, dX$$

But the last expression above is just the expectation of the likelihood ratio for hypothesis one divided by hypothesis zero, assuming hypothesis zero to be true. This means that the best critical region is such that this likelihood ratio is larger for all points inside $\omega$ than outside $\omega$, with the constraint that the "size" of the region is $\alpha$.

In the general case, the observable space of $X$ is many-dimensional and the Neyman-Pearson test consists of finding the hypercontour of constant likelihood ratio which divides this space into two parts such that the part corresponding to larger values of the likelihood ratio has an integrated probability (test size) of $\alpha$. This can be a very lengthy calculation. The usual simplification is to consider not the whole observable space $X$, but a one-dimensional test statistic $t(X)$ as introduced earlier. Then the test is no longer Neyman-Pearson and not necessarily optimal, depending on how good a choice of test statistic was made.

2.4 COMPOSITE HYPOTHESES.

In practice, real experiments often give rise to situations where both parameter estimation and hypothesis testing must be performed simultaneously. That is, the hypotheses to be tested are not completely defined in advance, but have free parameters whose values must be estimated. Such a hypothesis is called a composite hypothesis. The mathematical theory of composite hypothesis testing is not as well developed as that of simple hypothesis testing. The general techniques which can be found are valid only asymptotically (and the asymptotic limit may be very high!) or for certain related "families" of hypotheses. In many real cases, the only way to obtain realistic confidence levels is by resorting to Monte Carlo simulation.
2.4.1 The maximum likelihood ratio.

The most important general tool for studying composite hypotheses is again based on the likelihood function. It is the maximum likelihood ratio, defined as the ratio of the likelihood functions for the two hypotheses, each one maximized with respect to all the free parameters. If there were no free parameters, this would correspond to the Neyman-Pearson test, which is optimal and calculable. Unfortunately, the effect of the free parameters makes the test not necessarily optimal, and in general one does not even know how to calculate \( \alpha \) exactly.

Let us therefore restrict ourselves to the relatively common case where \( H_0 \) and \( H_1 \) have the same free parameters except that \( H_1 \) has some parameters free whose values are fixed for \( H_0 \). For example:

\[
\begin{align*}
H_0: & \quad \theta_1 \text{ free, } \theta_2 = c \\
H_1: & \quad \theta_1 \text{ free, } \theta_2 \text{ free}
\end{align*}
\]

The maximum likelihood ratio is:

\[
\lambda = \frac{\max f(X|\theta_1, \theta_2 = c)}{\max f(X|\theta_1, \theta_2)}
\]

where \( X \) is all the data, and the maximization is with respect to \( \theta_1 \) in the numerator and with respect to both \( \theta_1 \) and \( \theta_2 \) in the denominator. Since the denominator is maximized over the same parameter as the numerator and in addition one more, clearly \( 0 \leq \lambda \leq 1 \). The difficulty in knowing the expected distribution of \( \lambda \) arises from the fact that since the value of \( \theta_2 \) will in general be different for the two hypotheses, the value of \( \theta_1 \) will also be different due to the maximization, and the effect of the maximization will depend on the correlation between \( \theta_1 \) and \( \theta_2 \).

Asymptotically (for large amounts of data, or small errors on \( \theta \)) the distribution of \( \lambda \) is known, so this is normally used in evaluating its significance. Namely, the quantity \( -2 \ln \lambda \) is distributed as \( \chi^2(r) \) where there are \( r \) more free parameters in the denominator of \( \lambda \) than in the numerator. Thus one evaluates the maximum likelihood ratio for the data, and looks up the value of \( -2 \ln \lambda \) in a table of chi-square with \( r \) degrees of freedom, which gives the significance level \( \alpha \). A small value of \( \alpha \) is evidence against the more restrictive hypothesis \( H_0 \) and in favor of the more general hypothesis \( H_1 \).

2.5 Small-Sample Problems.

The maximum likelihood ratio test and significance level calculation as defined above is widely used since it is nearly the only real statistical tool for this situation. It is however notoriously unreliable for small samples of data. Unfortunately, it is hard to know exactly when a data sample is small, since it depends strongly on the hypotheses and distributions involved in ways which are not obvious.
A famous example of this problem is the attempt to determine whether a peak in a distribution of effective mass is a single peak or a split peak. In several examples it has been crucial to establish the significance of the evidence for split peaks because simple peaks would correspond to elementary particle states easily accommodated by the existing quark theory whereas split peaks would have implied resonances with exotic properties requiring a quite different approach to the whole theory. The importance of the problem triggered extensive statistical studies which revealed a marked tendency for data generated with a simple-peak hypothesis to fit the split-peak hypothesis better than the simple-peak model. This bias in the testing procedure can only be evaluated by Monte Carlo simulation, drawing samples from a known hypothesis and fitting them in the same way as the experimental data.

3. GOODNESS-OF-FIT.

In this chapter we consider the significance with which we can accept or reject a hypothesis, without specifying any alternative hypotheses. As for hypothesis testing, we will have a critical region \( \omega \) such that for all data \( X \in \omega \) we will reject the null hypothesis. As before, we can find the probability \( \alpha \) of rejecting \( H_0 \) when it is true:

\[
\int_{\omega(\alpha)} f(x|H_0) \, dx = \alpha
\]

However we can no longer evaluate the probability of accepting \( H_0 \) when it is false, since this would depend on the alternative hypothesis which is not specified. We have therefore a measure of the significance of evidence against \( H_0 \) (if the fit is bad) but the significance of the evidence in favor of \( H_0 \) coming from a good fit cannot be quantified.

3.1 CHOOSING A GOODNESS-OF-FIT TEST

Since there is no alternative hypothesis, we cannot calculate \( \beta \) and cannot know the power of a test. This in principle deprives us of any means of comparing goodness-of-fit tests since a test is better than another only if it is more powerful at rejecting unwanted hypotheses. Since the unwanted hypothesis is not specified, it appears that we cannot have a realistic basis for choosing a test. Indeed we can expect some tests to be sensitive to certain kinds of deviations from the null hypothesis, and other tests to be sensitive to others. With this in mind, there are at least two approaches which will help in choosing a test.
The first approach is rather intuitive, but can be made more rigorous with some use of statistical information theory (which we do not discuss explicitly in these notes). The idea is to make sure somehow that the test makes use of all the information relative to the hypothesis being tested, and that it does not have any arbitrary procedures or parameters which would affect the value of a independently of the data and the hypothesis. We will see examples of how this is used below.

The second approach is to invoke a class of alternative hypotheses which allow us to estimate the power without being too specific about the alternatives. For example one can define the local power of a test as its power against infinitesimal deviations from the null hypothesis. Still this is not as general as it may seem since even infinitesimal deviations may take different forms, but it is usually possible in this way to get a good comparison of tests under rather general conditions.

3.2 DISTRIBUTION-FREE TESTS.

The calculation of the confidence level for a given test involves an integration of a probability density function over a region which may be many-dimensional. Although such numerical calculations may not be too difficult when performed on modern computers, it has traditionally been necessary, and even today is still desirable, to avoid this calculation by means of a distribution-free test. Such a test involves a test statistic $t(X)$ whose distribution is known (under the null hypothesis) independent of the distribution of the $X$. Many distribution-free tests have been found, and the appropriate distributions of their test statistics are either known analytically or, more frequently, tabulated so that the user can simply calculate $t$ and read the corresponding significance level from a table.\footnote{Traditionally all tests were distribution-free, so the term was not even used, it being assumed. Very recently we see a new kind of test being used (see 3.5), in which the confidence level must be recomputed for each case, something that would not have been thinkable without modern computers.}

A well-known example of a distribution-free test is the chi-square test of goodness-of-fit of a probability density $g(x)$ to density estimated experimentally in the form of a histogram. If the number of events observed in the $i$th histogram bin is $n_i$, the value of $x$ in the middle of the bin is $x_i$, and the density $g(x)$ is normalized to the total number of events observed, then the test statistic for goodness-of-fit is:

$$t = \sum_i \frac{(g(x_i) - n_i)^2}{n_i}$$
This is often called the chisquare function, because under the null hypothesis (that the histogram really comes from \( g(x) \)), the quantity \( t \) should be distributed like a \( \chi^2 \) variable with \( n \) degrees of freedom if the number of bins is \( n \). This is distribution-free since the expected distribution of \( t \) does not depend on \( g(x) \). If we fit another function with different data we would still use the same table of \( \chi^2 \) to obtain the level of significance \( \alpha \).

3.3 COMPARING TWO ONE-DIMENSIONAL DISTRIBUTIONS.

In this section we will show what criteria may be used to choose a goodness-of-fit test by comparing two such tests for compatibility of one-dimensional distributions. Both tests, the chisquare test mentioned above and the Kolmogorov test, may be used either to compare two experimental samples of events or to compare an experimental sample with a completely-defined theoretical probability density.

3.3.1 The Chisquare test.

This test is defined just above for the fit of a one-dimensional experimental sample to a known curve \( g(x) \). It is easily generalised to compare two experimental distributions:

\[
t = \sum_{i} \frac{(m_i - n_i)^2}{m_i + n_i}
\]

We notice that this test requires the grouping of observations into bins (called by statisticians 'data classes') and does not prescribe exactly how this is to be done. We are in principle free to make as many bins as we want, and place the boundaries where we want. This arbitrariness, which clearly will affect the value of \( t \) and probably also its properties as a test statistic, is one of the undesirable features pointed out above.

In view of the popularity of the chisquare test, the subject of optimal binning has been the object of considerable study. We summarize here some results of these studies:

1. In order to increase the local power of the test, there should be as many bins as possible.

2. The upper limit on the number of bins comes from the requirement that \( t \) follow a \( \chi^2 \) distribution under the null hypothesis, which is only true for a 'large' number of events per bin, where the Poisson distribution becomes approximately Gaussian. Opinions vary as to how many events are needed for this, but most studies indicate that there should be very few bins of less than ten events and no bins of less than five events.
3. The distribution of bin boundaries is usually chosen equally spaced for practical reasons, but all studies indicate that it is better statistically if bins are chosen to be equally probable, that is, approximately the same number of events should fall in each bin.

4. Additional experiment-dependent considerations, such as the accuracy with which $x$ can be measured, may be important.

Another apparent source of arbitrariness is the exponent 2 in the expression for $t$. Any other positive non-zero exponent would give rise to an admissible test, although the expected distributions for $x^3$, $x^4$, etc. would have to be recalculated. In fact it can be shown using the theory of information that $x^2$ is indeed optimal whenever the deviations in each bin are Gaussian (here they are in fact Poisson which is approximately Gaussian). The 'square' in 'chisquare' is therefore not at all arbitrary, and is optimal for the usual case. One could, however, imagine cases where the measurements were not Gaussian-distributed and where a different test statistic would be better.

3.3.2 The Kolmogorov test.

We now turn to a somewhat different kind of test, based on what the statisticians call the order statistics, which are nothing but the experimental observations ordered by increasing $x$-value. This allows us to form the cumulative distribution of the data $S(x)$ as follows: The distribution starts at zero for $x=-\infty$, and increases by an amount $1/N$ at each point $x$ where an experimental point $x_i$ has been observed. [N is the total number of points observed, so $S(\infty)=1$.] We can use the Kolmogorov test either to compare two experimental distributions $S_1(x)$ and $S_2(x)$, containing respectively $N_1$ and $N_2$ events; or to compare $S_1(x)$ with a continuous known distribution of which the probability density function is $f(x)$, and whose integral, the cumulative distribution function is $F(x)$.

The Kolmogorov test statistic is a measure of the distance between the two distributions being compared. This measure is simply the largest distance (maximized with respect to $x$) between the two cumulative distributions:

$$D = \sqrt{N} \max |S(x) - F(x)|$$

where $S$ has $N$ events,

or

$$D = \sqrt{(N_1N_2/\max(N_1+N_2)) \max |S_1(x) - S_2(x)|}$$

It turns out that this test statistic is asymptotically distribution-free (that is, under the null hypothesis, the expected distribution of $D$ is independent of $S$ and $F$ for large enough $N$.), and as written here it is correctly normalized to be also asymptotically independent of $N_1$, $N_2$, and $N_2$. The significance level $\alpha$ can be calculated from formulas or the table given in Eadie et al., page 270, or calculated by the CERN Program Library subroutine PROBKL.
Notice that this test involves no arbitrary binning, and is still very easy to calculate, although for very large data samples the ordering of the data may be longer than the histogramming required for the chisquare test.

3.3.3 The Smirnov-Cramer-Von Mises test.

This test is very similar to the Kolmogorov test described above, except that the measure of the distance between two cumulative distributions is taken to be the integrated squared distance instead of the maximum:

\[ W = N \int_{-\infty}^{+\infty} [S(x)-F(x)]^2 f(x) \, dx \]

\[ = N \int_{0}^{1} [S(x)-F(x)]^2 \, dF(x) \]

The corresponding formula for comparing two experimental distributions is somewhat more complicated and is given in Eadie et al, page 269, as are formulas and tables for determining the significance level \( \alpha \).

The computational complexity of this test is clearly somewhat greater than for the Kolmogorov test, which probably explains why it is less popular. However it has the distinct advantage that the test is exactly distribution-free for all values of \( N \), although the significance level is independent of \( N \) only for 'large' \( N \) (in fact \( N \geq 3 \) is enough!). It is also more appealing (and probably more powerful against most alternatives) because it is really a function of all the data and not just the maximum distance.

Because it is free of binning, is sensitive to all the data, and is exactly distribution-free, the Smirnov-Cramer-Von Mises test is generally considered the most powerful goodness-of-fit test for one-dimensional data.

3.4 Comparing Multidimensional Distributions.

When the data are more than one-dimensional, goodness-of-fit testing becomes considerably more difficult.

3.4.1 Chisquare in \( d \) dimensions.

In principle, the chisquare test for goodness-of-fit is dimension-free, in the sense that it can be defined independently of the dimensionality of the data. One merely compares the expected number of events in each data class (bin) with the actual number. The dimensionality of the bin does not enter into the theory.

- 33 -
In practice, however, multidimensional bins cause not only computational problems (when the boundary of the space is curved) but the number of events required to have a minimum number per bin becomes enormous, unless the number of bins is reduced to a very small number per dimension. The reason is that the number of bins increases exponentially with dimensionality. It is therefore clear that for multidimensional data we should prefer a test which does not require binning.

3.4.2 Kolmogorov-type tests in \( d \) dimensions.

It is very appealing to try to extend tests based on order statistics to higher dimensionalities, since these tests (Kolmogorov, Smirnov-Cramer-Von Mises) use the data points as measured without binning. Such attempts however meet with several difficulties, both practical and theoretical, and such tests have not to my knowledge yet been used with success. The first difficulty is with the order statistics themselves, which lose some of their nice properties in higher dimensions, although they can still be defined in a straightforward way:

\[ S(X_1, X_2, \ldots, X_n) = \text{number of points } p \text{ such that } p_1 < X_1, p_2 < X_2, \ldots, p_n < X_n \]

For example, they now depend on the orientation of the axes in the \( d \)-dimensional space. Another difficulty is the computational complexity, since the definition of \( F \) now requires multidimensional integration, and other multidimensional difficulties arise in defining the test statistic. Perhaps the most important barrier is that straightforward extension of both the tests given above are no longer distribution-free in many dimensions, so that one has no easy way to determine \( \alpha \).

If we are willing to give up distribution-free testing, the more recent permutation methods given in the next section are promising for both one- and many-dimensions, and for both large and small sample sizes.

3.5 Permutation tests for comparing two point sets.

We wish to test the hypothesis that two sets of points are random samples from the same underlying distribution, where we do not know the underlying distribution. We wish to find a test valid for one or more-dimensional points, and the test should not involve binning. Two things are needed:

1. A test statistic \( \Delta \) which will be a measure of the distance between the point sets.
2. A way to measure the significance of the value of \( \Delta \) for the two point sets, giving the significance level \( \alpha \).
The permutation method allows us to measure the level of significance for any distance function \( d \), without knowing the underlying distribution, assuming of course the null hypothesis. One first calculates the distance \( d_{12} \) between point sets one and two. Then the significance level is found as follows. Put both samples together to form a single sample of \( N_1+N_2 \) points. Under the null hypothesis, this will also be distributed like sample 1 or sample 2, and any (random) partitioning of this pooled sample into samples of \( N_1 \) and \( N_2 \) points should yield sets with the same distribution. We therefore make many different partitionings, each time choosing \( N_1 \) points from the total \( N_1+N_2 \) points, and calculate \( d \) for each of these partitionings. If \( N_1 \) and \( N_2 \) are small enough, we can actually enumerate all the partitionings possible. Whether all partitionings are exhausted or only a random sample is used, the resulting distribution of the values of \( d \) can be used in an obvious way to determine the significance of the actual value \( d_{12} \) corresponding to the observed data points. The significance level \( \alpha \) is simply the proportion of all \( d \)-values lying above the value \( d_{12} \).

Now that the significance level can be determined for any distance measure \( d \), we return to the question of defining such a distance. Since the procedure outlined above is valid for any measure \( d \), we are free to choose a measure with physical significance for the situation at hand. For example, suppose we have treated a sample of mice in a certain way, and we wish to know if those mice are bigger than the ones in a sample which was not treated. We could take as the distance measure the average weight or length of the treated mice minus the average of the untreated. If the treatment was expected to affect only half the treated mice, we could take the difference between the average weight of the heaviest treated mice and the heaviest untreated mice, etc.

In spite of this extraordinary freedom in choosing a test statistic, it is still difficult to find good distance measures for multivariate data. A general procedure is given in Friedman (1974) using the \( k \)-nearest-neighbor concept. This is probably the best general method in use, although it suffers from at least two elements of arbitrariness:

1. The optimal number of nearest neighbors \( k \) is not generally known.

2. It is somewhat dependent on the metric of the space in order to find the nearest neighbors. However, it is hard to imagine any measure of the distance between two point sets which does not depend on the definition of the distance between two points.

5 The total number of partitions of \( N_1+N_2 \) points into two samples of \( N_1 \) and \( N_2 \) points is:

\[
P_{12} = \frac{(N_1+N_2)!}{N_1!N_2!}
\]

This increases very fast with \( N_1 \) and \( N_2 \); for example, the number of ways of dividing a sample of 10 into two samples of 5 is only 252, but the number of ways of dividing a sample of 20 into two samples of 10 is 184756.
If a good distance measure $\Delta$ is available for the physical problem at hand, the permutation technique gives a good way to evaluate its significance. The method is completely non-parametric, does not involve any attempt to actually estimate the point densities in either set, and is always exactly valid for the data sample at hand, no matter how large or small it is. On the other hand, the computation required is large by traditional standards, since one is in fact recalculating $\Delta$ each time instead of using a distribution-free test and a standard table. By modern computing standards however, this is a small price to pay for the advantages gained.

4. CONCLUSIONS.

We have seen how to define and evaluate the statistical significance of experimental data in the three different contexts of parameter estimation, hypothesis testing, and goodness-of-fit testing. In a logical development of the subject it is easy to keep these different contexts distinct and avoid confusion, but in solving real problems it may not be so clear exactly what question is being asked. This is partly because one is often asking several questions at once, partly because the different techniques are, after all, related, and partly because the same functions of the data are used -- in different ways -- to answer different questions. For this reason we shall conclude these lectures by pointing out the relationships, both similarities and differences, between the various problems and methods of solution.

4.1 CHISQUARE.

The ubiquitous chisquare function of Pearson is the source of much confusion because it can arise in many contexts. It is a measure of the distance between a set of observations and a model, namely the sum of squares of deviations of the observations from the model, each deviation normalized by its standard deviation. The model may contain unknown parameters, in which case the chisquare function may be used to estimate these parameters and their uncertainties. In this case, the chisquare function is considered as a function of these free parameters (although it is also of course a function of the data) and minimized with respect to these parameters. Parameter uncertainties are estimated by finding the change in parameter value required to produce a given change in chisquare, as described in these lectures. The actual value of chisquare is not used in this case, since it is assumed that the model is correct.

The same function can however be used to test goodness-of-fit. In this case there are no free parameters and chisquare is considered as a function of the data. Now only the value of chisquare is used, and compared with a table of values to find the confidence level $\alpha$. To confuse things further, the table is also referred to as a chisquare table, and it gives the integral of a function also called the chisquare function, integrated from the value obtained experimentally to infinity, thereby
giving the probability, under the null hypothesis, of obtaining a value of chisquare greater than that actually found.

4.2 LIKELIHOOD.

The likelihood function can also be used in different ways, to do parameter estimation or hypothesis testing. It cannot however be used effectively for goodness-of-fit testing, essentially because the actual value of the likelihood function has no statistical interpretation. Only differences in log-likelihood are meaningful. We have seen their interpretation in parameter and interval estimation. The interpretation in hypothesis-testing is really the same, since here we are concerned with the likelihood ratio between two hypotheses, and the logarithm of this ratio is of course the difference in the logarithms of the two likelihood values corresponding to the two hypotheses.

Therefore, using the likelihood function to determine the significance of the statement that the parameter \( \theta = \theta_0 + \sigma \) is exactly equivalent to using the likelihood ratio to test hypothesis \( H_0: \theta = \theta_0 \) against \( H_1: \theta = \theta_0 + \sigma \).
REFERENCES

Clopper, C.J. and Pearson, E.S., Biometrika 2 (1934) 404

Eadie, W.T., Drijard, D., James, F.E., Roos, M., and Sadoulet, B.
Statistical Methods in Experimental Physics. North-Holland,
Amsterdam (1971)

Friedman, J.H., Data Analysis Techniques for High Energy Particle
Physics, Lectures presented at the CERN School of Computing,

James, F., and Roos, M., MINUIT - A system for function minimization and
analysis of the parameter errors and correlations. Comp. Phys.
Comm., 10 (1975) 343-367.

James, F. and Roos, M., Errors on ratios of small numbers of events,
Nuclear Physics B172 (1980) 475-480

James, F., Interpretation of the shape of the likelihood function around