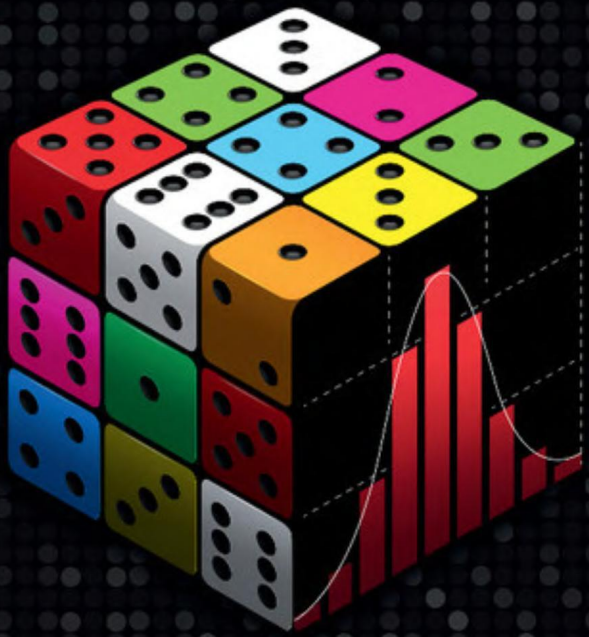
A Certain Uncertainty

Nature's Random Ways



MARK P. SILVERMAN

A CERTAIN UNCERTAINTY: NATURE'S RANDOM WAYS

MARK P. SILVERMAN
Trinity College, Connecticut

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107032811

First published 2014

Printing in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Silverman, Mark P., author.
A certain uncertainty : nature's random ways / Mark P. Silverman, G.A. Jarvis Professor of Physics, Trinity College, Connecticut.
pages cm
Includes bibliographical references.
ISBN 978-1-107-03281-1 (Hardback)
1. Statistical physics. 2. Mathematical physics. I. Title.
QC174.8.S545 2014
530.15′95–dc23 2014004090

ISBN 978-1-107-03281-1 Hardback

Additional resources for this publication at www.cambridge.org/silverman

Contents

How is it possible that mathematics, which is indeed a product of human thought independent of all experience, accommodates so well the objects of reality?

Here, in my view, is a short answer: In so far as mathematical statements concern reality, they are not certain, and in so far as they are certain, they do not refer to reality.

—*Albert Einstein*[1]

[1] Albert Einstein, from the lecture "Geometrie und Erfahrung" [Geometry and Experience] given in Berlin on 27 January 1921. (Translation from German by M. P. Silverman.)

Do you pay a power company each month for use of electric energy? Are you confident that the meter readings are accurate and that you are being charged correctly? Before answering the second question, perhaps you should read the chapter detailing the statistical analysis of my own electric energy consumption.

Do you enjoy sports, in particular ball games of one kind or another? Then you may be intrigued by my analysis of the ways in which a baseball can move if struck appropriately – or, perhaps of more practical consequence, how I inferred that a certain prominent US ballplayer was probably enhancing his performance with drugs long before the media became aware of it.

Are you concerned about global climate change? Then my statistical study of the climate *under ground* will give you a perspective on what is likely to be the most serious consequence to occur soonest – a consequence that has rarely been given public exposure.

And if you are a scientist yourself – especially a physicist – then you may be utterly astounded, as I was initially, to learn of persistent claims in the peer-reviewed physics literature of processes that, had they actually occurred, would turn nuclear physics (if not, in fact, all laws of physics) upside down. You should therefore find particularly interesting the chapter that describes my experiments and analyses that lay these extraordinary claims to rest.

The foregoing abbreviated descriptions should not disguise the fact that – as mentioned at the outset – this book is a *technical* narrative. The book can be read, I suppose, simply for the stories, skipping over the lines of mathematics. However, if your goal is to develop some proficiency in the use of probability and statistical reasoning, then you will want to follow the analyses carefully. I start the book with basic principles of probability and show every step to the conclusions reached in the detailed explanations of the empirical studies. (Some of the detailed calculations are deferred to appendices.)

A textbook, in which material is laid out in a "linear" progression of topics, may teach statistics more efficiently – but *this* book teaches the application of statistical reasoning *in context* – i.e. the use of principles as they are needed to solve specific problems. This means there will be a certain redundancy – but that is a *good* thing. In many years as a teacher, I have found that an important part of retention and mastery is to encounter the same ideas more than once but in different applications and at increasing levels of sophistication.

Virtually every standard topic of statistical analysis is encountered in this book, as well as a number of topics you are unlikely to find in any textbook. Furthermore, the book is written from the perspective of a "practical physicist", not a mathematician or statistician – and, where useful, my viewpoint is offered, schooled by some five decades of experimentation and analysis, concerning issues over which confusion or controversy have arisen in the past: for example, issues relating to sample size and uncertainty, use and significance of chi-square tests and *P*-values, the class

boundaries of histograms, the selection of Bayesian priors, the relationship between principles of maximum likelihood and maximum entropy, and others.

As a final point, it should be emphasized that this book is not merely a "statistics book". Rather, the subject matter at root is statistical *physics*. Every chapter, apart from the first, involves some experimental aspect, whether measured in a laboratory, simulated on a computer, or observed in the world at large. The themes of the narratives concern physical processes from widely different reaches of physics: dynamics of discrete particles, dynamics of fluids, dynamics of heat flow, statistical mechanics of bosons and fermions, creation of non-classical forms of light, transformations of radioactive nuclei, and more. In the process of solving particular problems, there arise – and I will answer – profound questions that are rarely encountered in physics textbooks. Consider thermodynamics, for example. Why is the chemical potential of black-body radiation zero? Is it zero for *all* kinds of photons? Is it zero because the photon is massless? Would a massless neutrino have a zero chemical potential? Read this book and find out.

What background do you need to read this book? Clearly, the more mathematics and physics you know beforehand, the more of the technical details you will be able to understand. An undergraduate physics major should be able to read all of it by the time he or she graduates. In fact, some of the content comes from the physics lectures I give at an undergraduate institution. A person with a knowledge of calculus should be able to read most of it. But anyone with an interest in probability, statistics, and physics should be able to take away something useful and thought-provoking from just the text.

That concludes the short answer, the long answer, and the objectives stated in the first paragraph of the Preface – if you read it.

Note regarding figures: Color figures for this book are available at the Cambridge University Press website www.cambridge.org/silverman.

Mark P. Silverman

Acknowledgments

I would like to thank my son Chris for his invaluable help in formatting the text of many of the figures in the book, for designing the beautiful cover of the book, and for his advice on the numerous occasions when my computers or software suddenly refused to co-operate. It is also a pleasure to acknowledge my long-time colleague, Wayne Strange, whose participation in our collaborative efforts to explore the behavior of radioactive nuclei was essential to the successful outcome of that work.

I very much appreciate the efforts of Dr. Simon Capelin, Elizabeth Horne, Samantha Richter, and Elizabeth Davey of Cambridge University Press to find practical solutions to a number of seemingly insurmountable problems in bringing this project to fruition. And I am especially grateful to my copy-editor, Beverley Lawrence, for her thorough reading and perceptive comments and advice.

1

Tools of the trade

> It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
>
> —*Pierre-Simon Laplace*[1]

1.1 Probability: The calculus of uncertainty

All measurements and observations, forecasts and inferences, are subject to uncertainty. These uncertainties reflect a lack of precise knowledge arising from the limitations of one's time, which restricts the amount of data that can be collected, or instrumentation, which determines the resolution with which signals or information can be acquired, or the fundamental laws of nature, which give rise to intrinsically random processes whose exact outcomes cannot be predicted irrespective of the apparatus and observation time. Although a well-ordered world governed by deterministic laws with no uncertainties may seem desirable at times, such a world will never be – and, in any event, would make for a rather dull place indeed.

To deal with the vagaries of nature one ordinarily must turn to the principles of mathematics bearing on probability and statistics. I will make no attempt to define probability. For one thing, innocuous as the subject may sound, it has spawned two schools of thought whose members have gone after one another (in a manner of speaking) like Crips and Bloods. So, from a practical standpoint, I would rather not begin a book with remarks likely to inflame any group of readers. Second, and more to the point, probability is a sufficiently basic concept that, in trying to capture its meaning in a few words, one ends up using tautological expressions like "chance" or "odds" or "likelihood" that do not really explain anything. The latter term, in fact, is not even a synonym, but is quite distinct from probability as will become apparent later when we encounter Bayes' theorem or make use of the method of maximum likelihood.

[1] Quoted by Mark Kac, "Probability" in *The Mathematical Sciences* (MIT Press, Cambridge, 1969) 239.

The second rule (1.2.3), although called Bayes' theorem, is a logical consequence of the laws of probability accepted by frequentists and Bayesians alike. It is regularly used in the sciences to relate $P(H|D)$, the probability of a particular hypothesis or model, given known data, to $P(D|H)$, the more readily calculable probability that a process of interest produces the known data, given the adoption of a particular hypothesis. In this way, Bayes' theorem is the basis for scientific inference, used to test or compare different explanations of some phenomenon.

The parts of Eq. (1.2.3), relabeled as

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}, \tag{1.2.5}$$

are traditionally identified as follows. $P(H)$ is the "prior" probability; it is what one believes about hypothesis H before doing an experiment or making observations to acquire more information. $P(D|H)$ is the "likelihood" function of the hypothesis H. $P(H|D)$ is the "posterior" probability. The flow of terms from right to left is a mathematical representation of how science progresses. Thus, by doing another experiment to acquire more data – let us refer to the outcomes of the two experiments as D_1 and D_2 – one obtains the chain of inferences

$$P(H|D_2D_1) = \frac{P(D_2|D_1H)P(D_1|H)P(H)}{P(D_2D_1)} \tag{1.2.6}$$

with the new posterior on the left and the sequential acquisition of information shown on the right.

As an example, consider the problem of inferring whether a coin is two-headed (i.e. biased) or fair without being able to examine it – i.e. to decide only by means of the outcomes of tosses. Before any experiment is done, it is reasonable to assign a probability of ½ to both hypotheses: (a) H_0, the coin is fair; (b) H_1, the coin is biased. Thus

$$\text{ratio of priors:} \qquad \frac{P(H_0)}{P(H_1)} = 1.$$

Suppose the outcome of the first toss is a head h. Then the posterior relative probability becomes

$$\text{first toss:} \qquad \frac{P(H_0|h)}{P(H_1|h)} = \frac{P(h|H_0)P(H_0)}{P(h|H_1)P(H_1)} = \frac{(\frac{1}{2})(\frac{1}{2})}{(1)(\frac{1}{2})} = \frac{1}{2}.$$

Let the outcome of the second toss also be h. Assuming the tosses to be independent of one another, we then have

$$\text{second toss:} \qquad \frac{P(H_0|h_2,h_1)}{P(H_1|h_2,h_1)} = \frac{P(h_2|h_1,H_0)P(h_1|H_0)P(H_0)}{P(h_2|h_1,H_1)P(h_1|H_1)P(H_1)} = \frac{(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})}{(1)(1)(\frac{1}{2})} = \frac{1}{4}.$$

It is evident, then, that the ratio of posteriors following n consecutive tosses resulting in h would be

$$n\text{th toss:} \qquad \frac{P(H_0|h_n \ldots h_1)}{P(H_1|h_n \ldots h_1)} = \frac{1}{2^n}.$$

Thus, although without direct examination one could not say with 100% certainty that the coin was biased, it would be a good bet (odds of H_0 over H_1: 1:4096) if 12 tosses led to straight heads.

It is important to note, however, that unlikely events can and do occur. No law of physics prevents a random process from leading to 12 straight heads. Indeed, the larger the number of trials, the more probable it will be that a succession of heads of any specified length will eventually turn up. In the nuclear decay experiments we consider later in the book, the equivalent of 20 h in a row occurred.

The probability of an outcome can be highly counter-intuitive if thought about in the wrong way. Consider a different application of Bayes' theorem. Suppose the probability of being infected with a particular disease is 5 in 1000 and your diagnostic test comes back positive. This test is not 100% reliable, however, but let us say that it registers accurately in 95% of the trials. By that I mean that it registers positive (+) if a person is sick (s) and negative (–) if a person is not sick ($\bar{s}$). What is the probability that you are sick?

From the given information and the rules of probability, we have the following numerical assignments.

Probability of infection $P(s) = 0.005$
Probability of no infection $P(\bar{s}) = 0.995$
Probability of correct positive: $P(+|s) = 0.95$
Probability of false negative $P(-|s) = 1 - P(+|s) = 0.05$
Probability of correct negative $P(-|\bar{s}) = 0.95$
Probability of false positive $P(+|\bar{s}) = 1 - P(-|\bar{s}) = 0.05$.

Then from Bayes' theorem it follows that the probability of being sick, given a positive test, is

$$P(s|+) = \frac{P(+|s)P(s)}{P(+|s)P(s) + P(+|\bar{s})P(\bar{s})} = \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.95)(0.995)} = 0.087$$

or 8.7%, which is considerably less worrisome than one might have anticipated on the basis of the high reliability of the test. Bayes' theorem, however, takes account as well of the low incidence of infection.

1.3 Probability density function and moments

In the investigation of stochastic[2] (i.e. random) processes, the physical quantity being measured or counted is often represented mathematically by a random variable.

[2] The world "stochastic" derives from a Greek root for "to aim at", referring to a guess or conjecture.

A random variable is a quantity whose value at each observation is determined by a probability distribution. For example, the number of radioactive nuclei decaying within some specified time interval is a discrete random variable; the length of time between two successive decays is a continuous random variable. Once the probability distribution is known – or at least approximated – the probability for any outcome (or combination of outcomes) can be calculated, as well as any statistical moments (provided they exist).

If we let X stand for a discrete random variable whose set of realizable values $\{x_i\ i = 1,2,\ldots N\}$ are the possible outcomes to an experiment with corresponding probability distribution $\{p_i\}$, then the probability that the experiment leads to *some* outcome in the set is the normalization or completeness requirement $P = \sum_{i=1}^{N} p_i = 1$. The average – i.e. mean value – of some function of the outcomes, $f(X)$, is expressed symbolically by angular brackets

$$\langle f(X)\rangle = \sum_{i=1}^{N} f(x_i)p_i. \tag{1.3.1}$$

Thus the nth moment of the distribution of X is defined to be

$$\mu_n \equiv \langle X^n\rangle = \sum_{i=1}^{N} x_i^n p_i. \tag{1.3.2}$$

Several particularly significant moments or combinations of moments include:

$$\text{mean:} \qquad \mu_X \equiv \mu_1 = \langle X\rangle = \sum_{i=1}^{N} x_i p_i, \tag{1.3.3}$$

$$\text{variance:} \quad \text{var}(X) \equiv \sigma_X^2 = \left\langle (X-\mu_X)^2 \right\rangle = \mu_2 - \mu_1^2, \tag{1.3.4}$$

from which the standard deviation σ_X is calculated. We also have

$$\text{skewness:}\ Sk_X \equiv \left\langle \left(\frac{X-\mu_X}{\sigma_X}\right)^3 \right\rangle = \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{\sigma_X^3}, \tag{1.3.5}$$

which is a measure of the asymmetry of a probability distribution about its center, and

$$\text{kurtosis:}\quad K_X \equiv \left\langle \left(\frac{X-\mu_X}{\sigma_X}\right)^4 \right\rangle = \frac{\mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4}{\sigma_X^4}, \tag{1.3.6}$$

which is a measure of the degree of flatness of a distribution near its center. It is ordinarily not necessary to go beyond the fourth moment in applying statistics to experimental distributions.

With regard to notation, the subscript X designating the random variable of interest may be omitted from the symbols for statistical functions where no confusion results.

To a continuous random variable X is associated a probability density function (pdf) $p(x)$, such that the probability that X lies within the range $(x, x + dx)$ is $p(x)dx$. The normalization requirement and moments of X are now given by integrals rather than sums:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \qquad m_n = \int_{-\infty}^{\infty} x^n p(x)dx. \tag{1.3.7}$$

The range of integration can always be taken to span the full real axis by requiring, if necessary, the pdf to vanish for specific segments. Thus, if X is a non-negative-valued random variable, then one defines $p(x) = 0$ for $x < 0$.

The cumulative distribution function (cdf) $F(x)$ – sometimes referred to simply as the distribution – is the probability $\Pr(X \leq x)$, which, geometrically, is the area under the plot of the pdf up to the point x:

$$\Pr(X \leq x) \equiv F(x) = \int_{-\infty}^{x} p(x')dx'. \tag{1.3.8}$$

It therefore follows by use of Leibnitz's equation from elementary calculus

$$\frac{d}{dx}\int_{a(x)}^{b(x)} F(x, y)dy = \frac{db}{dx}F(x, b) - \frac{da}{dx}F(x, a) + \int_{a(x)}^{b(x)} \frac{\partial F(x, y)}{\partial x}dy \tag{1.3.9}$$

that differentiation of the cdf yields the pdf: $p(x) = dF/dx$. This is a practical way to obtain the pdf, as we shall see later, under circumstances where it is easier to determine the cdf directly.

1.4 The binomial distribution: "bits" [*Bin*(1, *p*)] and "pieces" [*Bin*(*n*, *p*)]

The binomial distribution, designated $Bin(n, p)$, is perhaps the most widely encountered discrete distribution in physics, and it plays an important role in the research described in this book. Consider a binomial random variable X with two outcomes per trial:

$$X = \begin{cases} \text{success} \equiv 1 \text{ with probability } p \\ \text{failure} \equiv 0 \text{ with probability } q = 1 - p. \end{cases} \tag{1.4.1}$$

The number of distinct ways of getting k successes in n independent trials, which is represented by the random variable $Y = X_1 + X_2 + \cdots + X_n$, where each subscript

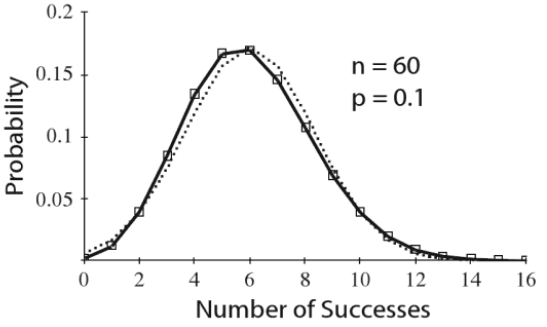


Fig. 1.1 Probability of x successes out of n trials for binomial distribution (solid) $Bin(n, p) = Bin(60, 0.1)$ and corresponding approximate normal distribution (dotted) $N(\mu, \sigma^2) = N(6,5.4)$.

labels a trial, is the coefficient of $p^k q^{n-k}$ in the binomial expansion $(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$ with combinatorial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Thus, the binomial probability function can be written in the form

$$P(x|n,p) = \binom{n}{p} p^x q^{n-x} \qquad (n \geq x \geq 0), \tag{1.4.2}$$

which shows explicitly the two parameters of the distribution. It is then straightforward, albeit somewhat tedious, to calculate from (1.3.2) the statistical quantities

$$\mu = np \qquad \text{var} = npq \qquad Sk = \frac{(q-p)}{\sqrt{npq}} \qquad K = \frac{3(n-2)pq + 1}{npq} \tag{1.4.3}$$

and others as needed. If the probability of obtaining either outcome is the same ($p = q = \frac{1}{2}$), the distribution is symmetric and the skewness vanishes. For $p < q$ the skewness is positive, which means the distribution skews to the right as shown in Figure 1.1. In the limit of infinitely large n, the kurtosis approaches 3, which is the value for the standard normal distribution (to be considered shortly). A distribution with high kurtosis is more sharply peaked than one with low kurtosis; the tails are "fatter" (in statistical parlance), signifying a higher probability of occurrence of outlying events.

In calculating statistical moments with the binomial probability function, the trick to performing the ensuing summations is to transform them into operations on the binomial expression $(p + q)^n$ whose numerical value is 1. For illustration, consider the steps in calculation of the mean

$$\langle X \rangle = \sum_{x=0}^{n} \binom{n}{x} x p^x q^{n-x} = p \frac{d}{dp} \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = p \frac{d}{dp} (p+q)^n = np(p+q)^{n-1} \xrightarrow{q=1-p} np$$

where only in the final step does one actually substitute the value of the sum: $p + q = 1$. For higher moments, one applies $p\frac{d}{dp}$ the requisite number of times. There is a

Table 1.1 *Distribution of outcomes of two dice*

y_i	(x_1, x_2)	$\Omega(y_i)$	$P(y_i) = \Omega(y_i)/\Omega$
2	(1,1)	1	1/36
3	(1,2), (2,1)	2	2/36 = 1/18
4	(1,3), (3,1), (2,2)	3	3/36 = 1/12
5	(1,4), (4,1), (3,2), (2,3)	4	4/36 = 1/9
6	(1,5), (5,1), (2,4), (4,2), (3,3)	5	5/36
7	(1,6), (6,1), (2,5), (5,2), (3,4), (4,3),	6	6/36 = 1/6
8	(2,6), (6,2), (3,5), (5,3), (4,4)	5	5/36
9	(3,6), (6,3), (4,5), (5,4)	4	4/36 = 1/9
10	(4,6), (6,4), (5,5)	3	3/36 = 1/12
11	(5,6), (6,5)	2	2/36 = 1/18
12	(6,6)	1	1/36
	Total	**36**	$\sum_{i=1} P(y_i) = 1$

$$\Omega(n_1, n_2, \ldots n_r|n) = \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r}. \tag{1.6.3}$$

(The symbol Ω is often used to represent "multiplicity" in statistical physics.) Note, however, that the the first two factors can be reduced in the following way

$$\binom{n}{n_1}\binom{n-n_1}{n_2} = \frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{k_2!(n-n_1-n_2)!} = \frac{n!}{n_1!n_2!(n-n_1-n_2)!}. \tag{1.6.4}$$

This pattern carries through for all subsequent factors, and by induction one obtains

$$\Omega(n_1, n_2, \cdots n_r\,|n) = \frac{n!}{n_1!n_2!\cdots n_r!} = \binom{n}{n_1 \ldots n_r}. \tag{1.6.5}$$

As an illustration useful to the discussion of histograms later, consider a game in which two dice are tossed simultaneously. Each die has six faces with outcomes $x_i = i$ (i = 1,2,...6). The outcomes of two dice are then $y_i = i$ (i = 2,3,...12). What is the probability of each outcome y_i, assuming the dice to be unbiased? Since there are $\Omega = 6 \times 6 = 36$ possible outcomes, the probability that a toss of two dice yields a particular value of y is the ratio of the number of ways to achieve y – i.e. the multiplicity $\Omega(y)$ – to the overall multiplicity Ω: $P(y_i) = \Omega(y_i)/\Omega$. By direct counting, we obtain Table 1.1.

If we were to cast the two dice 100 times, what would be the expected outcome in each category defined by the value y_i, and what fluctuations about the expected values would be considered reasonable? We would therefore want to know the theoretical means and variances in order to ascertain whether the dice were in fact unbiased. To determine means, variances and other statistics directly from a

Table 1.2 *Expected outcomes of* 100 *tosses of two unbiased dice*

y_i	n_i	σ_{n_i}
2	2.78	1.64
3	5.56	2.29
4	8.33	2.76
5	11.11	4.14
6	13.89	3.46
7	16.67	3.73
8	13.89	3.46
9	11.11	3.14
10	8.33	2.76
11	5.56	2.29
12	2.78	1.64
Total	**100.00**	

multinomial probability function is cumbersome; we will do this rigorously and efficiently by an alternative procedure later. However, a simple and intuitive way to answer the two questions is to recognize that each y-category in Table 1.1 may for the purposes of these questions be considered as the outcome of a binomial random variable because the result of a toss either falls into a specific category y_i or it does not. Thus, we deduce from relations (1.4.3) that the mean frequency of occurrence and variance of each category can be expressed as

$$n_i = nP(y_i) \qquad \sigma_{n_i}^2 = nP(y_i)(1 - P(y_i)), \tag{1.6.6}$$

as summarized in Table 1.2.

A plot of the frequency of outcomes (theoretical or observed) of this hypothetical experiment with two dice as a function of class constitutes a histogram. To know whether a set of observed frequencies is in accord or not with the expected values can be ascertained through various statistical tests to be described later in conjunction with actual experiments.

It is to be noted that the frequencies in a multinomial distribution are not all independent because they must sum to the fixed number n of trials. Thus, one would expect an *anti*-correlation (or negative correlation) between any pair of frequencies since an increase in one must result on average in a decrease in the other. How such correlations are to be calculated will also be taken up shortly.

Let us turn next to several continuous distributions of wide usage in physics.

1.7 The Gaussian distribution: measure of normality

The Gaussian or normal distribution, symbolically designated $N(\mu, \sigma^2)$, is quite likely the most widely encountered distribution employed in the service of science,

engineering, economics, and any other field of study where random phenomena are involved. The principal underlying reason for this – not always justified in the application – is the mathematical proposition known as the Central Limit Theorem (CLT), which shows the normal distribution to be the limiting form of numerous other probability distributions used to model the behavior of random phenomena. In particular, the normal distribution is most often employed as the "law of errors" – i.e. the distribution of fluctuations in some measured quantity about its mean. It has been written in jest (perhaps) that physicists believe in the law of errors because they think mathematicians have proved it, and that mathematicians believe in the law of errors because they think physicists have established it experimentally. There is some truth to the first assertion in that the Gaussian distribution emerges from a general principle of reasoning (referred to as the principle of maximum entropy) which addresses the question: Given certain information about a random process, what probability distribution describes the process in the most unbiased (i.e. least speculative) way? We will examine this question later. Suffice it to say at this point that the normal distribution does indeed apply widely, but, when it does not, one can be led astray with disastrous consequences by drawing conclusions from it.

The Gaussian distribution of a continuous random variable X whose values span the real axis takes the form

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/\sigma^2} \quad (\infty \geq x \geq -\infty). \tag{1.7.1}$$

By evaluating the moments of X one can show after a not insignificant amount of labor that the parameters μ and σ^2 are respectively the mean and variance. From the symmetry of $P(x|\mu, \sigma)$ about the mean, it follows that the skewness is identically zero. Evaluation of the fourth moment leads to a kurtosis of 3.

One can transform any Gaussian distribution to standard normal form $N(0, 1)$ by defining the new dimensionless random variable $Z = (X - \mu)/\sigma$. The cumulative distribution function (often represented by Φ) then takes the form

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du, \tag{1.7.2}$$

which is related to the error function

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-u^2} du \tag{1.7.3}$$

in the following way

$$\Phi(z) - \Phi(-z) = \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right). \tag{1.7.4}$$

As an academic physicist I am regularly asked by students whether I "grade on a curve". However, few students actually understand what grading on a curve means. The "curve" is the bell-shaped standard normal pdf, and to grade on it, strictly speaking, means to partition the area under the curve into four segments ($z \geq 1$), ($1 > z \geq 0$), ($0 > z \geq -1$), ($-1 > z$), such that the passing grades (A, B, C, D) will have (approximate) relative frequencies of 15%, 35%, 35%, 15%. For example, if I assign "A" to a student whose test score is $X \geq \mu + \sigma$, then

$$\Pr\left(\frac{x-\mu}{\sigma} \geq 1\right) = \Pr(z \geq 1) = \frac{1}{\sqrt{2\pi}} \int_1^\infty e^{-u^2/2} du = 0.159.$$

Thus, if test scores were normally distributed, I would expect about 15% of the class to receive a grade of A. Such an assumption might hold for a class of large enrollment (perhaps 50 or more), but not for small-enrollment classes. If I graded on a curve in an advanced physics class of six bright students, there would be one A, two Bs, two Cs, one D – and a great deal of dissatisfaction.

1.8 The exponential distribution: Waiting for Godot

The negative exponential distribution, symbolized by $E(\lambda)$, is interpretable as a distribution of waiting times between occurrences of random events – although it appears in other contexts in physics as we shall see. If X is a random variable whose realizations span the positive real axis, then the exponential pdf takes the form

$$P(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0). \end{cases} \tag{1.8.1}$$

Using the pdf to calculate the moments of X, one can show that $\langle X^n \rangle \equiv \mu_n = n!/\lambda^n$, from which follow the statistics

$$\mu = 1/\lambda \quad \sigma^2 = 1/\lambda^2 \quad Sk = 2/\lambda^3 \quad K = 9/\lambda^4. \tag{1.8.2}$$

The significance of the parameter λ is seen to be the inverse of the mean waiting time, which is equivalent to a frequency or rate. Though continuous, the exponential distribution has a direct connection to the discrete Poisson distribution in which the same parameter λ represents the intrinsic decay rate of a system. For example, if the number of occurrences of some phenomenon in a fixed window of observation time t is described by a Poisson distribution with parameter $\Lambda = \lambda t$, then the probability that 0 events will be observed in that time interval is $P_{\text{Poi}}(0|\lambda t) = e^{-\lambda t}$, and therefore the probability that at least 1 event will be observed in the time interval is the cumulative probability $F_{\text{Poi}}(t) = \Pr(X \leq t) = 1 - e^{-\lambda t}$. The derivative of $F_{\text{Poi}}(t)$ with respect to time

$$\frac{dF_{\text{Poi}}(t)}{dx} = P_{\exp}(t|\lambda) = \lambda e^{-\lambda t} \tag{1.8.3}$$

then gives the pdf of an exponential distribution of waiting times.

A significant attribute revealed by the variance of the exponential distribution is that the fluctuation ($\sim\sigma$) about the mean is of the order of the size of the signal ($\sim\mu$) itself. This will be seen to have important experimental consequences when we examine the physics of nuclear decay. The skewness and kurtosis of the exponential distribution bear no resemblance at all to those of the normal distribution and there is no limiting case in which the former reduce to the latter.

Another attribute of considerable interest is that the exponential distribution is the only continuous distribution with complete lack of memory. If the waiting times of a sample of decaying particles are described by an exponential distribution, then in a manner of speaking (to be understood statistically) the particles never get old so long as they have not yet decayed. To see this, suppose the particles were all created at time 0. Then the probability that there is no decay before time t is given by the integral

$$\Pr(X > t) = \int_t^\infty \lambda e^{-\lambda x} dx = e^{-\lambda t}. \tag{1.8.4}$$

Now let us suppose that T units of time have passed, and we seek the conditional probability that there is no decay before time $t + T$ given that there was no decay before time T

$$\Pr(X > t + T|X > T) = \frac{e^{-\lambda(t+T)}}{e^{-\lambda T}} = e^{-\lambda t}. \tag{1.8.5}$$

The probability is the same independent of the passage of time following creation of the particles. Note, in obtaining the preceding result we used the definition of conditional probability: $P(A|B) = P(AB)/P(B)$. As applied to the case of waiting times, the numerator $P(AB)$ is the probability that the waiting time is longer than both $t + T$ and T. But clearly if the first condition is satisfied, then the second must also be, and so in this case $P(AB) = P(A)$.

The lack of memory displayed by the exponential distribution has a discrete counterpart in the geometric distribution $P_{\text{geo}}(k|p) = pq^{k-1}$ in which an event occurs precisely at the kth trial (with probability p) after having failed to occur $k - 1$ times (with probability $q = 1 - p$). The probability of an eventual occurrence is 100%

$$\Pr(X \geq 1) = \sum_{k=1}^{\infty} q^{k-1} p = p \sum_{k=0}^{\infty} q^k = \frac{p}{1-q} = \frac{p}{p} = 1, \tag{1.8.6}$$

and the mean time between events is $1/p$

Then the mgf of $S_n = \sum_{i=1}^{n} a_i X_i$, with constant coefficients a_i, is deduced by the chain of steps below

$$g_{S_n}(t) = \langle e^{S_n t} \rangle = \left\langle e^{t\sum_{i=1}^{n} a_i X_i} \right\rangle = \prod_{i=1}^{n} \langle e^{a_i t X_i} \rangle = \prod_{i=1}^{n} g_{X_i}(a_i t) \underset{\text{iid}}{\longrightarrow} (g_X(at))^n, \tag{1.10.1}$$

where the third equality is permitted because the random variables are independent. Recall: If A and B are independent, then $\langle AB \rangle = \langle A \rangle \langle B \rangle$. The arrow above shows the reduction of $g_{S_n}(t)$ in the case of independent identically distributed (iid) random variables all combined with the same coefficient a.

Two widely occurring special cases are those involving the sum ($a_1 = a_2 = 1$) or difference ($a_1 = -a_2 = 1$) of two iid random variables for which (1.10.1) yields

$$g_{X_1+X_2}(t) = g_X(t)^2 \qquad\qquad g_{X_1-X_2}(t) = g_X(t) g_X(-t). \tag{1.10.2}$$

Another useful set of relations comes from evaluating the variance of the general linear superposition S_n by differentiating $\ln g_{S_n}(t) = \sum_{i=1}^{n} \ln g_{X_i}(a_i t)$

$$\begin{aligned} \left.\frac{d \ln g_{S_n}(t)}{dt}\right|_{t=0} &= \left.\sum_{i=1}^{n} \frac{a_i g'_{X_i}(a_i t)}{g_{X_i}(a_i t)}\right|_{t=0} \quad \Rightarrow \mu_{S_n} = \sum_{i=1}^{n} a_i \mu_i \\ \left.\frac{d^2 \ln g_{S_n}(t)}{dt^2}\right|_{t=0} &= \sum_{i=1}^{n} a_i^2 \left(\frac{g_{X_i}(a_i t) g''_{X_i}(a_i t) - g'_{X_i}(a_i t)^2}{g_{X_i}(a_i t)^2} \right)\Bigg|_{t=0} \quad \Rightarrow \sigma_{S_n}^2 = \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2. \end{aligned} \tag{1.10.3}$$

Another special case of particular utility is the equivalence relation for a normal variate X

$$N(\mu, \sigma^2) = \mu + \sigma N(0, 1), \tag{1.10.4}$$

which will be demonstrated later in the chapter.

A situation may arise – I have encountered it often – in which the mgf of some random variable X is a fairly complicated function of its argument and therefore does not correspond to any of the tabulated forms of known distributions. A useful procedure in that case may be to expand the mgf in a Taylor series to obtain an expression of the form

$$g(t) = e^{\sum_{n=0}^{\infty} a_n t^n}, \tag{1.10.5}$$

which is *not* to be confused with a structure like $\left\langle e^{t\sum_{i=1}^{n} a_i X_i} \right\rangle$ and does not necessarily correspond to a linear superposition of random variables. (For example, it may arise

from nonlinear operations.) An examination of the first few sequential derivatives of (1.10.5)

$$\begin{aligned} g^{(1)}|_0 &= a_1 \\ g^{(2)}|_0 &= 2a_2 + a_1 \\ g^{(3)}|_0 &= 6a_3 + 6a_1a_2 + a_1^3 \\ g^{(4)}|_0 &= 24a_4 + 24a_1a_3 + 12a_2^2 + 12a_2a_1^2 + a_1^4 \\ g^{(5)}|_0 &= 120a_5 + 120a_1a_4 + 120a_2a_3 + 60a_1a_2^2 + 20a_2a_1^3 + a_1^5 \end{aligned} \tag{1.10.6}$$

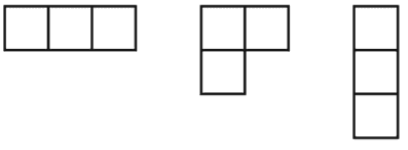
reveals a pattern that suggests a systematic way of calculating the moments of the distribution (and subsequently an approximation to the pdf if so desired). The form of the nth derivative is $n!$ times the sum over all partitions of the integer n weighted by a divisor $k!$ for each term in the partition that occurs k times. A partition of a positive integer n is a set of positive integers that sum to n. We can represent a particular partition $n = \sum_{j=1}^{n} j\alpha_j$ by the notation $\{1^{\alpha_1}2^{\alpha_2}3^{\alpha_3}\ldots n^{\alpha_n}\}$.

Consider, for example, $n = 3$. There are three ways to satisfy the integer relation $k + 2l + 3m = 3$, namely

$$3 = (3+0+0) = (2+1+0) = (1+1+1) \quad \Rightarrow \quad \{3\},\{2,1\},\{1^3\},$$

which leads to the weighted sum $3!\left(a_3 + a_2a_1 + \frac{a_1^3}{3!}\right)$ for the entry $g^{(3)}|_0$ in (1.10.6).

There is a graphical technique to construct the partitions of an integer relatively quickly by means of diagrams known as Young's tableaux. Each term in a partition is represented by a horizontal row of square boxes of length equal to the term; the boxes are stacked vertically, starting with the longest row. Thus, considering again the three partitions of $n = 3$, we have the three diagrams



(3) (2,1) (1^3)

The preceding ideas were drawn from the theory of symmetric groups,[3] which tells us that the total number $r(n)$ of partitions of an integer n is the coefficient of x^n in the power series expansion of Euler's generating function

$$E(x) = \prod_{j=1}^{\infty}\left(1 - x^j\right)^{-1} = 1 + x + 2x^2 + 3x^3 + 5x^4 + 7x^5 + 11x^6 + \cdots. \tag{1.10.7}$$

Examination of the first few terms verifies what could be easily determined by drawing the Young's tableaux. Should one need to know $r(n)$ for large n, there is

[3] J. S. Lomont, *Applications of Finite Groups* (Academic Press, New York, 1959) 258–261.

an asymptotic approximation derived by the renowned mathematicians G. H. Hardy and S. Ramanujan

$$r(n) \simeq \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{2n/3}}. \tag{1.10.8}$$

1.11 Binomial moment-generating function

As an illustration, let us re-examine the binomial distribution (coin-toss problem) from the vantage of its mgf. Define a binary random variable X whose value is 1 with probability p if the outcome is a head h or 0 with probability $q = 1 - p$ if the outcome is a tail t. Such a variable is termed a Bernoulli random variable, provided p remains constant for all trials. Then

$$g_X(t) = \langle e^{Xt} \rangle = pe^t + qe^0 = pe^t + q. \tag{1.11.1}$$

If the coin is tossed n times – or n coins are tossed independently and simultaneously once – the outcome is describable by a random variable $Y = \sum_{i=1}^{n} X_i$ whose mgf follows immediately from relation (1.10.1)

$$g_Y(t) = (pe^t + q)^n. \tag{1.11.2}$$

It is now a straightforward matter of taking derivatives – either of the mgf or its natural log – to confirm the previously given mean, variance, skewness, and kurtosis of the binomial distribution. For example:

$$\begin{aligned} \left.\frac{dg_X}{dt}\right|_{t=0} &= \left[npe^t(pe^t + q)^{n-1}\right]_{t=0} = np \\ \left.\frac{d^2 \ln g_X}{dt^2}\right|_{t=0} &= \left[npe^t(pe^t + q)^{-1} - n(pe^t)^2(pe^t + q)^{-2}\right]_{t=0} = npq. \end{aligned} \tag{1.11.3}$$

After the third or fourth derivative, the procedure becomes tedious to do by hand, but symbolic mathematical software (like *Maple* or *Mathematica*) can generate higher moments nearly instantly.

Although we arrived at the binomial mgf by starting with probabilities p and q of the Bernoulli random variable X and then calculating the generating function for the composite random variable Y, we could equally well have begun with the binomial probability function (1.4.2) and calculate the expectation value directly:

$$g_Y(t) = \langle e^{Yt} \rangle = \sum_{y=0}^{n} e^{yt} \binom{n}{y} \frac{p^y q^{n-y}}{n!} = \sum_{y=0}^{n} \binom{n}{y} \frac{(pe^t)^y q^{n-y}}{n!} = (pe^t + q)^n. \tag{1.11.4}$$

If, however, we already have the mgf from the procedure leading to (1.11.2), but do not know the binomial probability function, we can derive it from the mgf by a method to be demonstrated shortly.

A point worth noting about the procedure leading to Eq. (1.11.2) is that the sum of the "elemental" Bernoulli random variables (the Xs) produces a random variable Y which is also governed by a binomial distribution – or symbolically: $\underbrace{Bin(1,p) + \cdots + Bin(1,p)}_{n \text{ terms}} = Bin(n,p)$. From the mathematical form of the binomial mgf, one can see generally that the addition of independent random variables of type $Bin(n, p)$ and $Bin(m, p)$ generates a random variable of type $Bin(n + m, p)$. There are relatively few distributions that have the property that a sum of two random variables of a particular kind produces a random variable of the same kind. Moreover, as is easily demonstrated, this property does not hold for the difference of two binomial random variables. If $Y = X_1 - X_2$, where the two variates are independent and of type $Bin(n, p)$, then

$$g_Y(t) = (pe^t + q)^n (pe^{-t} + q)^n = [1 + 2pq(\cosh t - 1)]^n \tag{1.11.5}$$

in which the second equality was obtained after some algebraic manipulation employing the identity $p + q = 1$. The resulting mgf differs from that of a binomial random variable and, in fact, does not correspond to any of the standard types ordinarily tabulated in statistics books. Nevertheless, knowing the mgf, one can calculate from it all the moments of the difference of two independent binomial random variables of like kind. Although knowledge of the mgf affords a means to determine the probability function – and we shall examine shortly how to do this – in the present case it is better to proceed differently. We seek the probability $\Pr(X_1 - X_2 = z)$ that the difference is equal to some fixed value $n \geq z \geq -n$. This can be expressed by the suite of probability statements

$$\begin{aligned} \Pr(X_1 - X_2 = z) &= \sum_{x_2=0}^{n} \Pr(X_1 = x_2 + z | X_2 = x_2) \Pr(X_2 = x_2) \\ &= \sum_{x_2=0}^{n} P_{\text{Bin}}(x_2 + z) P_{\text{Bin}}(x_2), \end{aligned} \tag{1.11.6}$$

where the second equality is permissible because X_1 and X_2 are independent. The symbol $P_{\text{Bin}}(x)$ is an abbreviated representation of the complete probability function (1.4.2). It then follows upon substitution of the binomial probability functions that

$$\begin{aligned} \Pr(X_1 - X_2 = z) &= \sum_{y=0}^{n} \left[\binom{n}{y+z} p^{y+z} q^{n-y-z} \right] \left[\binom{n}{y} p^y \, q^{n-y} \right] \\ &= \left(\frac{p}{q}\right)^z \sum_{y=0}^{n-z} \left[\binom{n}{y+z} \binom{n}{y} (p^2)^y (q^2)^{n-y} \right]. \end{aligned} \tag{1.11.7}$$

Note that the upper limit to the sum over the dummy index y must be $n - z$ since the first coefficient vanishes when its lower index exceeds the upper index. The expression in (1.11.7) can be reduced to closed form in terms of a hypergeometric function ${}_2F_1$

$$\Pr(X_1 - X_2 = z) = \binom{n}{z} p^z (1-p)^{2n-z} \, {}_2F_1\left(-n, z-n, z+1, \left(\frac{p}{1-p}\right)^2\right) \tag{1.11.8}$$

but the derivation is beyond the intent of this chapter.[4]

1.12 Poisson moment-generating function

The moment generating function of a Poisson random variable X of mean value μ is also readily obtained

$$g_X(t) = \langle e^{Xt} \rangle = e^{-\mu} \sum_{x=0}^{n} e^{xt} \frac{\mu^x}{x!} = \sum_{y=0}^{n} \frac{(\mu e^t)^x}{x!} = e^{\mu(e^t - 1)}, \tag{1.12.1}$$

and leads to

$$\left.\frac{d \ln g_X(t)}{dx}\right|_{t=0} = \left.\frac{d^2 \ln g_X(t)}{dx^2}\right|_{t=0} = \mu,$$

which confirms the equality of $\langle X \rangle$ and $\mathrm{var}(X)$. Moreover, if X_1 and X_2 are independent Poisson random variables of respective means μ_1 and μ_2, then the mgf of their sum $Y = X_1 + X_2$

$$g_Y(t) = g_{X_1}(t) g_{X_2}(t) = e^{(\mu_1 + \mu_2)(e^t - 1)}$$

immediately establishes the fact that Y is a Poisson random variable of mean $\mu_Y = \mu_1 + \mu_2$.

If we had not used the mgf, we could have still arrived at the same conclusion by a method of reasoning based on summing over conditional probabilities, but it is a more cumbersome procedure:

$$\begin{aligned}
\Pr(X_1 + X_2 = y) &= \sum_{x_1=0}^{y} \Pr(X_2 = y - x_1 | X_1 = x_1) \Pr(X_1 = x_1) \\
&= \sum_{x_1=0}^{y} P_{\mathrm{Poi}}(y - x_1 | \mu_2) P_{\mathrm{Poi}}(x_1 | \mu_1) \\
&= \sum_{x_1=0}^{y} \left[e^{-\mu_2} \frac{\mu_2^{y-x_1}}{(y-x_1)!} \right] \left[e^{-\mu_1} \frac{{\mu_1}^{x_1}}{(x_1)!} \right] \\
&= \frac{e^{-(\mu_1+\mu_2)}}{y!} \sum_{x=0}^{y} \frac{y!}{x!(y-x)!} \mu_1^x \mu_2^{y-x} = \frac{e^{-(\mu_1+\mu_2)}}{y!} \sum_{x=0}^{y} \binom{y}{x} \mu_1^x \mu_2^{y-x} \\
&= \frac{e^{-(\mu_1+\mu_2)}}{y!} (\mu_1 + \mu_2)^y.
\end{aligned} \tag{1.12.2}$$

[4] Hypergeometric functions occur in the solution of second-order differential equations that describe a variety of physical system. One of the most important examples is the radial part of the wave function of the electron in a hydrogen atom (i.e. the Coulomb problem).

The moment generating function, in which t now stands for the set of r dummy variables $(t_1 \ldots t_r)$, is the expectation

$$g(t) = \langle e^{Nt} \rangle = n! \sum_{\{n_i\}} (e^{n_1 t_1} \ldots e^{n_r t_r}) \frac{p_1^{n_1} \cdots p_r^{n_r}}{n_1! \cdots n_r!} \tag{1.13.2}$$

subject to $\sum_i^r n_i = n$. Rearrangement of the preceding expression leads to a form recognizable as a multinomial expansion

$$g(t) = (p_1 e^{t_1} \ldots p_r e^{t_r})^n. \tag{1.13.3}$$

The set of probabilities $\{p_i\}$ are not all independent because $p_r = 1 - \sum_{i=1}^{r-1} p_i$. The factor e^{t_r}, the equivalent of which is absent in the generating function of a binomial distribution, was included for symmetry to permit all classes to be handled equivalently.

In most instances it is considerably simpler to work with the generating function than to carry out complex summations with the multinomial probability function. For example, by differentiating Eq. (1.13.3) we immediately obtain the means, variances, and covariances of the random variables $\{N_i\}$ representing the frequencies of each class:

$$\left.\begin{aligned} \langle N_i \rangle &= \left.\frac{\partial g}{\partial t_i}\right|_{t=0} = np_i \\ \langle N_i^2 \rangle &= \left.\frac{\partial^2 g}{\partial t_i^2}\right|_{t=0} = np_i + n(n-1)p_i^2 \\ \langle N_i N_j \rangle &= \left.\frac{\partial^2 g}{\partial t_i \partial t_j}\right|_{t=0} = n(n-1)p_i p_j \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} \operatorname{var}(N_i) \equiv \sigma_i^2 &= \langle N_i^2 \rangle - \langle N_i \rangle^2 = np_i(1-p_i) \\ \operatorname{cov}(N_i, N_j) &= \langle N_i N_j \rangle - \langle N_i \rangle \langle N_j \rangle = -np_i p_j. \end{aligned}\right. \tag{1.13.4}$$

A dimensionless measurement of the degree of correlation between outcomes in two classes is provided by the correlation coefficient

$$\rho_{ij} = \frac{\operatorname{cov}(N_i, N_j)}{\sigma_i \sigma_j} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}. \tag{1.13.5}$$

As noted before, the negative sign in the covariance or correlation coefficient signifies that on average the change in one frequency results in an opposite change in another frequency because of the constraint on the sum of all frequencies. The binomial distribution, where $p_2 = 1 - p_1$, provides an illustrative special case; Eq. (1.13.5) leads to $\rho_{12} = -1$, i.e. 100% anti-correlation, as would be expected.

A multinomial distribution can arise sometimes in unexpected ways. Consider the following situation, which will be of interest to us later when we examine means of judging the credibility of models (also referred to as hypothesis testing) with particular focus on examining the properties of radioactive decay. Suppose a random

process has generated K independent Poisson variates $\{N_k = Poi\,(\mu_k)\; k = 1 \ldots K\}$. The probability of getting the sequence of outcomes $\{n_1, n_2, \ldots n_K\}$ is then

$$\Pr\left(\{n_k\}|\{\mu_k\}\right) = \prod_{k=1}^{K} e^{-\mu_k}\frac{\mu_k^{n_k}}{n_k!} = e^{-\mu}\prod_{k=1}^{K}\frac{\mu_k^{n_k}}{n_k!}, \tag{1.13.6}$$

where $\mu = \sum_{k=1}^{K}\mu_k$. If, however, a constraint were imposed on the outcomes such that their sum must take a fixed value $\sum_{k=1}^{K} n_k = n$, then the *conditional* probability of obtaining the outcomes would be

$$\Pr\left(\{n_k\}\middle|\{\mu_k\}, \sum_{k=1}^{K} n_k = n\right) = \frac{P_{\mathrm{Poi}}(\{n_k\}|\{\mu_k\})}{P_{\mathrm{Poi}}\left(\sum_{k=1}^{K} n_k = n\middle|\mu\right)} = \frac{e^{-\mu}\prod_{k=1}^{K}\left(\frac{\mu_k^{n_k}}{n_k!}\right)}{e^{-\mu}\left(\frac{\mu^n}{n!}\right)} = n!\prod_{k=1}^{K}\frac{(\mu_k/\mu)^{n_k}}{n_k!}, \tag{1.13.7}$$

which is seen to be a multinomial probability function with parameters $p_k = \mu_k / \mu$. The substitution of the Poisson probability function for $\Pr\left(\sum_{k=1}^{K} n_k = n\middle|\mu\right)$ is justified because the sum of K independent Poisson variates is itself a Poisson random variable.

1.14 Gaussian moment-generating function

The moment generating function of the normal or Gaussian distribution is of particular significance in the statistical analysis of physical processes. Besides generating the moments of the distribution, it provides a reliable means of ascertaining how well an unknown probability distribution may be approximated by a normal one. Designate, as before, X to be a Gaussian random variable with mean μ and variance σ^2. Calculation of the mgf then leads to the integral

$$g(t) = \langle e^{xt}\rangle = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty} e^{xt}e^{-(x-\mu)^2/2\sigma^2}dx, \tag{1.14.1}$$

which is most easily evaluated by (a) transforming the integration variable to a dimensionless variable $z = (x - \mu)/\sigma$ said to be in standard normal form, (b) completing the "square" in the exponent, and (c) recognizing the normalization of the resulting Gaussian integral

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-\sigma t)^2/2}dz = 1 \tag{1.14.2}$$

to obtain the expression

$$g(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \tag{1.14.3}$$

We will make frequent use of this function throughout the book.

Using the mgf (1.14.3), we can easily demonstrate the equivalence relation (1.10.4). Define the random variable $X = a + bY$ where a and b are constants and $Y = N(0, 1)$ is a standard normal variate. Since a and Y are independent, the mgf of X is expressible as a product

$$g_X(t) = g_a(t) g_{bY}(t) = e^{at} g_Y(bt). \tag{1.14.4}$$

In going from the first equality to the second the mgf of a constant is simply

$$g_a(t) = \langle e^{\mathrm{at}} \rangle = e^{at}, \tag{1.14.5}$$

and the mgf of a constant times a random variable Y takes the form

$$g_{bY}(t) = \langle e^{bY\,t} \rangle = \left\langle e^{Y(bt)} \right\rangle = g_Y(bt). \tag{1.14.6}$$

However, for $Y = N(0,1)$, the mgf (1.14.3) applied to relation (1.14.6) yields $g_Y(bt) = e^{\frac{1}{2}b^2t^2}$. Thus, the product of the factors in (1.14.4) leads to

$$g_X(t) = e^{at} e^{\frac{1}{2}b^2t^2} = e^{at + \frac{1}{2}b^2t^2} \tag{1.14.7}$$

which identifies X as a normal random variable. Setting $a = \mu$ and $b = \sigma^2$ yields precisely relation (1.10.4).

One of the applications of the mgf is to establish the conditions for progressive approximation of one distribution by another. For example, the mgf of a binomial random variable $Bin(n, p)$ is $g_{\mathrm{Bin}}(t) = (pe^t + q)^n = (1 + p(e^t - 1))^n$. Expansion of $\ln g_{\mathrm{Bin}}(t) = n \ln(1 + p(e^t - 1))$ in powers of $(e^t - 1)$, which may be regarded as a small quantity since t is ultimately set to zero in calculations with the mgf, yields the Taylor series[6]

$$\ln g_{\mathrm{Bin}}(t) = np(e^t - 1) - \frac{1}{2}np^2(e^t - 1)^2 + \cdots.$$

In the limit that $p \to 0$ and $n \to \infty$ so that the product $np \to \mu$, we can truncate the preceding expansion after the first term to obtain a limiting form of the mgf

$$g_{\mathrm{Bin}}(t) \to e^{\mu(e^t - 1)} = g_{\mathrm{Poi}}(t), \tag{1.14.8}$$

which identifies a Poisson distribution of mean μ.

Next, consider expansion of $\ln g_{\mathrm{Bin}}(t)$ in powers of t

$$\ln g_{\mathrm{Bin}}(t) = np\left(t + \frac{1}{2}t^2 + \cdots\right) - np^2\left(\frac{1}{2}t^2 + \cdots\right) \to npt + \frac{1}{2}np(1-p)t^2 + \cdots$$

[6] Recall that: $\ln(1 + x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \cdots$.

taking care to include all contributions of the same order in t. For vanishing p, but $np \gg 1$, we truncate the expansion after the quadratic term to obtain the limiting form

$$g_{\text{Bin}}(t) \rightarrow e^{npt+\frac{1}{2}npqt^2} = g_{\text{Gaus}}(t), \tag{1.14.9}$$

recognizable as the mgf of a Gaussian distribution with mean $\mu = np$ and variance $\sigma^2 = npq$, where $q = 1 - p \approx 1$.

In summary, one can say that the "shape" of the probability curve of a binomial distribution approaches in form that of a Poisson distribution for low p and large n leading to a mean np of arbitrary magnitude. If np is much greater than 1, however, the shape – formed by a continuous curve connecting the discrete points of the binary (or Poisson) distribution – takes on the symmetrical shape of a Gaussian distribution with mean and variance equal to np.

1.15 Central Limit Theorem: why things seem mostly normal

It often occurs in science that one encounters random variables whose probability distributions are not known. This is particularly the case when the quantity being sought is inferred from more elemental randomly varying quantities. Then, even if the probability distributions of the elemental variables are known, it may be very difficult to calculate exactly the distribution of the composite quantity. For example, consider the traditional experiment in introductory physics labs to measure the acceleration g of freefall at the surface of the Earth. This requires timing a vertically falling object and marking the intermediate locations as a function of time. The data comprise measurements of time intervals and spatial intervals with random experimental errors of measurement whose distributions are not *a priori* known. The standard statistical procedure of error propagation analysis lets one estimate a mean value and standard deviation of g, but, without knowledge of the underlying probability distribution, it is not possible to interpret the significance of these statistical quantities. This is not merely an academic problem confined to instructional labs, but an issue that can have potentially serious consequences in the real world, particularly in science, medicine, and engineering.

The Central Limit Theorem of statistics often provides a workable solution by elucidating the circumstances under which a combination of random variables of different distributions together form a quantity distributed for all practical purposes like a Gaussian variate. Consider, as an illustration, the special case of a random variable $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ interpretable as the mean of n independent, identically distributed measurements $\{X_i \, i = 1 \ldots n\}$ each with mgf $g_X(t)$. From Eq.(1.10.1), the mgf of $\overline{X}$ takes the form $g_{\overline{X}}(t) = \left[g_X\left(\frac{t}{n}\right)\right]^n$, the natural log of which can be expressed in terms of the moments of X by expanding $g_X(t)$ in a Taylor series about $t = 0$

$$\ln g_{\overline{X}}(t) = n\ln g_X\left(\frac{t}{n}\right) = n\ln\left(1+\sum_{k=1}^{\infty}\mu_k\frac{(t/n)^k}{k!}\right) \equiv n\ln(1+\varepsilon(t)). \tag{1.15.1}$$

Here $\mu_k = \frac{d^k g_X(t)}{dt^k}\big|_{t=0}$ is the kth moment of X and the term $\varepsilon(t)$ is to be regarded as a small quantity since t will eventually be set to 0. A Taylor series expansion of the logarithm

$$\ln g_{\overline{X}}\left(\frac{t}{n}\right) = n\left[\varepsilon(t) - \frac{1}{2}\varepsilon(t)^2 + \frac{1}{3}\varepsilon(t)^3 - \cdots\right], \tag{1.15.2}$$

followed by arrangement of all terms in increasing powers of t, then leads to an expression

$$\begin{aligned}\ln g_{\overline{X}}\left(\tfrac{t}{n}\right) &= \mu_1 t + \left(\mu_2 - \mu_1^2\right)\frac{t^2}{2n} + \left(\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3\right)\frac{t^3}{6n^2} + \cdots \\ &= \mu_1 t + \frac{\sigma_X^2}{2n}t^2 + \frac{\left\langle (X-\mu_1)^3\right\rangle}{6n^2}t^3 + \cdots\end{aligned} \tag{1.15.3}$$

in increasing moments about the mean of X. If the number of observations n, which appears in the denominator of each term to a power of one less than the corresponding moment, is sufficiently large that terms beyond the second moment can be neglected, the truncated series is of the form of a Gaussian mgf of mean $\mu_{\overline{X}} = \mu_1$ and variance

$$\sigma_{\overline{X}}^2 = \sigma_X^2/n. \tag{1.15.4}$$

If the condition that the variables $\{X_i\}$ be identically distributed is relaxed, then the foregoing analysis carries through in the same way, albeit with some extra summations, leading to a Gaussian distribution $N\left(\mu_{\overline{X}}, \sigma_{\overline{X}}^2\right)$ with parameters

$$\mu_{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n}\mu_{X_i} \qquad \sigma_{\overline{X}}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_{X_i}^2. \tag{1.15.5}$$

It is worth noting explicitly that the only requirement on the distributions of the original variables $\{X_i\}$ is the existence of first and second moments. This modest requirement is usually met by the distributions one is likely to encounter in physics although the Cauchy distribution, which appears in spectroscopy as the Lorentzian lineshape, is an important exception. A Cauchy distribution has a median, but the mean, variance, and higher moments do not exist.

A significant outcome of the foregoing calculation is that the standard deviation of the mean of n observations is *smaller* than the standard deviation of a single observation by the factor $\sqrt{n}$. This statistical prediction is the justification for repetition and combination of measurements in experimental work. Perhaps it is intuitively obvious to the reader that the greater the number of measurements taken, the greater would be the precision of the result, but historically this was not at all

Table 1.3 *Outcome of Poisson RNG with* $\mu = 100$ *(1600 bins per bag)*

Bag No. i	Mean $\bar{x}$	Std Dev. $s_{\bar{x}}$	Bag No. i	Mean $\bar{x}$	Std Dev. $s_{\bar{x}}$
1	99.6	99.0	9	99.6	97.5
2	99.9	100.7	10	100.1	100.0
3	100.1	100.7	11	99.9	99.7
4	100.4	101.5	12	99.9	96.7
5	100.1	100.8	13	99.9	105.1
6	100.0	104.1	14	100.1	105.6
7	100.4	101.0	15	100.3	101.5
8	100.1	98.5	16	100.0	100.6

expectations and empirical outcomes, we find excellent agreement with the principles outlined above.

THEORY	EMPIRICAL
$\sigma_X = \sqrt{100} = 10$	$s_X = 10.040$
$\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}} = \dfrac{10}{40} = 0.250$	$s_{\overline{X}} = 0.251$
$\sigma_{\overline{Y}} = \dfrac{\sigma_X}{\sqrt{nb}} = \dfrac{10}{160} = 0.0625$	$s_{\overline{Y}} = \dfrac{s_X}{\sqrt{nb}} = 0.0628$

A final point (for the moment) in regard to Eq. (1.15.4) or Eq. (1.15.5) is that the expression for variance of the mean is a general property of variances irrespective of the Central Limit Theorem. Without the CLT, however, we would not necessarily know what to do with this information. The theorem tells us, for example, that, if the process generating the particle counts can be approximated by a Gaussian distribution, then we should expect about 68.3% of the bins to contain counts that fall within a range $\pm s_{\overline{X}}$ about the observed mean $\bar{x}$.

1.16 Characteristic function

The characteristic function (cf) of a statistical distribution is closely related to the moment generating function (mgf) when the latter exists and can be used in its place when the mgf does not exist. It is a complex-valued function defined by

$$h_X(t) = \left\langle e^{iXt} \right\rangle = g_X(it), \tag{1.16.1}$$

where $i = \sqrt{-1}$ is the unit imaginary number. For a random variable X characterized by a pdf $p_X(x)$, the characteristic function takes the form

$$h_X(t) = \int_{-\infty}^{\infty} e^{iXt} p_X(x) dx \tag{1.16.2}$$

which is recognizable as the Fourier transform of the pdf. In this capacity lies its primary utility, for it permits one to calculate the probability density (or probability function) by an inverse transform

$$p_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iXt} h_X(t)dt, \tag{1.16.3}$$

which cannot always be done so straightforwardly by means of the mgf itself. One can, of course, also calculate moments of a distribution by expansion of $h_X(t)$ in a Taylor series about $t = 0$ to obtain an alternating progression of real and imaginary valued quantities, but I have found little advantage to using it this way when $g_X(t)$ is available.

As an illustration of the inverse problem of determining the pdf from the cf, consider the standard normal distribution for which the generating function is $g_X(t) = e^{t^2/2}$ and therefore $h_X(t) = e^{-t^2/2}$. The probability density then follows from the integral

$$\begin{aligned} p_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}e^{-t^2/2}dt = \frac{e^{-x^2/2}}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(t^2+2ixt-x^2\right)}dt \\ &= \frac{e^{-x^2/2}}{\sqrt{2\pi}}\left[\underbrace{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(t+ix)^2}dt}_{=1}\right] = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \end{aligned} . \tag{1.16.4}$$

The calculation is easily extended to the case of an arbitrary Gaussian distribution $N(\mu, \sigma^2)$ at the expense of a few more algebraic manipulations in completing the square in the exponential.

The method can also be applied to calculate the probability function of a discrete distribution (as an alternative procedure to using a probability generating function). Consider, for example, a binomial distribution $Bin(n, p)$ for which the mgf was found to be $g_X(t) = (pe^t + q)^n$. The cf is then $h_X(t) = (pe^{it} + q)^n$ and implementation of the transform (1.16.3) is accomplished through the following steps: (a) binomial expansion of the terms in parenthesis, (b) collection of factors containing the integration variable and reversal of the order of summation and integration, (c) "collapse" of the summation by means of a δ function:

$$\begin{aligned} p_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}\left(pe^{it}+q\right)^n dt &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}\sum_{k=0}^{n}\binom{n}{k}\left(pe^{it}\right)^k q^{n-k}dt \\ &= \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\underbrace{\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(k-x)t}dt}_{\delta(k-x)} \\ &= \binom{n}{x}p^x q^{n-x}. \end{aligned} \tag{1.16.5}$$

The last step bears some comment. A Dirac delta function $\delta(x)$ is technically not a function, but a mathematical structure with numerous representations whose value is zero everywhere except where its argument is zero, at which point its value is infinite; yet the area under the delta function (that is, the integral of the delta function over the real axis) is 1. The object was introduced into physics by P. A. M. Dirac to the horror of mathematicians (or so I have read) but eventually was legitimized by Laurent Schwarz in a theory of generalized functions (referred to as distribution theory although the concept of distribution is unrelated to that in statistics). Ordinarily, the delta function has meaning only in an integral where it serves to "sift" out selected values of the argument of the integrand – for example: $\int_{-\infty}^{\infty} f(x)\delta(x-a)dx = f(a)$. One gets a sense of how this occurs from the integral representation

$$\delta(x) = \underset{K\to\infty}{Lim}\left(\frac{1}{2\pi}\int_{-K}^{K} e^{ixt}dt\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ixt}dt \tag{1.16.6}$$

identified in (1.16.5) by the horizontal bracket. The second equality expresses the familiar form one usually sees for the representation of the delta function. If the argument is not zero, then the integrand oscillates wildly with average value of 0. The proof that the foregoing representation satisfies the property of unit area is best accomplished by means of contour integration in the complex plane and will not be given here. To perform that integral rigorously, however, one must employ the correct representation of $\delta(x)$ as a limiting process expressed in the first equality.

In the calculation (1.16.5) of the binomial probability function, the Dirac delta function causes the right side of the equation to vanish for all values of the discrete summation index k except for $k = x$. It is therefore assuming the role of the discrete Kronecker delta δ_{kx}, which by definition equals 1 if $k = x$ and zero otherwise. There is no inconsistency here, however, because the inverse transform of the characteristic function is a probability *density*, and the Dirac delta function, which in general is a dimensioned quantity (with dimension equal to the reciprocal dimension of the integration variable) is required for the left-hand side of (1.16.5) to be a density, even though it is defined only for discrete values of x. In short, the method works, and we shall not worry about mathematical refinements to make the analysis more elegant, only to end up with the same result.

1.17 The uniform distribution

An idea of how rapidly the compounding of non-normal probability distributions can approach normality may be gleaned from examining the extreme case of the uniform distribution $U(a, b)$, in which the probability density of a random variable X

$$p_X(x|b,a) = \begin{cases} \dfrac{1}{b-a} & (b \geq x \geq a) \\ 0 & \text{otherwise} \end{cases} \tag{1.17.1}$$

is constant over the entire interval within which the variable can fall. The value of the constant is the reciprocal of the interval, as determined by the completeness relation. Use of pdf (1.17.1) leads to the moment-generating function

$$g_X(t) = \langle e^{xt} \rangle = (b-a)^{-1} \int_a^b e^{xt} dx = \frac{e^{bt} - e^{at}}{(b-a)t}. \tag{1.17.2}$$

The uniform distribution is perhaps one of very few distributions where it is considerably easier to determine statistical moments directly by integrating the pdf than by differentiating the mgf. Performing the integrations, we obtain

$$\begin{aligned} \mu_X = \langle X \rangle = \frac{1}{2}(b-a) \quad & \sigma_X^2 = \left\langle (X-\mu_X)^2 \right\rangle = \frac{1}{12}(b-a)^2 \\ \left\langle X^2 \right\rangle = \frac{1}{3}\left(b^2 + ab + a^2\right) \quad & Sk = \left\langle (X-\mu_X)^3 \right\rangle \Big/ \sigma_X^3 = 0 \\ & K = \left\langle (X-\mu_X)^4 \right\rangle \Big/ \sigma_X^4 = \frac{9}{5} = 1.8. \end{aligned} \tag{1.17.3}$$

Since the distribution is symmetric (being constant over the entire interval), the skewness is expected to vanish. The kurtosis turns out to be a number independent of the interval boundaries and much smaller than 3 (the value for a normal distribution) signifying a comparatively broader peak about the center, which is one way of looking at a completely flat distribution.

The difficulty with using the mgf for a uniform variate is that substitution of $t = 0$ into $g_X(t)$ and its derivatives leads to an indeterminate expression 0/0. In such cases, we must apply L'Hôpital's rule from elementary calculus to differentiate separately the numerator and denominator (more than once, if necessary) before taking the limit. Consider, for example, calculation of the mean

$$\begin{aligned} \mu_X = \frac{dg_X(t)}{dt}\Big|_{t=0} &= \left[\frac{be^{bt} - ae^{at}}{(b-a)t} - \frac{e^{bt} - e^{at}}{(b-a)t^2}\right]\Bigg|_{t=0} \\ &= \frac{b^2 - a^2}{(b-a)} - \left[\frac{be^{bt} - ae^{at}}{2(b-a)t}\right]\Bigg|_{t=0} = \frac{b^2 - a^2}{(b-a)} - \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{b+a}{2}. \end{aligned} \tag{1.17.4}$$

To avoid indeterminacy, the numerator and denominator of the second term in the second line had to be differentiated twice. Clearly, use of the mgf to determine moments of the uniform distribution is a tedious procedure to be avoided if possible. However, there are other uses, more pertinent to our present focus, in which the mgf is indispensable.

Suppose we want to determine the statistical properties of a random variable $Y = \sum_{i=1}^{n} X_i$, which is a sum of n independent random variables each distributed

uniformly over the unit interval, i.e. $X_i = U(0,1)$. Y, therefore, spans the range $(n \geq Y \geq 0)$. The mgf of Y – and correspondingly the characteristic function $h_Y(t) = g_Y(it)$ – are immediately deducible from (1.10.1)

$$g_Y(t) = \left(\frac{e^t - 1}{t}\right)^n \qquad \Rightarrow \qquad h_Y(t) = \left(\frac{e^{it} - 1}{it}\right)^n. \tag{1.17.5}$$

Although at this point we do not have the pdf of Y, we can determine the moments from the derivatives of $g_Y(t)$

$$\left.\begin{aligned} \langle Y \rangle &\equiv \mu_Y = \frac{n}{2} \\ \langle Y^2 \rangle &= \frac{n^2}{4} + \frac{n}{12} \\ \langle Y^3 \rangle &= \frac{n^3}{8} + \frac{n^2}{8} \\ \langle Y^4 \rangle &= \frac{n^4}{16} + \frac{n^3}{8} + \frac{n^2}{48} - \frac{n}{120} \end{aligned}\right\} \Rightarrow \begin{aligned} \sigma_Y^2 &= \frac{n}{12} \\ Sk &= 0 \\ K &= 3 - \frac{6}{5n}. \end{aligned} \tag{1.17.6}$$

As expected, the skewness vanishes and the kurtosis approaches 3 in the limit of infinite n. Moreover, expansion of $\ln g_Y(t)$ to order t^3 leads to an approximate mgf of Gaussian form

$$g_Y(t) \approx e^{\frac{n}{2}t + \frac{1}{2}\left(\frac{n}{12}\right)t^2} = e^{\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2} \tag{1.17.7}$$

in accordance with the Central Limit Theorem.

The CLT, however, does not tell us how rapidly a distribution approaches normal form. To ascertain this, we need the pdf $p_Y(y)$, which the characteristic function in (1.17.5) allows us to determine, by means of the Fourier transform,

$$p_Y(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} h_Y(t)e^{-iyt}dt = \frac{1}{(n-1)!}\sum_{0}^{[y]}(-1)^k\binom{n}{k}(y-k)^{n-1}. \tag{1.17.8}$$

I have used the symbol $[y]$ in the upper limit of the sum above to represent the greatest integer less than or equal to y. Recall that Y is a continuous random variable over the interval 0 to n, but the numbers in the binomial coefficient must be integers.

The calculation leading from the first equality to the second in (1.17.8) is most easily performed by contour integration in the complex plane and will be left to an appendix. To verify that $p_Y(y)$ satisfies the completeness relation, we calculate the cumulative distribution function

$$F_Y(y) = \int_0^y p_Y(y')dy' = \frac{1}{n!}\sum_{k=0}^{[y]}(-1)^k\binom{n}{k}(y-k)^n. \tag{1.17.9}$$

Completeness follows from the binomial identity

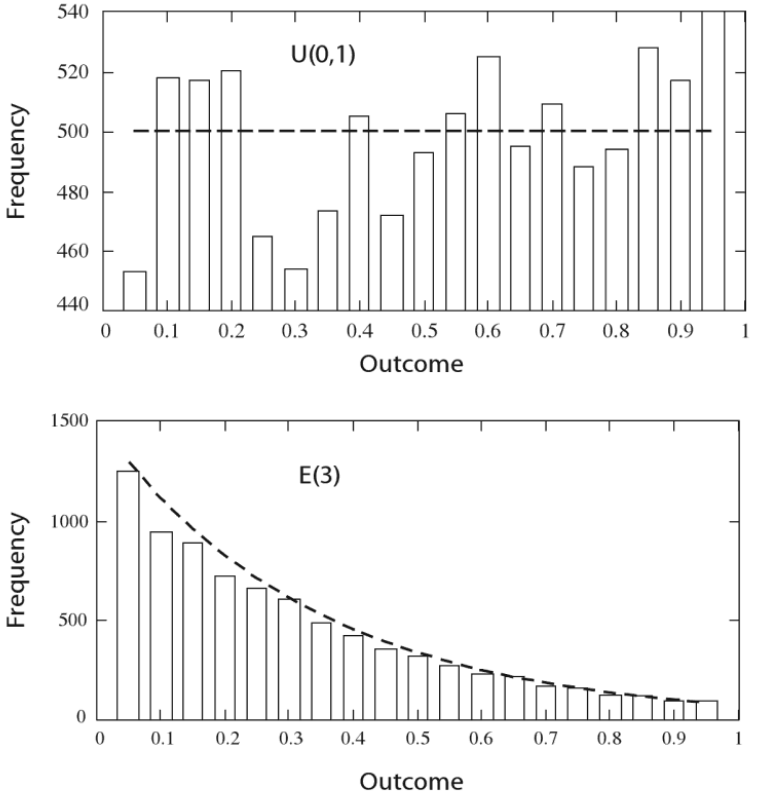


Fig. 1.3 Top panel: histogram of 10 000 samples from a $U(0, 1)$ random number generator. Lower panel: histogram of exponential variates $E(\lambda)$ generated by transformation (1.17.15) with parameter $\lambda = 3$. Dashed curves are theoretical densities.

To start with, consider a standard normal random variable $Z = N(0,1)$, for which the probability density is $p_Z(z) = (2\pi)^{-1/2} e^{-z^2/2}$. Under a transformation $W = Z^2$, the new pdf can be deduced by the following chain of steps

$$\int_0^\infty p_W(w)\, dw = \int_{-\infty}^{\infty} p_Z(z)\, dz = 2\int_0^\infty p_Z(z)\, dz = 2\int_0^\infty p_Z(z(w))\left|\frac{dz}{dw}\right| dw, \tag{1.18.1}$$

leading to

$$p_W(w) = \frac{2p_Z(z(w))}{|\frac{dw}{dz}|} = \frac{2(2\pi)^{-1/2} e^{-w/2}}{2^{1/2} w^{1/2}} = \frac{1}{2\sqrt{\pi}}\left(\frac{w}{2}\right)^{-1/2} e^{-w/2}, \tag{1.18.2}$$

which is identifiable as the pdf of a chi-square random variable of one degree of freedom, or, symbolically $W = \chi_1^2$. From the pdf above, the corresponding mgf, $g_W(t) = (1 - 2t)^{-1/2}$ is derivable by algebraically manipulating the integral occurring

in the expectation $\langle e^{Wt} \rangle$ into the form of the gamma function $\Gamma(1) = 1$. (See Eqs. (1.12.10) and (1.12.11).)

Given the mgf for a single variate Z^2, it follows immediately that the superposition of k independent random variables, $W = \sum_{i=1}^{k} Z_i^2$, each the square of a standard normal random variable, yields the mgf

$$g_W(t) = (1 - 2t)^{-k/2} \tag{1.18.3}$$

of a chi-square random variable of k degrees of freedom. We will take up the concept of degrees of freedom at the appropriate point, but for the present let us focus on the properties of the distribution, designated symbolically by χ_k^2.

From the derivatives of the mgf (1.18.3) one finds that the first four moments of a χ_k^2 random variable are

$$\begin{array}{ll} \mu_1 = k & \mu_3 = k^3 + 6k^2 + 8k \\ \mu_2 = k^2 + k & \mu_4 = k^4 + 12k^3 + 44k^2 + 48k \end{array} \tag{1.18.4}$$

and therefore

$$\text{mean } = k \qquad \text{var} = 2k \qquad Sk = \sqrt{\frac{8}{k}} \qquad K = 3 + \frac{12}{k}. \tag{1.18.5}$$

With increasing k, the skewness of the distribution function approaches 0 and the kurtosis approaches that of a standard normal variate.

The inverse Fourier transform of the characteristic function $h_W(t) = g_W(it)$ yields the pdf

$$p_W(w|k) = \frac{1}{2\Gamma\left(\frac{k}{2}\right)} \left(\frac{w}{2}\right)^{\frac{k}{2}-1} e^{-w/2}, \tag{1.18.6}$$

but this calculation, like that of the integral encountered in the previous section, also entails contour integration in the complex plane, and the demonstration will be left to an appendix. Figure 1.4 shows the variation in χ_k^2 density function (1.18.6) for a set of low degrees of freedom ($k = 1$–5) (upper panel) and a set of relatively high degrees of freedom ($k = 58$–66 in intervals of 4) (lower panel). For $k = 1$, the pdf is infinite at the origin although the area under the curve is of course finite. For $k = 2$, the curve is a pure exponential, as can be seen from the expression in (1.18.6). As k increases beyond 2, the plot approaches (although with slow convergence) the shape of a Gaussian pdf with mean k and variance $2k$.

Although ubiquitously used in its own right to test how well a set of data is accounted for by a theoretical expression, the chi-square pdf can also be considered a special case of a more general class of gamma distribution $Gam(\lambda, k)$ with defining probability density

$$p_X(x|\lambda, \kappa) = \frac{\lambda^\kappa}{\Gamma(\kappa)} \; x^{\kappa-1} \, e^{-\lambda x} \quad [(\lambda, \kappa) > 0]. \tag{1.18.7}$$

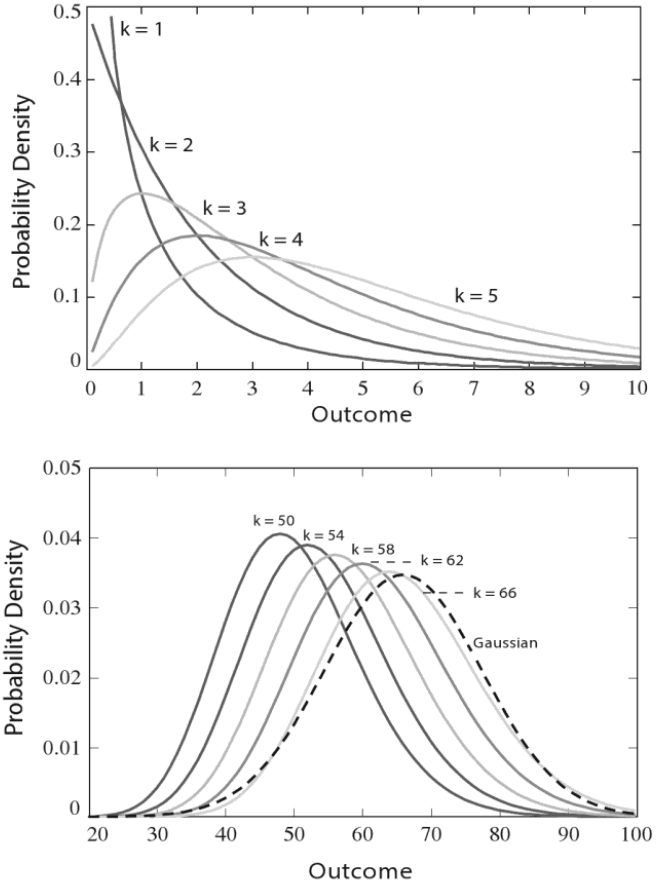


Fig. 1.4 Probability density of χ^2_k (solid) for low k (top panel) and high k (bottom panel). The dashed plot is the density of a normal variate $N(k, 2k)$ for $k = 66$.

and moment generating function

$$g_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\kappa} \quad (t < \lambda). \tag{1.18.8}$$

Looked at in this light – e.g. by comparison of mgfs – a chi-square random variable χ^2_k is a gamma random variable $Gam\left(\lambda = \frac{1}{2}, \kappa = \frac{k}{2}\right)$.

1.19 Student's *t* distribution

The "*t* distribution", published anonymously in 1908 by William Gossett under the pseudonym of "Student" (because his employer, the Guinness Brewery in Dublin, did

not permit employees to publish scientific papers), is the distribution of a random variable T constructed to be the ratio of a standard normal variate $U = N(0, 1)$ and an independent normalized chi-square variate $V^2 = \chi^2_d$ of d degrees of freedom. Specifically, one defines T by

$$T \equiv \frac{U}{\sqrt{V^2/d}} = \frac{U\sqrt{d}}{V}. \tag{1.19.1}$$

The motivation for this peculiar arrangement of random variables arises from its statistical application in testing the mean of a sample against a hypothesized mean of a normal distribution or in comparing two or more sample means to infer whether or not they are statistically equivalent to the mean of the same parent population. We will employ the t distribution in this way later in the book.

When testing a sample mean $\bar{x}$ against the theoretical mean μ of a parent population, it is often the case that the population variance σ^2 is not known although the variance s^2 of a sample of size n has been determined. One could, of course, estimate σ^2 by s^2 in implementing the test with a normal distribution, but the error incurred by this approximation can be significant for samples of small size. The Central Limit Theorem validates the ubiquitous occurrence of a normal distribution in the limit of a large (technically, infinite) number of samples. When used to make statistical inferences on small samples, however, the normal distribution gives probabilities that are too small because the tails of the distribution fall off (exponentially) too fast. In other words, the normal distribution can underestimate the probability of occurrence of outlying events that ordinarily have a low probability but which, when they occur, can prove catastrophic. The t distribution allows one to sidestep the problem of an unknown population variance in the following way.

If $X = N(\mu, \sigma^2)$ is a normal variate for which values $\bar{x}$ and s^2 have been obtained for the mean and variance by a random sample of size n, then the quantity

$$u = \frac{\bar{x} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \Rightarrow \quad U = N(0, 1) \tag{1.19.2}$$

is a realization of a random variable $U = N(0, 1)$. It is also demonstrable that the quantity

$$v^2 = \frac{ns^2}{\sigma^2} \quad \Rightarrow \quad V^2 = \chi^2_{n-1} \tag{1.19.3}$$

is a realization of an *independent* chi-square random variable $V^2 = \chi^2_{n-1}$. It may seem surprising that the distributions of s^2 and $\bar{x}$ are independent of one another since both quantities are calculated from the same set of data, but this demonstration – of both the independence and the type of distribution – can be found in advanced statistics books.[9] From (1.19.2) and (1.19.3) it follows that the ratio

[9] P. G. Hoel, *Introduction to Mathematical Statistics* (Wiley, New York, 1947) 136–138.

$$t = \frac{u\sqrt{n-1}}{v} = \frac{(\bar{x}-\mu)\sqrt{n-1}}{s} \xrightarrow[\mu=0]{} \frac{\bar{x}\sqrt{n-1}}{s} \tag{1.19.4}$$

does not contain the unknown population variance ... or the population mean, as well, if the parent population is hypothesized to have a mean of 0, a situation characterizing a "null test" (e.g. a test that some process has produced no effect distinguishable from pure chance).

The derivation of the pdf $p_T(t)$ from the component pdfs

$$\begin{aligned} p_U(u) &= \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \\ p_{V^2}(v^2) &= \frac{(v^2)^{\frac{d-2}{2}} e^{-v^2/2}}{2^{d/2}\Gamma(d/2)} \end{aligned} \tag{1.19.5}$$

proceeds easily if one ignores the constant factors – i.e. just designates all constant factors by a single symbol c – and focuses attention only on the variables. In a subsequent chapter I discuss the distribution of products and quotients of random variables more generally, but for the present the solution can be worked out by a straightforward transformation of variables. The idea is to

(a) start with the joint probability distribution $f_{UV^2}(u, v^2) = p_U(u)p_{V^2}(v^2)$,
(b) transform to a new probability distribution $f_{TV}(t, v)$ where $t = u\sqrt{d}/v$,
(c) integrate over v to obtain the marginal distribution $p_T(t)$ of t alone, and
(d) determine the normalization constant c from the completeness relation $\int f_T(t)dt = 1$.

Execution of steps (a) and (b) by means of the transformation

$$f_{TV}(t, v) = f_{UV}(u, v)\left|\frac{\partial(u, v)}{\partial(t, v)}\right| = \frac{f_{UV}(u, v)}{\left|\frac{\partial t}{\partial u}\right|} = \frac{v f_{UV}(u, v)}{\sqrt{d}} \tag{1.19.6}$$

leads to

$$f(t, v) = cv^d e^{-\frac{v^2}{2}\left(1+\frac{t^2}{d}\right)} \tag{1.19.7}$$

which by step (c) results in the marginal probability density

$$f(t) = c\left(1 + \frac{t^2}{d}\right)^{-\left(\frac{d+1}{2}\right)}. \tag{1.19.8}$$

The integral in step (d) is not elementary, but can be worked out by means of contour integration in the complex plane with use of the residue theorem. This calculation, deferred to an appendix, leads to the density

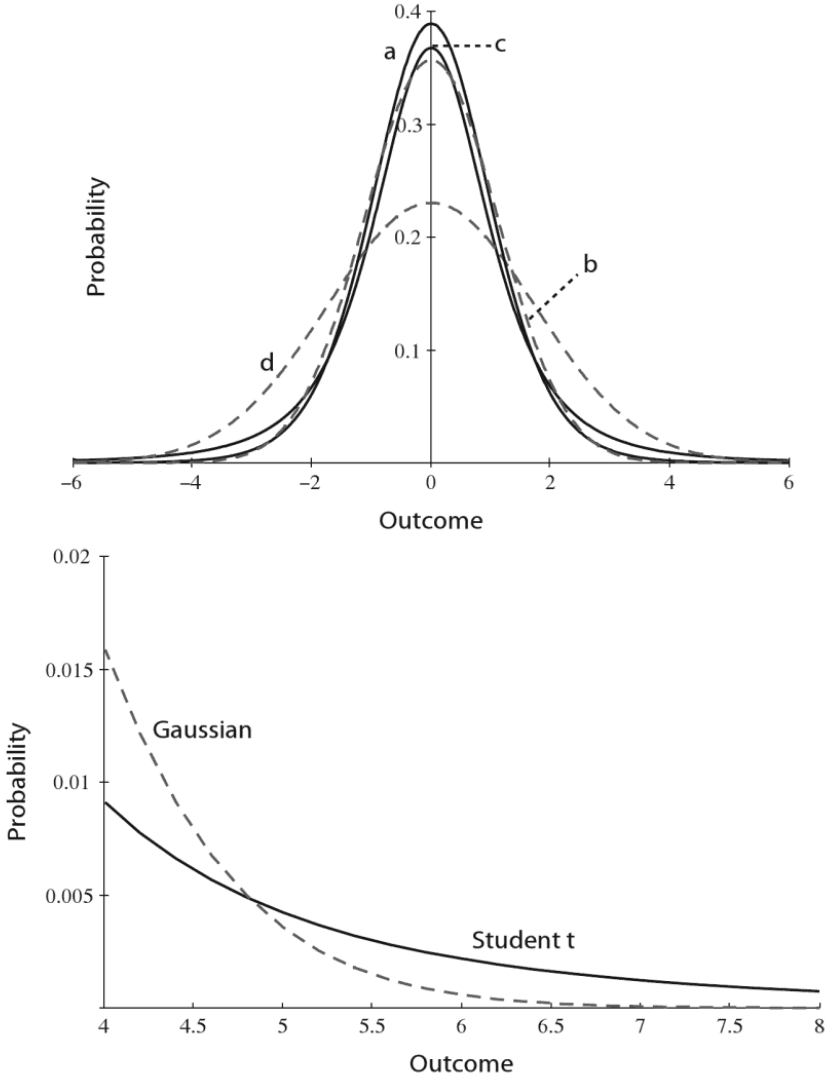


Fig 1.5 Top panel: Student t (solid) and Gaussian (dashed) densities for degrees of freedom $d = 10$ (plots (a), (b)) and $d = 3$ (plots (c), (d)). Bottom panel: tails of the Student t (solid) and Gaussian (dashed) densities for $d = 3$.

1.21 The principle of maximum entropy

Entropy, together with energy, constitutes one of the two pillars upon which the discipline of equilibrium thermodynamics – the study (broadly speaking) of the transformation of energy – rests. Einstein had once remarked upon the robust nature of thermodynamics in that if our theoretical understanding of the quantum structure of matter should ever fail entirely, the principles of thermodynamics would remain

valid and unaffected. This is so because thermodynamics is a consistent body of macroscopic relationships not tied to an underlying model of matter. That attribute is both its strength and its limitation.

The objective of a subject as vast in scope and application as thermodynamics is not easily reduced to a few words, but the following statement by Herbert Callen comes as close as any I have seen: "The basic problem of thermodynamics is the determination of the equilibrium state that eventually results after the removal of internal constraints in a closed composite system."[11] And how is one to determine that equilibrium state? The solution lies in the concept of entropy, a function of the extensive (i.e. size-dependent) variables of the system, which is itself additive over constituent subsystems. In the absence of an internal constraint, the values assumed by the extensive variables are those that maximize the entropy over the manifold of all equilibrium states which might have been realized while the constraints were in place. From this "entropy maximum postulate" plus a few definitions and some empirical relations (equations of state) describing how matter behaves, unfolds the mathematically elegant structure of equilibrium thermodynamics.

There is, however, a more fundamental statistical way to view the content of thermodynamics. It is, again in Callen's words[12], "the study of the macroscopic consequences of myriads of atomic coordinates, which, by virtue of the statistical averaging, do not appear explicitly in a macroscopic description of a system." From this statistical perspective, the concept of entropy is detached from the workaday measurable quantities of heat, work, temperature, and the like, and becomes instead a measure of the distribution of the elemental constitutents of a physical system over their available states. It is frequently said that entropy is a measure of order (or disorder) in a system – the greater the order, the lower the entropy – but this is an ambiguous relationship at best since there is no thermodynamic or statistical mechanical "order" function. Moreover, examples can be adduced that refute the association.[13]

In a thoroughly statistical treatment – which physicists generally refer to as "statistical mechanics" or "statistical thermodynamics", depending on emphasis – expressions for the mean values and fluctuations of macroscopic thermal quantities are derived from the characteristic energies (energy "eigenvalues") of the particles (nuclei, atoms, molecules ...) of the system and the probability distribution of the particles over their energy states (referred to as occupation probabilities). Out of this grand scheme, which *does* depend on our understanding of the atomic structure of matter, emerges a most remarkable expression for entropy

$$S = -k_B \sum_i p_i \ln p_i, \tag{1.21.1}$$

[11] H. B. Callen, *Thermodynamics* (Wiley, New York, 1960) 24. [12] Callen, *op. cit.* p. 7.
[13] K. G. Denbigh, "Note on Entropy, Disorder, and Disorganization", *The British Journal for the Philosophy of Science* **40** (1989) 323–332.

where the sum is over all states of the system. Apart from a universal constant (Boltzmann's constant k_B) chosen so that corresponding statistically and thermodynamically derived quantities agree, S depends explicitly only on the occupation probabilities. Implicitly, S is also a function of measurable physical properties of the system because the equilibrium probabilities themselves depend in general on the energy eigenvalues, the equilibrium temperature, and the chemical potential (which itself may be a function of temperature, volume, and number of particles in the system). Nevertheless, the connection between entropy and probability is striking. One can in fact interpret the expression for S as proportional to the expectation value of the logarithm of the occupation probability.

The identical expression, made dimensionless and stripped of all ties to heat, work, and energy, was proposed by Claude Shannon in 1948 as a measure of the uncertainty in information transmitted by a communications channel.[14] This was the key advance that, nearly ten years later, permitted Ed Jaynes, in one of the most fruitful and far-reaching reversals of reasoning I have seen, to develop an alternative way[15] of understanding and deriving all of equilibrium statistical mechanics from the concept of entropy as expressed by Shannon's information function

$$H = -\sum_i p_i \ln p_i. \tag{1.21.2}$$

As Jaynes described it:

> Previously, one constructed a theory based on the equations of motion, supplemented by additional hypotheses of ergodicity, metric transitivity, or equal *a priori* probabilities, and the identification of entropy was made only at the end, by comparison of the resulting equations with the laws of phenomenological thermodynamics. Now, however, we can take entropy as our starting concept, and the fact that a probability distribution maximizes the entropy subject to certain constraints becomes the essential fact which justifies use of that distribution for inference.

The significance of Jaynes' perspective was the realization that the *structure* of statistical mechanics did not in any way depend on the details of the physics it described. Rather, it was a consequence of a general form of pure mathematical reasoning that could be employed on countless problems totally unrelated to thermodynamics. In particular, this mode of reasoning – subsequently termed the principle of maximum entropy (PME) – can be used to answer Question I: What is the most unbiased probability distribution that takes account of known information but makes no further speculations or hypotheses? We have seen how the Central Limit Theorem explains the apparently ubiquitous occurrence of the normal distribution. The PME, as will be demonstrated, provides another reason.

[14] C. E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal **27** (1948) 379–423, 623–656.

[15] E. T. Jaynes, "Information Theory and Statistical Mechanics", *Physical Review* **106** (1957) 620–630; "Information Theory and Statistical Mechanics II, *Physical Review* **108** (1957) 171–190.

1.22 Shannon entropy function

Before examining the PME, it is instructive to see how the Shannon (or statistical) entropy function (1.21.2) satisfies the properties one would expect of both entropy, which is an extensive physical quantity, and probability. If A and B are two independent physical systems, then the total entropy of the combined system is additive: $H = H_A + H_B$. By contrast, if $p_A(i)$ is the probability of occurrence of state i in system A and $p_B(j)$ the probability of occurrence of state j in system B, then the probability that the two independent states occur simultaneously is multiplicative: $p(i, j) = p_A(i)p_B(j)$.

That the statistical entropy of the combined system behaves this way may be seen as follows

$$\begin{aligned} H &= -\sum_{i,j} p(i,j)\, ln(p(i,j)) = -\sum_{i,j} p_A(i)\, p_B(j)\, ln\,(p_A(i)\, p_B(j)) \\ &= -\underbrace{\sum_{j} p_B(j)}_{=1} \sum_{i} p_A(i)\, ln\, p_A(i) - \underbrace{\sum_{i} p_A(i)}_{=1} \sum_{j} p_B(j)\, ln\, p_B(j) \\ &= H_A + H_B, \end{aligned} \tag{1.22.1}$$

where the completeness relation was used to reduce the sums above the horizontal brackets to unity. No other functional form has this property.

1.23 Entropy and prior information

To implement the PME to find an unknown probability distribution in a specific problem one maximizes H subject to constraints posed by any prior information about the system being studied. In the simplest cases, each constraint is introduced as an algebraic expression multiplied by an unknown factor known as a Lagrange multiplier. The entire procedure is actually a fairly routine application of a branch of mathematics known as the calculus of variations. Whereas in standard calculus one finds the maximum or minimum values of a function, in the calculus of variations one seeks a *function* that yields the extremum of a "*functional*".

1.23.1 No prior information

Consider first the simplest case of a discrete system with n states $\{x_i\ i = 1 \ldots n\}$ (or, equivalently, a stochastic process with n possible outcomes per trial), each with a probability of occurrence p_i. If we have no prior information at all about the probability distribution, other than that it must satisfy the completeness relation $\sum_{i=1}^{n} p_i = 1$, then the most unbiased entropy functional we can write takes the form

$$H = -\sum_{i=1}^{n} p_i \ln p_i - \lambda \left(1 - \sum_{i=1}^{n} p_i\right) \tag{1.23.1}$$

in which λ is a Lagrange multiplier. Seeking the extremum of H by setting the derivative $\partial H/\partial p_j$ (for all j) to zero, leads to the uniform distribution $p_j = e^{-(1+\lambda)}$, which, upon substitution into the completeness relation, gives $p_j = 1/n$. In other words, if nothing is known beforehand about the system or process, then the most unbiased distribution is one in which all outcomes are equally probable. This choice, made intuitively (rather than derived systematically from an overarching principle) by early developers of probability theory such as Laplace and Bayes, has been termed the "principle of insufficient reason" or "principle of indifference".

There are subtle, yet profound, issues connected with the question of how to frame mathematically the proposition that one knows nothing about a system (... what exactly is "nothing"? ...) that have led to much of the fireworks between Bayesians and frequentists. For now, let us sidestep the matter and examine a problem at the next level of complexity.

1.23.2 Prior information is a single mean value

Consider the same system as before except that now, in addition to the completeness relation, we have as prior information the mean value F of some function $f(x)$ of the states

$$F = \langle f \rangle = \sum_{i=1}^{n} p_i f(x_i) \equiv \sum_{i=1}^{n} p_i f_i. \tag{1.23.2}$$

Finding the extremum of the entropy functional

$$H = -\sum_{i=1}^{n} p_i \ln p_i - \lambda_0 \left(1 - \sum_{i=1}^{n} p_i\right) - \lambda_1 \left(1 - \sum_{i=1}^{n} p_i f_i\right), \tag{1.23.3}$$

which now contains two Lagrange multipliers, one for each constraint, leads to an exponential distribution

$$p_j = e^{-(1+\lambda_0)} e^{-\lambda_1 f_j} = \frac{e^{-\lambda_1 f_j}}{Z(\lambda_1)}, \tag{1.23.4}$$

where the second equality, obtained by substitution of the first expression into the completeness relation, displays the so-called partition function

$$Z(\lambda_1) = \sum_{i=1}^{n} e^{-\lambda_1 f_i}. \tag{1.23.5}$$

The value of the Lagrange multiplier λ_1 is determined (implicitly) from the second constraint

and to probabilities

$$p_1 = \left(\frac{f_2 - F}{f_2 - f_1}\right) \qquad p_2 = \left(\frac{F - f_1}{f_2 - f_1}\right). \tag{1.23.19}$$

Note that once the partition function is expressed in terms of the mean values of observables, then one cannot calculate moments, as in Eq. (1.23.7), simply by taking derivatives of Z with respect to the Lagrange multipliers. In that case, the straightforward thing to do is construct the moment-generating function, which in the present case becomes

$$g(t) = \left\langle e^{ft} \right\rangle = p_1 e^{f_1 t} + p_2 e^{f_2 t} \tag{1.23.20}$$

and readily generates the moments

$$\begin{aligned} \langle f \rangle &= F \\ \sigma^2 &= \left\langle f^2 \right\rangle - \langle f \rangle^2 = (f_2 - F)(F - f_1). \end{aligned} \tag{1.23.21}$$

1.23.5 Prior information is mean and variance

As a final illustration of the maximum entropy principle, consider the original system again where now our prior information comprises the completeness relation and both the first (μ_1) and second (μ_2) moments of the observable quantity, which is itself the variable X. The three equations of constraint are embedded in the entropy functional by means of three Lagrange multipliers, leading to

$$H = -\sum_{i=1}^{n} p_i \ln p_i - \lambda_0 \left(1 - \sum_{i=1}^{n} p_i\right) - \lambda_1 \left(\mu_1 - \sum_{i=1}^{n} p_i x_i\right) - \lambda_2 \left(\mu_2 - \sum_{i=1}^{n} p_i x_i^2\right). \tag{1.23.22}$$

However, this is not the most convenient form in which to find the extremum. Often (perhaps even most often) the analyst's interest is in moments about the mean. There is no loss of generality, then, in defining the Lagrange multipliers differently in order to rewrite the entropy functional in a way that reflects that interest

$$H = -\sum_{i=1}^{n} p_i \ln p_i - \lambda_0 \left(1 - \sum_{i=1}^{n} p_i\right) - \lambda'_1 \left(0 - \sum_{i=1}^{n} p_i (x_i - \mu)\right) - \frac{1}{2} \lambda'_2 \left(\sigma^2 - \sum_{i=1}^{n} p_i (x_i - \mu)^2\right). \tag{1.23.23}$$

For notational simplicity I dropped the subscript 1 from the label of the first moment and combined the prior information to form a variance $\sigma^2 = \mu_2 - \mu_1^2$. Since the sum in the second bracket vanishes identically (by virtue of the expression in the first bracket) irrespective of the probability distribution, it provides no new information and therefore one loses nothing in simply setting λ'_1 to zero. The procedure to maximize the reduced entropy functional

$$H = -\sum_{i=1}^{n} p_i \ln p_i - \lambda_0 \left(1 - \sum_{i=1}^{n} p_i\right) - \frac{1}{2}\lambda'_2 \left(\sigma^2 - \sum_{i=1}^{n} p_i (x_i - \mu)^2\right) \tag{1.23.24}$$

immediately yields a discrete probability distribution

$$p_j \propto e^{-\lambda'_2 (x_j - \mu)^2 / 2} \rightarrow p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2 / 2\sigma^2} \tag{1.23.25}$$

which, when transformed to an appropriately normalized continuous distribution, becomes the normal distribution $N(\mu, \sigma^2)$.

In summary, illustrations of the principle of maximum entropy show that

(a) a *uniform* distribution (principle of indifference) results when one has no prior information beyond the requirement that the total probability is unity;
(b) an *exponential* distribution, such as those that occur in statistical physics (e.g. Maxwell–Boltzmann, Fermi–Dirac, Bose–Einstein), results when the prior information consists of the mean values of functions of some stochastic quantity; and
(c) a *Gaussian* or *normal* distribution results when the prior information consists of the first and second moments (or the first moment and variance) of some stochastic quantity.

Under the assumed conditions in each case, the use of any other probability distribution would imply that either more information was known at the outset or that the analyst has incorporated into the analysis an element of unjustified speculation.

1.24 Method of maximum likelihood

Two principal tasks of statistics are to test hypotheses and to estimate physical quantities from data. Let us suppose that the data – referred to in statistics as the sample – are the outcomes of n independent observations, each regarded as an independent, identically distributed (iid) random variable $X_i (i = 1 \ldots n)$ with probability density (or in the discrete case a probability function) $f(x|\theta)$. In many cases it is the parameter (or set of parameters) θ upon which the pdf depends, that is to be estimated. The task of estimation, then, is to extract from the statistics of a sample the "true" values of quantities characteristic of the full population. This population may be a real one as, for example, in the census of a nation in which the total number of people is generally too large for each person to be queried; hence a representative random sample of people is selected for questioning. However, a set of repeated measurements of the mass of an elementary particle can be imagined to be a sample drawn from a hypothetical infinitely large population (or "ensemble") of potential measurements executed under equivalent conditions.

The ensemble mode of thinking is the point of view of orthodox statistics and the basis of statistical mechanics as developed by J. Willard Gibbs, which is the approach ordinarily taught in statistical mechanics courses. There is an alternative point of

view based on Bayes' theorem, which dispenses with the philosophical encumbrance of ensembles and focuses exclusively on the data to hand, not those that did not materialize. This divergence of thought constitutes one of the battlefronts in the probability wars alluded to at the beginning of the chapter. Estimates based on the two approaches do not always turn out to be the same. (Indeed, estimates made by different orthodox procedures, do not necessarily turn out to be the same either.) Philosophy aside, the differences between orthodox and Bayesian estimates derive principally from what one does with the likelihood function. I will come back to this point later in the chapter.

From the orthodox perspective, the likelihood function of n independent random variables is defined as their joint probability density. Thus, if $\{x_i \; i = 1 \ldots n\}$ is a realization of the set of random variables introduced above, the corresponding likelihood function would be

$$L(\theta|\{x_i\}) = f(x_1|\theta)\, f(x_2|\theta) \ldots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{1.24.1}$$

where, in the general case, θ may stand for a set of parameters. The method of maximum likelihood (ML), due primarily to geneticist and statistician R. A. Fisher[16], may be expressed somewhat casually as follows: The best estimate (usually) of the parameter θ is the value $\hat{\theta}$ that maximizes the likelihood $L(\theta|\{x_i\})$. This immediately raises the question of what is meant by "best".

It is said that a spoken language has many words of varying nuances for something of particular importance in the culture of the people who speak the language. If that is true, then the concept of "estimate" is to a statistician what the perception of "snow" is to an Eskimo (... or perhaps to a meteorologist). To start with, the statistician distinguishes between an "estimator" Θ, which is a random variable used to estimate some quantity, and the "estimate" θ, which is a value that the estimator can take. The orthodox statistician considers the quantity to be estimated to have a fixed, but unknown, value, whereas the estimates of the estimator are governed by some probability density function of supposedly finite mean and variance. The goal of estimation is therefore to find an estimator whose expectation value yields the sought-for parameter with the least uncertainty possible. With those points in mind:

- An estimator is "unbiased" if its expectation value $\langle\Theta\rangle$ equals the estimated parameter θ.
- An estimator is "close" if its distribution is concentrated about the true value of the parameter with small variance.
- An estimator is "consistent" if the value of the estimation gets progressively closer to the estimated parameter as the sample size increases.

[16] R. A. Fisher, "Theory of Statistical Estimation", *Proceedings of the Cambridge Philosophical Society* **22** (1925) 700–725.

- An estimator is "minimum-variance unbiased" if the variance of its pdf is the lowest of all unbiased estimators. There is, in fact, a lower bound, known as the Cramér–Rao theorem, to the variance of an estimator that meets certain reasonable conditions regarding existence of the first and second derivatives of the logarithm of the likelihood function.
- An estimator is "asymptotically normal" if its pdf approaches that of a normal distribution with increasing sample size.
- An estimator is deemed "efficient" if, among a set of consistent, asymptotically normal estimators of the same quantity, it has the minimum variance.
- And last (for our purposes), but of particular utility, is sufficiency, a concept also due to Fisher. A statistic S is "sufficient" in regard to an unknown parameter if it condenses the data (i.e. the sample) so as to contain all the information that the sample can provide for estimation of that parameter. In other words, having the single sufficient statistic, one cannot learn anything further about the unknown parameter by knowing the individual values of the sample or by seeking other estimators. Clearly, it is desirable that an estimator be a function of sufficient statistics.

With this basic vocabulary, one can say of ML estimators that some are uniformly minimum-variance unbiased, while others are not; that a sequence of ML estimators is consistent and asymptotically normal with a variance equal to the Cramér–Rao lower bound; and that, if a sufficient statistic exists for the parameter to be estimated (which is not always the case), the ML estimator must be a function of it. All in all, for large sample size the ML estimate of θ is about as good as one may hope to find – although there may be others just as good.

From the perspective of a practical physicist, an especially attractive feature of the ML method is the facility with which it delivers both the estimate and its uncertainty. Noting that it is often easier to work with the logarithm of a sequential product of functions (as in Eq. (1.24.1)) and that a function and its log are maximized at the same point, we consider

$$\mathcal{L} \equiv \ln L = \sum_{i=1}^{n} \ln(f(x_i|\theta)), \tag{1.24.2}$$

a quantity that some statisticians have termed the "support function", but which I will refer to simply as the log-likelihood. In the general case of m parameters $\{\theta_1 \ldots \theta_m\}$ one must then solve the set of equations

$$\frac{\partial L}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial f(x_i|\theta)/\partial \theta_j}{f(x_i|\theta)} = 0 \qquad (j = 1 \ldots m). \tag{1.24.3}$$

The variance of each ML estimate and covariance of pairs of estimates are given by the elements of a covariance matrix $\mathbf{C} = -\mathbf{H}^{-1}$, where $C_{jj} = \sigma^2_{\theta_j}$, $C_{jk} = \mathrm{cov}\,(\theta_j, \theta_k)$ are derived from the second derivatives of the log-likelihood

$$(\mathbf{H})_{jk} \equiv H_{jk} = \left(\frac{\partial^2 \mathcal{L}}{\partial\theta_j \partial\theta_k}\right)_{\hat{\theta}} \qquad (j,k = 1 \ldots m). \tag{1.24.4}$$

The symbol $\hat{\theta}$ appended to the bracket signifies that the second derivatives are to be evaluated by substitution of the ML values of the parameters $\{\hat{\theta}_j\}$.

The preceding method for estimating uncertainty of the parameters follows straightforwardly from the structure and interpretation of the log-likelihood function expanded in a Taylor series about the ML values of its argument. For simplicity, consider the example of two parameters:

$$\begin{aligned}
\mathcal{L} \equiv \ln L(\theta_1,\theta_2) &= \mathcal{L}\left(\hat{\theta}_1,\hat{\theta}_2\right) + \sum_{i=1}^{2}\left(\frac{\partial\mathcal{L}}{\partial\theta_i}\right)_{\hat{\theta}}\left(\theta_i - \hat{\theta}_i\right) + \frac{1}{2}\sum_{i,j=1}^{2}\left(\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}\right)_{\hat{\theta}}\left(\theta_i - \hat{\theta}_i\right)\left(\theta_j - \hat{\theta}_j\right) + \cdots \\
&= \mathcal{L}\left(\hat{\theta}_1,\hat{\theta}_2\right) + \frac{1}{2}\sum_{i,j=1}^{2} H_{ij}\left(\theta_i - \hat{\theta}_i\right)\left(\theta_j - \hat{\theta}_j\right) + \cdots \\
&= \mathcal{L}\left(\hat{\theta}_1,\hat{\theta}_2\right) + \frac{1}{2}\mathbf{U}^T\mathbf{H}\mathbf{U} + \cdots \;=\; \mathcal{L}\left(\hat{\theta}_1,\hat{\theta}_2\right) - \frac{1}{2}\mathbf{U}^T\mathbf{C}^{-1}\mathbf{U} + \cdots.
\end{aligned} \tag{1.24.5}$$

In the first line of the expansion, the term involving a sum over first derivatives of $\mathcal{L}$ vanishes by virtue of the ML maximization procedure. The second line shows the reduced expression with matrix elements of $\mathbf{H}$ substituted for the second derivatives of $\mathcal{L}$. The third line shows the equivalent expression in terms of the parameter vector

$$\mathbf{U} = \begin{pmatrix} \theta_1 - \hat{\theta}_1 \\ \theta_2 - \hat{\theta}_2 \end{pmatrix} \tag{1.24.6}$$

(and its transpose $\mathbf{U}^T$) and the inverse of the covariance matrix $\mathbf{C}$

$$\mathbf{C} \equiv \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \tag{1.24.7}$$

where the correlation coefficient is defined by $\rho \equiv \rho_{12} = \frac{\mathrm{cov}\left(\hat{\theta}_1,\hat{\theta}_2\right)}{\sigma_1\sigma_2}$. The matrices $\mathbf{H}$ and $\mathbf{C}$ are related as follows

$$\mathbf{H} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} = -\mathbf{C}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} -1/\sigma_1^2 & \rho/\sigma_1\sigma_2 \\ \rho/\sigma_1\sigma_2 & -1/\sigma_2^2 \end{pmatrix}. \tag{1.24.8}$$

Upon neglect of derivatives higher than second, the likelihood function then becomes proportional to the negative exponential of a quadratic form

$$L\left(\hat{\theta}_1,\hat{\theta}_2|D\right) \propto e^{-\frac{1}{2}\mathbf{U}^T\mathbf{C}^{-1}\mathbf{U}}, \tag{1.24.9}$$

which is recognized as a multivariable Gaussian function of the ML parameters $\left(\hat{\theta}_1,\hat{\theta}_2\right)$ and data D. For a single variable, the exponential (1.24.9)

the negative inverse of which yields the covariance matrix whose elements constitute the variances of the ML parameters

$$\mathrm{var}(\hat{\mu}) = \frac{\hat{\sigma}^2}{n} \tag{1.24.23}$$

$$\mathrm{var}\left(\hat{\sigma}^2\right) = \frac{2\hat{\sigma}^4}{n} \tag{1.24.24}$$

with zero covariance. This means that the ML estimators derived above are independent, asymptotically normal random variables of the forms

$$\Theta_1 = N\left(\hat{\mu}, \frac{\hat{\sigma}^2}{n}\right) \qquad \Theta_2 = N\left(\hat{\sigma}^2, \frac{2\hat{\sigma}^4}{n}\right). \tag{1.24.25}$$

Note, as pointed out previously, that the variance of the mean is smaller than the variance of a single observation by the factor n [a relation also contributing to Eq. (1.24.21)].

The property of normality and the variance (1.24.23) of the ML estimator $\overline{X}$ are actually valid statements irrespective of the size n of the sample. However, the exact variance of the ML estimator S'^2 can be shown to be $2\sigma^4(n-1)/n^2$, which asymptotically reduces to the expression in (1.24.24). The explanation for this is that the exact distribution of the variance of the sample mean, $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, which is proportional to a form $\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma/\sqrt{n}}\right)^2$ constructed to be the sum of the squares of n standard normal random variables, is not Gaussian, but a chi-square distribution χ^2_{n-1}. There are $n-1$, rather than n, degrees of freedom because the sample mean $\overline{X}$ is itself calculated from the data and, once known, signifies that only $n-1$ of the set of variates $\{X_i\}$ are independent.

One last point of interest in regard to the variances of the ML estimates for μ and σ^2 is to see how they compare with the lower bound of the Cramér–Rao theorem, which can take either of the two forms below for an estimate of a function $\tau(\theta)$.

$$\mathrm{var}(\tau(\theta))_{\mathrm{CR}} = \frac{(d\tau/d\theta)^2}{n\left\langle(\partial \log f(X|\theta)/\partial\theta)^2\right\rangle_\theta} = \frac{-(d\tau/d\theta)^2}{n\left\langle\partial^2 \log f(X|\theta)/\partial\theta^2\right\rangle_\theta}. \tag{1.24.26}$$

Since $\tau(\theta) = \theta$ in this case, the derivative in the numerator becomes 1. Given a Gaussian pdf with natural logarithm

$$\ln f\left(x|\mu,\sigma^2\right) = -\frac{1}{2}\ln\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi), \tag{1.24.27}$$

the first equality of (1.24.26) reduces to the expressions

$$\mathrm{var}(\mu) = \frac{1}{n\left\langle \frac{(X-\mu)^2}{\sigma^4} \right\rangle} = \frac{\sigma^2}{n} \tag{1.24.28}$$

$$\mathrm{var}(\sigma^2) = \frac{1}{n\left\langle \left(-\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4} \right)^2 \right\rangle} = \frac{4\sigma^4/n}{\left\langle 1 - 2\left(\frac{X-\mu}{\sigma}\right)^2 + \left(\frac{X-\mu}{\sigma}\right)^4 \right\rangle} = \frac{2\sigma^4}{n}, \tag{1.24.29}$$

where use was made of the expectations $\langle Z^2 \rangle = 1$ and $\langle Z^4 \rangle = 3$ of the standard normal variable $Z = (X - \mu)/\sigma$. Comparison with (1.24.23) shows that the ML variances of the Gaussian parameters are as small as theoretically possible. The same minimum variances would have been obtained had we used the second equality in (1.24.26).

1.25 Goodness of fit: maximum likelihood, chi-square, and *P*-values

Suppose we have made n observations of some randomly varying quantity X that at each observation could take any one of K values $\{A_k \; k = 1 \ldots K\}$. We have, therefore, a multinomial distribution of frequencies $\{n_k\}$ of outcomes sorted into K classes with the constraint $\sum_{k=1}^{K} n_k = n$ and probability function

$$\Pr(\{n_k\}|\{p_k\}) = n! \prod_{k=1}^{K} \left(\frac{p_k^{n_k}}{n_k!} \right) \tag{1.25.1}$$

for the totality of n trials. In general, apart from the completeness relation $\sum_{k=1}^{K} p_k = 1$, we might not know the probability p_k for an outcome to take the value A_k, but we can do two things: (a) estimate the maximum likelihood (ML) probabilities from the frequency data, and (b) make a theoretical model of the random process that has generated the data. Consider first the ML estimate.

In the case of a large sample size n, the log-likelihood function of the multinomial expression (1.25.1) can be written and simplified as shown below

$$\begin{aligned} \mathcal{L} = \ln L &= \ln\left(n! \prod_{k=1}^{K} \frac{p_k^{n_k}}{n_k!} \right) = \sum_{k=1}^{K} n_k \ln p_k - \sum_{k=1}^{K} \ln n_k! + \ln n! \\ &= \sum_{k=1}^{K} n_k \ln p_k - \sum_{k=1}^{K} n_k \ln n_k + n \ln n, \end{aligned} \tag{1.25.2}$$

where we have approximated the natural log of a factorial $n!$ by the two largest terms [$\ln n! \sim n \ln n - n$] in Stirling's approximation

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \cdots\right). \tag{1.25.3}$$

To maximize $\mathcal{L}$ with respect to the set of parameters $\{p_k\}$ given only the completeness relation, we introduce a single Lagrange multiplier to form the functional

$$\mathcal{L}' = \sum_{k=1}^{K} n_k \ln p_k + \lambda\left(1 - \sum_{k=1}^{K} p_k\right) \tag{1.25.4}$$

with omission of all terms not containing the parameters since they would vanish anyway from the ML equations

$$\frac{\partial \mathcal{L}'}{\partial p_k} = \frac{n_k}{p_k} - \lambda = 0 \qquad (k = 1 \ldots K). \tag{1.25.5}$$

Substitution of the solution $\hat{p}_k = n_k/\lambda$ into the completeness relation leads to $\lambda = n$ and therefore to the ML estimates

$$\hat{p}_k = \frac{n_k}{n}. \tag{1.25.6}$$

It is worth stressing that the set of probability parameters $\{\hat{p}_k\}$ arrived at by the foregoing procedure give the largest value to the likelihood function (1.25.2); no alternative set of probabilities yield a larger value.

Suppose now we were to model the random process by some probability function $f(x|\theta)$, which depends on parameters θ whose values may be unknown at the outset. Let $f_k \equiv f(A_k|\theta)$ be the hypothesized probability that an observation results in the outcome A_k. We need some way to estimate the optimum set of parameters for the given model – referred to in statistical parlance as the "null hypothesis" – and then ascertain whether the model credibly accounts for the data. As before, a suitable way to do this would be to calculate the ML estimates $\hat{\theta}$ of the parameters and then compare the likelihood of the model $L\big(\{n_k\}|\{\hat{f}_k\}\big)$ to the maximum likelihood $L(\{n_k\}|\{\hat{p}_k\})$ attainable by any alternative model. Substitution of the ML estimates $\{\hat{p}_k\}$ of Eq. (1.25.6) into Eq. (1.25.1) leads to a relatively simple expression for the ratio of the two likelihood functions

$$\frac{L_0}{L_{\max}} \equiv \frac{L(\{n_k\}|\{\hat{f}_k\})}{L(\{n_k\}|\{\hat{p}_k\})} = \prod_{k=1}^{K}\left[\frac{\hat{f}_k}{(n_k/n)}\right]^{n_k} = n^n \prod_{k=1}^{K}\left[\frac{\hat{f}_k}{n_k}\right]^{n_k} \tag{1.25.7}$$

because the products of factorials in numerator and denominator cancel. The log of the ratio then yields a relation

$$\ln\left(\frac{L_0}{L_{\max}}\right) = n\ln n + \sum_{k=1}^{K} n_k \ln\left(\frac{\hat{f}_k}{n_k}\right) = \sum_{k=1}^{K} n_k\left[\ln n + \ln\left(\frac{\hat{f}_k}{n_k}\right)\right] = -\sum_{k=1}^{K} n_k \ln\left(\frac{n_k}{n\hat{f}_k}\right) \tag{1.25.8}$$

from which one can calculate how "likely" the null hypothesis is in comparison to the maximum likelihood.

An advantage to the use of the likelihood ratio for comparison of two hypotheses or models is that it is invariant under a transformation of parameters. For example, if you wanted to test whether the parameter θ_1 or θ_2 characterized a set of data believed to be drawn from a distribution with pdf $\propto e^{-x/\theta^2}$, the likelihood ratio would be the same if, instead, you transformed the distribution by $\phi = \theta^2$ and then tested for parameters ϕ_1 and ϕ_2. The example is a trivial one, but the conclusion still holds in the general case of more complicated transformations of a multi-component parameter vector. The reason for the invariance is that the likelihood ratio is a value at a point, rather than an integral over a range.

That same asset can become a disadvantage, however, to using Eq. (1.25.8) for inference because the distribution function associated with the likelihood ratio in specific cases may be difficult or impossible to determine – and so to say that one model is 50% as likely as another does not tell us how *probable* either is. The "power" of a statistical test of inference is defined to be the probability of rejecting a hypothesized model when it is correct – i.e. when the parameters of the model are the "true" but unknown parameters of the distribution from which the data were obtained. A test is the more powerful if it can reject the null hypothesis with a lower probability of making a false judgment. In a significance test of a model, an ideal power function would be 0 if the parameters of the model corresponded to the true parameters, and 1 otherwise. In general, the likelihood function is not a probability but a conditional probability density, a fact that is a virtue to some and a liability to others.

With the adoption of a few approximations and some algebraic rearrangements, the final expression in (1.25.8) can be worked into a form with a known distribution irrespective of the null hypothesis. To see this, start by

(a) adding and subtracting 1 in the argument of the logarithm,
(b) adding and subtracting $n\hat{f}_k$ in the pre-factor, and
(c) dividing and multiplying the entire summand by $n\hat{f}_k$

$$\ln\left(\frac{L_0}{L_{\max}}\right) = -\sum_{k=1}^{K} n\hat{f}_k \left(\frac{n_k - n\hat{f}_k + n\hat{f}_k}{n\hat{f}_k}\right) \ln\left(1 + \frac{n_k}{n\hat{f}_k} - 1\right)$$
$$= -\sum_{k=1}^{K} n\hat{f}_k (1 + \Delta_k) \ln(1 + \Delta_k) \tag{1.25.9}$$

so as to express the log-likelihood ratio in terms of a quantity $\Delta_k \equiv \frac{n_k - n\hat{f}_k}{n\hat{f}_k}$, expected to be small if the null hypothesis is credible. Expanding (1.25.9) in a Taylor series in Δ_k

$$\ln\left(\frac{L_0}{L_{\max}}\right) = -\sum_{k=1}^{K} n\hat{f}_k \left(\Delta_k + \frac{1}{2}\Delta_k^2 - \frac{1}{6}\Delta_k^3 + \cdots\right), \tag{1.25.10}$$

recognizing that the linear term vanishes identically

$$\sum_{k-1}^{K} n\hat{f}_k \Delta_k = \sum_{k=1}^{K} \left(n_k - n\hat{f}_k\right) = n - n\sum_{k-1}^{K} \hat{f}_k = n - n = 0, \tag{1.25.11}$$

and truncating after the quadratic term, we obtain an expression

$$\mathcal{L} \equiv \ln\left(\frac{L_0}{L_{\max}}\right) \approx -\frac{1}{2}\sum_{k=1}^{K} \frac{\left(n_k - n\hat{f}_k\right)^2}{n\hat{f}_k} \equiv -\frac{1}{2}\,\chi_d^2 \tag{1.25.12}$$

identified as a sum of K chi-square random variables of some number d of degrees of freedom to be specified momentarily.

The justification for the interpretation derives from unstated assumptions that (a) the number of observations n and classes K are both reasonably large (with $n \gg K$), in which case (b) the probability $\hat{f}_k$ of a particular outcome A_k is fairly small and approximately Poissonian, whereupon (c) $n\hat{f}_k$ is an acceptable measure of the variance of frequency N_k whose realization is the observed n_k. We have seen previously in Eq. (1.13.7) that a multinomial distribution – such as we have begun with in (1.25.1) – results from the conditional probability of observing K independent Poisson variates whose sum is a fixed quantity, a connection first pointed out by Fisher.

If these assumptions hold, then $-2\mathcal{L}$ in Eq. (1.25.12) corresponds to a sum of the squares of K standard normal variates $Z_k^2 = (N_k - \langle N_k \rangle)^2/\sigma_{N_k}^2$, which, if all are independent, would be equivalent to a chi-square variate of K degrees of freedom. However, because the frequencies N_k are constrained to sum to n, only $K - 1$ can be independent. Moreover, if the data were used to estimate the parameters $\{\theta_j \;\; j = 1 \ldots m\}$, then the number of degrees of freedom would be reduced by 1 for each estimate. We may therefore take the statistic

$$Q_K \equiv \sum_{k=1}^{K} \frac{\left(n_k - n\hat{f}_k\right)^2}{n\hat{f}_k} \Rightarrow \chi^2_{d=K-1-m} \tag{1.25.13}$$

chi-square distribution (i.e. the *P*-value) or (b) a ratio of the ordinates (i.e. point-value) of the probability density of the statistic?

1.25.2 No significance to high P?

In the diametrically opposite situation where a significance test of a model has resulted in a value of χ^2_d considerably *less* than the expected value d for the assumed number of degrees of freedom, the corresponding P value is close to 1. Does this mean we are to reject a model because it accounts for the observations too closely? The situation has engendered a variety of replies from statisticians, generally to the effect that in nearly every instance the investigators have done something "wrong"[21] – for example, to have made numerical errors in computation or to have biased their data inadvertently or intentionally – and therefore the results are "*too good to be true*".[22] A different interpretation, principally by Edwards,[23] is that the chi-square test is "essentially a test concerning the overall variance of a model" in contrast to the mean. According to Edwards

> The crucial question the experimenter must ask himself before applying χ^2 is 'if I get a very small value, will it make me suspicious about my null hypothesis?' If the answer is 'no', then his interest is in means and not variances, and the χ^2 test is inappropriate.

A low value of χ^2, therefore, according to Edwards would not be indicative of a fit that is too good; rather, it would suggest that a model leading to a variation smaller than Poissonian would be better.

1.25.3 No significance to any P since the whole χ^2 business is arbitrary?

Statisticians have long remarked upon the fact that the number of classes and their boundaries are arbitrary choices at the disposal of the investigator and that different choices can result in radically different values of χ^2 and P for the same data set. How, then, can a test of significance be significant if you can get any desired outcome? In the administration of a chi-square test, class boundaries are ordinarily chosen so that all class intervals are equal with the consequence that the number of samples in each class diminishes the further the class value is from the mean. As a step towards rendering the chi-square test less arbitrary, some statisticians have proposed defining classes of unequal widths with boundaries calculated to lead to equal frequencies.[24] However, this modified procedure has its own difficulties.

[21] W. G. Cochran, "The χ^2 Test of Goodness of Fit", *The Annals of Mathematical Statistics* **23** (1952) 337.
[22] G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* (Griffin, London, 1940) 423.
[23] A. W. F. Edwards, *Likelihood* (Johns Hopkins, Baltimore, 1992) 188. Original Cambridge edition 1972.
[24] H. B. Mann and A. Wald, "On the choice of the number of class intervals in the application of the chi square test", *Annals of Mathematical Statistics* **13** (1942) 306–317.

1.25.4 Why bother with χ^2 anyway since all models would fail if the sample is large enough?

The claim has been made that, in testing a null hypothesis which is not expected to be exactly true, but credible to a good approximation, the hypothesis will always fail a chi-square test applied to a sufficiently large sample of experimental data. Phrased provocatively, one statistician wrote[25]

> I make the following dogmatic statement, referring for illustration to the normal curve: 'If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large—for instance, on the order of 200,000—the chi-square P will be small beyond any usual limit of significance.'

The conclusion, therefore, cited by a second acquiescing statistician,[26] was "What is the point of applying a chi-square test to a moderate or small sample if we already know that a large sample would show P highly significant?". Recall that a highly significant P means that we can with justification reject the null hypothesis – so in a sense this criticism is the opposite of the third, which ascribes no significance to P.

Before adding my own two cents, first an admission: I have selectively quoted comments from statisticians so as to frame their remarks in the most confrontational way to highlight issues that I believe really are important and deserve careful attention. No statistician, however – at least none whose papers I have read – actually recommended discarding the chi-square test. No experimental physicist would in any event do that because the test is far too useful and easily implemented (... and required for publication).

Much of the confusion that may accompany use of a chi-square test can be avoided by keeping in mind that the original test statistic followed a multinomial distribution (1.25.1) from which the chi-square statistic arose in consequence of three approximations: (1) Stirling's approximation of factorials; (2) Taylor expansion of a natural logarithm; and (3) substitution of a continuous integral for a discrete summation. So long as each expectation nf_k of the tested model $f(x|\theta)$ is reasonably large, the reduction is reasonably valid, and the "chi-square" statistic (1.25.13) is distributed as χ^2_d to good approximation. If necessary, one may combine classes to achieve a suitable expectation, which for satisfactory testing should be no fewer than about 5–10 as a rule of thumb. There was nothing in the derivation, as far as I can see, that subsequently restricted the chi-square test of significance to the *variance* of a model to the exclusion of all other attributes.

[25] J. Berkson, "Some difficulties of interpretation encounered in the application of the chi-square test", *Journal of the American Statistical Association* **33** (1938) 526–536.
[26] W. G. Cochran, *op cit*. p. 336.

The arbitrariness of classes and boundaries arises only in testing the significance of a continuous distribution, for in the case of a discrete distribution where specific objects are counted (e.g. photons, electrons, phone calls ... whatever), there is a natural, irreducible assignment of classes whereby each class differs in integer value from the one that comes before or after by one unit. This may not be the most practical choice for every test, since it may require a very large sample size, but conceptually, at least, it establishes a non-arbitrary standard.

In the case where data arising from a discrete distribution have been approximated by (or transformed into) continuous random variables, there is a simple procedure for avoiding a ridiculously large and statistically unwarranted chi-square. Statisticians have pointed this out long ago,[27] but, unaware of their papers, I discovered it for myself in testing a distribution of counts from a radioactive source. The experience makes for a lesson worth relating. The counts, which were all integers believed on theoretical grounds to be Poisson variates, decreased (on average) in time as the experiment progressed because of the diminishing sample of nuclei. In the next chapter I will discuss in detail the statistics of nuclear decay. For now, however, suffice it to say that a standard procedure in the analysis of nuclear data is to remove the negative trend line in order to examine the variation in counts as if the population of radioactive nuclei were infinite. In de-trending the data, however, the transformed numbers were no longer integers. Sorted into 90 classes, the data were tested for goodness of fit by a Poisson distribution of known mean, leading to an astounding result of $\chi^2_{89} > 1600$, where a number around 90 was expected. A previous test on the original (not de-trended) data had given highly satisfactory results. What went wrong?

The 90 classes $\{A_k\ k = 1...90\}$ were labeled by the number of counts obtained in a specified window of time (one bin of data); thus $A_1 = 150, A_2 = 151, A_3 = 152$, etc. In the test on the de-trended data, the frequency of outcomes x for $k + 1 > x \geq k$ was compared with the Poisson probability for A_k – and this gave a very high chi-square, suggesting that the null hypothesis (namely, the data were Poisson variates) was untenable. However, if the class values were shifted by 0.5, so that the central value of each class was an integer – i.e. $k + \frac{1}{2} > x \geq k - \frac{1}{2}$, the chi-square of the de-trended data became 85.14 for 90 classes, corresponding to $P = 0.596$, which was entirely reasonable.

One must likewise be aware of the circumstances under which a discrete distribution is approximated by a continuous one. Return to the previous example where data originated as integer counts of particles from a sample of radioactive nuclei. The mean number of counts $\bar{x}$ per bin being much larger than 1, the hypothesized Poisson distribution $Poi\ (\mu)$, with population mean μ estimated by the sample mean $\bar{x}$, should have been well approximated by a normal distribution $N(\bar{x}, \bar{x})$. However, a chi-square test of the goodness of fit of $N(0, 1)$ to the data in standard normal form

[27] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics Vol. 2: Inference and Relationship* (Griffin, London, 1961) 508–509.

$z = (x - \bar{x})/\sqrt{\bar{x}}$ led to so high a value of χ^2 that the presumed model would have been unambiguously rejected. Again, what went wrong?

The problem in this instance lay not with locations of class boundaries, but with the widths of class intervals. The transformed data z are not integers, but neither are they continuously distributed. Since the values of the counts x are always integer, the values of z can have a minimum separation of $\bar{x}^{-1/2}$. Thus, if one makes the bin width smaller than that minimum, there can result numerous bins of 0 count, which causes failure of the chi-square test. With adequately sized bin widths, a value of chi-square and associated P-value were obtained that did not justify rejection of the null hypothesis. Note that there was nothing intrinsically wrong with applying the test to a continuous distribution so long as one took steps to insure that the data being tested actually were continuously distributed. Nor does the fact that I could get either a high P or low P by changing the size of the bins imply that the test outcomes were "arbitrary" and therefore meaningless. On the contrary, the low P-value resulted from executing the test under conditions that were inappropriate in two related ways: (a) testing goodness of fit of a continuous distribution to quasi-discrete data which resulted in (b) violation of an approximation leading to the chi-square statistic (i.e. no "empty" bins).

The same suite of investigations convinced me that the assertion that any model "fitted to a body of data representing ... quantities in the physical world" would fail a chi-square test, given a sufficiently large (e.g. $> 200\,000$) number of observations was entirely without foundation. If the model is a "true" representation of the body of data – i.e. the model captures the essential features of the stochastic process that generates the data – then a chi-square test can yield a respectable P-value for any sample size. In testing, for example, 1 000 000 standard normal variates, sorted into 400 classes, for goodness of fit to $N(0, 1)$, I have obtained $\chi^2_{399} = 419$, giving $P = 0.236$.

However – and here is a point of critical importance that all too often seems to have been overlooked in the confused wrangle over the meaning or worth of P-values – the quantity P is itself a random variable. As a cumulative probability [see Eq. (1.25.14)] P is governed by a uniform distribution [see (1.17.12)] with mean $\frac{1}{2}$ and variance $\frac{1}{12}$. Therefore, obvious though it may be to state this, one should not expect too much from a single P, any more than is to be expected from a single nuclear count or the reply of a single respondent to a poll. That does not mean that either P or χ^2 is not useful. Rather, if an inference to be made is important, then it is incumbent upon the investigator to collect sufficient data – even if that means more time-consuming experiments and fewer publications – to determine how the P or χ^2 is distributed. If discrepancies between the hypothetical model (null hypothesis) and the data are due to pure chance, then, although a range of P values from low to high will be obtained from numerous experimental repetitions, they should nevertheless follow a uniform distribution. By contrast, if a proposed model is a poor one, the P-values should nearly all be low.

Table 1.4 *Chi-square test of Poisson variates*

Statistics	χ^2_{89}	P
Mean	87.03	0.541
Standard Error	2.22	0.046
Median	84.83	0.605
Standard Deviation	15.69	0.322
Skewness	0.172	−0.219
Kurtosis	−0.203	−1.376
Minimum	53.82	0.0062
Maximum	125.87	0.999
Count	50	50

Consider, as an illustration of the preceding homily, a suite of chi-square tests that were performed on 50 samples of nuclear decay data, each sample comprising one million bins of data, presumed to be independent, Poisson-distributed variates (the null hypothesis) sorted into 90 classes. As shown in Table 1.4, the 50 chi-square tests yielded the following statistics on both χ^2_{89} and P.

Note that a minimum $P_{\min} = 0.0062$ was obtained without there being any justification for rejecting the null hypothesis; that a maximum $P_{\max} = 0.999$ was obtained without any computational errors having been made or my having lied about the results; that the sample mean $\overline{P} = 0.541$ and standard error (standard deviation of the mean) $s_{\overline{P}} = 0.046$ are in excellent agreement with their respective theoretical values $\langle P \rangle = 0.500$, $\sigma_{\langle P \rangle} = \sqrt{\frac{1/12}{50}} = 0.041$. The upper and lower panels of Figure 1.6 respectively show histograms of the observed χ^2_{89} and P-values sorted into 10 bins with the theoretically expected results superposed. This outcome of a series of 50 chi-square tests can itself be tested for significance by a chi-square test (where now we have nine degrees of freedom). The results

$$\begin{aligned} \text{test on chi-square} \qquad & \chi^2_{\text{obs}} = 10.33; P = 0.324 \\ \text{test on } P \qquad & \chi^2_{\text{obs}} = 9.8; P = 0.367 \end{aligned}$$

support the null hypothesis that distribution of chi-square values arose through pure chance. Had I performed only a single test (rather than 50) of the Poisson variates and obtained a particularly low or high P-value, statisticians (e.g. those writing cautionary philosophical commentaries) would have had grave doubts about the randomness of the nuclear decays. And yet, because P is distributed uniformly (see lower panel of Figure 1.6), a P-value is just as likely to fall between 0.0 and 0.1 as between 0.4 and 0.5.

The lesson in all this – if there is one – is that ambiguous or troubling outcomes to chi-square tests often stem from insufficient data, a problem that can be solved by experiment, not philosophy.

Table 1.5 *Extreme order statistics for n variates U(0,1)*

Statistic	Density	Expectations	Theory ($n = 50$)	Observed ($n = 50$)
Y_1	$f_{Y_1}(y) = n\,(1-y)^{n-1}$	$\langle Y_1 \rangle = \frac{1}{n+1}$	0.0196	0.006 17
		$\langle Y_1^2 \rangle = \frac{2}{(n+1)(n+2)}$	0.000 75	
		$\sigma_{Y_1} = \sqrt{\frac{n}{(n+1)^2(n+2)}}$	0.0192	
	$z_1 = \lvert(y_{\min} - \langle Y_1 \rangle)/\sigma_{Y_1}\rvert = \lvert -0.699\rvert < 1$			
Y_n	$f_{Y_n}(y) = ny^{n-1}$	$\langle Y_n \rangle = \frac{n}{n+1}$	0.9804	0.999
		$\langle Y_n^2 \rangle = \frac{n}{n+2}$	0.9615	
		$\sigma_{Y_n} = \sqrt{\frac{n}{(n+1)^2\,(n+2)}}$	0.0192	
	$z_n = \lvert(y_{\max} - \langle Y_n \rangle)/\sigma_{Y_n}\rvert = 0.969 < 1$			

The probability density corresponding to the general expression (1.26.3) is given by the derivative $f_{Y_i}(y) = dF_{Y_i}(y)/dy$, which can be calculated by either (a) a straightforward, plodding method that calls for tenacity and careful attention to detail, or (b) a quick, simple method that calls for insight. Both ways are instructive and lead to

$$f_{Y_i}(y) = \frac{n!}{(i-1)!(n-i)!} F(y)^{i-1}(1-F(y))^{n-i} f(y). \tag{1.26.6}$$

The details are left to an appendix. The pdfs of the extreme order statistics, however, can be calculated directly and easily from (1.26.4) and (1.26.5).

Consider the circumstance, pertinent to tests of significance, where variates $\{X_i\}$ are distributed uniformly as $U(0, 1)$, in which case the cdf is simply $F(x) = x$. The pdf and first two moments of the lowest and highest order statistics may then be summarized in Table 1.5 above. Returning to the example in the previous section of the 50 chi-square variates and corresponding P-values, one sees from Table 1.5 that the observed lowest and highest Ps fall within one standard deviation of the predicted expectations. Statistical principles, more than intuition and hunches, provide a better guide for judging whether extreme events are too extreme to have occurred by chance.

1.27 Bayes' theorem and the meaning of ignorance

The use of Bayes' theorem for estimation and inference is ordinarily regarded as an alternative to the maximum likelihood method. However, just as the chi-square test of significance and least-square method of estimation can be regarded as reductions

of the maximum likelihood method to special cases, I prefer to think of the maximum likelihood method itself as a particular application of Bayes' theorem. For one thing, this is a "friendlier" perspective in discussing the matter with other colleagues, since the use of Bayes' theorem has been the source of much contention in the theory of statistical inference. But more importantly, it is basically accurate to do so since Bayes' theorem, without the accumulated emotional overburden, is an uncontested fundamental principle in probability theory and therefore a starting point for nearly all methods of statistical estimation and inference.

Recall the structure of Bayes' theorem, Eq. (1.2.5). Given a set of experimental data D and various models (hypotheses) H_i proposed to account for the data, then

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)} = \frac{P(D|H_i)P(H_i)}{\sum_{i=1} P(D|H_i)P(H_i)}. \tag{1.27.1}$$

As discussed earlier,

(1) $P(H_i)$ is the prior probability of a model based on whatever initial information may be pertinent;
(2) $P(D|H_i)$ is the likelihood, i.e. the conditional probability of obtaining the experimental results given a particular model; and
(3) $P(H_i|D)$ is the posterior probability of a particular model after the results of the experiment have been taken into account.

In comparing two models H_1, H_2, one way to use Bayes' theorem would be to evaluate the ratio

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)} \tag{1.27.2}$$

and select the model leading to the larger posterior probability.

Different models are usually distinguished by the choice and numerical values of a set of parameters θ, whereupon Bayes' theorem can be written to show this functional dependence explicitly:

(a) $P(\theta|D) \propto P(D|\theta)P(\theta)$ for a discrete parameter or
(b) $P(d\theta|D) \propto P(D|\theta)p(\theta)d\theta$ for a continuous parameter with density $p(\theta)$.

A problem of inference ("which hypothesis?") then reduces at least in part to a problem of estimation ("which parameter?"). There are various, not-necessarily equivalent, ways to make this estimate. For example, estimate the parameter θ by

(i) the value $\hat{\theta}$ that maximizes the posterior probability, i.e. the mode of the posterior probability function

$$\left.\frac{dP(\theta|D)}{d\theta}\right|_{\theta=\hat{\theta}} = 0, \tag{1.27.3}$$

or

(ii) the mean value $\langle\theta\rangle$

$$\langle\theta\rangle = \frac{\int_{-\infty}^{\infty} \theta P(D|\theta)p(\theta)d\theta}{\int_{-\infty}^{\infty} P(D|\theta)p(\theta)d\theta}, \tag{1.27.4}$$

or

(iii) the root-mean-square (rms) value of $\theta_{\rm rms}$

$$\theta_{\rm rms} = \sqrt{\left\langle (\theta - \langle\theta\rangle)^2 \right\rangle}, \tag{1.27.5}$$

or

(iv) the value $\tilde{\theta}$ that minimizes the "squared error"

$$\frac{d\left\langle (\theta - \tilde{\theta})^2 \right\rangle}{d\tilde{\theta}} = 0, \tag{1.27.6}$$

the solution of which works out to be $\langle\theta\rangle$

$$\frac{d}{d\tilde{\theta}}\left\langle (\theta - \tilde{\theta})^2 \right\rangle = -2\langle\theta\rangle + 2\tilde{\theta} = 0 \Rightarrow \tilde{\theta} = \langle\theta\rangle\ . \tag{1.27.7}$$

The impediment to using these expressions, however, and the flashpoint for much of the contention over Bayesian methods of inference, is the prior probability $p(\theta)$. In particular, what functional form does $p(\theta)$ take to represent the condition of *no* prior information about θ – i.e. the state of "ignorance".[28] It is to be stressed – and this is another critical point whose misunderstanding has been the source of much contentious discussion in the past – that the prior does *not* assign probability to the *value* of the unknown parameter, which is *not* a random variable, but to our prior *knowledge* of that parameter. There have been other potentially divisive issues as well, such as repudiation by some statisticians of the very idea that the probability of a hypothesis makes any sense, but I will dispense with all that here. From my own perspective as a practical physicist, any set of non-negative numbers summing to unity and conforming to the rules of probability theory can be considered legitimate probabilities, whether they arose from frequencies or not. The essential is that the set of numbers be testable, reproducible (statistically), and help elucidate the problem being investigated.

[28] "Ignorance" derives from a root word meeting "not to know" and, as used in statistics, does not carry the vernacular connotations of stupidity or incompetence.

It would seem, at first, that the logical course of action would be to assume a uniform distribution for unknown parameters in those instances where one has no prior information about them. There are difficulties with this course, however. The most serious is that the estimate then depends on an arbitrary choice of how the model is parameterized. For example, if the random variables of a model are believed to be generated by a pdf of the form $p(x|\theta) \propto e^{-x^2/\theta^2}$, and one assumes a uniform distribution $p(\theta) =$ constant for the prior, then one cannot assume the transformed parameter $\varphi = \theta^2$ in the pdf $p(x|\varphi) \propto e^{-x^2/\varphi}$ to be uniformly distributed as well because

$$p(\varphi) = \frac{p(\theta(\varphi))}{|d\varphi/d\theta|} = \frac{\text{constant}}{2\theta} \propto \varphi^{-1/2}. \tag{1.27.8}$$

And yet an analyst, having no more prior information about φ than about θ, could have begun the analysis by assuming φ to be uniformly distributed. Clearly, then, there is a logical inconsistency here somewhere, since the same state of prior knowledge should lead to the same posterior estimate no matter how one chooses to label the parameters of a model.

The maximum likelihood (ML) method provides a way around the problem of priors by disregarding them and basing the estimate on the mode of the likelihood, i.e. the maximum of the conditional probability $P(D|\theta)$. The method is invariant to a transformation of parameters since, by the chain rule of calculus,

$$0 = \frac{d}{d\theta}P(D|\theta) = \frac{d}{d\varphi}P(D|\theta(\varphi))\frac{d\varphi}{d\theta}, \tag{1.27.9}$$

and therefore $\frac{d}{d\varphi}P(D|\varphi) = 0$ if $\frac{d}{d\theta}P(D|\theta) = 0$, which leads to the same point estimate whether the model is formulated in terms of θ or φ.

A secondary difficulty with assuming that a parameter about which no prior information is known is distributed uniformly is that Bayes' theorem then leads to some odd results in comparison with corresponding ML estimates. For example, consider the set of observations $\{x_i\ i = 1\ldots n\}$ believed to have arisen from a Poisson process with unknown parameter θ. As worked out previously, the parameter dependence of the likelihood function is $e^{-n\theta}\theta^{n\bar{x}}$, maximization of which gives the ML estimate $\hat{\theta} = \bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$, the mean value of the observations, a reasonable result. Contrast this with the Bayes' estimate obtained by calculating the expectation $\langle\theta\rangle$ under assumption of a uniform prior $p(\theta) =$ constant:

$$\theta_{\mathrm{B}} \equiv \langle\theta\rangle = \left.\frac{\int_0^\infty \theta e^{-n\theta}\theta^{n\bar{x}}d\theta}{\int_0^\infty e^{-n\theta}\theta^{n\bar{x}}d\theta}\right|_{\text{Uniform Prior}} = \left.\frac{\int_0^\infty e^{-n\theta}\theta^{n\bar{x}+1}d\theta}{\int_0^\infty e^{-n\theta}\theta^{n\bar{x}}d\theta}\right|_{\text{Uniform Prior}} = \frac{1}{n}\frac{\Gamma(n\bar{x}+2)}{\Gamma(n\bar{x}+1)} = \bar{x} + \frac{1}{n}. \tag{1.27.10}$$