

PRINCIPLES OF DATA ANALYSIS

Supplementary Material for Courses in Physics Experiments

F.C. CHEN

Physics Department

The Chinese University of Hong Kong

September 1969

CONTENTS

FOREWORD	page ii
I INTRODUCTION	1
1.1 The Interpretation of Physical Measurements	
1.2 The Graphical Method	
1.3 Experimental Errors	
II RANDOM ERRORS	6
2.1 The Distribution of Random Errors*	
2.2 The Dispersion Indices	
2.3 The Probable Error	
2.4 The Error of an Error	
2.5 How to Round Off Numbers	
III THE PROPAGATION OF ERRORS	11
3.1 The Functions of Observables	
3.2 The Basic Formula for Error Propagation*	
3.3 Some Useful Corollaries	
IV CURVE FITTING	15
4.1 The Method of Least-Squares	
4.2 The Linear Least-Squares Fit	
4.3 Errors in the Linear Least-Squares Fit	
V THE POISSON DISTRIBUTION	19
5.1 Characteristics of the Distribution Function	
5.2 Applications to Particle Counting	
NOTES	25
APPENDIX THE METHOD OF MAXIMUM LIKELIHOOD	28
SUMMARY OF FORMULAE	34

FOREWORD

In preparing these notes for you, students taking courses in physics experiments, we have kept two objects in mind : to give a brief but self-contained account of the basic ideas and principles underlying the statistical analysis of experimental data, and to discuss, often without proof, the most important formulae used in data analysis. Because of the limitation in space, we have no choice but to emphasize practical applications, keeping all the explanations terse and relegating the more involved arguments and manipulations to footnotes and the Appendix. With some examples given in class instructions, you should thus be able to learn and apply the formulae summarized at the end of these Notes in a very short time. However, we do hope that you will resist the strong temptation of blindly applying these formulae, and perhaps may even be induced to do some further reading (see References) to fill up the many gaps that we leave between assumptions and results, and to venture into unexplored areas, such as the t -, F - and χ^2 - tests, the correlation of variables, etc.

The mathematics required has been kept to a minimum : rudiments of classical probability theory, simple algebra, and elementary calculus, nearly all of which you should have mastered in high school. In particular, the first three Chapters without the asterisked sections form the core material which even First Year students can go through and make use of without difficulty. The rest may take some more maturity to appreciate and hence be delayed until the Second Year. In any case, none of the things discussed in these Notes need be memorized or would ever appear as examination questions. The only thing important is to know how to use them intelligently.

It is a pleasure to thank Dr. N.N. Chan and Dr. Martin H.S. Lau for reading parts of the manuscript and for many interesting and illuminating criticisms, suggestions and discussions. We also thank Miss C. Lee for her careful and patient typing of the manuscript.

F.C.C.

September 1969.

REFERENCES

1. Lyman G. Parratt : "Probability and Experimental Errors in Science"
(John Wiley and Sons, New York, 1961)

This is a book with the experimental scientists in mind, and strikes a good balance between mathematics and physical considerations. Also contains numerous examples, anecdotes and problems. Highly recommended for further reading.

2. B.M. Shchigolev : "Mathematical Analysis of Observations"
(ILLIFFE Books Ltd., London, 1965)

3. E.B. Mode : "Elements of Probability and Statistics"
(Prentice-Hall, Inc., Englewood Cliffs, N.J., U.S.A., 1966)

Both are mainly concerned with mathematical statistics, and provide more rigorous and general treatments of the topics here covered.

I. INTRODUCTION

1.1 The Interpretation of Physical Measurements

Suppose one makes n separate measurements of a physical quantity x , with results x_1, x_2, \dots, x_n , which are usually not all the same. Then two questions would immediately arise: what is the "best result" for x and "how good" is this result?

There are several possibilities in choosing the best x :

- (A) The MEAN VALUE $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$,
- (B) The MEDIAN, which is the "middle" value for an odd number of measurements ordered in magnitude and the interpolated middle value otherwise, and
- (C) The MODE, (the most probable value), which is simply the most frequently measured value.

These three answers (which are known as LOCATION INDICES) are usually (but not always) the same. For physicists the mean value \bar{x} is the most commonly used, because it is said to have the "highest statistical efficiency".

The "goodness" of the best result, (which is henceforth taken to be \bar{x}) can be gauged in two ways:

- (A) ACCURACY, which reflects how close the result \bar{x} is to the hypothetical true value x_T , and
- (B) PRECISION, which reflects the uncertainty we have in assigning \bar{x} as the best result.

A quantitative estimate of the goodness of the result, with both of these criteria taken into account, is known as the EXPERIMENTAL ERROR.

As we will show, a careful determination of the error of \bar{x} may be quite involved. However, it is very simple to make a quick, rough estimate: the scale uncertainty of the measuring instrument (e.g. ± 0.05 cm for a meter stick calibrated to 0.1 cm) reflects the accuracy and the "range" in x_i (which is the difference between the largest and the smallest measured values) divided by n reflects the precision. The larger of the two is the error in x . In other words,

$$\delta\bar{x} \sim \max \left\{ (x_{\max} - x_{\min})/n, \text{scale uncertainty} \right\}, \quad (1.1)$$

and

$$\text{Best result} = \bar{x} \pm \delta\bar{x}. \quad (1.2)$$

1.2 The Graphical Method

When measurements are made with some parameters being varied at the same time, one may either

- (A) infer the physical quantity of interest from each measurement and then take the average, or
- (B) plot all the results on a graph (most commonly a straight line) and then infer the desired result from the characteristics (e.g. the slope or intercept) of the best fitted curve.

While most of you may be used to the more direct method (A), method (B) is usually much more superior. It should be used whenever possible.

Let us give a few simple examples :

- (A) Suppose the period τ_i of a simple pendulum is measured, with the length set at different values l_i . To obtain the gravitational acceleration g from these measurements one can of course find g_i from

$$g_i = \left(\frac{2\pi}{\tau_i}\right)^2 l_i, \quad (i = 1, 2, \dots, N)$$

and take the average $\bar{g} = \frac{1}{N} \sum g_i$. A far better way, however, is to plot

$y = (2\pi/\tau)^2$ against $x = 1/l$, so that $y = gx$. Each determination gives one point on the graph, and the slope of the best straight line going through all the points is g .

- (B) One wants to find the EMF ϵ and internal resistance R_i of a battery by connecting it to a variable external resistance R and recording the resultant current at different settings of R . If we set $y = R$ and $x = \frac{1}{i}$ then

$$y = \epsilon x - R_i$$

so that the slope of the fitted straight line is the EMF and the negative intercept the internal resistance.

- (C) Yet another example is the exponential decay in time of any quantity :

$$Q = Q_0 e^{-\frac{t}{\tau}}$$

in which Q is measured from time to time. Plotting $y = \ln Q$ against the time t yields a straight line

$$y = -\frac{1}{\tau} t + \ln Q_0,$$

where the negative inverse of the slope is the decay time τ and the intercept is $\ln Q_0$.

We must delay until chapter 4 before venturing to touch upon the many important questions connected with the graphical method, e.g. why is it "good" ? What is the precise mathematical criterion and procedure to fit "the best" curve ? What are the errors in such a fit ? etc. For the moment we presume that only straight lines are encountered and visual inspections suffice to give the best fit. As for errors in the slope and intercept, they can be visually estimated by varying the fit to see the "tolerances" in these quantities.

1.3 Experimental Errors

There are many reasons why non-trivial physical measurements are never exact : the instruments used are not completely accurate or precise, fluctuations in room temperature, pressure or air current cause unpredictable changes, the quantity being measured is not exactly defined, etc. With all of these taken into account as best as possible, the EXPERIMENTAL ERROR is an estimated range of value centered at the experimental result such that we are confident the "true" physical result probably falls within the range. (Fig.1)

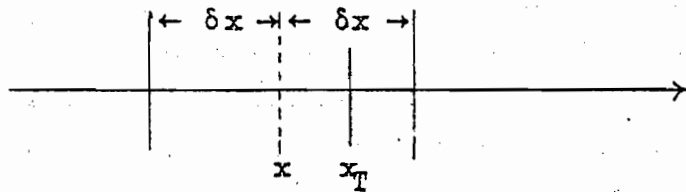


Fig. 1 x_T is the "true" result, x the experimental result and δx the experimental error.

We usually consider the error of a measurement from three different aspects, which, however, are quite closely related in most actual situations:

(A) The INTRINSIC ERROR δ_I

is caused either by the limit of the precision of the instruments used or by the lack of precise definition of the quantity measured. For instance, the thickness of a piece of crystal cannot be determined to better than the smallest division on the measuring micrometer, and, even if an extremely sensitive instrument is used, still cannot be measured to within, say, 1\AA , because it is not defined down to intermolecular distance.

It is common to set intrinsic error arising from instrumental imprecision at half of the smallest division on the calibrated scale (the SCALE UNCERTAINTY). There is a lot of flexibility to this, however : you may cautiously reduce it by visual interpolation, or you may want to increase it when using the instrument (ESPECIALLY ELECTRICAL INSTRUMENTS : METERS,

OSCILLOSCOPES, BRIDGES, etc.) on very sensitive ranges. Intrinsic errors due to the inherent nature of things are usually negligible in comparison. They must be given separate individual physical consideration otherwise.

(B) The SYSTEMATIC ERROR δ_S

reflects the estimate of necessary corrections (to the result) which have not been actually made. Among the many sources of systematic errors the following are most common :

- (a) Persistent HUMAN BIAS, such as parallax in reading a meter or delay in the reflex action of pushing a button, etc.,
- (b) Long term DRIFTS in experimental conditions, such as the gradual rise of room temperature, change in line voltage, wearing down of the moving parts of an instrument, etc.,
- (c) Faults in the CALIBRATION, ALIGNMENT or even CONSTRUCTION of instruments, and
- (d) NEGLECTED PHYSICAL EFFECTS, such as the heat loss (or gain) to (or from) surroundings during an experiment in calorimetry, the terrestrial magnetic field which distorts the field in a solenoid, air resistance and friction which slows down a moving cart on an air-track, etc.

From this by no means exhaustive list it is evident that no general rules apply to systematic errors. In most cases it suffices to make order-of-magnitude estimates to show that they are negligible when compared to other errors. In more difficult situations (which you may well encounter) careful individual physical considerations are necessary for a good estimate of their magnitudes. The complete determination or elimination of them sometimes may even necessitate additional separate experiments.

(C) The RANDOM ERROR δ_R

is caused by a great many very small but untraceable FLUCTUATIONS, such as the minute and erratic changes in temperature, atmospheric pressure, cosmic ray strength, human reaction and judgment, etc., which constantly affect the measurement in an unpredictable way. That is why repeated measurements give slightly different results even under almost identical experimental conditions. Even though, by its very nature, the random error in any individual measurement defies analysis, yet the DISTRIBUTION OF RANDOM ERRORS is subject to the powerful method of statistical analysis.

Their average size can therefore be reasonably determined from the deviations of each result from the average, as will be discussed in the next chapter. The random error usually decreases as the number of measurements increases, but will not become much smaller than the intrinsic error if we have not overestimated the latter. After all, the uncertainty in visual interpolation is also a random effect.

How does one obtain the OVERALL EXPERIMENTAL ERROR δ_E from all these? Clearly both δ_I and δ_R have the same origin, and the larger of the two should reflect the random effect. The systematic error δ_S is something different, however, and must be ADDED ON VECTORIALLY unless negligibly small. In other words,

$$\delta_E^2 = [\max(\delta_I, \delta_R)]^2 + \delta_S^2,$$

or

$$\delta_E \sim \max(\delta_I, \delta_R). \quad (\delta_S \ll \max(\delta_I, \delta_R)) \quad (1.3)$$

Since δ_I cannot be reduced by repeated measurements it is clear that there is a lower limit to δ_E :

$$\delta_E \geq (\delta_I^2 + \delta_S^2)^{\frac{1}{2}}, \quad (1.4)$$

and one should only make an enough number of measurements to ensure that δ_R is well below δ_I . Any further measurement is then just a waste of effort.

Last but not least, it is OBVIOUS that errors have nothing to do with BLUNDERS (or mistakes), such as misreading meter scales, wrong setting of instruments, errors in transcription of data, etc. Once a blunder is discovered, the data should be immediately discarded without hesitation.

II. RANDOM ERRORS

2.1 The Distribution of Deviations

To analyze random errors, it is necessary to assume some knowledge of their origin. For this purpose we consider the following naive but plausible model : in any physical measurement a great number (n) of hypothetical "elementary errors" can each change the result x from its true value x_T by the same minute amount Δx , which has a fixed probability p (taken to be $\frac{1}{2}$) of being positive. The resultant deviation $\delta x = x - x_T$ is the algebraic sum of all such elementary errors :

$$\begin{aligned}\delta x &= k\Delta x + (n - k)(-\Delta x) \\ &= (2k - n)\Delta x,\end{aligned}\tag{2.1}$$

where k is the number of elementary errors that happen to be positive. In a number of measurements the probability of finding a particular k can be shown to be

$$B(k; n, p) = \frac{n!}{k!(n - k)!} p^k (1 - p)^{n - k},\tag{2.2}$$

which is the well-known BINOMIAL DISTRIBUTION. In the limit $n \rightarrow \infty$ it reduces to the NORMAL (or GAUSSIAN) DISTRIBUTION :

$$B \rightarrow G(k; \sigma', \bar{k}) = \frac{1}{\sqrt{2\pi} \sigma'} e^{-\frac{(k - \bar{k})^2}{2\sigma'^2}},\tag{2.3a}$$

where $\bar{k} = np$ and $\sigma' = \sqrt{np(1 - p)}$.

Since δx is related to k linearly, the distribution of the former can now be easily obtained from Eqs. (2.1) and (2.3a)^[1] :

$$G(\delta x, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(\delta x)^2}{2\sigma^2}},\tag{2.3}$$

where $\sigma = \sqrt{n}\Delta x$ can be shown to be the standard deviation (see next section) of the distribution. If $\Delta x \rightarrow 0$ and $n \rightarrow \infty$ (but σ remaining finite in

the limit) we may regard δx in Eq. (2.3) as a continuous variable. Since $x = x_T + \delta x$, Eq. (2.3) also gives the distribution of the measured values of x . As expected, this is a dumbbell-shaped curve centered at x_T (Fig. 2).

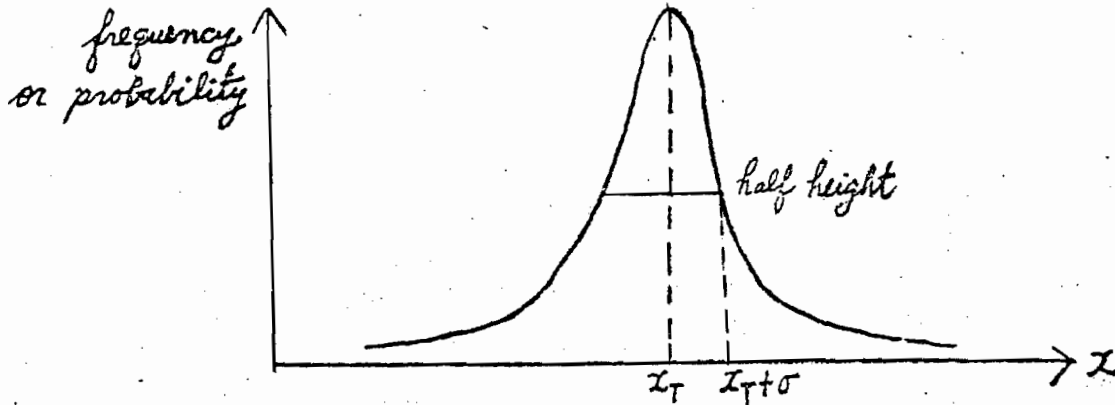


Fig. 2 The Normal Distribution Curve. The peak is at x_T , and the half-width at half height of peak is approximately σ .

By virtue of a certain very powerful CENTRAL LIMIT THEOREM, the normal distribution holds even if the elementary errors are not of the same size, or if n is not large but the sizes of Δx are normally distributed themselves. It is thus the result of a very general and plausible model, and, even though not exact, fits the ACTUAL DISTRIBUTIONS OF a wide class of experimental deviations fairly well^[2]. As will be seen, it is also a most CONVENIENT ASSUMPTION, which we adopt in most of our following discussions.

In the limit $n \rightarrow \infty$, $p \rightarrow 0$ but np remaining moderate (i.e. $k^2 + (np)^2 \ll n$), the binomial distribution reduces to the POISSON DISTRIBUTION, which gives the probability of k successes out of n individually unlikely events. This is an extremely important topic in modern experimental physics and will be taken up in Chapter V.

2.2 The Dispersion Indices

Without referring to the distribution of a set of data x_1, x_2, \dots, x_n (average value \bar{x}), we can use the various DISPERSION INDICES to indicate how they spread, e.g.

- (A) The RANGE (of. Section 1.1), i.e. $\max(x_i) - \min(x_i)$,

(B) The MEAN DEVIATION, $\overline{\Delta x} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$, and

(C) The STANDARD DEVIATION :

$$s = \sqrt{\overline{\Delta x^2}} = \left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{\frac{1}{2}} \quad (2.4)$$

(D) The SKEWNESS and KURTOSIS, which we do not discuss.

Among the dispersion indices, the standard deviation is said to be the most STATISTICALLY EFFICIENT^[3], which means that, for several sets of n measurements each, the standard deviations for different sets will agree among themselves much better than the range or the mean deviation. It is therefore the most commonly used dispersion index among physical scientists. Eq. (2.4) is, however, an underestimate of the standard deviation of x_i from the TRUE MEAN VALUE x_T , defined as $\lim_{n \rightarrow \infty} \bar{x}$. We therefore use the ESTIMATED PARENT STANDARD DEVIATION (or VARIANCE σ^2) :

$$\sigma = \left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{\frac{1}{2}}, \quad (2.5)$$

where the replacement of n by $(n-1)$ in the denominator reflects the fact that \bar{x} differs from x_T . If we assume a normal distribution for x_i , then Eq. (2.5) gives the MOST LIKELY ESTIMATE of the parent standard deviation^[4]

$$\sigma_p, \text{ defined as } \left[\frac{1}{n} \sum_{i=1}^n (x_i - x_T)^2 \right]^{\frac{1}{2}}.$$

2.3 The Probable Error

The PROBABLE ERROR δx is defined such that any further SINGLE measurement has a $\frac{1}{2}$ probability of falling within the range $(x - \delta x, x + \delta x)$. It thus has a direct and simple physical interpretation not shared by the dispersion indices.

How does one find δx ? For a Normal Distribution it is proportional to the stand deviation^[5] :

and

$$\begin{aligned} \delta x &= 0.6745 \sigma \\ &= 0.6745 \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \quad (2.6)$$

However, it is the mean value \bar{x} which is of the greatest interest. As will be shown in the next chapter and in the appendix as well, the PROBABLE ERROR OF THE MEAN $\delta \bar{x}$ is

$$\begin{aligned} \delta \bar{x} &= \delta x / \sqrt{n} \\ &= 0.6745 \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}} \end{aligned} \quad (2.7)$$

This is what we use as the random error of the mean value of a whole set of measurements. If another set of measurements are made to yield a new mean value \bar{x}' then there is a 50-50 chance \bar{x}' will be within the range $(\bar{x} - \delta \bar{x}, \bar{x} + \delta \bar{x})$.

The standard deviation of the mean can be defined in analogy to Eq. (2.7) and is also used sometimes.

2.4 The Error of An Error

In Section 2.2 we mentioned the spread of errors in connection with statistical efficiency. The question therefore arises : how "good" are the estimates given by Eqs (2.5) - (2.7) ? How are the deviations of errors from the "true error" distributed, and what is the error of an error ?

For normal distributions the answer is quite simple : the deviations of errors from their mean will tend to a normal distribution for a sufficiently large number of data, and the PROBABLE ERROR OF A PROBABLE ERROR is [6]

$$\delta(\delta x) = 0.6745 \delta x / \sqrt{2n}, \quad (2.8)$$

where n is the number of measurements. Obviously δx may be placed by $\delta \bar{x}$ on both sides of Eq. (2.8). The fractional probable error of δx is

$$\frac{\delta(\delta x)}{\delta x} = \frac{0.6745}{\sqrt{2n}} \quad (2.9)$$

For $n < 6$ the percentage error in δx is more than 20%. Consequently there is no sense in putting down δx with more than one significant figure. Two significant figures may be used if $n \geq 6$, but there is no rigid criterion. In any case, it is senseless to attach more than two significant figures to δx (or $\bar{\delta x}$) under any circumstances. For instance, a 5% precision in δx requires $n \sim 92$, but then the intrinsic error is most likely larger than the random error and it would also have an error of, say, 10%.

2.5 How to Round Off Numbers

In a physical statement 1.0 cm and 1.0000 cm are quite different things, because the former implies an implicit error of, say, ± 0.1 cm and the latter, ± 0.0001 cm. It is thus important to round off the number of significant figures consistently in all data. The following self-explanatory rules would be helpful :

- (A) Take all readings down to the last digit shown by the instrument. Add one (but never more) digit by visual interpolation if judged desirable.
- (B) The mean value should have the same number of digits as the individual readings, or at most one more, depending on the experimental error. This rule may not apply to particle counting rates (see Ch. V).
- (C) In any further calculations, carry all (or as many as practicable) the significant figures of the rounded off primary data, and for any constants (such as $\sqrt{2}$ or π) use at least one more significant figure than the most accurately determined physical quantity. Round off the final result according to its error (see next Chapter).
- (D) The error should always occupy the same position as the last one (or two, as the case may be) digit(s) of a measured result.

III. THE PROPAGATION OF ERRORS

3.1 The Functions of Observables

Very often the physical quantity of interest y is found indirectly as a function f of the directly measurable quantities $x^{(e)}$ ($e = 1, 2, \dots, m$). If n sets of measurements are made, each yielding a set of data $x_i^{(e)}$ ($i = 1, 2, \dots, n$), then the most common procedure of obtaining a final result y_f is by substituting the average values $\bar{x}^{(e)}$ (the "-" signifies averaging over the measurements: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, etc.) into f :

$$y_f = f(\bar{x}^{(1)}, \bar{x}^{(2)}, \dots, \bar{x}^{(m)}). \quad (3.1)$$

However, from a statistical point of view the result obtained by averaging over f is more convenient and satisfactory:

$$\bar{y} = \overline{f(x^{(1)}, x^{(2)}, \dots, x^{(m)})}. \quad (3.1A)$$

By making a Taylor expansion of each term on the right of Eq. (3.1A) about the average values $\bar{x}^{(e)}$ and using the fact that $(x^{(e)} - \bar{x}^{(e)})$ always vanishes, one can show that the two definitions (3.1) and (3.1A) agree to first-order approximation. We will therefore use y_f and \bar{y} interchangeably provided the dispersions in $x^{(e)}$ are not too large.

3.2 The Basic Formula of Error Propagation

Once y_f or \bar{y} has been determined, one is immediately faced with the problem of deducing its error from those of $x^{(e)}$. Let all of $x^{(e)}$ be INDEPENDENT VARIABLES, so that the probability distribution of one does not affect that of the other. The overall distribution is then just the product of the individual distributions:

$$p(x^{(1)}, x^{(2)}, \dots, x^{(m)}) = p(x^{(1)})p(x^{(2)}) \dots p(x^{(m)}). \quad (3.2)$$

As a consequence, the $x^{(e)}$ are also UNCORRELATED [7]:

$$\overline{x^{(k)} x^{(e)}} = \bar{x}^{(k)} \bar{x}^{(e)}, \quad (k, e = 1, 2, \dots, m) \quad (3.3)$$

Now suppose a number of HYPOTHETICAL measurements of all $x^{(e)}$ are made, each yielding a set of results $x^{(e)}$ ($e = 1, 2, \dots, m$) and hence a value for y . How would each y deviate from \bar{y} ? In first-order approximation, elementary calculus together with Eq. (3.1) leads to

$$y - \bar{y} = \sum_{e=1}^m \frac{\partial f(\bar{x}^{(1)}, \dots, \bar{x}^{(m)})}{\partial x^{(e)}} (x^{(e)} - \bar{x}^{(e)}). \quad (3.4)$$

We square both sides of Eq. (3.4) and take the average over ALL the hypothetical measurements :

$$\begin{aligned} (\delta y)^2 &= \overline{(y - \bar{y})^2} \\ &= \sum_{e=1}^m \left(\frac{\partial f}{\partial x^{(e)}} \right)^2 \overline{(x^{(e)} - \bar{x}^{(e)})^2} \\ &\quad + \sum_{\substack{k, e \\ k \neq e}} \frac{\partial f}{\partial x^{(k)}} \frac{\partial f}{\partial x^{(e)}} \overline{(x^{(k)} - \bar{x}^{(k)})(x^{(e)} - \bar{x}^{(e)})}, \quad (3.5) \end{aligned}$$

where we note that all differential coefficients are evaluated at the means $\bar{x}^{(e)}$ and hence need not be averaged. With the help of Eq. (3.3) it is simple to show that the second term always vanishes :

$$\begin{aligned} \overline{(x^{(k)} - \bar{x}^{(k)})(x^{(e)} - \bar{x}^{(e)})} &= \overline{x^{(k)} x^{(e)}} - \bar{x}^{(k)} \bar{x}^{(e)} - \bar{x}^{(k)} \bar{x}^{(e)} + \bar{x}^{(k)} \bar{x}^{(e)} \\ &= 0. \end{aligned}$$

We then arrive at the BASIC FORMULA FOR ERROR PROPAGATION :

$$\boxed{(\delta y)^2 = \sum_{e=1}^m \left(\frac{\partial f}{\partial x^{(e)}} \right)^2 (\delta x^{(e)})^2,} \quad (3.6)$$

in which both $\delta x^{(e)}$ and δy refer to parent standard deviations, so that from a finite number of actual measurements one should use Eq. (2.5) to compute the $\delta x^{(e)}$ on the right.

So far our derivation is quite independent of the kind of distributions of $x^{(e)}$. If normal distributions are assumed, then we can multiply both sides of Eq. (3.6) by $(0.6745)^2$ and reinterpret $\delta x^{(e)}$ and δy as probable errors [8].

We stress again that $x^{(e)}$ must all be INDEPENDENT (or at least UNCORRELATED) for Eq. (3.6) to apply. It is NOT VALID OTHERWISE.

3.3 Some Useful Corollaries

(A) The Probable Error of the Mean

Since the mean value is defined by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

direct application of Eq. (3.6) yields

$$\begin{aligned} \delta \bar{x} &= \left[\sum_{i=1}^n \frac{1}{n^2} (\delta x_i)^2 \right]^{\frac{1}{2}} \\ &= \delta x / \sqrt{n}, \end{aligned}$$

where δx is the probable error of any single measurement x_i . This is then a proof of Eq. (2.7).

(B) Linear Combinations of Variables

$$y = \sum_{\ell=1}^m c_{\ell} x^{(\ell)},$$

where c_{ℓ} are constant coefficients.

$$\delta y = \left[\sum_{\ell=1}^m (c_{\ell} \delta x^{(\ell)})^2 \right]^{\frac{1}{2}}.$$

(C) Product of Powers of Variables

$$y = \prod_{\ell=1}^m (x^{(\ell)})^{p_{\ell}} = (x^{(1)})^{p_1} (x^{(2)})^{p_2} \dots (x^{(m)})^{p_m},$$

where p_{ℓ} are constant powers.

$$(\delta y/y)^2 = \sum_{\ell=1}^m p_{\ell}^2 \left(\frac{\delta x_{\ell}}{x_{\ell}} \right)^2 .$$

(D) In particular, if

$$y = a + b$$

then

$$(\delta y)^2 = \sqrt{(\delta a)^2 + (\delta b)^2} ;$$

and if

$$y = ab$$

or

$$y = a/b,$$

then

$$\frac{\delta y}{y} = \sqrt{\left(\frac{\delta a}{a} \right)^2 + \left(\frac{\delta b}{b} \right)^2} .$$

IV CURVE FITTING

4.1 The Method of Least-Squares

In applying the graphical method to data analysis (cf. Section 1.2) it is necessary to fit a curve of the form $y = f(x; a, b, c, \dots)$ to n experimentally determined points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. How can the physical parameters a, b, c, \dots be found to give the best curve?

We make the following simplifying assumptions :

- (A) x_i ($i = 1, 2, \dots, n$) are the pre-selected locations of the variable x and are quite accurately determined, so that they have negligible error.
- (B) The deviations of y_i ($i = 1, 2, \dots, n$) along the y -axis from the hypothetical best curve form a normal distribution.
- (C) All y_i are determined with approximately equal accuracy.

Under these assumptions it is possible to show^[9] that the MOST LIKELY curve satisfies the LEAST-SQUARES CRITERION, i.e. a, b, c, \dots should be chosen to minimize the mean-square deviation S of y_i from the curve (Fig. 3) :

$$S = \frac{1}{n} \sum_{i=1}^n [y_i - f(x_i; a, b, c, \dots)]^2 \quad (4.1)$$

and

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = \frac{\partial S}{\partial c} = \dots = 0. \quad (4.2)$$

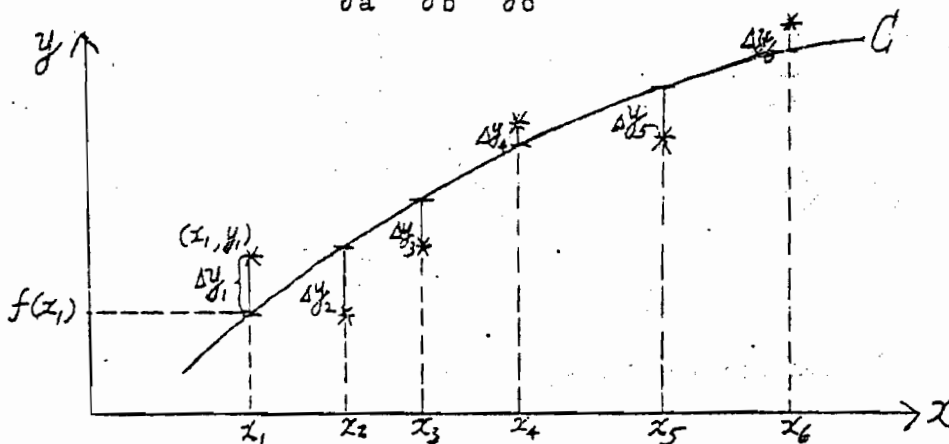


Fig. 3. In a Least-Squares Fit to 6 data points the parameters of the curve C are chosen so that $\frac{1}{6} \sum_{i=1}^6 (\Delta y_i)^2$ is minimized.

Excepting very pathological cases, the necessary conditions (4.2) are usually also sufficient for determining a, b, c, In particular, this can be proved to be always true in fitting a straight line.

If the error δy_i differs very much from point to point (say, more than 100%) then it would be necessary to weigh each term in Eq. (4.1) by an additional factor $(\delta y_i)^{-2}$ so as to emphasize the more accurate points and hence maintain maximum likelihood. The assumptions (A) and (B) are of course never strictly valid, but can be verified to hold true to a fair degree of approximation in most actual cases to justify the use of the least-squares method, which is, as you will see, very convenient.

4.2 The Linear Least-Squares Fit

We now apply criterion (4.2) to fitting a straight line (i.e. finding the slope a and intercept b)

$$y = f(x; a, b) = ax + b \tag{4.3}$$

to n points $(x_i, y_i) (i = 1, 2, \dots, n)$. Starting with

$$S = \frac{1}{n} \sum_{i=1}^n (y_i - ax_i - b)^2, \tag{4.4}$$

the minimization conditions leads to

$$\begin{cases} Xa + nb = Y \\ Ra + Xb = Z \end{cases}, \tag{4.5}$$

where

$$\begin{aligned} X &= \sum_{i=1}^n x_i, & Y &= \sum_{i=1}^n y_i \\ R &= \sum_{i=1}^n x_i^2, & Z &= \sum_{i=1}^n x_i y_i \end{aligned} \tag{4.6}$$

The solution for a and b is :

$$a = (nZ - XY)/\Delta , \quad (4.7)$$

$$b = (RY - XZ)/\Delta , \quad (4.8)$$

where

$$\Delta = nR - X^2 \quad . \quad (\text{always } \geq 0) \quad (4.9)$$

The actual computation of \bar{X} , \bar{Y} , \bar{Z} and \bar{R} is a very tedious and time-consuming task, which however must be done with high accuracy, since there is usually strong cancelling between all the opposing terms.

It is interesting to note that, as easily verified, the solution (4.7) - (4.9) is always such that $\sum_{i=1}^n (y_i - ax_i - b) = 0$, which provides a

simple check on the correctness of the computation. Of course, desk calculators or computers should be used whenever possible. As a matter of fact, computer subroutines on least-squares fits are usually available in most program libraries.

If weigh factors $\omega_i = (\delta y_i)^{-2}$ are used, then the same result as Eq. (4.7) - (4.9) are obtained, but with \bar{X} , \bar{Y} , \bar{Z} , \bar{R} , and n replaced by $\sum \omega_i x_i$, $\sum \omega_i y_i$, $\sum \omega_i x_i y_i$, $\sum \omega_i x_i^2$ and $\sum \omega_i$, respectively.

In the special case of fitting a straight line going through the origin, $y = ax$, Eq. (4.7) reduces to $a = Y/\bar{X}$ or $a = Z/\bar{R}$, both of which should give the same result. The special case of fitting for $a = 0$ is just the same as taking the mean value of y_i with Eq. (4.8) reduced to $b = \bar{Y}$ as expected.

4.3 Errors in the Linear Least-Squares Fit

The errors in a and b can be found from Eq. (3.6) and (4.7) - (4.9):

$$\begin{aligned} (\delta a)^2 &= \sum_i \left(\frac{\partial a}{\partial y_i} \right)^2 (\delta y_i)^2 \\ &= \frac{1}{\Delta^2} \left[\sum_i (-X + nx_i)^2 \right] (\delta y)^2 \\ &= \frac{nX^2 - 2n\bar{X} \cdot X + n\bar{R}}{\Delta^2} (\delta y)^2 , \end{aligned}$$

or

$$\delta a = \sqrt{\frac{n}{\Delta}} \delta y, \quad (4.10)$$

and, similarly,

$$\delta b = \sqrt{\frac{R}{\Delta}} \delta y. \quad (4.11)$$

In this derivation we again neglect the errors in x_i and assume all y_i to have the same error δy , which is the probable error deduced from the parent standard DEVIATION of y_i from the straight line [8, 9] :

$$\delta y = 0.6745 \left[\frac{\sum_i (y_i - ax_i - b)^2}{n - 2} \right]^{\frac{1}{2}} \quad (4.12)$$

where the $(n - 2)$ in the denominator reflects the fact that a and b are but the best estimates of the true values, so that there are only $(n - 2)$ degrees of freedom left for the n points. We need only ONE determination of y_i for each x_i to obtain δy this way. The EXPERIMENTAL ERROR Δy_i (from one or MORE determinations) should agree approximately with Eq. (4.12). If that is not true then either y does not really depend on x linearly or there must be very large errors in the estimate of Δy_i .

V. THE POISSON DISTRIBUTION

5.1 Characteristics of the Distribution Function

In modern physics one very often wants to ask the question : if a microscopic entity (atom, molecule, particle, etc.) has a very small probability p (say, $\sim 10^{-8}$) of doing something (e.g. jump from one state to another) within a certain time, then, out of a macroscopic entity containing a large number n (say, $\sim 10^{10}$) of these, how many will do it within the same time interval ?

Obviously the AVERAGE NUMBER OF SUCCESSES (i.e. of those who have "done it") μ , obtained by repeated counting over the same time interval, is simply

$$\mu = np. \quad (5.1)$$

The DISTRIBUTION of the "number of successes" k is, strictly speaking, a binomial distribution, but, in the limit $n \rightarrow \infty$, $p \rightarrow 0$, and if $k^2 + \mu^2 \ll n$, reduces to the POISSON DISTRIBUTION^[10]

$$P(k; \mu) = \frac{\mu^k e^{-\mu}}{k!} \quad (5.2)$$

which is NORMALIZED (i.e. $\sum_{k=0}^n P(k; \mu) = 1$), and is most easily understood

as the PROBABILITY OF OBTAINING k success in a measurement of n trials

(Fig. 4)

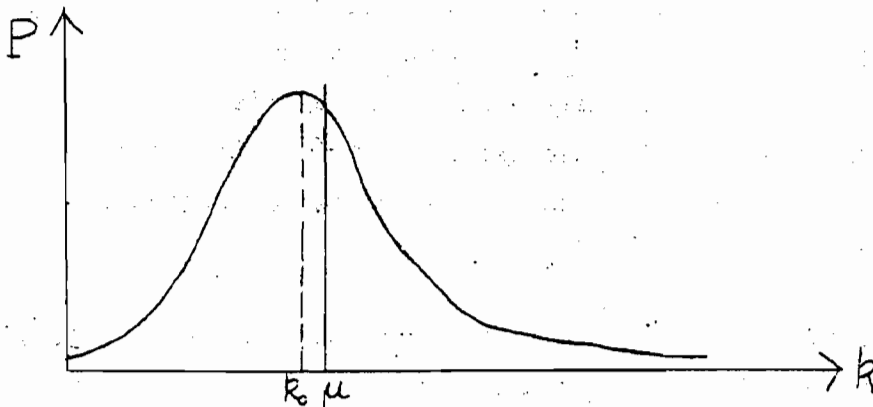


Fig. 4. A rough sketch of the smoothed Poisson distribution. Note the asymmetry and the fact that the peak position k_0 is always below the mean value μ ($\mu - 1 < k_0 < \mu$).

The Poisson distribution differs from the Normal Distribution in two respects : (a) its variable k is a DISCRETE INTEGER, and (b) k being a random variable, the mean μ also determines the standard deviation of k :

$$\sigma_k = \sqrt{\mu} . \quad (5.3)$$

The standard deviation of the mean is

$$\sigma_\mu = \sqrt{\mu/n} , \quad (5.4)$$

where n is the number of measurements. It is seen that Eq. (5.4) reduces to Eq. (5.3) in case $n = 1$, and so is applicable even if only one measurement is made^[11].

Eq. (5.3) gives the best estimates of the parent standard deviation. One may, however, also compute the SAMPLE STANDARD DEVIATION directly :

$$\sigma_k^{(s)} = \left[\frac{1}{n} \sum_{l=1}^n (k_l - \mu)^2 \right]^{\frac{1}{2}} , \text{ which should tend to } \sigma_k \text{ as } n \rightarrow \infty . \text{ This provides}$$

an independent check of σ_k , and can be experimentally verified.

The probable error δ is not related to σ by a constant coefficient:

$$\delta = c(\mu)\sigma ,$$

where $c(20) = 0.575$, $c(200) = 0.640$, $c(\infty) = 0.6745$. It is common to fix $c(\mu)$ at 0.6745 and disregard the slight overestimation of δ . Furthermore, the Poisson distribution is not symmetrical, so that the mean value μ is not the same as the mode or the median. The probable error is therefore not symmetrical with respect to μ . This asymmetry can be disregarded for large μ , because the "skewness" of the distribution decreases as $1/\sqrt{\mu}$. All in all, except for very small μ (say < 20 or 30), σ and δ can still be treated in roughly the same manner as in Chapter 2.

If μ becomes large ($\mu \gg 1$), then the curve in Fig. 4 gradually becomes symmetrical. The Poisson distribution then reduces^[12] to a PARTICULAR Gaussian distribution peaked at μ and with standard deviation $\sigma = \sqrt{\mu}$, as can be shown by applying Stirling's formula to Eq. (5.2) and then Taylor-expand $\ln P(k)$ at the point $k = \mu$, resulting in

$$\lim_{\substack{\mu \rightarrow \infty \\ k \rightarrow \infty}} P(k; \mu) = \frac{e^{-\frac{(k-\mu)^2}{2\mu}}}{\sqrt{2\pi\mu}}, \quad (|k-\mu|^2 \sim \mu)$$

which further justifies our treatment of the probable error for $\mu \gg 1$.

Another interesting feature is that the "sum" of two Poisson distributions remains Poissonian, so that, for instance, the source count, background count and combined count in radioactive counting all follow Poisson distributions.

5.2 Applications to Particle Counting

(A) The Standard Deviation in the Counting Rate

It is important to remember that $\sigma = \sqrt{\mu}$ refers to the standard deviation in the number of counts, but not to counting rates. If N counts are recorded by a particle detecting system in m seconds then the counting rate is

$$R = \left(\frac{N}{m} \pm \frac{\sqrt{N}}{m} \right) \text{ counts/sec.}, \quad (5.5)$$

NOT $N/m \pm \sqrt{N/m}$ counts/sec.

(B) Different Schemes of Counting

Since σ_N is independent of the number of measurement, the accuracy in the counting rate depends only on the grand total number of counts (N) taken, and is independent of the numbers of measurements in which these counts are made.

Suppose m measurements yield N_1, N_2, \dots, N_m counts, respectively, each taken in 1 sec. The average rate is

$$R = \frac{1}{m} \sum_{i=1}^m N_i.$$

The standard deviation in R is, according to Eq. (3.6),

$$\begin{aligned}\sigma_R^2 &= \sum_{l=1}^m \frac{1}{m^2} \sigma_l^2 \\ &= \frac{1}{m^2} \sum_{l=1}^m N_l,\end{aligned}$$

or

$$\sigma_R = \sqrt{N}/m, \quad (5.6)$$

where $N = \sum_{l=1}^m N_l$ is the total number of counts, and we have used the fact that $\sigma_l^2 = N_l$.

If we regard all the counts as taken in one single measurement, then, clearly as given by Eq. (5.5), σ_R is again \sqrt{N}/m , in agreement with Eq. (5.6). However, taking the counts in several consecutive measurements has the advantage of revealing long term drifts in either the source or the background, if any.

(C) Counting in the Presence of Background

Suppose in time t_b we record N_b "background" counts and then, with the particle source put in, N_o counts are recorded in time t_o . The source rate R_s is obviously

$$R_s = \frac{N_o}{t_o} - \frac{N_b}{t_b}. \quad (5.7)$$

The standard deviation σ_s follows from Eqs (3.6) and (5.3) :

$$\begin{aligned}\sigma_s &= [(\sigma_o^2/t_o^2) + (\sigma_b^2/t_b^2)]^{1/2} \\ &= \sqrt{(R_o/t_o) + (R_b/t_b)},\end{aligned} \quad (5.8)$$

where R_o and R_b are N_o/t_o and N_b/t_b , respectively.

(D) The Optimum Division of Counting Time

If a total of T secs is available for counting, what fraction λ should be allotted to background counting so as to minimize σ_s and give the best result? To find λ we differentiate Eq. (5.8) with respect to σ_s^2 , with $t_b = \lambda T$ and $t_o = (1 - \lambda)T$, and set the result to 0:

$$\frac{\partial}{\partial \lambda} \frac{1}{T} \left(\frac{R_o}{1 - \lambda} + \frac{R_b}{\lambda} \right) = 0,$$

which yields

$$\lambda = \frac{\sqrt{R_b}}{\sqrt{R_o} + \sqrt{R_b}} \quad (5.9)$$

or simply

$$\frac{t_b}{t_o} = \sqrt{\frac{R_b}{R_o}} \quad (5.10)$$

If this scheme is adopted then the minimized σ_s is

$$\sigma_s = \frac{\sqrt{R_o} + \sqrt{R_b}}{\sqrt{T}} \quad (\text{Optimum}) \quad (5.11)$$

(E) In Case of Heavy Background

If we try to count a very weak source in the presence of intense background, how much time T is necessary for a meaningful result? Clearly $R_s \gg \sigma_s$ is a sensible criterion, which, in the approximation $R_b \gg R_s$, leads to the requirement

$$T \gg \frac{4R_b}{R_s} \quad (5.12)$$

when the optimum time Division described above is used. If both R_b and R_s are given in seconds then we need $T \gg 400$ secs or $T \sim 1$ hr. for measuring a source one tenth as strong as background.

(F) Time Interval Distributions

The Poisson distribution can also be applied to the TIME INTERVALS between successive particle counts. If the average counting rate is μ then the average time interval is

$$\overline{\Delta t} = \frac{1}{\mu}, \quad (5.13)$$

which has a standard deviation equal to itself: $\sigma(\Delta t) = \Delta t$. Due to the asymmetry of the Poisson distribution there are more intervals (63%) shorter than the average value.

All counting equipments are limited in that, during a characteristic time interval τ_R (the RESOLVING or DEAD TIME) immediately after a count, it is insensitive to any counts received. Since short intervals between counts are more probable, the dead time artificially reduces $\sigma(\Delta t)$ by rejecting all $\Delta t < \tau_R$.

For an observed counting rate R_o we must make corrections for the "lost counts" as follows :

$$R \sim R_o(1 + R_o\tau_R), \quad (5.14)$$

in case $R_o\tau_R \ll 1$, or if the loss counts do not lengthen the dead time; and

$$R_o \sim R e^{-R\tau_R}, \quad (5.15)$$

if the dead time must be reckoned from the last count received.

NOTES

- (1) Applying STIRLING'S FORMULA $n! \rightarrow \sqrt{2\pi n} (n/e)^n$ (as $n \rightarrow \infty$) to the factorials in Eq. (2.1), one obtains the discrete Gaussian distribution

$$G_{k_0}(k; \sigma) = \exp[-(k - k_0)^2 / (2\sigma^2)] / (\sqrt{2\pi} \sigma), \text{ where } \sigma = \sqrt{np(1-p)} \text{ and}$$

$k_0 = np$. We take $p = \frac{1}{2}$ as may be expected, so that $\sigma = \sqrt{n}/2$, $k_0 = n/2$.

The probability of finding $k' = 2k - n$ is $p(k') = G((k' + n)/2; \sigma)(dk/dk')$
 $= \exp[-k'^2 / (2\sigma'^2)] / (\sqrt{2\pi} \sigma') = G(k'; \sigma')$, centered at $k' = 0$ and

$\sigma' = 2\sigma = \sqrt{n}$. It is then obvious that Eq. (2.3) follows, with $\sigma = \sqrt{n} \Delta x$.

We note that n and Δx cannot and need not be separately determined, so that the elementary errors are really of a hypothetical nature.

- (2) The simplest way to see if a set of data forms a Normal Distribution is to plot $y = \ln(n/n_0)$ against $x = (\delta x)^2$, so that $y = (-1/2\sigma^2)x - \ln(\sqrt{2\pi}\sigma)$, where n_0 is the total number of measurements and n the number of results which deviates from the mean by δx within some discrete interval. The data points should fit a straight line (slope = $-1/2\sigma^2$, intercept = $-\ln(\sqrt{2\pi}\sigma)$) if they form a normal distribution. The choice of the size of intervals along the δx -axis can affect the result, and, for a sufficiently large set of data, it is most convenient just to "quantize" δx by its last digit. It turns out that most experimental results would fit quite well within one or two standard deviations ($\delta x < 2\sigma$), but then get bad in the "tail region", which is very sensitive to small fluctuations.

For a more systematic way of gauging whether a mathematical distribution fits the data well, one usually turns to the χ^2 -test, which gives a single numerical estimate of the goodness of the fit. The interested readers are referred to Parratt, Section 4-8, p. 180.

- (3) The relative Statistical Efficiency is gauged by the inverse of the number of measurements which are necessary to yield a certain statistical precision (as judged from the standard deviation) in the measured quantity. It can be shown that the MEAN as a location index is the most efficient in general, respective of the distribution of data. For Normal Distributions, the standard deviation is considerably more efficient than either the mean deviation or the range, especially for small n . It is also known to be the most efficient among all quadratic estimators for a whole class of distributions (i.e. those with kurtosis = 3, which includes the normal distribution as a special case.)

(4) The basic problem of estimating the characteristic parameters of a distribution from a finite set of experimental data is discussed in the Appendix on the method of maximum likelihood. Nearly all the formulae we give in Chapters 2, 4 and 5 should be regarded as THE BEST ESTIMATES of the hypothetical parent distributions.

(5) This is obtained by looking for δx so that the probability of x satisfying $0 \leq |x - x_T| \leq \delta x$ is 50%, i.e. one requires

$$\int_{-\delta x}^{\delta x} dy \exp[-y^2/(2\sigma^2)]/(\sqrt{2\pi}\sigma) = \text{erf}(\delta x/\sigma) = 0.5. \text{ By looking up a table}$$

of the error function $\text{erf}(z) (\equiv 2(2\pi)^{-1/2} \int_0^z \exp(-t^2/2) dt)$, one finds that $\text{erf}(0.6745) \sim 0.5$, thus $\delta x = 0.6745\sigma$.

(6) For an unbiased estimate one should replace n by $(n - 1)$ in Eq. (2.8). See discussions connected with Eq. (A7).

(7) The proof is very simple : by definition

$$\overline{x^{(k)} x^{(l)}} = \int p(x^{(k)} x^{(l)}) dx^{(k)} dx^{(l)} = \int p(x^{(k)}) p(x^{(l)}) dx^{(k)} dx^{(l)}$$

$$= \int p(x^{(k)}) dx^{(k)} \int p(x^{(l)}) dx^{(l)} = \bar{x}^{(k)} \bar{x}^{(l)}. \text{ The converse, however, is}$$

not true in general : statistical independence is a stronger condition than non-correlation.

(8) The function of normally distributed variables is in general NOT a normal distribution, as can be readily verified. It is therefore not really correct to reinterpret δy as the probable error this way. We may, however, regard $\delta y = 0.6745\sigma_y$ as an operating definition, always bearing in mind that it may not give the correct 50-50% limit. An investigation of the distribution of y would be very complicated, and is rarely attempted in ordinary experimental situations. It does come to be of paramount importance, however, in the analysis of high-energy particles, especially in connection with the detection of resonances.

- (9) See, for instance, Parratt, Section 3-5, p.135.
- (10) The proof again depends on a straightforward application of Stirling's formula.
- (11) In contrast, the parent standard deviation σ blows up in a Normal Distribution if $n = 1$ (Eq. (2.5)). This stresses the difference between a normal and a Poisson distribution : in the former the location of the mean is completely independent of the "width" σ , which must be estimated from a number of measurements, while in the latter the location also determines the dispersion, and the distribution cannot be "translated" along the k axis.
- (12) We apply Stirling's formula to the expression for P :

$$z(k) = \ln P(k, \mu) \rightarrow k \ln(\mu/k) + (k - \mu) - \frac{1}{2} \ln(2\pi k).$$

The peak is given by

$$z'(k) = \ln(\mu/k) - \frac{1}{2k} = 0,$$

or $k_0 = \mu$ if we neglect $\frac{1}{2k}$. The second derivative at k_0 is

$$z''(k_0) = \left[-\frac{1}{k} + \frac{1}{2k^2} \right]_{k_0} \approx -\frac{1}{k_0},$$

and

$$z(k_0) = -\frac{1}{2} \ln(2\pi\mu).$$

Consequently

$$\begin{aligned} z(k) &\approx z(k_0) + \frac{1}{2} z''(k_0)(k - k_0)^2 \\ &\approx -\frac{1}{2} \ln(2\pi\mu) - \frac{1}{2\mu}(k - \mu)^2. \end{aligned}$$

Hence

$$P(k, \mu) = \exp(z) \approx \exp[-(k - \mu)^2 / (2\mu)] / \sqrt{2\pi\mu},$$

as we wish to prove.

(i) The Maximum Likelihood Criterion

Physical Measurements always produce a finite amount of information only and hence can never determine the characteristics of a distribution function or a curve with complete exactitude. The principle which enables us to bridge the gap between physical data and mathematical formulae is known as the METHOD OF MAXIMUM LIKELIHOOD*, which is really the keystone to nearly everything that we have said in these notes.

The statement of this principle is extremely simple : once a distribution function ϕ has been chosen to describe a set of data, the characteristic parameters of ϕ should be chosen to maximize the probability that such data will be actually observed from events distributed according to ϕ .

(ii) Estimation of μ and σ in a Normal Distribution

Let us take the normal distribution as an example. We assume the Gaussian function

$$G(x, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[- \frac{(x - \mu)^2}{2\sigma^2} \right] \quad (2.3)$$

describes the data points x_i ($i = 1, 2, \dots, n$). The probability of observing one point x in the interval dx being $G(x, \sigma)dx$, the probability of observing n points x_i in intervals dx_i is therefore

* First used in special applications by Gauss in the 19th century, this method was systematically developed by R.A. Fisher between 1912 and 1935, just the years when quantum mechanics was being born. It is now the most important approach to "real" statistical problems.

$$L(\sigma, \mu) dv' = \prod_{i=1}^n G(x_i, \sigma) dx_i$$

$$= \frac{1}{(\sqrt{2\pi} \sigma)^n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right] \prod_{i=1}^n dx_i, \quad (A')$$

where L is known as the LIKELIHOOD FUNCTION and $dv' = \prod_{i=1}^n dx_i$ is the n -dimensional volume element.

We now require σ and μ to be so chosen that L , or equivalently $\ln L$, is a maximum :

$$\ln L = -n[\ln\sqrt{2\pi} + \ln \sigma] - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Hence

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

$$= 0,$$

or

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$= \bar{x},$$

which is our best and unbiased estimate for μ (cf. Section 1.1). Next we set

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2$$

$$= 0,$$

or

$$\sigma = \left[\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \right]^{\frac{1}{2}},$$

which, as already mentioned, is a BIASED estimate of σ because the solution depends on an estimated quantity μ and it can be easily shown that any deviation of μ from the true mean always increases σ .

To obtain an unbiased estimate, it is necessary to consider a single quantity containing both μ and σ . To do this we define the rectangular coordinates z_i of an n -dimensional space S_n :

$$z_i = (x_i - \mu)/\sigma \quad (i = 1, 2, \dots, n) \quad (A2)$$

The radial distance R of a point (z_1, z_2, \dots, z_n) in S_n from the origin is

$$R^2 = \sum z_i^2 \quad (A3)$$

and the volume element dv of a hypersphere centered at the origin in S_n is

$$\begin{aligned} dv &= \prod dz_i \quad (A4) \\ &= C_n R^{n-1} dR, \end{aligned}$$

where we have expressed dv in terms of R , and C_n is a characteristic constant (e.g. $C_3 = 4\pi$, $C_2 = 2\pi$). Substituting Eqs. (A2)-(A4) into Eq. (A1) and considering L as a function of R only, we obtain

$$L(R)dR = (2\pi)^{-\frac{n}{2}} C_n R^{n-1} e^{-\frac{1}{2}R^2} dR. \quad (A5)$$

The best estimate for R is obtained from

$$\begin{aligned} \frac{\partial \ln L(R)}{\partial R} &= \frac{\partial}{\partial R} \left((n-1) \ln R - \frac{1}{2} R^2 \right) \\ &= \frac{n-1}{R} - R \\ &= 0, \end{aligned}$$

or

$$R^2 = n - 1,$$

which, together with definitions (A2) and (A3), immediately yields the previously given result

$$\sigma = \left[\frac{1}{n-1} \sum (x_i - \mu)^2 \right]^{\frac{1}{2}}$$

Strictly speaking, really we have only shown that $\Sigma(x_i - \mu)^2 / (n - 1)$ is the best unbiased estimate of σ^2 , which does not at all mean the square root of it is necessarily the unbiased estimate of $\sqrt{\sigma^2} = \sigma$ (it is not !). However, we can well neglect the difference, which turns out to be extremely small.

(iii) Precision in Estimates Based on Maximum Likelihood

By assuming* a Gaussian shape for the likelihood function L , it is possible to obtain a general formula for the standard deviation of the best estimate g_b :

$$L(g) = C \exp\left[-\frac{(g - g_b)^2}{2\sigma_g^2}\right],$$

so that, as can be easily verified :

$$\sigma_g = \left[-\frac{\partial^2}{\partial g^2} \ln L(g)\right]^{-\frac{1}{2}}. \quad (A6)$$

Direct application of Eq. (A6) to μ in Eq.(A1) gives us the standard deviation of the mean :

$$\begin{aligned} \sigma_\mu &= \left[\frac{\partial^2}{\partial \mu^2} \frac{1}{2\sigma^2} \Sigma(x_i - \mu)^2\right]^{-\frac{1}{2}} \\ &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

which agrees with Eq. (2.7).

* This in general is of course not true. In particular, it does not hold for $L(\sigma)$ given by Eq. (A1). We may nevertheless approximate the distribution of g near the optimum value g_{op} by such a curve, and thus ESTIMATE its standard deviation by Eq. (A6).

In exactly the same way, we can obtain the standard deviation of the standard deviation :

$$\begin{aligned}\sigma_{\sigma} &= \left\{ -\frac{\partial^2}{\partial \sigma^2} \left[-n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right] \right\}^{-\frac{1}{2}} \\ &= \left\{ \frac{\partial}{\partial \sigma} \left[\frac{n}{\sigma} - \frac{1}{\sigma^3} \sum (x_i - \mu)^2 \right] \right\}^{-\frac{1}{2}} \\ &= \left(-\frac{n}{\sigma^2} + \frac{3}{\sigma^4} n\sigma^2 \right)^{-\frac{1}{2}} \\ &= \frac{\sigma}{\sqrt{2n}} \quad , \quad (\text{Biased})\end{aligned}$$

in agreement with formula (2.8). Here we are actually making a biased estimate of σ_{σ} , and hence set $\sum (x_i - \mu)^2 = n\sigma^2$. To obtain an unbiased estimate (again qualified as previously mentioned in (ii)) we have, from Eq. (A5),

$$\begin{aligned}\sigma_R &= \left[-\frac{\partial^2}{\partial R^2} \ln L(R) \right]^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{2}} \quad .\end{aligned}$$

σ_R can be related to σ_{σ} and σ_{μ} through Eqs. (A2), (A3) and (3.6), yielding

$$\begin{aligned}\sigma_{\sigma} &= \frac{\sigma_R^3}{\sum (x_i - \mu)^2} \sigma_R \\ &= \frac{\sigma^3 \sqrt{n-1}}{(n-1)\sigma^2} \frac{1}{\sqrt{2}} \quad ,\end{aligned}$$

or

$$\sigma_{\sigma} = \frac{\sigma}{\sqrt{2(n-1)}} \quad . \quad (\text{"unbiased"}) \quad (A7)$$

This is the more exact formula that one should use for small n, even though it would hardly make any difference in practice.

(iv) Treatment of the Poisson Distribution

The Poisson Distribution presents a particularly simple example for the application of the maximum likelihood method. For m measurements k_1, k_2, \dots, k_m the likelihood function L is

$$\begin{aligned}
 L(\mu) &= \prod_{i=1}^m \frac{\mu^{k_i} e^{-\mu}}{(k_i)!} \\
 &= \frac{\mu^K e^{-m\mu}}{\prod_{i=1}^m k_i!} \quad (A8)
 \end{aligned}$$

where $K = \sum_{i=1}^m k_i$. So the most likely value for the mean is given by

$$\begin{aligned}
 \frac{\partial}{\partial \mu} \ln L(\mu) &= \frac{\partial}{\partial \mu} [K \ln \mu - m\mu - \sum_{i=1}^m (\ln k_i!)] \\
 &= \frac{K}{\mu} - m \\
 &= 0,
 \end{aligned}$$

which yields the expected result $\mu = K/m = \bar{k}$. The standard deviation of the mean is

$$\begin{aligned}
 \sigma_{\mu} &= \left\{ -\frac{\partial^2}{\partial \mu^2} [K \ln \mu - m\mu - \sum_{i=1}^m (\ln k_i!)] \right\}^{-\frac{1}{2}} \\
 &= (K/\mu^2)^{-\frac{1}{2}} \\
 &= \sqrt{\frac{\mu}{m}}
 \end{aligned}$$

which agrees with Eq. (5.4). Since only a single unknown parameter μ is involved, both estimates are not biased.

(v) Conclusions

In the same manner, the method of maximum likelihood can be extended to treat the problem of curve fitting, thus "proving" the least-squares criterion for normal distributions, and yield the results given by Eqs. (4.7)-(4.9), and (4.10)-(4.12). These formulae should all be understood as the most "probable values" obtained from estimates based on all the available data.

The method of maximum likelihood is of course not unlimited in its applications. In particular, to apply it one must first guess or assume the precise form of the parent distribution, which may not be straightforward. In addition, the likelihood function L becomes so complicated in more realistic problems that the proper maximization of L can be impossibly difficult. One may then have to settle for a less satisfactory but more manageable approach.

A SUMMARY OF IMPORTANT FORMULAE

(1) (Experimental Error)²
 = [max(Random Error, Intrinsic Error)]² + (Systematic Error Estimate)²

(2) The Best Location Index
 = MEAN VALUE $\bar{x} = \frac{1}{n} \sum x_i$
 The Best Dispersion Index
 = PARENT STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

(3) For Normal Distributions
 Probable Error of one measurement
 $\delta x = 0.6745 \sigma$
 Probable Error of the Mean

$$\delta \bar{x} = \frac{\delta x}{\sqrt{n}}$$

 The Probable Error of the Probable Error

$$\delta(\delta x) = 0.6745 \frac{\delta x}{\sqrt{2n}}$$

(4) The Formula of Error Propagation
 If $y = f(x, y, z, \dots, \omega)$,
 then

$$(\delta y)^2 = \left(\frac{\partial f}{\partial x}\right)^2 (\delta x)^2 + \left(\frac{\partial f}{\partial y}\right)^2 (\delta y)^2 + \dots + \left(\frac{\partial f}{\partial \omega}\right)^2 (\delta \omega)^2$$

(5) The Least-Squares Fit for a Straight Line $y = ax + b$ to (x_i, y_i) ,
 $i = 1, 2, \dots, n$.

$a = (nZ - XY)/\Delta$, $b = (RY - XZ)/\Delta$,
 $\Delta = nR - X^2$,

$\delta a = \sqrt{n/\Delta} \delta y$, $\delta b = \sqrt{R/\Delta} \delta y$,

$X = \sum x_i$, $Y = \sum y_i$,

$Z = \sum x_i y_i$, $R = \sum x_i^2$

$\delta y = 0.6745 [(n-2)^{-1} \sum (y_i - ax_i - b)^2]^{\frac{1}{2}}$

(6) The Binomial Distribution for k successes out of n trials:

$$B(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

The Normal Distribution ($\Delta x \rightarrow 0, n \rightarrow \infty$)

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\bar{x})^2/2\sigma^2]$$
,

where σ is the standard deviation.

The Poisson Distribution ($p \rightarrow 0, n \rightarrow \infty$)

$$P(k; \mu) = \mu^k e^{-\mu} / k!$$
,

$$\bar{k} = \sum_0^n P(k; \mu) k = \mu = np$$

$$\sigma_k = \sqrt{\mu}$$

$$\sigma_\mu = \sqrt{\mu/m}$$

(*) Excel 的標準誤差不算 0.6745 所以結果是大了 1.48

$$= \frac{1}{0.6745}$$