Experimental High Energy Physics Group

HEN-**343**
October 17, 2020

Statistical Methods in Data Analysis

W. J. Metzger

Faculteit der Natuurwetenschappen,
Wiskunde en Informatica

Radboud Universiteit Nijmegen

Nijmegen, The Netherlands

These notes were initially prepared for a third-year course for physics and chemistry students at the Katholieke Universiteit Nijmegen (since September, 2004, known as Radboud Universiteit Nijmegen, officially translated to ungrammatical English as Radboud University Nijmegen).

The course was given for the first time in the Spring semester of 1991. The author would like to thank the students who followed the course then for their cheerful forbearance under the sometimes chaotic appearance of successive installments and multiple versions of these notes. Since then these notes have undergone numerous revisions and expansions to the joy or dismay of later students.

The author would appreciate any comments which could result in an improvement of the course. In particular, he would like to know of any errors in the equations or text.

Contents

1 Introduction **1**
1.1 Language . . . 3
1.2 Computer usage . . . 4
1.3 Some advice to the student . . . 4

I Probability **7**

2 Probability **9**
2.1 First principles . . . 9
2.1.1 Probability—What is it? . . . 9
2.1.2 Sampling . . . 10
2.1.3 Probability density function (p.d.f.) . . . 11
2.1.4 Cumulative distribution function (c.d.f.) . . . 12
2.1.5 Expectation values . . . 12
2.1.6 Moments . . . 13
2.2 More on Probability . . . 15
2.2.1 Conditional Probability . . . 15
2.2.2 More than one r.v. . . . 16
2.2.3 Correlation . . . 17
2.2.4 Dependence and Independence . . . 20
2.2.5 Characteristic Function . . . 22
2.2.6 Transformation of variables . . . 23
2.2.7 Multidimensional p.d.f. – matrix notation . . . 25
2.3 Bayes' theorem . . . 26
2.4 Probability—What is it?, revisited . . . 27
2.4.1 Mathematical probability (Kolmogorov) . . . 27
2.4.2 Empirical or Frequency interpretation (von Mises) . . . 27
2.4.3 Subjective (Bayesian) probability . . . 28
2.4.4 Are we frequentists or Bayesians? . . . 29

3 Some special distributions **31**
3.1 Bernoulli trials . . . 31

3.2 Binomial distribution 32
3.3 Multinomial distribution 33
3.4 Poisson distribution 34
3.5 Exponential and Gamma distributions 39
3.6 Uniform distribution 41
3.7 Gaussian or Normal distribution 41
3.8 Log-Normal distribution 43
3.9 Multivariate Gaussian or Normal distribution 44
3.10 Binormal or Bivariate Normal p.d.f. 46
3.11 Cauchy (Breit-Wigner or Lorentzian) p.d.f. 49
3.12 The χ^2 p.d.f. 50
3.13 Student's t distribution 53
3.14 The F-distribution 55
3.15 Beta distribution 55
3.16 Double exponential (Laplace) distribution 56
3.17 Weibull distribution 56

4 Real p.d.f.'s **57**
4.1 Complications in real life 57
4.2 Convolution 58

5 Central Limit Theorem **61**
5.1 The Theorem 61
5.2 Implications for measurements 62

II Monte Carlo **65**

6 Monte Carlo **67**
6.1 Random number generators 67
6.1.1 True random number generators 68
6.1.2 Pseudo-random number generators 68
6.2 Monte Carlo integration 69
6.2.1 Crude Monte Carlo 69
6.2.2 Hit or Miss Monte Carlo 71
6.2.3 Buffon's needle, a hit or miss example 72
6.2.4 Accuracy of Monte Carlo integration 72
6.2.5 A crude example in 2 dimensions 74
6.2.6 Variance reducing techniques 76
6.3 Monte Carlo simulation 79
6.3.1 Weighted events 80
6.3.2 Rejection method 80
6.3.3 Inverse transformation method 81
6.3.4 Composite method 83

6.3.5 Example 83
6.3.6 Gaussian generator 85

III Statistics **89**

7 Statistics—What is it/are they? **91**

8 Parameter estimation **93**
8.1 Introduction 93
8.2 Properties of estimators 96
8.2.1 Bias 96
8.2.2 Consistency 99
8.2.3 Variance, efficiency 99
8.2.4 Interpretation of the Variance 101
8.2.5 Information and Likelihood 102
8.2.6 Minimum Variance Bound 105
8.2.7 Efficient estimators—the Exponential family 108
8.2.8 Sufficient statistics 112
8.3 Substitution methods 113
8.3.1 Frequency substitution 113
8.3.2 Method of Moments 115
8.3.3 Descriptive statistics 116
8.3.4 Generalized method of moments 116
8.3.5 Variance of moments 117
8.3.6 Transformation of the covariance matrix under a change of parameters 118
8.4 Maximum Likelihood method 120
8.4.1 Principle of Maximum Likelihood 120
8.4.2 Asymptotic properties 126
8.4.3 Change of parameters 128
8.4.4 Maximum Likelihood *vs.* Bayesian inference 129
8.4.5 Variance of maximum likelihood estimators 130
8.4.6 Summary 136
8.4.7 Extended Maximum Likelihood 137
8.4.8 Constrained parameters 140
8.5 Least Squares method 142
8.5.1 Introduction 142
8.5.2 The Linear Model 145
8.5.3 Derivative formulation 151
8.5.4 Gauss-Markov Theorem 152
8.5.5 Examples 153
8.5.6 Constraints in the linear model 156

8.5.7 Improved measurements through constraints 158
8.5.8 Linear Model with errors in both x and y 158
8.5.9 Non-linear Models 161
8.5.10 Summary 162
8.6 Estimators for binned data 162
8.6.1 Minimum Chi-Square 162
8.6.2 Binned maximum likelihood 164
8.6.3 Comparison of the methods 165
8.7 Practical considerations 166
8.7.1 Choice of estimator 166
8.7.2 Bias reduction 169
8.7.3 Variance of estimators—Jackknife and Bootstrap 170
8.7.4 Robust estimation 172
8.7.5 Detection efficiency and Weights 174
8.7.6 Systematic errors 178

9 Confidence intervals 183
9.1 Introduction 183
9.2 Confidence belts 186
9.3 Confidence bounds 187
9.4 Normal confidence intervals 188
9.4.1 σ known 188
9.4.2 σ unknown 188
9.5 Binomial confidence intervals 189
9.6 Poisson confidence intervals 190
9.6.1 Large N 190
9.6.2 Small N — Confidence bounds 191
9.6.3 Background 192
9.7 Use of the likelihood function or χ^2 193
9.8 Fiducial intervals 193
9.9 Credible (Bayesian) intervals 194
9.10 Discussion of intervals 195
9.11 Measurement of a bounded quantity 195
9.12 Upper limit on the mean of a Poisson with background 197

10 Hypothesis testing 199
10.1 Introduction 199
10.2 Basic concepts 199
10.3 Properties of tests 203
10.3.1 Size 203
10.3.2 Power 204
10.3.3 Consistency 204
10.3.4 Bias 205

10.3.5 Distribution-free tests 205
10.3.6 Choice of a test 206
10.4 Parametric tests 207
10.4.1 Simple Hypotheses 207
10.4.2 Simple H_0 and composite H_1 209
10.4.3 Composite hypotheses—same parametric family 211
10.4.4 Composite hypotheses—different parametric families 217
10.5 And if we are Bayesian? 218
10.6 Goodness-of-fit tests 220
10.6.1 Confidence level or p-value 220
10.6.2 Relation between Confidence level and Confidence Intervals . 221
10.6.3 The χ^2 test 222
10.6.4 Use of the likelihood function 222
10.6.5 Binned data 223
10.6.6 Run test 227
10.6.7 Tests free of binning 229
10.6.8 But use your eyes! 231
10.7 Non-parametric tests 233
10.7.1 Tests of independence 233
10.7.2 Tests of randomness 235
10.7.3 Two-sample tests 235
10.7.4 Two-Gaussian-sample tests 237
10.7.5 Analysis of Variance 240

IV Bibliography 247

V Exercises 253

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

—J. Clerk Maxwell

Chapter 1

Introduction

Statistics is a tool useful in the design, analysis and interpretation of experiments. Like any other tool, the more you understand how it works the better you can use it.

The fundamental laws of classical physics do not deal with statistics, nor with probability. Newton's law of gravitation $F = G\frac{Mm}{r^2}$ contains an exponent 2 in the denominator—exactly 2, not 2.000 ± 0.001. But where did the 2 come from? It came from analysis of many detailed and accurate astronomical observations of Tycho Brahe and others.

In "statistical" physics you have such a complicated situation that you treat it in a "statistical" manner, although I would prefer to make a distinction between statistics and probability and call it a probabilistic manner. In quantum mechanics the probability is intrinsic to the theory rather than a mere convenience to get around complexity.

Thus in studying physics you have no need of statistics, although in some subjects you do need probability. But when you *do* physics you need to know what measurements really mean. For that you need statistics.

Using probability we can start with a well defined problem and calculate the chance of all possible outcomes of an experiment. With probability we can thus go from theory to the data.

In statistics we are concerned with the inverse problem. From data we want to infer something about physical laws. Statistics is sometimes called an art rather

than a science, and there is a grain of truth in it. Although there are standard approaches, most of the time there is no "best" solution to a given problem. Our most common tasks for statistics fall into two categories: parameter estimation and hypothesis testing.

In parameter estimation we want to determine the value of some parameter in a model or theory. For example, we observe that the force between two charges varies with the distance r between them. We make a theory that $F \sim r^{-\alpha}$ and want to determine the value of α from experiment.

In hypothesis testing we have an hypothesis and we want to test whether that hypothesis is true or not. An example is the Fermi theory of β-decay which predicts the form of the electron's energy spectrum. We want to know whether that is correct. Of course we will not be able to give an absolute yes or no answer. We will only be able to say how confident we are, *e.g.*, 95%, that the theory is correct, or rather that the theory predicts the correct shape of the energy spectrum. Here the meaning of the 95% confidence is that if the theory is correct, and if we were to perform the experiment many times, 95% of the experiments would appear to agree with the theory and 5% would not.

Parameter estimation and hypothesis testing are not completely separate topics. It is obviously nonsense to estimate a parameter if the theory containing the parameter does not agree with the data. Also the theory we want to test may contain parameters; the test then is whether values for the parameters exist which allow the theory to agree with the data.

Although the main subject of this course is statistics, it should be clear that an understanding of statistics requires understanding probability. We will begin therefore with probability. Having had probability, it seems only natural to also treat, though perhaps briefly, Monte Carlo methods, particularly as they are often useful not only in the design and understanding of an experiment but also can be used to develop and test our understanding of probability and statistics.

There are a great many books on statistics. They vary greatly in content and intended audience. Notation is by no means standard. In preparing these lectures I have relied heavily on the following sources (sometimes to the extent of essentially copying large sections):

- R. J. Barlow,[1] a recent text book in the Manchester series. Most of what you need to know is in this book, although the level is perhaps a bit low. Nevertheless (or perhaps therefore), it is a pleasure to read.

- Siegmund Brandt,[2] a good basic book at a somewhat higher level. Unfortunately, the `FORTRAN` sample programs it contains are rather old-fashioned. There is an emphasis on matrix notation. There is also a German edition.

- A. G. Frodesen, O. Skjeggestad, and H. Tøfte,[3] also a good basic text (despite the words "particle physics" in the title) at a higher level. Unfortunately, it is out of print.

- W. T. Eadie *et al.*,[4] or the second edition of this book by F. James[5], a book at an advanced level. It is difficult to use if you are not already fairly familiar with the subject.

- G. P. Yost,[6] the lecture notes for a course at Imperial College, London. They are somewhat short on explanation.

- Glen Cowan,[7] a recent book at a level similar to these lectures. In fact, had this book been available I probably would have used it rather than writing these notes.

Other books of general interest are those of Lyons,[8] Meyer,[9] and Bevington.[10] A comprehensive reference for almost all of probability and statistics is the three-volume work by Kendall and Stuart[11]. Since the death of Kendall, volumes 1 and 2 (now called 2a) are being kept up to date by others,[12,13] and a volume (2b) on Bayesian statistics has been added.[14] Volume 3 has been split into several small books, "Kendall's Library of Statistics", covering many specialized topics. Another classic of less encyclopedic scope is the one-volume book by Cramér[15].

1.1 Language

Statistics, like physics, has it own specialized terminology with words whose meaning differs from the meaning in everyday use or the meaning in physics. An example is the word estimate. In statistics "estimate" is used where the physicist would say "determine" or "measure", as in parameter estimation. The physicist or indeed ordinary people tend to use "estimate" to mean little more than "guess" as in "I would estimate that this room is about 8 meters wide." We will generally use the statisticians' word.

Much of statistics has been developed in connection with population studies (sociology, medicine, agriculture, *etc.*) and industrial quality control. One cannot study the entire population; so one "draws a sample". But the population exists.

In experimental physics the set of all measurements (or observations) forms the "sample". If we make more measurements we increase the size of the sample, but we can never attain the "population". The population does not really exist but is an underlying abstraction. For us some terminology of the statisticians is therefore rather inappropriate. We therefore sometimes make substitutions like the following:

"demographic" terminology	"physics" terminology
sample	data (set)
draw a sample	observe, measure
sample of size N	N observations, N measurements
population	observable space
population mean	parent mean
	= mean of the underlying distribution
population variance, *etc.*	parent variance, *etc.*
sample mean	sample mean = mean of the data
	= experimental mean or average

We will just say "mean" when it is clear from the context whether we are referring to the parent or the sample mean.

1.2 Computer usage

In this day and age, data analysis without a computer is inconceivable. While there exist (a great many) programs to perform statistical analyses of data, they tend to be difficult to learn and/or limited in what they can do. Their use also tends to induce a cook-book mentality. Consequently, we shall not use them, but will write our own programs (in `FORTRAN` or `C`). In this way we will at least know what we are doing. Subroutines will be provided for a few conceptually simple, but technically complicated, tasks.

Data is often presented in a histogram (1 or 2 dimensional). Computer packages to do this will be available.

As an aid to understanding it is often useful to use random numbers, *i.e.*, perform simple Monte Carlo (*cf.* Part II). On a computer there is generally a routine which returns a "pseudo-random" number. What that actually is will be treated in section 6.1.2. An example of such use is to generate random numbers according to a given distribution, *e.g.*, uniformly between 0 and 1, and then to histogram some function of these numbers.

Parameter estimation (chapter 8) is often most conveniently done by numerically finding the maximum (or minimum) of some function. Computer programs to do this will also be available.

1.3 Some advice to the student

The goal of this course is *not* to provide a cook book of statistical data analysis. Instead, we aim for some understanding of statistical techniques, of which there are many. Lack of time will preclude rigorous proof (or sometimes any proof) of results. Moreover, we will introduce some theoretical concepts, which will not seem

immediately useful, but which should put the student in a better position to go beyond what is included in this course, as will almost certainly be necessary at some time in his career. Further, we will point out the assumptions underlying, and the limitations of, various techniques.

A major difficulty for the student is the diversity of the questions statistical techniques are supposed to answer, which results in a plethora of methods. Moreover, there is seldom a single "correct" method, and deciding which method is "best" is not always straightforward, even after you have decided what you mean by "best".

A further complication arises from what we mean by "probability". There are two major interpretations, "frequentist" (or "classical") and "Bayesian" (or "subjective"), which leads to two different ways to do statistics. While the emphasis will be on the classical approach, some effort will go into the Bayesian approach as well.

While there are many questions and many techniques, they are related. In order to see the relationships, the student is strongly advised not to fall behind.

Finally, some advice to astronomers which is equally valid for physicists:

> *Whatever* your choice of area, make the choice to live your professional life at a high level of statistical sophistication, and not at the level—basically freshman lab. level—that is the unfortunate common currency of most astronomers. Thereby we will all move forward together.
>
> —*William H. Press*[16]

Part I

Probability

"La théorie des probabilités n'est que
le bon sens reduit au calcul."
—P.-S. de Laplace, "Mécanique Céleste"

Chapter 2

Probability

2.1 First principles

2.1.1 Probability—What is it?

We begin by taking the "frequentist" approach. A given experiment is assumed to have a certain number of possible outcomes or events E. Suppose we repeat the identical experiment N times and find outcome E_i N_i times. We define the probability of outcome E_i to be

$$P(E_i) = \lim_{N\to\infty} \frac{N_i}{N} \tag{2.1}$$

We can also be more abstract. Intuitively, probability must have the following properties. Let Ω be the set of all possible outcomes.

Axioms:

1. $P(\Omega) = 1$ The experiment must have an outcome.
2. $0 \leq P(E)$, $E \in \Omega$
3. $P(\cup E_i) = \sum P(E_i)$, for any set of disjoint $E_i, E_i \in \Omega$
   (Axiom of Countable Additivity)

It is straightforward to derive the following theorems:

1. $P(E) = 1 - P(E^*)$, where $\Omega = E \cup E^*$, E and E^* disjoint.
2. $P(E) \leq 1$

3. $P(\emptyset) = 0$, where $\emptyset$ is the null set.

4. If E_1,$E_2 \in \Omega$ and not necessarily disjoint, then
$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

A philosopher once said, "It is necessary
for the very existence of science that
the same conditions always produce the same results."
Well, they do not. —Richard P. Feynman

2.1.2 Sampling

We restrict ourselves to experiments where the outcome is one or more real numbers, X_i. Repetition of the experiment will not always yield the same outcome. This could be due to an inability to reproduce *exactly* the initial conditions and/or to a probabilistic nature of the process under study, *e.g.*, radioactive decay. The X_i are therefore called **random variables (r.v.)**, *i.e.*, variables whose values cannot be predicted exactly. Note that the word 'random' in the term 'random variable' does not mean that the allowed values of X_i are equiprobable, contrary to its use in everyday speech. The set of possible values of X_i, which we have denoted Ω, is called the sample space. A r.v. can be

- discrete: The sample space Ω is a set of discrete points. Examples are the result of a throw of a die, the sex of a child (F=1, M=2), the age (in years) of students studying statistics, names of people (Marieke=507, Piet=846).

- continuous: Ω is an interval or set of intervals. Examples are the frequency of radiation from a black body, the angle at which an electron is emitted from an atom in β-decay, the height of students studying statistics.

- a combination of discrete and continuous.

An experiment results thus in an outcome which is a set of real numbers X_i which are random variables. They form a *sampling* of a parent 'population'. Note the difference between the sample, the sample space and the population:

- The **population** is a list of all members of the population. Some members of the population may be identical.

- The **sample space** is the set of all possible results of the experiment (the sampling). Identical results are represented by only one member of the set.

- The **sample** is a list of the results of a particular experiment. Some of the results may be identical. How often a particular result, *i.e.*, a particular member of the sample space, occurs in the sample should be approximately proportional to how often that result occurs in the population.

The members of the population are equiprobable while the members of the sample space are not necessarily equiprobable. The sample reflects the population which is derived from the sample space according to some probability distribution, usually called the parent (or underlying) probability distribution.

2.1.3 Probability density function (p.d.f.)

Conventionally, one uses a capital letter for the experimental result, *i.e.*, the sampling of a r.v. and the corresponding lower case letter for other values of the r.v. Here are some examples of probability distributions:

- the throw of a die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The probability distribution is $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$, which gives a parent population of $\{1, 2, 3, 4, 5, 6\}$. An example of an experimental result is $X = 3$.

- the throw of a die having sides marked with one 1, two 2's, and three 3's. The sample space is $\Omega = \{1, 2, 3\}$. The probability distribution is $P(1) = \frac{1}{6}$, $P(2) = \frac{1}{3}$, $P(3) = \frac{1}{2}$. The parent population is $\{1, 2, 2, 3, 3, 3\}$. An experimental result is $X = 3$ (maybe).

In the discrete case we have a probability function, $f(x)$, which is greater than zero for each value of x in Ω. From the axioms of probability,

$$\sum_{\Omega} f(x) = 1$$

$$P(A) \equiv P(X \in A) = \sum_{A} f(x) \quad , \quad A \subset \Omega$$

For a continuous r.v., the probability of any exact value is zero since there are an infinite number of possible values. Therefore it is only meaningful to talk of the probability that the outcome of the experiment, X, will be in a certain interval. $f(x)$ is then a probability density function (p.d.f.) such that

$$P(x \leq X \leq x + \mathrm{d}x) = f(x)\,\mathrm{d}x \ , \qquad \int_{\Omega} f(x)\,\mathrm{d}x = 1 \tag{2.2}$$

Since most quantities of interest to us are continuous we will usually only treat the continuous case unless the corresponding treatment of the discrete case is not obvious. Usually going from the continuous to the discrete case is simply the replacement of integrals by sums. We will also use the term p.d.f. for $f(x)$ although

in the discrete case it is really a probability rather than a probability density. Some authors use the term 'probability law' instead of p.d.f., thus avoiding the misleading (actually wrong) use of the word 'density' in the discrete case. However, such use of the word 'law' is misleading to a physicist, *cf.* Newton's second law, law of conservation of energy, *etc.*

2.1.4 Cumulative distribution function (c.d.f.)

The cumulative distribution function (c.d.f.) is the probability that the value of a r.v. will be $\leq$ a specific value. The c.d.f. is denoted by the capital letter corresponding to the small letter signifying the p.d.f. The c.d.f. is thus given by

$$F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x' = P(X \leq x) \tag{2.3}$$

Clearly, $F(-\infty) = 0$ and $F(+\infty) = 1$.

Properties of the c.d.f.:

- $0 \leq F(x) \leq 1$
- $F(x)$ is monotone and not decreasing.
- $P(a \leq X \leq b) = F(b) - F(a)$
- $F(x)$ discontinuous at x implies

  $$P(X = x) = \lim_{\delta x \to 0} \left[F(x + \delta x) - F(x - \delta x)\right] \text{ , } i.e.\text{, the size of the jump.}$$

- $F(x)$ continuous at x implies $P(X = x) = 0$.

The c.d.f. can be considered to be more fundamental than the p.d.f. since the c.d.f. is an actual probability rather than a probability density. However, in applications we usually need the p.d.f. Sometimes it is easier to derive first the c.d.f. from which you get the p.d.f. by

$$f(x) = \frac{\partial F(x)}{\partial x} \tag{2.4}$$

2.1.5 Expectation values

Consider some single-valued function, $u(x)$ of the random variable x for which $f(x)$ is the p.d.f. Then the expectation value of $u(x)$ is defined:

$$E[u(x)] = \int_{-\infty}^{+\infty} u(x)\, f(x)\,\mathrm{d}x \tag{2.5}$$

$$= \int_{-\infty}^{+\infty} u(x)\,\mathrm{d}F(x) \quad , \qquad f(x) \text{ continuous} \tag{2.6}$$

Properties of the expectation value:

- If k is a constant, then $E[k] = k$
- If k is a constant and u a function of x, then $E[ku] = kE[u]$
- If k_1 and k_2 are constants and u_1 and u_2 are functions of x, then $E[k_1u_1 + k_2u_2] = k_1E[u_1] + k_2E[u_2]$, *i.e.*, E is a linear operator.

Note that some books, *e.g.*, Barlow[1], use the notation $\langle u(x) \rangle$ instead of $E[u(x)]$.

2.1.6 Moments

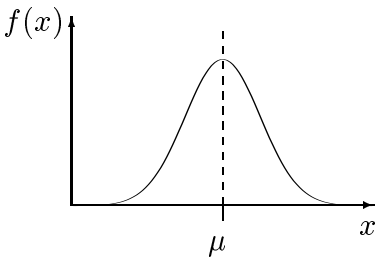
Moments are certain special expectation values. The mth moment is defined (think of the moment of inertia) as

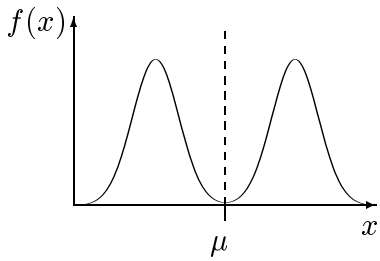
$$E[x^m] = \int_{-\infty}^{+\infty} x^m f(x)\,\mathrm{d}x \tag{2.7}$$

The moment is said to exist if it is finite. The most commonly used moment is the (population or parent) **mean**,

$$\mu \equiv E[x] = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x \tag{2.8}$$

The mean is often a good measure of location, *i.e.*, it frequently tells roughly where the most probable region is, but not always.





In statistics we will see that the sample mean, $\overline{x}$, the average of the result of a number of experiments, can be used to estimate the parent mean, μ, the mean of the underlying p.d.f.

Central moments are moments about the mean. The mth central moment is defined as

$$E[(x-\mu)^m] = \int_{-\infty}^{+\infty} (x-\mu)^m f(x)\,\mathrm{d}x \tag{2.9}$$

If μ is finite, the first central moment is clearly zero. If $f(x)$ is symmetric about its mean, all odd central moments are zero.

The second central moment is called the **variance**. It is denoted by $V[x]$, σ_x^2, or just σ^2.

$$\sigma_x^2 \equiv V[x] \equiv E\left[(x-\mu)^2\right] \tag{2.10}$$

$$= E\left[x^2\right] - \mu^2 \tag{2.11}$$

The square root of the variance, σ, is called the standard deviation. It is a measure of the spread of the p.d.f. about its mean.

Since all symmetrical distributions have all odd central moments zero, the odd central moments provide a measure of the asymmetry. The first central moment is zero. The third central moment is thus the lowest order odd central moment. One makes it dimensionless by dividing by σ^3 and defining the **skewness** as

$$\gamma_1 \equiv \frac{E\left[(x-\mu)^3\right]}{\sigma^3} \tag{2.12}$$

This is the definition of Fisher, which is the most common. However, be aware that other definitions exist, *e.g.*, the Pearson skewness,

$$\beta_1 \equiv \left(\frac{E\left[(x-\mu)^3\right]}{\sigma^3}\right)^2 = \gamma_1^2 \tag{2.13}$$

The sharpness of the peaking of the p.d.f. is measured by the **kurtosis** (also spelled curtosis). There are two common definitions, the Pearson kurtosis,

$$\beta_2 \equiv \frac{E\left[(x-\mu)^4\right]}{\sigma^4} \tag{2.14}$$

and the Fisher kurtosis,

$$\gamma_2 \equiv \frac{E\left[(x-\mu)^4\right]}{\sigma^4} - 3 = \beta_2 - 3 \tag{2.15}$$

The -3 makes $\gamma_2 = 0$ for a Gaussian. For this reason, it is somewhat more convenient, and is the definition we shall use. A p.d.f. with $\gamma_2 > 0$ (< 0) is called **leptokurtic** (**platykurtic**) and is less (more) peaked than a Gaussian, *i.e.*, having higher (lower) tails.

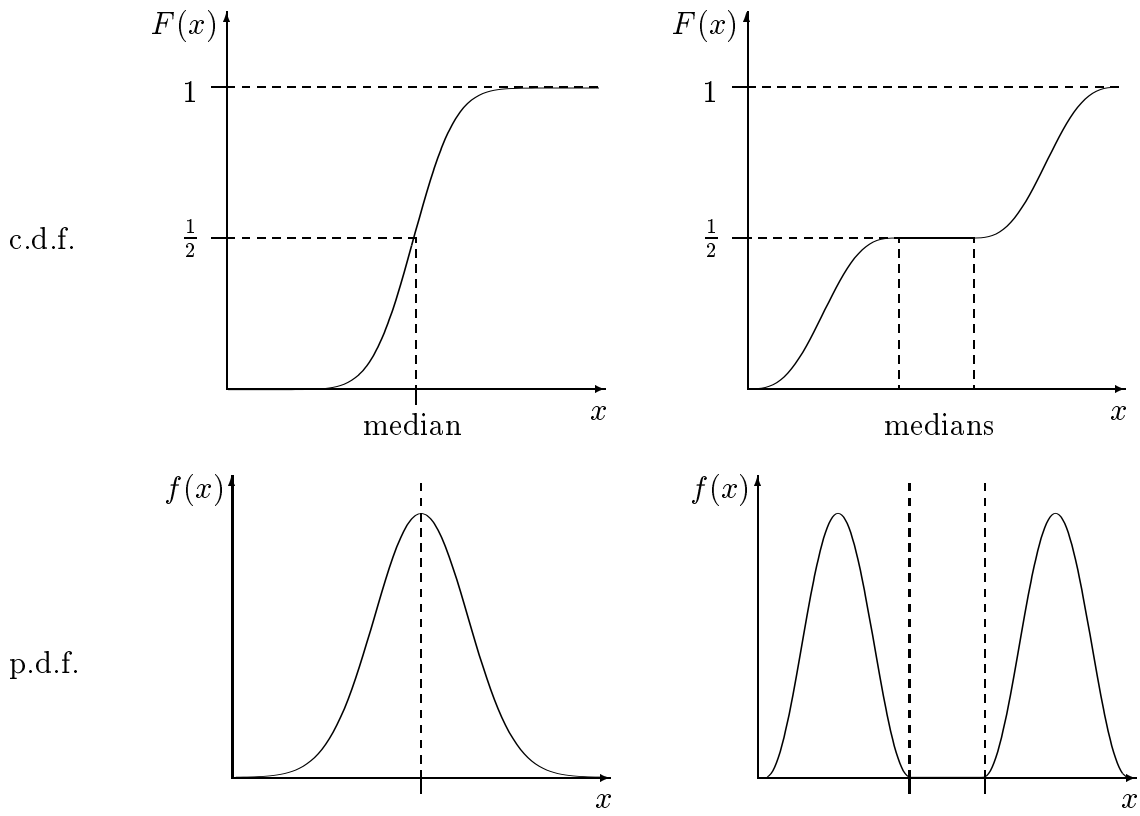
Moments are often normalized in some other way than we have done with γ_1 and γ_2, *e.g.*, with the corresponding power of μ:

$$c_k \equiv \frac{E\left[x^k\right]}{\mu^k} \quad ; \quad r_k \equiv \frac{E\left[(x-\mu)^k\right]}{\mu^k} \tag{2.16}$$

It can be shown that if all central moments exist, the distribution is completely characterized by them. In statistics we can estimate each parent moment by its sample moment (*cf.* section 8.3.2) and so, in principle, reconstruct the p.d.f.

Other attributes of a p.d.f.:

- **mode**: The location of a maximum of $f(x)$. A p.d.f. can be multimodal.
- **median**: That value of x for which the c.d.f. $F(x) = \frac{1}{2}$. The median is not always well defined, since there can be more than one such value of x.



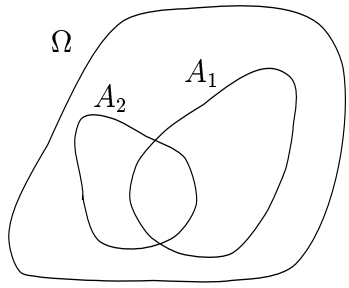
"If any one imagines that he knows something,
he does not yet know as he ought to know."
—1 Corinthians **8***:2*

2.2 More on Probability

2.2.1 Conditional Probability

Suppose we restrict the set of results of our experiment (observations or events) to a subset $A \subset \Omega$. We denote the probability of an event E given this restriction by $P(E \mid A)$; we speak of "the probability of E given A." Clearly this 'conditional' probability is greater than the probability without the restriction, $P(E)$ (unless of course A^*, the complement of A, is empty). The probability must be renormalized such that the probability that the condition is fulfilled is unity. The conditional probability should have the following properties:

$$\begin{aligned} P(A \mid A) &= 1 \qquad\qquad \text{renormalization} \\ P(A_2 \mid A_1) &= P(A_1 \cap A_2 \mid A_1) \end{aligned}$$



While the probability changes with the restriction, ratios of probabilities must not:

$$\frac{P(A_1 \cap A_2 \mid A_1)}{P(A_1 \mid A_1)} = \frac{P(A_1 \cap A_2)}{P(A_1)}$$

These requirements are met by the definition, assuming $P(A_1) > 0$,

$$P(A_2 \mid A_1) \equiv \frac{P(A_1 \cap A_2)}{P(A_1)} \tag{2.17}$$

If $P(A_1) = 0$, $P(A_2 \mid A_1)$ makes no sense. Nevertheless, for completeness we define $P(A_2 \mid A_1) = 0$ if $P(A_1) = 0$.

It can be shown that the conditonal probability satisfies the axioms of probability.

It follows from the definition that

$$P(A_1 \cap A_2) = P(A_2 \mid A_1)\, P(A_1)$$

If $P(A_2 \mid A_1)$ is the same for all A_1, *i.e.*, A_1 and A_2 are independent, then

$$\begin{aligned} P(A_2 \mid A_1) &= P(A_2) \\ \text{and} \quad P(A_1 \cap A_2) &= P(A_1)\, P(A_2) \end{aligned}$$

2.2.2 More than one r.v.

Joint p.d.f.

If the outcome is more than one r.v., say X_1 and X_2, then the experiment is a sampling of a joint p.d.f., $f(x_1, x_2)$, such that

$$P(x_1 < X_1 < x_1 + \mathrm{d}x_1\,,\, x_2 < X_2 < x_2 + \mathrm{d}x_2) = f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \tag{2.18}$$

$$P(a < X_1 < b\,,\, c < X_2 < d) = \int_a^b \mathrm{d}x_1 \int_c^d \mathrm{d}x_2\, f(x_1, x_2) \tag{2.19}$$

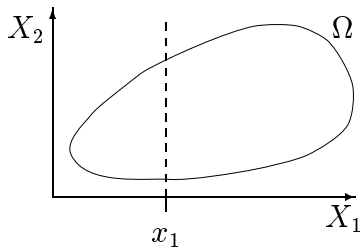
Marginal p.d.f.

The marginal p.d.f. is the p.d.f. of just one of the r.v.'s; all dependence on the other r.v.'s of the joint p.d.f. is integrated out:

$$f_1(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2)\,\mathrm{d}x_2 \tag{2.20}$$

$$f_2(x_2) = \int_{-\infty}^{+\infty} f(x_1, x_2)\,\mathrm{d}x_1 \tag{2.21}$$

Conditional p.d.f.



Suppose that there are two r.v.'s, X_1 and X_2, and a space of events Ω.

Choosing a value x_1 of X_1 restricts the possible values of X_2. Assuming $f_1(x_1) > 0$, then $f(x_2 \mid x_1)$ is a p.d.f. of X_2 given $X_1 = x_1$.

In the discrete case, from the definition of conditional probability (eq. 2.17), we have

$$\begin{aligned} f(x_2 \mid x_1) \equiv P(X_2 = x_2 \mid X_1 = x_1) &= \frac{P\left(X_2 = x_2 \cap X_1 = x_1\right)}{P(X_1 = x_1)} \\ &= \frac{P(X_2 = x_2, X_1 = x_1)}{P(X_1 = x_1)} = \frac{f(x_1, x_2)}{f_1(x_1)} \end{aligned}$$

The continuous case is, analogously,

$$f(x_2 \mid x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} \tag{2.22}$$

Note that this conditional p.d.f. is a function of only one r.v., x_2, since x_1 is fixed. Of course, a different choice of x_1 would give a different function. A conditional probability is then obviously calculated

$$P(a < X_2 < b \mid X_1 = x_1) = \int_a^b f(x_2 \mid x_1)\,\mathrm{d}x_2 \tag{2.23}$$

This may also be written $P(a < X_2 < b \mid x_1)$.

We can also compute conditional expectations:

$$E[u\,(x_2) \mid x_1] = \int_{-\infty}^{+\infty} u(x_2) f(x_2 \mid x_1)\,\mathrm{d}x_2 \tag{2.24}$$

For example, the conditional mean, $E[x_2 \mid x_1]$,
or the conditional variance, $E[(x_2 - E[x_2 \mid x_1])^2 \mid x_1]$.

The generalization to more than two variables is straightforward, *e.g.*,

$$\begin{aligned} f(x_2, x_4 \mid x_1, x_3) &= \frac{f(x_1, x_2, x_3, x_4)}{f_{13}(x_1, x_3)} \\ \text{where} \quad f_{13}(x_1, x_3) &= \int\!\!\int f(x_1, x_2, x_3, x_4)\,\mathrm{d}x_2\,\mathrm{d}x_4 \end{aligned}$$

2.2.3 Correlation

When an experiment results in more than one real number, *i.e.*, when we are concerned with more than one r.v. and hence the p.d.f. is of more than one dimension, the r.v.'s may not be independent. Here are some examples:

- Let $A =$'It is Sunday', $B =$'It is raining'. The probability of rain on Sunday is the same as the probability of rain on any other day. A and B are independent. But if $A =$'It is December', the situation is different. The probability of rain in December is not the same as the probability of rain in all other months. A and B are correlated.

- If you spend 42 hours each week at the university, the probability that at a randomly chosen moment your head is at the university is $1/4$. Similarly, the probability that your feet are at the university is $1/4$. The probability that both your head and your feet are at the university is also $1/4$ and not $1/16$; the locations of your head and your feet are highly correlated.

- Abram and Lot were standing at a road junction. The probability that Lot would take the left-hand road was $1/2$. The probability that Abram would take the left-hand road was also $1/2$. But the probability that they both would take the left-hand road was zero.[17]

- The Fermi theory allows us to calculate the energy spectrum of the particles produced in β-decay, *e.g.*, $\mathrm{n} \to \mathrm{p}\mathrm{e}^-\overline{\nu}_\mathrm{e}$, from which we can calculate the probability that the proton will have more than, say $3/4$, of the available energy. We can also calculate the probability that the electron will have more than $3/4$ of the available energy. But the probability that both the electron and the proton will have more than $3/4$ of the available energy is zero. The energies of the electron and the proton are not independent. They are constrained by the law of energy conservation.

Given a two-dimensional p.d.f. (the generalization to more dimensions is straightforward), $f(x, y)$, the mean and variance of X, μ_X and σ_X^2 are given by

$$\mu_X = E[X] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$
$$\sigma_X^2 = E\left[(X-\mu_X)^2\right]$$

A measure of the dependence of X on Y is given by the **covariance** defined as

$$\begin{aligned}\mathrm{cov}(X,Y) &\equiv E[(X-\mu_X)(Y-\mu_Y)] && (2.25)\\ &= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y && (2.26)\end{aligned}$$

From the covariance we define a dimensionless quantity, the **correlation coefficient**

$$\rho_{XY} \equiv \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} \tag{2.27}$$

If $\sigma_X = 0$, then $X \equiv \mu_X$ and consequently $E[XY] = \mu_X E[Y] = \mu_X\mu_Y$, which means that $\mathrm{cov}(X,Y) = 0$. In this case the above definition would give ρ indeterminate, and we define $\rho_{XY} = 0$.

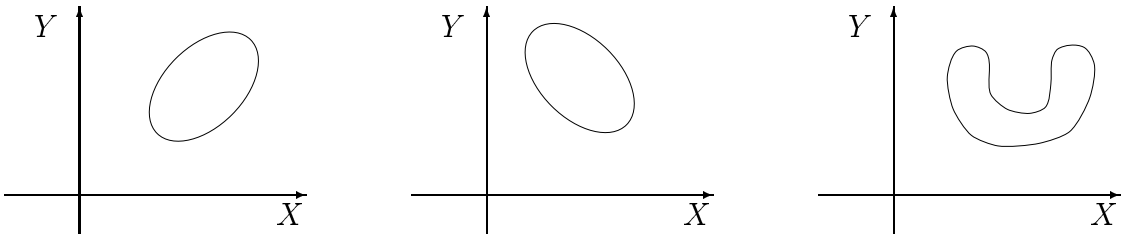
It can be shown that $\rho^2 \leq 1$, the equality holding if and only if X and Y are linearly related. The proof is left to the reader (exercise 7).

Note that while the mean and the standard deviation scale, the correlation coefficient is scale invariant, *e.g.*,

$$\mu_{2X} = 2\mu_X \qquad \text{and} \qquad \sigma_{2X} = 2\sigma_X$$

$$\rho_{2X,Y} = \frac{\operatorname{cov}(2X, Y)}{\sigma_{2X}\sigma_Y} = \frac{2\operatorname{cov}(X, Y)}{2\sigma_X\sigma_Y}$$

The correlation coefficient ρ_{XY} is a measure of how much the variables X and Y depend on each other. It is most useful when the contours of constant probability density, $f(x, y) = k$, are roughly elliptical, but not so useful when these contours have strange shapes:



$\rho > 0$ $\qquad$ $\rho < 0$ $\qquad$ $\rho \approx 0$

In the last case, even though X and Y are clearly related, $\rho \approx 0$. This can be seen as follows:

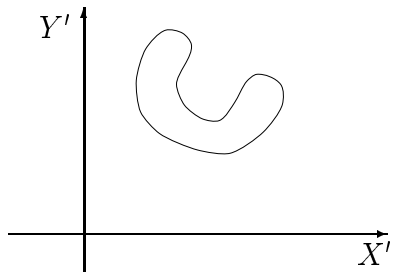
$$\begin{aligned} E\left[(X - \mu_X) \mid y\right] &= \int (x - \mu_X) f(x \mid y)\,\mathrm{d}x \\ &= \int (x - \mu_X) \frac{f(x, y)}{f_Y(y)}\,\mathrm{d}x \\ &= 0 \quad \text{for all } y \end{aligned}$$

Thus, the mean value of X is independent of y. Then,

$$\begin{aligned} \operatorname{cov}(X, Y) &= E\left[(X - \mu_X)(Y - \mu_Y)\right] \\ &= \int (y - \mu_Y) \underbrace{\int (x - \mu_X) f(x, y)\,\mathrm{d}x}_{=0}\,\mathrm{d}y \\ &= 0 \end{aligned}$$

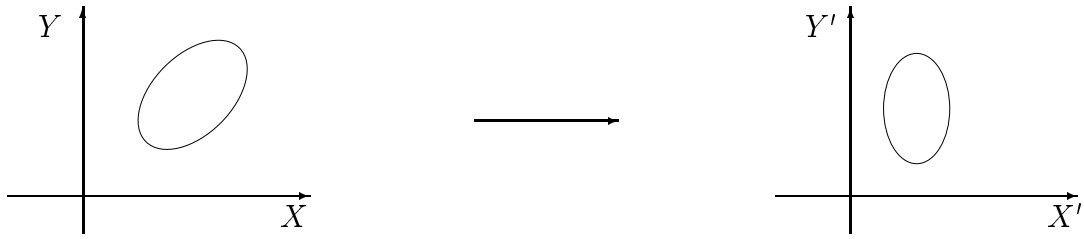
Consequently, $\rho_{XY} = 0$.

However, if we change variables, *e.g.*, by rotating, ρ, *i.e.*, $\rho_{X'Y'}$, will no longer be 0.



$\rho > 0$

Also in the elliptical case, such a change in variables can make $\rho = 0$.



$\rho > 0$ $\rho = 0$

In fact, it is always possible (also in n dimensions) to remove the correlation by a change of variables (*cf.* section 2.2.7).

The correlation coefficient, ρ, measures the average linear change in the marginal p.d.f. of one variable for a specified change in the other variable. This can be 0 even when the variables clearly depend on each other. This occurs when a change in one variable produces a change in the marginal p.d.f. of the other variable but no change in its average, only in its shape. Thus zero correlation does *not* imply independence.

2.2.4 Dependence and Independence

We know from the definitions of conditional and marginal p.d.f.'s that

$$\begin{aligned} f(x_1, x_2) &= f(x_2 \mid x_1) f_1(x_1) \qquad (2.28) \\ \text{and} \quad f_2(x_2) &= \int f(x_1, x_2)\, \mathrm{d}x_1 \\ \text{Hence} \quad f_2(x_2) &= \int f(x_2 \mid x_1) f_1(x_1)\, \mathrm{d}x_1 \end{aligned}$$

Now suppose that $f(x_2 \mid x_1)$ does not depend on x_1, *i.e.*, is the same for all x_1. Then

$$\begin{aligned} f_2(x_2) &= f(x_2 \mid x_1) \underbrace{\int f_1(x_1)\, \mathrm{d}x_1}_{=1,\ \text{normalization}} \\ &= f(x_2 \mid x_1) \end{aligned}$$

Substituting this in (2.28) gives

$$f(x_1, x_2) = f_1(x_1) f_2(x_2)$$

The joint p.d.f. is then just the product of the marginal p.d.f.'s. We take this as the definition of independence:

$$\begin{array}{lcl} \text{r.v.'s } X_1 \text{ and } X_2 \text{ are independent} & \equiv & f(x_1, x_2) = f_1(x_1) f_2(x_2) \\ \text{r.v.'s } X_1 \text{ and } X_2 \text{ are dependent} & \equiv & f(x_1, x_2) \neq f_1(x_1) f_2(x_2) \end{array}$$

We can easily derive two theorems:

Theorem: X_1 and X_2 are independent r.v.'s with joint p.d.f. $f(x_1, x_2)$ if and only if $f(x_1, x_2) = g(x_1)h(x_2)$ with $g(x_1) \geq 0$ and $h(x_2) \geq 0$ for all $x_1, x_2 \in \Omega$.

$\Longrightarrow$ From the definition of independence, f can be written as the product of the marginal p.d.f.'s, which fulfill the requirement of being positive for all $x_1, x_2 \in \Omega$.

$\Longleftarrow$ Assume $f(x_1, x_2) = g(x_1)h(x_2)$ with g and h positive. Then the marginal distributions are

$$f_1(x_1) = \int g(x_1)\, h(x_2)\, \mathrm{d}x_2 = g(x_1) \int h(x_2)\, \mathrm{d}x_2 = c\, g(x_1)$$

$$\text{and} \qquad f_2(x_2) = \int g(x_1)\, h(x_2)\, \mathrm{d}x_1 = h(x_2) \int g(x_1)\, \mathrm{d}x_1 = d\, h(x_2)$$

$$\text{Hence,} \qquad f(x_1, x_2) = g\, h = \frac{1}{cd} f_1(x_1) f_2(x_2)$$

And, since f_1 and f_2 are normalized to 1, $cd = 1$. *Q.E.D.*

Note that g and h do not have to be the marginal p.d.f.'s; the only requirement is that their product equal the product of the marginals.

Theorem: If X_1 and X_2 are independent r.v.'s with marginal p.d.f.'s $f_1(x_1)$ and $f_2(x_2)$, then for functions $u(x_1)$ and $v(x_2)$, assuming all E's exist,

$$E[u(x_1)\, v(x_2)] = E[u(x_1)]\, E[v(x_2)]$$

$\Longrightarrow$ From the definition of expectation, and since X_1 and X_2 are independent,

$$\begin{aligned} E[u(x_1)\, v(x_2)] &= \int\int u(x_1)\, v(x_2)\, f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\ &= \int u(x_1)\, f_1(x_1)\, \mathrm{d}x_1 \int v(x_2)\, f_2(x_2)\, \mathrm{d}x_2 \\ &= \int\int u(x_1)\, \underbrace{f(x_1) f(x_2)}_{=f(x_1,x_2)}\, \mathrm{d}x_1\, \mathrm{d}x_2 \int\int v(x_2)\, \underbrace{f(x_2) f(x_1)}_{=f(x_1,x_2)}\, \mathrm{d}x_2\, \mathrm{d}x_1 \\ &= E[u(x_1)]\, E[v(x_2)] \end{aligned}$$

A consequence of this last theorem is that X_1, X_2 independent implies

$$\mathrm{cov}(x_1, x_2) \equiv E[(x_1 - \mu_1)(x_2 - \mu_2)] = E[x_1 - \mu_1]\, E[x_2 - \mu_2] = 0$$

But the converse is not true.

2.2.5 Characteristic Function

So far we have only considered real r.v.'s. But from two real r.v.'s we can construct a complex r.v., $Z = X + \imath Y$ with expectation $E[Z] = E[X] + \imath E[Y]$

The characteristic function of the p.d.f. $f(x)$ is defined as the expectation of the complex quantity $e^{\imath tx}$, t real:

$$\phi(t) = E\left[e^{\imath tx}\right] = \begin{cases} \int_{-\infty}^{+\infty} e^{\imath tx} f(x)\,\mathrm{d}x & (X \text{ continuous}) \\ \sum_k e^{\imath t x_k} f(x_k) & (X \text{ discrete}) \end{cases} \tag{2.29}$$

For X continuous, $\phi(t)$ is the Fourier integral of $f(x)$.

The characteristic function completely determines the p.d.f., since by inverting the Fourier transformation we regain $f(x)$:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \phi(t) e^{-\imath xt}\,\mathrm{d}t \tag{2.30}$$

From the definition, it is clear that $\phi(0) = 1$ and $|\phi(t)| \leq 1$.

The cumulative distribution function, or indeed the probability for any interval $[x_{\min}, x]$, can also be found from $\phi(t)$:

$$\begin{aligned} F(x) = \int_{x_{\min}}^{x} f(x)\,\mathrm{d}x &= \int_{x_{\min}}^{x} \frac{1}{2\pi}\int_{-\infty}^{+\infty} \phi(t) e^{-\imath xt}\,\mathrm{d}t\,\mathrm{d}x \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \phi(t) \int_{x_{\min}}^{x} e^{-\imath xt}\,\mathrm{d}x\,\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \phi(t) \left(\frac{1}{-\imath t}\right)\left(e^{-\imath xt} - e^{-\imath x_{\min} t}\right)\,\mathrm{d}t \\ &= \frac{\imath}{2\pi}\int_{-\infty}^{+\infty} \frac{e^{-\imath xt} - e^{-\imath x_{\min} t}}{t}\,\phi(t)\,\mathrm{d}t \end{aligned}$$

In the discrete case, $f(x_k)$ is given by the difference in the probability of adjacent values of x,

$$\begin{aligned} f(x_k) &= F(x_k) - F(x_{k-1}) \\ &= \frac{\imath}{2\pi}\int_{-\infty}^{+\infty} \frac{e^{-\imath t x_k} - e^{-\imath t x_{k-1}}}{t}\,\phi(t)\,\mathrm{d}t \end{aligned}$$

The characteristic function is particularly useful in calculating moments. Differentiating $\phi(t)$ with respect to t and evaluating the result at $t = 0$ gives

$$\left.\frac{\mathrm{d}^q \phi(t)}{\mathrm{d}t^q}\right|_{t=0} = \int_{-\infty}^{+\infty} (\imath x)^q e^0 f(x)\,\mathrm{d}x = \imath^q E[x^q]$$

The characteristic function can also be written in terms of the moments by means of a Taylor expansion.

$$\begin{aligned} \phi(t) = E\left[e^{\imath tx}\right] &= E\left[\sum_{r=0}^{\infty} \frac{(\imath tx)^r}{r!}\right] \\ &= \sum_{r=0}^{\infty} \frac{(\imath t)^r}{r!} E[x^r] \end{aligned} \tag{2.31}$$

Some authors prefer, especially for discrete r.v.'s, to use the probability generating function instead of the characteristic function. It is in fact the same thing, just replacing $e^{\imath t}$ by z:

$$G(z) = E\left[z^x\right] = \begin{cases} \int_{-\infty}^{+\infty} z^x f(x)\,\mathrm{d}x \\ \sum_k z^{x_k} f(x_k) \end{cases}$$

The moments are then found by differentiating with respect to z and evaluating at $z = 1$,

$$\begin{aligned} G'(1) &= \left.\frac{\mathrm{d}G(z)}{\mathrm{d}z}\right|_{z=1} = \left.\int_{-\infty}^{+\infty} x z^{x-1} f(x)\,\mathrm{d}x\right|_{z=1} &&= E[x] \\ G''(1) &= \left.\frac{\mathrm{d}^2G(z)}{\mathrm{d}z^2}\right|_{z=1} = \left.\int_{-\infty}^{+\infty} x(x-1) z^{x-2} f(x)\,\mathrm{d}x\right|_{z=1} &&= E[x(x-1)] = E[x^2] - E[x] \end{aligned}$$

Thus the variance is given by

$$V[x] = E\left[x^2\right] - (E[x])^2 = G''(1) + G'(1) - [G'(1)]^2$$

Another application of the characteristic function is to find the p.d.f. of sums of independent r.v.'s. Let x and y be r.v.'s. Then $w = x + y$ is also an r.v. The characteristic function of w is

$$\phi_w(t) = E\left[e^{\imath t w}\right] = E\left[e^{\imath t(x+y)}\right] = E\left[e^{\imath t x} e^{\imath t y}\right]$$

If x and y are independent, this becomes

$$\phi_w(t) = E\left[e^{\imath t x}\right] E\left[e^{\imath t y}\right] = \phi_x(t)\,\phi_y(t) \tag{2.32}$$

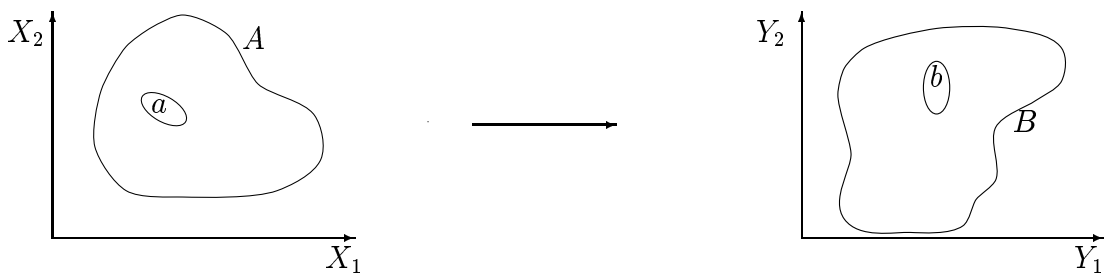
Thus the characteristic function of the sum of independent r.v.'s is just the product of the individual characteristic functions.

2.2.6 Transformation of variables

We will treat the two-dimensional case. You can easily generalize to N dimensions.

Continuous p.d.f.

Given r.v.'s X_1, X_2 from a p.d.f. $f(x_1, x_2)$ defined on a set A, we transform (X_1, X_2) to (Y_1, Y_2). Under this transformation the set A maps onto the set B.



Let $a \subset A$ be a small subset which the transformation maps onto $b \subset B$, *i.e.*,

$$(X_1, X_2) \in a \to (Y_1, Y_2) \in b \text{ such that } P(a) = P(b)$$

$$\text{Then } P\left[(Y_1, Y_2) \in b\right] = P\left[(X_1, X_2) \in a\right] = \int\int_a f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2$$

The transformation is given by

$$\begin{aligned} y_1 &= u_1(x_1, x_2) \\ y_2 &= u_2(x_1, x_2) \end{aligned}$$

The transformation must be one-to-one. Then a unique inverse transformation exists:

$$\begin{aligned} x_1 &= w_1(y_1, y_2) \\ x_2 &= w_2(y_1, y_2) \end{aligned}$$

(Actually the condition of one–to–one can be relaxed in some cases.) Assume also that all first derivatives of w_1 and w_2 exist. Then

$$\begin{aligned} P(a) &= P(b) \\ \int\int_a f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 &= \int\int_b f\left(w_1(y_1, y_2), w_2(y_1, y_2)\right) |J|\, \mathrm{d}y_1\, \mathrm{d}y_2 \end{aligned}$$

where J is the Jacobian determinant (assumed known from calculus) and the absolute value is taken to ensure that the probability is positive,

$$J = J\left(\frac{w_1, w_2}{y_1, y_2}\right) = \begin{vmatrix} \frac{\partial w_1}{\partial y_1} & \frac{\partial w_2}{\partial y_1} \\ \frac{\partial w_1}{\partial y_2} & \frac{\partial w_2}{\partial y_2} \end{vmatrix} \tag{2.33}$$

Hence the p.d.f. in (Y_1, Y_2) is the p.d.f. in (X_1, X_2) times the Jacobian:

$$g(y_1, y_2) = f\left(w_1\left(y_1, y_2\right), w_2(y_1, y_2)\right) |J| \tag{2.34}$$

Discrete p.d.f.

This is actually easier, since we can take the subsets a and b to contain just one point. Then

$$\begin{aligned} P(b) = P(Y_1 = y_1, Y_2 = y_2) = P(a) = P(X_1 = x_1 = w_1(y_1, y_2), X_2 = w_2(y_1, y_2)) \\ g(y_1, y_2) = f(w_1\left(y_1, y_2\right), w_2\left(y_1, y_2\right)) \end{aligned}$$

Note that there is no Jacobian in the discrete case.

2.2.7 Multidimensional p.d.f. – matrix notation

In this section we present the vector notation we will use for multidimensional p.d.f.'s. An n-dimensional random variable, *i.e.*, the collection of the n r.v.'s $x_1, x_2, \ldots, x_n$ is denoted by an n-dimensional column vector and its transpose by a row vector:

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \underline{x}^{\mathrm{T}} = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix} \tag{2.35}$$

If the r.v. $\underline{x}$ is distributed according to the p.d.f. $f(\underline{x})$, the c.d.f. is

$$F(\underline{x}) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(\underline{x})\, \mathrm{d}\underline{x} \qquad , \qquad \mathrm{d}\underline{x} = \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_n$$

The p.d.f. and the c.d.f. are related by

$$f(\underline{x}) = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F(\underline{x})$$

The moments about the origin of order $l_1, l_2, \ldots, l_n$ are

$$\mu_{l_1, l_2, \ldots, l_n} = E\left[x_1^{l_1}, x_2^{l_2}, \ldots, x_n^{l_n}\right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_1^{l_1} x_2^{l_2} \cdots x_n^{l_n}\, f(\underline{x})\, \mathrm{d}\underline{x}$$

The mean of a particular r.v., *e.g.*, x_2, is

$$\mu_2 = \mu_{010\ldots0}$$

These means can be written as a vector, the mean of $\underline{x}$:

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

Moments about the mean are

$$\lambda_{l_1, l_2, \ldots, l_n} = E\left[(x_1 - \mu_1)^{l_1} (x_2 - \mu_2)^{l_2} \ldots (x_n - \mu_n)^{l_n}\right]$$

The variances are, *e.g.*,

$$\sigma_1^2 = \sigma^2(x_1) = \lambda_{200\ldots00} = E\left[(x_1 - \mu_1)^2\right]$$

and the covariances

$$\sigma_{ij} = \mathrm{cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)] \qquad , \quad i \neq j$$

e.g., $$\mathrm{cov}(x_1, x_2) = \lambda_{1100\ldots00}$$

The variances and covariances may be written as a matrix, called the covariance (or variance) matrix:

$$\underline{V} = E\left[(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})^{\mathrm{T}}\right] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix} \tag{2.36}$$

$$= \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots \\ \vdots & \vdots & \ddots \\ \rho_{1n}\sigma_1\sigma_n & \rho_{2n}\sigma_2\sigma_n & \dots \end{pmatrix} \tag{2.37}$$

where ρ_{ij} is the correlation coefficient for r.v.'s x_i and x_j:

$$\rho_{ij} \equiv \frac{\sigma_{ij}}{\sigma_i\sigma_j} = \frac{\mathrm{cov}(x_i, x_j)}{\sqrt{\sigma_i^2\sigma_j^2}} \tag{2.38}$$

The covariance matrix is clearly symmetric ($\sigma_{ji} = \sigma_{ij}$). As is well known in linear algebra, it is always possible to find a unitary transformation, U, of the r.v. $\underline{x}$ to the r.v. $\underline{y} = \underline{U}\,\underline{x}$ such that the covariance matrix of $\underline{y}$, $\underline{V}\left[\underline{y}\right] = \underline{U}\,\underline{V}\,[\underline{x}]\,\underline{U}^{\mathrm{T}}$, is diagonal, which means that the y_i are uncorrelated.

2.3 Bayes' theorem

$A \cap B = B \cap A$. Hence, $P(A \cap B) = P(B \cap A)$. From the definition of conditional probability, eq. (2.17), $P(A \mid B) \equiv P(A \cap B)/P(B)$, it then follows that

$$P(A \mid B)P(B) = P(B \mid A)P(A) \tag{2.39}$$

This simple theorem* of Rev. Thomas Bayes[18] is quite innocuous. However it has far-reaching consequences in one interpretation of probability, as we shall see in the next section.

"When I use a word," Humpty Dumpty said in a rather scornful tone, "it means just what I choose it to mean—neither more nor less."
—Lewis Carroll, "Through the Looking Glass"

*Sometimes called the chain rule of probability, this theorem was first formulated by Rev. Bayes around 1759. The exact date is not known; the paper was published posthumously by his good friend Richard Price in 1763. Bayes' formulation was only for $P(A)$ uniform. The theorem was formulated in its present form by Laplace,[19] who was apparently unaware of Bayes' work. Laplace

2.4 Probability—What is it?, revisited

We have used mathematical probability, which is largely due to Kolmogorov, to derive various properties of probability. In our minds we have so far an idea of what probability means, which we refer to as the frequency approach. In this section we shall first review these two topics and then discuss another interpretation of the meaning of probability, which we shall call subjective probability.

2.4.1 Mathematical probability (Kolmogorov)

In this approach[21] we began with three axioms, from which we can derive everything. We can calculate the probability of any complicated event for which we know the *a priori* probabilities of its components. But this is simply mathematics. What probability really means requires a connection to the real world. As Bayes wrote,[22]

> It is not the business of the Mathematician to dispute whether quantities do in fact ever vary in the manner that is supposed, but only whether the notion of their doing so be intelligible; which being allowed, he has the right to take it for granted, and then to see what deductions he can make from that supposition... He is not inquiring how things are in matter of fact, but supposing things to be in a certain way, what are the consequences to be deduced from them; and all that is to be demanded of him is, that his suppositions be intelligible, and his inferences just from the suppositions he makes.

Bertrand Russel put it somewhat more succinctly:

> Mathematics is the only science where one never knows what one is talking about nor whether what is said is true.

2.4.2 Empirical or Frequency interpretation (von Mises)

In this approach, largely due to von Mises,[23] probability is viewed as the limit of the frequency of a result of an experiment or observation when the number of identical experiments is very large, *i.e.*,

$$P(x_i) = \lim_{N\to\infty} \frac{N_i}{N} \tag{2.40}$$

There are two shortcomings to this approach:

- $P(x_i)$ is not just a property of the experiment. It also depends on the "collec-

went on to apply[20] it to problems in celestial mechanics, medical statistics and even, according to some accounts, to jurisprudence.

tive" or "ensemble", *i.e.*, on the N repetitions of the experiment. For example, if I take a resistor out of a box of resistors, the probability that I measure the resistance of the resistor as 1 ohm depends not only on how the resistor was made, but also on how all the other resistors in the box were made.

- The experiment must be repeatable, under identical conditions, but with different outcomes possible. This is a great restriction on the number of situations in which we can use the concept of probability. For example, what is the probability that it will rain tomorrow? Such a question is meaningless for the frequentists, since the experiment cannot be repeated!

2.4.3 Subjective (Bayesian) probability

This approach attempts to extend the notion of probability to the areas where the experiment of the frequentists cannot be repeated. Probability here is a subjective "degree of belief" which can be modified by observations. This was, in fact, the interpretation of such pioneers in probability as Bayes and Laplace.

This approach takes Bayes' theorem (2.39), which we repeat here,

$$P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$$

and interprets A as a theory or hypothesis and B as a result or observation. $P(A)$ is then the probability that A is true, or, in other words, our "belief" in the theory. Then Bayes' theorem becomes

$$P(\text{theory} \mid \text{result})\, P(\text{result}) = P(\text{result} \mid \text{theory})\, P(\text{theory})$$

Then

$$P(\text{theory} \mid \text{result}) = \frac{P(\text{result} \mid \text{theory})}{P(\text{result})}\, P(\text{theory})$$

Here, $P(\text{theory})$ is our "belief" in the theory before doing the experiment, $P(\text{result} \mid \text{theory})$ is the probability of getting the result if the theory is true, $P(\text{result})$ is the probability of getting the result irrespective of whether the theory is true or not, and $P(\text{theory} \mid \text{result})$ is our belief in the theory after having obtained the result.

This seems to make sense. We see that if the theory predicts the result with high probability, *i.e.*, $P(\text{result} \mid \text{theory})$ big, then $P(\text{theory} \mid \text{result})$, *i.e.*, your belief in the theory after the result, will be higher than it was before, $P(\text{theory})$, and *vice versa*. However, if the result is likely even if the theory is not true, then your belief in the theory will not increase by very much since then $\frac{P(\text{result}|\text{theory})}{P(\text{result})}$ is not much greater than 1.

Suppose we want to determine some parameter of nature, λ, by doing an experiment which has outcome Z. Further, suppose we know the conditional p.d.f. to get Z given λ: $f(z \mid \lambda)$. Our prior, *i.e.*, before we do the experiment, belief about λ is given by $P_{\text{prior}}(\lambda)$. Now the probability of z, $P(z)$, is just the marginal p.d.f.:

$f_1(z) = \sum_{\lambda'} f(z \mid \lambda') P_{\text{prior}}(\lambda')$. Then by Bayes' theorem,

$$P_{\text{posterior}}(\lambda \mid z) = \frac{f(z \mid \lambda)}{f_1(z)} P_{\text{prior}}(\lambda) \tag{2.41}$$

Or, if λ is a continuous variable, which in physics is most often the case,

$$f_{\text{posterior}}(\lambda \mid z) = \frac{f(z \mid \lambda)}{f_1(z)} f_{\text{prior}}(\lambda) \tag{2.42}$$

where $f_1(z) = \int f(z \mid \lambda') f_{\text{prior}}(\lambda') \, \mathrm{d}\lambda'$.

Given $P_{\text{prior}}(\lambda)$ this is all OK. The problem here is: What is $P_{\text{prior}}(\lambda)$? By its nature this is not known. Guessing the prior probability is clearly subjective and unscientific. The usual prescription is

Bayes' Postulate: If completely ignorant about $P_{\text{prior}}(\lambda)$, take all values of λ as equiprobable.

There are objections to this postulate:

- If we are completely ignorant about $P(\lambda)$, how do we know $P_{\text{prior}}(\lambda)$ is a constant?
- A different choice of $P_{\text{prior}}(\lambda)$ would give a different $P_{\text{posterior}}$.
- If we are ignorant about $P(\lambda)$, we are also ignorant about $P(\lambda^2)$ or $P(\sqrt{\lambda})$ or $P(1/\lambda)$. Taking any of these as constant would imply a different $P_{\text{prior}}(\lambda)$, giving a different posterior probability.

These objections are usually answered by the assertion (supported by experience) that $P_{\text{posterior}}$ usually converges to about the same value after several experiments irrespective of the initial choice of P_{prior}.

2.4.4 Are we frequentists or Bayesians?

First we note that it is in the sense of frequencies that the word 'probability' is used in quantum mechanics and statistical physics. Turning to experimental results, in the physical sciences, most experiments are, in principle, repeatable and the problem can be stated to specify the "collective". So the frequentist interpretation is usually OK for us. Given the objections we have seen in the Bayesian approach, particularly that of subjectivity, most physicists today, like mathematicians starting in the mid-nineteenth century, would claim to be frequentists.

However in interpreting experimental results we often sound like Bayesians. For example, you measure the mass of the electron to be 520 ± 10 keV/c^2, *i.e.*, you measured 520 keV/c^2 with an apparatus with a resolution of 10 keV/c^2. You might then say "The mass of the electron is probably close to 520 keV/c^2." Or you might say "The mass of the electron is between 510 and 530 keV/c^2 with 68% probability.

But this is not the frequentist's probability—the experiment has not been repeated an infinite or even a large number of times. It sounds much more like a Bayesian probability: With a resolution, or 'error', of $\sigma = 10$ keV/c^2, the probability that we will measure a mass m when the true value is m_e is

$$P(m \mid m_\mathrm{e}) \propto e^{-(m-m_\mathrm{e})^2/2\sigma^2}$$

Then by Bayes' theorem, the probability that the true mass has the value m_e after we have measured a value m is

$$\begin{aligned} P(m_\mathrm{e} \mid m) &= \frac{P(m \mid m_\mathrm{e})}{P(m)} P_\mathrm{prior}(m_\mathrm{e}) \\ &\propto P(m \mid m_\mathrm{e}) \qquad \text{assuming } P_\mathrm{prior}(m_\mathrm{e}) = \text{const.} \\ &\propto e^{-(m-m_\mathrm{e})^2/2\sigma^2} \end{aligned}$$

In a frequentist interpretation of probability, the statement that the electron has a certain mass with a certain probability is utter nonsense. The electron has a definite mass: The probability that it has that mass is 1; the probability that it has some other value is 0. Our only problem is that we do not know what the value is. We can, nevertheless, make the statement "The mass of the electron is between 510 and 530 keV/c^2 with 68% confidence." Note that this differs from the Bayesian statement above by just one word. This will be discussed further in the sections on maximum likelihood (sect. 8.2.4) and confidence intervals (sect. 9), where what exactly we mean by the word confidence will be explained.*

"That's a great deal to make one word mean,"
Alice said in a thoughtful tone.
"When I make a word do a lot of work like that,"
said Humpty Dumpty, "I always pay it extra."
—Lewis Carroll, "Through the Looking Glass"

*Fisher[24], introducing his prescription for confidence intervals, had this scathing comment on Bayesian probability (referred to as inverse probability):

> I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet that is quite exactly the position in respect of inverse probability. Bayes, who seems to have first attempted to apply the notion of probability, not only to effects in relation to their causes but also to causes in relation to their effects, invented a theory, and evidently doubted its soundness, for he did not publish it during his life. It was posthumously published by Price, who seems to have felt no doubt of its soundness. It and its applications must have made great headway during the next 20 years, for Laplace takes for granted in a highly generalised form what Bayes tentatively wished to postulate in a special case.

Chapter 3

Some special distributions

We now examine some distributions which are frequently encountered in physics and/or statistics. We begin with discrete distributions.

3.1 Bernoulli trials

A Bernoulli trial is an experiment with two possible outcomes, *e.g.*, the toss of a coin. The random variable is the outcome of the experiment, k:

outcome	probability
'success', $k = 1$	p
'failure', $k = 0$	$q = 1 - p$

The p.d.f. is

$$f(k;p) = p^k q^{1-k} \tag{3.1}$$

Note that we use a semicolon to separate the r.v. k from the parameter of the distribution, p. This p.d.f. results in the moments and central moments:

$$\begin{aligned} E[k^m] &= 1 \cdot p \,+\, 0 \cdot (1-p) \,=\, p \\ E[(k-\mu)^m] &= \underbrace{(1-p)^m p}_{k=1} + \underbrace{(0-p)^m (1-p)}_{k=0} \end{aligned}$$

In particular,

$$\begin{aligned} \mu &= p \\ V[k] = E\left[k^2\right] - (E[k])^2 &= p - p^2 = p(1-p) \end{aligned}$$

3.2 Binomial distribution

The binomial distribution gives the probability of k successes (ones) in n independent Bernoulli trials each having a probability p of success. We denote this distribution by $B(k;n,p)$. The probability of k successes followed by $n-k$ failures is $p^k q^{n-k}$. But the order of the successes and failures is unimportant. There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ different permutations. Therefore the p.d.f. is given by

$$B(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.2}$$

It has the following properties:

$$\begin{aligned}
\mu &= E[k] = np && \text{(mean)}\\
\sigma^2 &= V[k] = np(1-p) && \text{(variance)}\\
\gamma_1 &= \tfrac{1-2p}{\sqrt{np(1-p)}} && \text{(skewness)}\\
\gamma_2 &= \tfrac{1-6p(1-p)}{np(1-p)} && \text{(kurtosis)}\\
\phi(t) &= [pe^{it} + (1-p)]^n && \text{(characteristic function)}
\end{aligned}$$

We will derive the first of these properties and leave the rest as exercises.

$$\begin{aligned}
\mu = E[k] &= \sum_{k=0}^{n} kB(k;n,p) = \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k}\\
&= \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}\\
&= np\sum_{k=1}^{n} k\frac{(n-1)!}{k(k-1)!(n-k)!} p^{k-1}(1-p)^{n-k} \qquad k=0 \text{ term is } 0\\
&= np\underbrace{\sum_{k'=0}^{n'} \frac{n'!}{k'!(n'-k')!} p^{k'}(1-p)^{n'-k'}}_{=[p+(1-p)]^{n'}=1} \qquad \text{with } n'=n-1,\ k'=k-1\\
&= np
\end{aligned}$$

Many distributions have a **reproductive property**, *i.e.*, the p.d.f. of the sum of two or more independent r.v.'s, each distributed according to the same p.d.f., is the same p.d.f. as for the individual r.v.'s although (usually) with different parameters.

Let X, Y be independent r.v.'s both distributed according to a binomial p.d.f. with parameter p. Thus

$$f(x,y) = B(x;n_x,p)B(y;n_y,p) = \binom{n_x}{x} p^x (1-p)^{n_x-x} \binom{n_y}{y} p^y (1-p)^{n_y-y}$$

What is then the p.d.f. of the r.v. $X+Y$? We change variables and, for convenience, introduce new parameters:

$$\begin{array}{lrclrcl}
\text{new variables} & Z_1 & = & X+Y & Z_2 & = & Y\\
\text{inverse transformation} & X & = & Z_1 - Z_2 & Y & = & Z_2\\
\text{new parameters} & n_{z_1} & = & n_x + n_y & n_{z_2} & = & n_y
\end{array}$$

The p.d.f. for the new variables is then

$$\begin{aligned} g(z_1, z_2) &= f(z_1 - z_2, z_2) \\ &= \binom{n_{z_1} - n_{z_2}}{z_1 - z_2} \binom{n_{z_2}}{z_2} p^{z_1} (1-p)^{n_{z_1} - z_1} \end{aligned}$$

The p.d.f. for $Z_1 = X + Y$ is the marginal of this. Hence we must sum over z_2.

$$g_1(z_1) = \sum_{z_2} g(z_1, z_2) = p^{z_1}(1-p)^{n_{z_1}-z_1} \sum_{z_2} \binom{n_{z_1} - n_{z_2}}{z_1 - z_2} \binom{n_{z_2}}{z_2}$$

For normalization the sum must be just $\binom{n_{z_1}}{z_1}$. Thus g_1 is also a binomial p.d.f.:

$$g_1(x+y) = B(z_1; n_{z_1}, p) = B(x+y; n_x + n_y, p)$$

3.3 Multinomial distribution

This is the generalization of the binomial distribution to more than two outcomes. Let there be m different outcomes, with probabilities p_i. Consider n experiments and let k_i denote the number of experiments having outcome i. The p.d.f. is then

$$M(k_1, k_2, \ldots, k_m; p_1, p_2, \ldots, p_m, n) = \frac{n!}{k_1! k_2! \ldots k_m!} p_1^{k_1} p_2^{k_2} \ldots p_m^{k_m} \tag{3.3}$$

subject to the conditions

$$\sum_{i=1}^{m} p_i = 1 \quad \text{and} \quad \sum_{i=1}^{m} k_i = n$$

We can write the multinomial p.d.f. in a more condensed form:

$$M(\underline{k}; \underline{p}, n) = n! \prod_{i=1}^{m} \frac{p_i^{k_i}}{k_i!} \tag{3.4}$$

An example of application of this p.d.f. is a histogram of m bins with a probability of p_i that the outcome of an experiment will be in the ith bin. Then for n experiments, the probability that the numbers of entries in the bins will be given by the k_i is given by the multinomial p.d.f.

To calculate expectation values we make use of the binomial p.d.f.: For a given bin, either an outcome is in it (probability p_i) or not (probability $1 - p_i = \sum_{j \neq i} p_j$). This is just the case of the binomial p.d.f. In other words, the marginal p.d.f. of the multinomial is the binomial. Hence,

$$\begin{aligned} \mu_i &= E[k_i] &= np_i \\ \sigma_i^2 &= V[k_i] &= np_i(1-p_i) \end{aligned}$$

Further, $\qquad \mathrm{cov}(k_i, k_j) = -np_ip_j \qquad i \neq j$
The correlation coefficient is then

$$\rho_{ij} = \frac{\mathrm{cov}(k_i, k_j)}{\sigma_i\sigma_j} = -\sqrt{\frac{p_i}{1-p_i}\frac{p_j}{1-p_j}}$$

The correlation comes about because n is fixed: $\sum k_i = n$. The k_i are thus not independent. If n is not fixed, *i.e.*, n is a r.v., the bin contents are not correlated. But then we do not have the multinomial p.d.f. but the Poisson p.d.f. for each bin.

The characteristic function of the multinomial p.d.f. is

$$\phi(t_2, t_3, \ldots, t_m) = \left(p_1 + p_2e^{it_2} + p_3e^{it_3} + \ldots + p_me^{it_m}\right)^n$$

3.4 Poisson distribution

This p.d.f. applies to the situation where we detect events but do not know the number of trials. An example is a radioactive source where we detect the decays but do not detect the non-decays. The events are counted as a function of some parameter x, *e.g.*, the time of a decay. The probability of an event in an interval Δx is assumed proportional to Δx.

Now make Δx so small that the probability of more than one event in the interval Δx is negligible. Consider n such intervals. Let λ be the probability of an event in the total interval $n\Delta x$. Assume $\lambda \neq \lambda(x)$. Then the probability of an event in Δx is $p = \lambda/n$. The probability of r events in the total interval, *i.e.*, r of the n subintervals contain one event, is given by the binomial p.d.f.

$$P(r;\lambda) = B\left(r; n, \frac{\lambda}{n}\right) = \frac{n!}{r!(n-r)!}\left(\frac{\lambda}{n}\right)^r\left(1-\frac{\lambda}{n}\right)^{n-r}$$

$$\begin{array}{llll} \text{Now} & \frac{n!}{(n-r)!} & = & n(n-1)(n-2)\ldots(n-r+1) \qquad r \text{ terms} \\ & & \approx & n^r \qquad \text{since } n >> r \\ \text{and} & \left(1-\frac{\lambda}{n}\right)^{n-r} & \approx & \left(1-\frac{\lambda}{n}\right)^n \underset{n\to\infty}{\longrightarrow} e^{-\lambda} \end{array}$$

Hence, we arrive at the expression for the Poisson p.d.f.:

$$P(r;\lambda) = \frac{e^{-\lambda}\lambda^r}{r!}$$

We can check that $P(r;\lambda)$ is properly normalized:

$$\sum_{r=0}^{\infty} P(r;\lambda) = e^{-\lambda}\sum_{r=0}^{\infty}\frac{\lambda^r}{r!} = e^{-\lambda}e^{\lambda} = 1$$

The mean is

$$\begin{aligned}
\mu = E[r] &= \sum_{r=0}^{\infty} r e^{-\lambda} \frac{\lambda^r}{r!} = \lambda e^{-\lambda} \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} \\
&= \lambda e^{-\lambda} \sum_{r'=0}^{\infty} \frac{\lambda^{r'}}{r'!} \qquad\qquad r' = r-1 \\
&= \lambda \sum_{r'=0}^{\infty} P(r';\lambda) \\
&= \lambda
\end{aligned}$$

Hence the Poisson p.d.f. is usually written

$$P(r;\mu) = \frac{e^{-\mu}\mu^r}{r!} \tag{3.5}$$

It gives the probability of getting r events if the expected number (mean) is μ.

Further, you can easily show that the variance is equal to the mean:

$$\sigma_r^2 = V[r] = \mu \tag{3.6}$$

Other properties:

$$\begin{aligned}
\gamma_1 &= \tfrac{E\left[(r-\mu)^3\right]}{\sigma^3} = \tfrac{\mu}{\mu^{3/2}} = \tfrac{1}{\sqrt{\mu}} && \text{(skewness)} \\
\gamma_2 &= \tfrac{E\left[(r-\mu)^4\right]}{\sigma^4} = \tfrac{3\mu^2+\mu}{\mu^2} - 3 = \tfrac{1}{\mu} && \text{(kurtosis)} \\
\phi(t) &= \textstyle\sum_{r=0}^{\infty} e^{\imath t r} P(r;\mu) = \sum_{r=0}^{\infty} e^{\imath t r} \frac{\mu^r}{r!} e^{-\mu} \\
&= e^{-\mu} \textstyle\sum_{r=0}^{\infty} \frac{(\mu e^{\imath t})^r}{r!} = e^{-\mu} \exp\left(\mu e^{\imath t}\right) \\
\phi(t) &= \exp\left[\mu\left(e^{\imath t} - 1\right)\right] && \text{(characteristic function)}
\end{aligned}$$

From the skewness we see that the p.d.f. becomes more symmetric as μ increases.

When calculating a series of Poisson probabilities, one can make use of the recurrence formula $P(r+1) = \frac{\mu}{r+1} P(r)$.

Reproductive property

The Poisson p.d.f. has a reproductive property: For independent r.v.'s X and Y, both Poisson distributed, the joint p.d.f. is

$$f(x,y) = \frac{\mu_x^x \mu_y^y e^{-\mu_x} e^{-\mu_y}}{x!y!} \qquad\qquad x, y = 0, 1, 2, 3, \ldots$$

To find the p.d.f. of $X+Y$ we change variables

new variables	$Z_1 = X + Y$	$Z_2 = Y$
inverse transformation	$X = Z_1 - Z_2$	$Y = Z_2$

The joint p.d.f. of the new variables is then

$$g(z_1, z_2) = \frac{\mu_x^{z_1 - z_2} \mu_y^{z_2} \, e^{-\mu_x} e^{-\mu_y}}{(z_1 - z_2)! z_2!}$$

The marginal p.d.f. for z_1 is (using the fact that $0 \leq z_2 \leq z_1$)

$$\begin{aligned} g_1(z_1) = \sum_{z_2=0}^{z_1} g(z_1, z_2) &= \frac{e^{-\mu_x - \mu_y}}{z_1!} \underbrace{\sum_{z_2=0}^{z_1} \frac{z_1!}{(z_1 - z_2)! z_2!} \mu_x^{z_1 - z_2} \mu_y^{z_2}}_{=(\mu_x + \mu_y)^{z_1} \text{ (binomial theorem)}} \\ &= \frac{(\mu_x + \mu_y)^{z_1} e^{-(\mu_x + \mu_y)}}{z_1!} \end{aligned}$$

which has the form of a Poisson p.d.f. *Q.E.D.* We rewrite it

$$g(x + y) = \frac{(\mu_x + \mu_y)^{x+y} e^{-(\mu_x + \mu_y)}}{(x + y)!}$$

The p.d.f. of the sum of two Poisson distributed random variables is also Poisson with μ equal to the sum of the μ's of the individual Poissons. This can also be easily shown using the characteristic function (exercise 12).

Examples

The Poisson p.d.f. is applicable when

- the events are independent, and
- the event rate is constant ($= \mu$).

We give a number of examples:

- Thus the number of raisins per unit volume in raisin bread should be Poisson distributed. The baker has mixed the dough thoroughly so that the raisins do not stick together (independent) and are evenly distributed (constant event rate).
- However, the number of zebras per unit area is not Poisson distributed (even in those parts of the world where there are wild zebras), since zebras live in herds and are thus not independently distributed.
- A classic example of Poisson statistics is the distribution of the number of Prussian cavalry soldiers kicked to death by horses.[25] In 10 different cavalry corps over 20 years there were 122 soldiers kicked to death by horses. The average is thus $\overline{k} = 122/200 = 0.610$ deaths/corps/year.

Assuming that the death rate is constant over the 20 year period and independent of corps and that the deaths are independent (not all caused by one killer horse) then the deaths should be Poisson distributed: the probability of k deaths in one particular corps in one year is $P(k;\mu)$. Since the mean of P is μ, we take the experimental average as an 'estimate' of μ. The distribution should then be $P(k;0.61)$ and we should expect $200 \times P(k;0.61)$ occurrences of k deaths in one year in one corps. The data:

number of deaths in 1 corps in 1 year k	actual number of times 1 corps had k deaths in 1 year	Poisson prediction $200 \times P(k;0.610)$
0	109	108.67
1	65	66.29
2	22	20.22
3	3	4.11
4	1	0.63
5	0	0.08
	200	200.00

The 'experimental' distribution agrees very well with the Poisson p.d.f. The reader can verify that the experimental variance, estimated by $\frac{1}{N}\sum(k_i - \overline{k})^2$, is 0.608, very close to the mean (0.610) as expected for a Poisson distribution.

- The number of entries in a given bin of a histogram when the (independent) data are collected over a fixed time interval, *i.e.*, when the total number of entries in the histogram is not fixed.

However, if the rate of the basic process is not constant, the distribution may not be Poisson, *e.g.*,

- The radioactive decay over a period of time significant compared with the lifetime of the source.
- The radioactive decay of a small amount of material.
- The number of interactions produced by a beam consisting of a small number of particles incident on a thick target.

In the first two examples the event rate decreases with time, in the third with position. In the last two there is the further restriction that the number of events is significantly restricted, as it can not exceed the number of atoms or beam particles, while for the Poisson distribution the number extends to infinity.

- The number of people who die each year while operating a computer is also

not Poisson distributed. Although the probability of dying while operating a computer may be constant, the number of people operating computers increases each year. The event rate is thus not constant.

The Poisson p.d.f. requires that the events be independent. Consider the case of a counter with a dead time of 1 μsec. This means that if a second particle passes through the counter within 1 μsec after one which was recorded, the counter is incapable of recording the second particle. Thus the detection of a particle is not independent of the detection of other particles. If the particle flux is low, the chance of a second particle within the dead time is so small that it can be neglected. However, if the flux is high it cannot be. No matter how high the flux, the counter cannot count more than 10^6 particles per second. In high fluxes, the number of particles detected in some time interval will not be Poisson distributed.

Radioactive decays – Poisson approximation of a Binomial

Let us examine the case of radioactive decays more closely. Consider a sample of n radioactive atoms. In a time interval T some will decay, others will not. There are thus two possibilities between which the n atoms are divided. The appropriate p.d.f. is therefore the binomial. The probability that r atoms decay in time T is thus

$$f(r) = B(r; n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \tag{3.7}$$

where p is the probability for one atom to decay in time T. Of course, p depends on the length of the time interval. In the following time interval n will be less but the value of p will remain the same. But if n is large and p small, then $n >> r$ and the change in n can be neglected. Then

$$\begin{aligned} \frac{n!}{(n-r)!} &= n(n-1)(n-2)\cdots(n-r+1) \qquad\qquad r \text{ terms} \\ &\approx n^r \end{aligned}$$

Also,

$$\begin{aligned} (1-p)^{n-r} &= 1 - p(n-r) + \frac{p^2}{2!}(n-r)(n-r-1) + \ldots \\ &\approx 1 - p(n-r) + \frac{p^2}{2!}(n-r)^2 + \ldots \\ &= e^{-p(n-r)} \approx e^{-pn} \end{aligned}$$

Substituting these approximations in (3.7) yields

$$f(r) = B(r; n, p) \approx \frac{(np)^r}{r!} e^{-np} = P(r; np)$$

which is a Poisson p.d.f. with $\mu = np$. This derivation is in fact only slightly different from our previous one; the approximations involved here are the same.

3.5 Exponential and Gamma distributions

Radioactive decays (again): As discussed in the previous section, the probability of r decays in time $\mathrm{d}t$ is given by the binomial p.d.f.:

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

where n is the number of undecayed atoms at the start of the interval. The probability that one atom decays is p, which of course depends on the length of the time interval, $\mathrm{d}t$. Now r is just the current value of $-\frac{\mathrm{d}n}{\mathrm{d}t}$, *i.e.*, the number of atoms which decay in $\mathrm{d}t$ equals the change in the number of undecayed atoms. Therefore,

$$E\left[\frac{\mathrm{d}n}{\mathrm{d}t}\right] = -E[r] = -np \tag{3.8}$$

Interchanging the order of the differentiation and the integration of the expectation operator yields

$$\frac{\mathrm{d}E[n]}{\mathrm{d}t} = -np$$

Identifying the actual number with its expectation,

$$\begin{aligned} \frac{\mathrm{d}n}{\mathrm{d}t} &= -np \\ n &= n_0 e^{-pt} \end{aligned} \tag{3.9}$$

Thus the number of undecayed atoms falls exponentially. From this it follows that the p.d.f. for the distribution of individual decay times (lifetimes) is exponential:

Exponential p.d.f.: Let $f(t)$ be the p.d.f. for an individual atom to decay at time t. The probability that it decays before time t is then $F(t) = \int_0^t f(t)\ \mathrm{d}t$. The expected number of decays in time t is

$$E[n_0 - n] = n_0 F(t) = n_0 \int_0^t f(t)\ \mathrm{d}t$$

Substituting for $E[n]$ from equation 3.9 and differentiating results in the exponential p.d.f.:

$$f(t; t_0) = \frac{1}{t_0} e^{-t/t_0} \tag{3.10}$$

which gives, *e.g.*, the probability that an individual atom will decay in time t. Note that this is a continuous p.d.f.

Properties:

$$\begin{aligned} \mu = E[t] &= t_0 \\ \sigma^2 = V[t] &= t_0^2 \end{aligned} \qquad \begin{aligned} \gamma_1 &= 2 \\ \gamma_2 &= 6 \\ \phi(x) &= [1 - \imath x t_0]^{-1} \end{aligned}$$

Since we could start timing at any point, in particular at the time of the first event, $f(t)$ is the p.d.f. for the time of the second event. Thus the p.d.f. of the time interval between decays is also exponential. This is the special case of $k = 1$ of the following situation:

Let us find the distribution of the time t for k atoms to decay. The r.v. $T = \sum_1^k t_i$ is the sum of the time intervals between k successive atoms. The t_i are independent. The c.d.f. for t is then just the probability that more than k atoms decay in time t:

$$F(t) = P(T \leq t) = 1 - P(T > t)$$

Since the decays are Poisson distributed, the probability of m decays in the interval t is

$$P(m) = \frac{(\lambda t)^m e^{-\lambda t}}{m!}$$

where $\lambda = 1/t_0$, and t_0 is the mean lifetime of an atom. The probability of $< k$ decays is then

$$P(T > t) = \sum_{m=0}^{k-1} \frac{(\lambda t)^m e^{-\lambda t}}{m!} = \int_{\lambda t}^{\infty} \frac{z^{k-1} e^{-z}}{(k-1)!} \, \mathrm{d}z$$

(The replacement of the sum by the integral can be found in any good book of integrals.) Substituting the gamma function, $\Gamma(k) = (k-1)!$, the c.d.f. becomes

$$F(t) = 1 - \int_0^{\lambda t} \frac{z^{k-1} e^{-z}}{\Gamma(k)} \, \mathrm{d}z$$

Changing variables, $y = z/\lambda$,

$$F(t) = \int_0^t \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma(k)} \, \mathrm{d}y$$

The p.d.f. is then

$$f(t; k, \lambda) = \frac{\mathrm{d}F}{\mathrm{d}t} = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)}, \quad t > 0, \tag{3.11}$$

which is called the **gamma distribution**. Some properties of this p.d.f. are

$$\begin{aligned} \mu = E[t] &= k/\lambda \\ \sigma^2 = V[t] &= k/\lambda^2 \end{aligned} \qquad \begin{aligned} \gamma_1 &= 2/\sqrt{k} \\ \gamma_2 &= 6/k \\ \phi(x) &= \left[1 - \tfrac{ix}{\lambda}\right]^{-k} \end{aligned}$$

Note that the exponential distribution, $f(t; 1, \lambda) = \lambda e^{-\lambda t}$, is the special case of the gamma distribution for $k = 1$. The exponential distribution is also a special case of the Weibull distribution (section 3.17).

Although in the above derivation k is an integer, the gamma distribution is, in fact, more general: k does not have to be an integer. For $\lambda = \frac{1}{2}$ and $k = \frac{n}{2}$, the gamma distribution reduces to the $\chi^2(n)$ distribution (section 3.12).

3.6 Uniform distribution

The uniform distribution (also known as the rectangular distribution),

$$f(x;a,b) = \frac{1}{b-a}\ , \qquad a \le x \le b \qquad \text{and } f(x) = 0\ ,\ \text{elsewhere} \tag{3.12}$$

is the p.d.f. of a r.v. distributed uniformly between a and b.
Properties:

$$\begin{aligned}
\mu = E[t] &= \textstyle\int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{b+a}{2} && \text{mean}\\
\sigma^2 = V[x] &= \textstyle\int_a^b \frac{x^2}{b-a}\,\mathrm{d}x - \mu^2 = \frac{(b-a)^2}{12} && \text{variance}\\
\gamma_1 &= \frac{E\left[(x-\mu)^3\right]}{\sigma^3} = 0 && \text{skewness}\\
\gamma_2 &= \frac{E\left[(x-\mu)^4\right]}{\sigma^4} - 3 = -1.2 && \text{kurtosis}\\
\phi(t) &= \frac{\sinh\left[\frac{1}{2}\imath t(b-a)\right]}{\imath t(b-a)} + \tfrac{1}{2}\imath t(b+a) && \text{characteristic function}
\end{aligned}$$

Round-off errors in arithmetic calculations are uniformly distributed.

3.7 Gaussian or Normal distribution

This is probably the best known and most used p.d.f.

$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{3.13}$$

Some books use the notation $N(x;\mu,\sigma)$. The Gaussian distribution is symmetric about μ, and σ is a measure of its width.

We name this distribution after Gauß, but in fact many people discovered it and investigated its properties independently. The French name it after Laplace, who had noted[26] its main properties when Gauß was only six years old. The first known reference to it, before Laplace was born, is by the Englishman A. de Moivre in 1733,[27] who, however, did not realize its importance and made no use of it. Its importance in probability and statistics (*cf.* section 8.5) awaited Gauß[28] (1809).

The origin of the name 'normal' is unclear. It certainly does not mean that other distributions are abnormal.

Properties: (The first two justify the notation used for the two parameters of the Gaussian.)

$$\begin{aligned}
\mu = E[x] &= \mu && \text{mean}\\
\sigma^2 = V[x] &= \sigma^2 && \text{variance}\\
\gamma_1 = \gamma_2 &= 0 && \text{skewness and kurtosis}\\
E[(x-\mu)^n] &= \begin{cases} 0, & n \text{ odd} \\ (n-1)!!\sigma^n = \frac{n!\sigma^n}{2^{n/2}(\frac{n}{2})!}, & n \text{ even} \end{cases} && \text{central moments}\\
&\quad \text{where } a!! \equiv 1\cdot 3\cdot 5\cdots a\\
\phi(t) &= \exp\left[\imath t\mu - \tfrac{1}{2}t^2\sigma^2\right] && \text{characteristic function}
\end{aligned}$$

When using the Gaussian, it is usually convenient to shift the origin, $x \to x' = x - \mu$ to obtain

$$N(x'; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x'^2/2\sigma^2} \tag{3.14}$$

We can also change the scale, $x \to z = (x - \mu)/\sigma$, defining a 'standard' variable, *i.e.*, a variable with $\mu = 0$ and $\sigma = 1$. Then we obtain the unit Gaussian (also called the unit Normal or standard Normal) p.d.f.:

$$N(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{3.15}$$

which has the cumulative distribution (c.d.f.)

$$\mathrm{erf}(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, \mathrm{d}x \tag{3.16}$$

which is called the error function or normal probability integral. The c.d.f. of $N(x; \mu, \sigma^2)$ is then $\mathrm{erf}\left(\frac{x-\mu}{\sigma}\right)$.

Some authors use the following definition of the error function instead of equation 3.16:

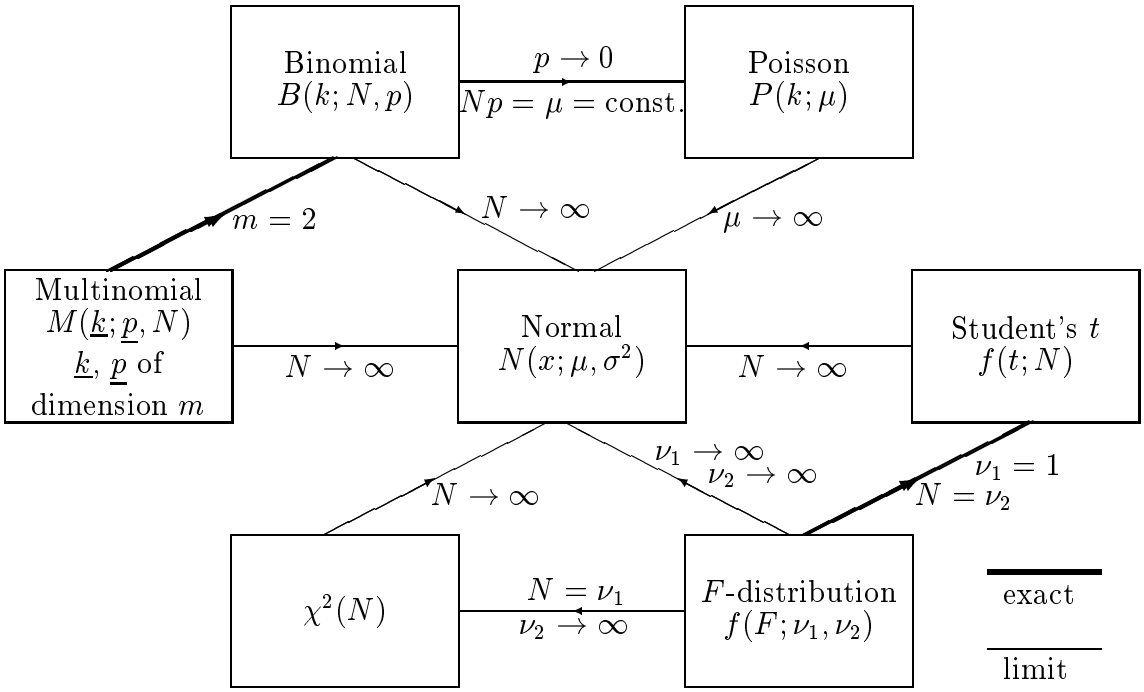
$$\phi(z) \equiv \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2}\, \mathrm{d}t \tag{3.17}$$

It is this definition which is used by the `FORTRAN` library function `ERF(Z)`. Our definition (3.16) is related to this definition by

$$\mathrm{erf}(z) = \frac{1}{2} + \frac{1}{2}\phi\left(\frac{z}{\sqrt{2}}\right) \tag{3.18}$$

The Gaussian as limiting case

The Gaussian distribution is so important because it is a limiting case of nearly all commonly used p.d.f.'s. This is a consequence of the Central Limit Theorem, which we will discuss shortly (*cf.* chapter 5). This relationship is shown for a number of distributions in the following figure:



Reproductive property

Since the Poisson p.d.f. has the reproductive property and since the Gaussian p.d.f. is a limit of the Poisson, it should not surprise us that the Gaussian is also reproductive: If X and Y are two independent r.v.'s distributed as $N(x;\mu_x,\sigma_x^2)$ and $N(y;\mu_y,\sigma_y^2)$ then $Z = X + Y$ is distributed as $N(z;\mu_z,\sigma_z^2)$ with $\mu_z = \mu_x + \mu_y$ and $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$. The proof is left as an exercise (exercise 19).

3.8 Log-Normal distribution

If an r.v., y, is normally distributed with mean μ and variance σ^2, then the r.v., $x = e^y$, is distributed as

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}\exp\left(-\frac{1}{2}\frac{(\log x-\mu)^2}{\sigma^2}\right) \tag{3.19}$$

As with the normal p.d.f., some authors consider σ, rather than σ^2, as the parameter of the p.d.f.
Properties:

$$\begin{aligned} E[x] &= \exp\left(\mu + \tfrac{1}{2}\sigma^2\right) && \text{mean} \\ V[x] &= \exp\left(2\mu+\sigma^2\right)\left[\exp\left(\sigma^2\right)-1\right] && \text{variance} \end{aligned}$$

Note that the parameters μ and σ^2 are *not* the mean and variance of the p.d.f. of x, but rather the parameters of the corresponding normal p.d.f. of $y = \log x$, $N(y;\mu,\sigma^2)$.

3.9 Multivariate Gaussian or Normal distribution

Consider n random variables x_i with expectations (means) μ_i, which we write as vectors:

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

The Gaussian is an exponential of a quadratic form in $(x - \mu)$. In generalizing the Gaussian to more than one dimension, we replace $(x - \mu)$ by the most general n-dimensional quadratic form which is symmetric about the point $\underline{\mu}$,

$$-\frac{1}{2}\,(\underline{x} - \underline{\mu})^{\mathrm{T}}\,\underline{A}\,(\underline{x} - \underline{\mu})$$

We have written the $-\frac{1}{2}$ explicitly in order that $A = \frac{1}{\sigma^2}$ in the one-dimensional case. Since we have constructed this to be symmetric about $\underline{\mu}$, we must have that $E[\underline{x}] = \underline{\mu}$. Hence, $E\left[\underline{x} - \underline{\mu}\right] = 0$, and

$$\int_{-\infty}^{+\infty} (\underline{x} - \underline{\mu}) \exp\left[-\frac{1}{2}(\underline{x} - \underline{\mu})^{\mathrm{T}}\underline{A}\,(\underline{x} - \underline{\mu})\right] \mathrm{d}\underline{x} = 0$$

By differentiating this with respect to $\underline{\mu}$ we get ($\underline{1}$ is the unit matrix)

$$\int_{-\infty}^{+\infty} \left[\underline{1} - (\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^{\mathrm{T}}\underline{A}\right] \exp\left[-\frac{1}{2}(\underline{x} - \underline{\mu})^{\mathrm{T}}\underline{A}\,(\underline{x} - \underline{\mu})\right] \mathrm{d}\underline{x} = 0$$

Therefore,

$$E\left[\underline{1} - (\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^{\mathrm{T}}\underline{A}\right] = 0$$

$$E\left[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^{\mathrm{T}}\right]\underline{A} = \underline{1}$$

This expectation is just the definition of the covariance matrix, $\underline{V}$ (equation 2.36). Hence $\underline{V}\,\underline{A} = \underline{1}$ or

$$\underline{A} = \underline{V}^{-1}$$

If the correlations between all the x_i, are zero, *i.e.*, if all $\rho_{ij}, i \neq j$, are zero, then $\underline{V}$ is diagonal with $V_{ii} = \sigma_i^2$. Then $\underline{A}$ is also diagonal with $A_{ii} = \frac{1}{\sigma_i^2}$ and

$$\begin{aligned}\exp\left[-\frac{1}{2}(\underline{x} - \underline{\mu})^{\mathrm{T}}\underline{A}(\underline{x} - \underline{\mu})\right] &= \exp\left[-\frac{1}{2}\left(\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} + \ldots\right)\right] \\ &= \exp\left[-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2}\right] \exp\left[-\frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right] \cdots\end{aligned}$$

The p.d.f. is thus just the product of n 1-dimensional Gaussians. Thus all $\rho_{ij} = 0$ implies that x_i and x_j are independent. As we have seen (sect. 2.2.4), this is not true of all p.d.f.'s.

It remains to determine the normalization. The result is

$$N\left(\underline{x};\underline{\mu},\underline{V}\right) = \frac{1}{(2\pi)^{n/2}\,|\underline{V}|^{1/2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})^{\mathrm{T}}\underline{V}^{-1}(\underline{x}-\underline{\mu})\right] \tag{3.20}$$

where $|\underline{V}|$ is the determinant of $\underline{V}$. This assumes that $\underline{V}$ is non-singular, *i.e.*, $|\underline{V}| \neq 0$. If $\underline{V}$ is singular, that means that two of the x_i are completely correlated, *i.e.*, $|\rho_{ij}| = 1$. In that case we can replace x_j by a function of x_i thus reducing the dimension by one.

Comparison of equations 3.13 and 3.20 shows that an n-dimensional Gaussian may be obtained from a 1-dimensional Gaussian by the following substitutions:

$$\begin{array}{rcl} x & \to & \underline{x} \\ \sigma^2 & \to & \underline{V} \\ \sigma & \to & |\underline{V}|^{1/2} \end{array} \qquad \begin{array}{rcl} \mu & \to & \underline{\mu} \\ \sigma^{-2} & \to & \underline{V}^{-1} \\ \frac{1}{\sqrt{2\pi}} & \to & \frac{1}{(2\pi)^{n/2}} \end{array}$$

These same substitutions are applicable for many (not all) cases of generalization from 1 to n dimensions, as we might expect since the Gaussian p.d.f. is so often a limiting case.

Multivariate Normal - summary:

p.d.f.	$N\left(\underline{x};\underline{\mu},\underline{V}\right)$	$=$	$\frac{1}{(2\pi)^{n/2}\,\lvert\underline{V}\rvert^{1/2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})^{\mathrm{T}}\underline{V}^{-1}(\underline{x}-\underline{\mu})\right]$	(3.20)
mean	$E[\underline{x}]$	$=$	$\underline{\mu}$	
covariance	$\mathrm{cov}(\underline{x})$	$=$	$\underline{V}$	
	$V[x_i]$	$=$	V_{ii}	
	$\mathrm{cov}(x_i, x_j)$	$=$	V_{ij}	
characteristic function	$\phi(\underline{t})$	$=$	$\exp\left[\imath t\underline{\mu} - \frac{1}{2}\underline{t}^{\mathrm{T}}\,\underline{V}\,\underline{t}\right]$	

Other interesting properties:

- Contours of constant probability density are given by

$$(\underline{x}-\underline{\mu})^{\mathrm{T}}\underline{V}^{-1}(\underline{x}-\underline{\mu}) = C\ , \qquad \text{a constant}$$

- Any section through the distribution, *e.g.*, at $x_i = \text{const.}$, gives again a multivariate normal p.d.f. It has dimension $n-1$. For the case $x_i = \text{const.}$, the covariance matrix $\underline{V_{n-1}}$ is obtained by removing the ith row and column from $\underline{V}^{-1}$ and inverting the resulting submatrix.

- Any projection onto a lower space gives a marginal p.d.f. which is a multivariate normal p.d.f. with covariance matrix obtained by deleting appropriate rows and columns of $\underline{V}$. In particular, the marginal distribution of x_i is

$$f_i(x_i) = N(x_i;\mu_i,\sigma_i^2)$$

- A set of variables, each of which is a linear function of a set of normally distributed variables, has itself a multivariate normal p.d.f.

We will now examine a special case of the multivariate normal p.d.f., that for two dimensions.

3.10 Binormal or Bivariate Normal p.d.f.

This is the multivariate normal p.d.f. for 2 dimensions. Using (x, y) instead of (x_1, x_2), we have

$$\underline{V} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

$$\underline{V}^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)} \begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$$

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}G}$$

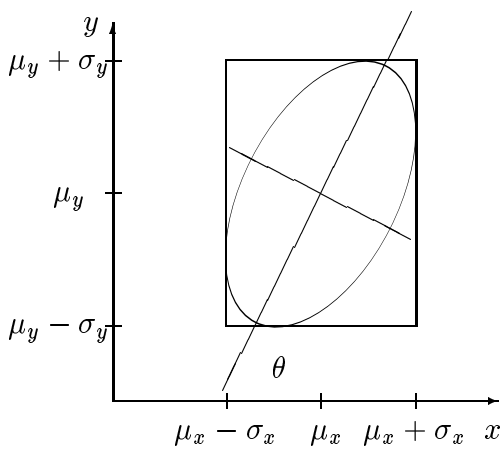
$$\text{where} \quad G = \frac{1}{(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \right]$$

Contours of constant probability density are given by setting the exponent equal to a constant. These are ellipses, called covariance ellipses.

For $G = 1$, the extreme values of the ellipse are at $\mu_x \pm \sigma_x$ and $\mu_y \pm \sigma_y$. The larger the correlation, the thinner is the ellipse, approaching 0 as $|\rho| \to 1$. (Of course in the limit of $\rho = \pm 1$, G is infinite and we really have just 1 dimension.)

The orientation of the major axis of the ellipse is given by

$$\tan 2\theta = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$



Note that $\theta = \pm 45°$ only if $\sigma_x^2 = \sigma_y^2$
$\theta = 0$ if $\rho = 0$

In calculating θ by taking the arctangent of the above equation, one must be careful of quadrants. If the arctangent function is defined to lie between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, then θ is the angle of the major axis if $\sigma_x > \sigma_y$; otherwise it is the angle of the minor axis. It is therefore more convenient to use an arctangent function defined between $-\pi$ and π such as the ATAN2(y,x) of some languages: $2\theta = \text{ATAN2}(2\rho\sigma_x\sigma_y, \sigma_x^2 - \sigma_y^2)$.

In the one-dimensional Gaussian the probability that x is within k standard deviations of μ is given by

$$P(\mu - k\sigma \leq x \leq \mu + k\sigma) = \int_{\mu - k\sigma}^{\mu + k\sigma} N(x; \mu, \sigma^2)\,\mathrm{d}x \tag{3.21}$$

which is an integral over the interval of x where $G \leq k$. In two dimensions this generalizes to the probability that (x, y) is within the ellipse corresponding to k standard deviations, which is given* by

$$P(G \leq k) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{G \leq k} \int e^{-\frac{1}{2}G}\,\mathrm{d}x\,\mathrm{d}y \tag{3.22}$$

Some values:

2-dimensional $P(G \leq k)$	k	1-dimensional $P(G \leq k) = P(\mu - k\sigma \leq x \leq \mu + k\sigma)$	2× 1-dimensional $P(\mu_x - k\sigma \leq x \leq \mu_x + k\sigma)$ $P(\mu_y - k\sigma \leq y \leq \mu_y + k\sigma)$
0.3934693	1	0.6826895	0.466065
0.6321206	2	0.9544997	0.911070
0.7768698	3	0.9973002	0.994608
0.8646647	4	0.9999367	0.999873
0.9179150	5	0.9999994	0.999999
0.9502129	6		

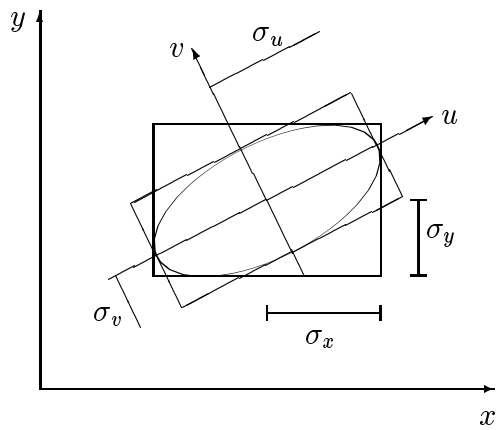
Note that the 2-dimensional probability for a given k is much less than the corresponding 1-dimensional probability. This is easily understood: the product of the two 1-dimensional probabilities is the probability that (x, y) is in the rectangle defined by $\mu_x - k\sigma_x \leq x \leq \mu_x + k\sigma_x$ and $\mu_y - k\sigma_y \leq y \leq \mu_y + k\sigma_y$. The ellipse is entirely within this rectangle and hence the probability of being within the ellipse is less than the probability of being within the rectangle.

Since the covariance matrix is symmetric, there exists a unitary transformation which diagonalizes it. In two dimensions this is the rotation matrix $\underline{U}$,

$$\underline{U} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

This matrix transforms (x, y) to (u, v):

$$\begin{pmatrix} u \\ v \end{pmatrix} = \underline{U} \begin{pmatrix} x \\ y \end{pmatrix}$$



*We will see in sect. 3.12 that G is a so-called χ^2 r.v. $P(G \leq k)$ can therefore also be found from the c.d.f. of the χ^2 distribution, tables of which, as well as computer routines, are readily available.

The new covariance matrix is $\underline{U}\,\underline{V}\,\underline{U}^{\mathrm{T}}$. Since
the transformation is unitary, areas are preserved (Jacobian $|J| = 1$). Hence,

$$P\left[(x,y) \text{ inside ellipse}\right] = P\left[(u,v) \text{ inside ellipse}\right]$$

The standard deviations of u, v are then found from the transformed covariance matrix. After some algebra we find

$$\sigma_u^2 = \frac{\sigma_x^2 \cos^2\theta - \sigma_y^2 \sin^2\theta}{\cos^2\theta - \sin^2\theta} \tag{3.23a}$$

$$\sigma_v^2 = \frac{\sigma_y^2 \cos^2\theta - \sigma_x^2 \sin^2\theta}{\cos^2\theta - \sin^2\theta} \tag{3.23b}$$

or

$$\sigma_u^2 = \frac{\sigma_x^2\sigma_y^2(1-\rho^2)}{\sigma_y^2\cos^2\theta - \rho\sigma_x\sigma_y \sin 2\theta + \sigma_x^2 \sin^2\theta} \tag{3.24a}$$

$$\sigma_v^2 = \frac{\sigma_x^2\sigma_y^2(1-\rho^2)}{\sigma_x^2\cos^2\theta + \rho\sigma_x\sigma_y \sin 2\theta + \sigma_y^2 \sin^2\theta} \tag{3.24b}$$

Or starting from the uncorrelated (diagonalized) variables (u,v), a rotation by θ to the new variables x, y will give

$$\sigma_x^2 = \sigma_u^2 \cos^2\theta + \sigma_v^2 \sin^2\theta \tag{3.25a}$$

$$\sigma_y^2 = \sigma_v^2 \cos^2\theta + \sigma_u^2 \sin^2\theta \tag{3.25b}$$

$$\rho = \sin\theta\cos\theta\,\frac{\sigma_u^2 - \sigma_v^2}{\sigma_x\sigma_y} \tag{3.25c}$$

Note that if ρ is fairly large, *i.e.*, the ellipse is thin, just knowing σ_x and σ_y would give a very wrong impression of how close a point (x, y) is to (μ_x, μ_y).

The properties stated at the end of the previous section, regarding the conditional and marginal distributions of the multivariate normal p.d.f. can be easily verified for the bivariate normal. In particular, the marginal p.d.f. is

$$f_x(x) = N(x; \mu_x, \sigma_x^2) \tag{3.26}$$

and the conditional p.d.f. is

$$\begin{aligned} f(y\,|\,x) = \frac{f(y,x)}{f_x(x)} &= \frac{1}{\sqrt{2\pi\sigma_y^2}\sqrt{1-\rho^2}}\ \exp\left\{-\frac{1}{2\sigma_y^2(1-\rho^2)}\left[y - \left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)\right]^2\right\} \\ &= N\left(y; \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), \sigma_y^2\left[1-\rho^2\right]\right) \end{aligned} \tag{3.27}$$

3.11 Cauchy (Breit-Wigner or Lorentzian) p.d.f.

The Cauchy p.d.f. is

$$C(x;\mu,\alpha) = \frac{1}{\pi\alpha}\frac{1}{1+(x-\mu)^2/\alpha^2} \tag{3.28}$$

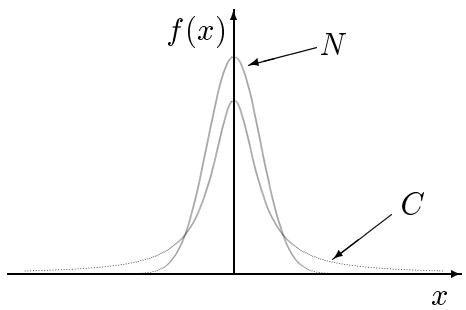
or in its 'standard' form with $\mu = 0$ and $\alpha = 1$,

$$C(x;0,1) = \frac{1}{\pi}\frac{1}{1+x^2} \tag{3.29}$$

It looks something like a Gaussian, but with bigger tails.

It is usually encountered in physics in a slightly different form as the Breit-Wigner (or Lorentz) function which gives the distribution of particles of mass m due to a resonance of mass M and width Γ:

$$f(m;M,\Gamma) = \frac{1}{2\pi}\frac{\Gamma}{(m-M)^2+(\frac{\Gamma}{2})^2}$$



M is the mode and Γ the full width at half maximum (FWHM) of the distribution.

The Cauchy p.d.f. is a pathological distribution. Let us try to calculate the mean:

$$\begin{aligned} E[x] &= \frac{1}{\pi\alpha}\int_{-\infty}^{+\infty}\frac{x}{1+\frac{(x-\mu)^2}{\alpha^2}}\,\mathrm{d}x = \frac{1}{\pi\alpha}\int_{-\infty}^{+\infty}\frac{(x-\mu)+\mu}{1+\frac{(x-\mu)^2}{\alpha^2}}\,\mathrm{d}x \\ &= \frac{\alpha}{\pi}\int_{-\infty}^{+\infty}\frac{z}{1+z^2}\,\mathrm{d}z + \frac{\mu}{\pi}\int_{-\infty}^{+\infty}\frac{1}{1+z^2}\,\mathrm{d}z\ , \quad z = \tfrac{x-\mu}{\alpha} \\ &= \frac{\alpha}{2\pi}\ln(1+z^2)\Big]_{-\infty}^{+\infty} + \frac{\mu}{\pi}\pi \ = \ +\infty - \infty + \mu \end{aligned}$$

which is indeterminate. The mean does not exist! However, noting that the p.d.f. is symmetric about μ, we can define the mean as

$$\lim_{L\to\infty}\int_{\mu-L}^{\mu+L} xC(x;\mu,\alpha)\,\mathrm{d}x = \mu$$

All higher moments are also divergent, and no such trick will allow us to define them. In actual physical problems the distribution is truncated, *e.g.*, by energy conservation, and the resulting distribution is well-behaved.

The characteristic function of the Cauchy p.d.f. is

$$\phi(t) = e^{-\alpha|t|+\imath\mu t}$$

The reproductive property of the Cauchy p.d.f. is rather unusual: $\overline{x} = \frac{1}{n}\sum x_i$ is distributed according to the identical Cauchy p.d.f. as are the x_i. (The proof is left as an exercise.)

3.12 The χ^2 p.d.f.

Let $\underline{x}$ be a vector of n independent r.v.'s, x_i, each distributed normally with mean μ_i and variance σ_i^2. Then the joint p.d.f. is

$$\begin{aligned} f(\underline{x};\underline{\mu},\underline{\sigma}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \\ &= \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \end{aligned}$$

The variable χ^2 is defined:

$$\chi^2(n) = \sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2 \tag{3.30}$$

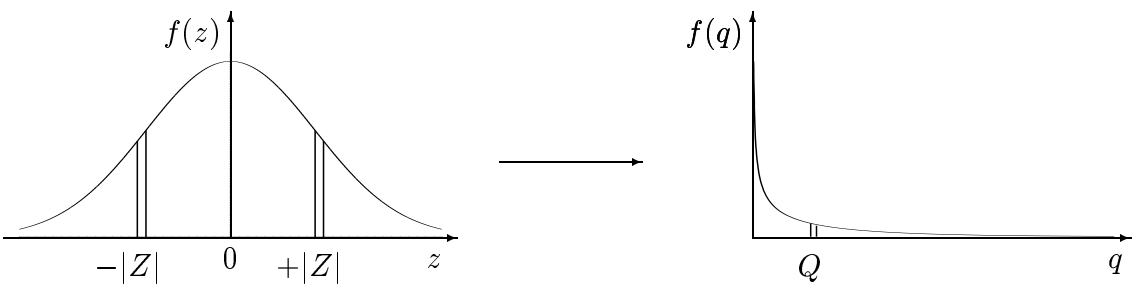
Being a function of r.v.'s, χ^2 is itself a r.v. The χ^2 has a parameter n, which is called the number of degrees of freedom (d.o.f.), since each of the r.v.'s, x_i, is free to vary independently of the others. Note that χ^2 is regarded as the variable, not the square of a variable; one does not usually refer to $\chi = \sqrt{\chi^2}$.

χ^2 with 1 d.o.f.

For example, for $n = 1$, letting $z = (x-\mu)/\sigma$, the p.d.f. for z is $N(z;0,1)$ and the probability that $z \le Z \le z + \mathrm{d}z$ is

$$f(z)\,\mathrm{d}z = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}\,\mathrm{d}z$$

Let $Q = Z^2$. (We use Q here instead of χ^2 to emphasize that this is the variable.) This is not a one-to-one transformation; both $+Z$ and $-Z$ go into $+Q$.



The probability that Q is between q and $q + \mathrm{d}q$ is the sum of the probability that Z is between z and $z + \mathrm{d}z$ around $z = \sqrt{q}$ and the probability that Z is between z and $z - \mathrm{d}z$ around $z = -\sqrt{q}$. Therefore, we must add the p.d.f. obtained from the $+Z \to q$ transformation to that obtained from the $-Z \to q$ transformation. The

Jacobians for these two transformations are (*cf.* section 2.2.6)

$$J_\pm = \frac{\mathrm{d}(\pm z)}{\mathrm{d}q} = \pm\frac{1}{2\sqrt{q}}$$

$$f(q)\,\mathrm{d}q = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}q}\,(|J_+| + |J_-|)\;\mathrm{d}q = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}q}\left(\frac{\mathrm{d}q}{2\sqrt{q}} + \frac{\mathrm{d}q}{2\sqrt{q}}\right) = \frac{1}{\sqrt{2\pi q}}\,e^{-\frac{1}{2}q}\,\mathrm{d}q$$

Now Q was just χ^2. Hence the p.d.f. for χ^2 with 1 d.o.f. is

$$\chi^2(1) = f(\chi^2;1) = \frac{1}{\sqrt{2\pi\chi^2}}\;e^{-\frac{1}{2}\chi^2} \tag{3.31}$$

It may be confusing to use the same symbol, χ^2, for both the r.v. and its p.d.f., but that's life!

χ^2 with 3 degrees of freedom

For $n = 3$, using standardized normal variables $z_i = \left(\frac{x_i-\mu_i}{\sigma_i}\right)$, let

$$R^2 = \chi^2 = z_1^2 + z_2^2 + z_3^2$$

The joint probability is then

$$g(z_1, z_2, z_3)\,\mathrm{d}z_1\,\mathrm{d}z_2\,\mathrm{d}z_3 = \frac{1}{(2\pi)^{3/2}}\,e^{-R^2/2}\,\mathrm{d}z_1\,\mathrm{d}z_2\,\mathrm{d}z_3$$

Think of R as the radius of a sphere in 3-dimensional space. Then, clearly, $\mathrm{d}z_1\,\mathrm{d}z_2\,\mathrm{d}z_3 = R^2\,\mathrm{d}R\,\mathrm{d}\cos\theta\,\mathrm{d}\phi$. To get the marginal p.d.f. for R, we integrate over $\cos\theta$ and ϕ, which gives a factor 4π. Hence, the probability that R is between R and $R+\,\mathrm{d}R$ is

$$f(R)\,\mathrm{d}R = \frac{2}{\sqrt{2\pi}}\,R^2\,e^{-R^2/2}\,\mathrm{d}R$$

Now $\chi^2 = R^2$. Hence, $\mathrm{d}\chi^2 = 2R\,\mathrm{d}R$ and $\mathrm{d}R = \mathrm{d}\chi^2/2\sqrt{\chi^2}$. Hence,

$$f(\chi^2;3)\,\mathrm{d}\chi^2 = \frac{2}{\sqrt{2\pi}}\,\chi^2 e^{-\chi^2/2}\frac{\mathrm{d}\chi^2}{2\sqrt{\chi^2}}$$

$$\chi^2(3) = f(\chi^2;3) = \frac{(\chi^2)^{1/2}}{\sqrt{2\pi}}\,e^{-\chi^2/2} \tag{3.32}$$

χ^2 with n degrees of freedom

For n degrees of freedom, the p.d.f. of χ^2 is

$$\chi^2(n) = f(\chi^2;n) = \frac{(\chi^2)^{\frac{n}{2}-1}\,e^{-\chi^2/2}}{\Gamma(\frac{n}{2})\,2^{n/2}} \tag{3.33}$$

Properties:

mean	$\mu = E[\chi^2(n)]$	$=$	n
variance	$V[\chi^2(n)] = \sigma^2_{\chi^2(n)}$	$=$	$2n$
mode (max.) at	$\chi^2(n)$	$=$	$\begin{cases} n-2 & n \geq 2 \\ 0 & n \leq 2 \end{cases}$
skewness	γ_1	$=$	$2\sqrt{\frac{2}{n}}$
kurtosis	γ_2	$=$	$12/n$
characteristic function	$\phi(t)$	$=$	$(1-2\imath t)^{-n/2}$

Reproductive property: Let χ_i^2 be a set of variables which are distributed as $\chi^2(n_i)$. Then $\sum \chi_i^2$ is distributed as $\chi^2(\sum n_i)$. This is obvious from the definition of χ^2: The variables χ_1^2 and χ_2^2 are, by definition,

$$\chi_1^2(n_1) = \sum_{i=1}^{n_1} z_i^2 \qquad \text{and} \qquad \chi_2^2(n_2) = \sum_{i=n_1+1}^{n_1+n_2} z_i^2$$

Hence, their sum is

$$\chi^2_{n_1+n_2} = \chi^2_{n_1} + \chi^2_{n_2} = \sum_{i=1}^{n_1+n_2} z_i^2$$

which from the definition is a χ^2 of $(n_1 + n_2)$ degrees of freedom.

Since the expectation of a $\chi^2(n)$ is n, the expectation of $\chi^2(n)/n$ is 1. The quantity $\chi^2(n)/n$ is called a "**reduced χ^2**".

Asymptotically (for large n), the χ^2 p.d.f. approaches the normal distribution with mean n and variance $2n$:

$$f(\chi^2; n) = \chi^2(n) \longrightarrow N(\chi^2; n, 2n) \tag{3.34}$$

A faster convergence occurs for the variable $\sqrt{2\chi^2}$:

$$f(\sqrt{2\chi^2}; n) = \chi^2(\chi^2; n)\sqrt{2\chi^2} \longrightarrow N(\sqrt{2\chi^2}; \sqrt{2n-1}, 1) \tag{3.35}$$

This approximation is good for n greater than about 30.

General definition of χ^2

If the n Gaussian variables are not independent, we can change variables such that the covariance matrix is diagonalized. Since this is a unitary transformation, it does not change the covariance ellipse $G = k$. In the diagonal case $G \equiv \chi^2$. Hence, $\chi^2 = G$ also in the correlated case. Thus we can take

$$\chi^2 = (\underline{x} - \underline{\mu})^{\mathrm{T}} \underline{V}^{-1} (\underline{x} - \underline{\mu}) \tag{3.36}$$

as the general definition of the random variable χ^2.

3.13 Student's t distribution

Consider an r.v., x, normally distributed with mean μ and standard deviation σ. Then $z = \frac{x-\mu}{\sigma}$ is normally distributed with mean 0 and standard deviation 1. In the normal p.d.f., the mean determines the origin and the standard deviation the scale. By transforming to the standard variable z, both dependences are removed.

In analyzing data we may not know the σ of the p.d.f. We may then remove the scale dependence by using the sample standard deviation, $\hat{\sigma}$, instead of the parent standard deviation. We may also not know the parent mean and will use the sample mean, $\bar{x}$, instead. For N independent x_i (*cf.* equations 8.3, 8.7),

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \qquad , \quad \text{using } \mu \tag{3.37a}$$

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2 \qquad , \quad \text{using } \bar{x} = \tfrac{1}{N}\textstyle\sum x_i \tag{3.37b}$$

In either case, $n\hat{\sigma}^2/\sigma^2$ is a $\chi^2(n)$, *i.e.*, is distributed according to the χ^2 distribution for $n = N - k$ degrees of freedom, where k is 0 if μ is used and is 1 if $\bar{x}$ is used, since in the latter case only $N-1$ of the terms in the sum are independent. This is discussed in more detail in section 8.2.1.

We now seek the p.d.f. for the r.v.

$$t = \frac{x-\mu}{\hat{\sigma}} = \frac{(x-\mu)/\sigma}{\sqrt{(n\hat{\sigma}^2/\sigma^2)/n}} = \frac{z}{\sqrt{\chi^2/n}} \tag{3.38}$$

Now z is a standard normal r.v. and χ^2 is a $\chi^2(n)$. A Student's t r.v. is thus the ratio of a standard normal r.v. to the square root of a reduced χ^2 r.v. The joint p.d.f. for z and χ^2 is then (equation 3.33)

$$f(z,\chi^2;n)\,\mathrm{d}z\,\mathrm{d}\chi^2 = N(z;0,1)\,\chi^2(\chi^2;n)\,\mathrm{d}z\,\mathrm{d}\chi^2 = \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,\frac{(\chi^2)^{\frac{n}{2}-1}\,e^{-\chi^2/2}}{\Gamma(\frac{n}{2})\,2^{n/2}}\,\mathrm{d}z\,\mathrm{d}\chi^2$$

where we have assumed that z and χ^2 are independent. This is certainly so if the x has not been used in determining $\hat{\sigma}$, and asymptotically so if n is large. Making a change of variable, we transform this distribution to one for t and χ^2:

$$f(t,\chi^2;n)\;\mathrm{d}t\,\mathrm{d}\chi^2 = \frac{1}{\sqrt{2\pi n}\,\Gamma(\frac{n}{2})\,2^{n/2}}(\chi^2)^{\frac{n-1}{2}}e^{-\frac{\chi^2}{2}(1+\frac{t^2}{n})}\;\mathrm{d}t\,\mathrm{d}\chi^2$$

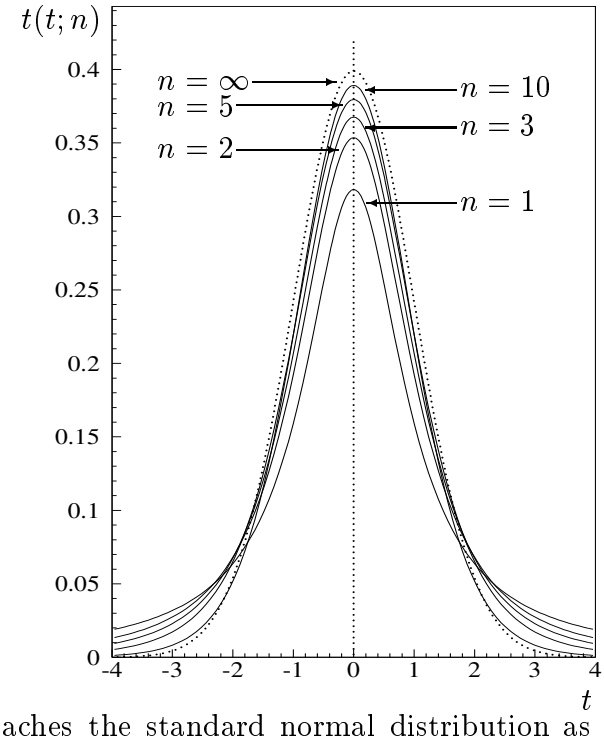
Integrating this over all χ^2, we arrive finally at the p.d.f. for t, called Student's t distribution,

$$t(n) = f(t;n) = \frac{1}{\sqrt{\pi n}}\,\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\,\frac{1}{(1+\frac{t^2}{n})^{(n+1)/2}} \tag{3.39}$$

Properties:

mean	$\mu = E[t]$	$=$	0	$,\ n > 1$
variance	$V[t] = \sigma_t^2$	$=$	$\frac{n}{n-2}$	$,\ n > 2$
skewness	γ_1	$=$	0	
kurtosis	γ_2	$=$	$\frac{6}{n-4}$	$,\ n > 4$
moments	μ_r	$=$	$\begin{cases} \frac{n^{2r}\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{n-r}{2}\right)}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} & , \quad r \text{ even and } r < n \\ 0 \quad , & \quad r \text{ odd and } r \leq n \\ \text{does not exist} & , \quad \text{otherwise.} \end{cases}$	

Student's t distribution is thus the p.d.f. of a r.v., t, which is the ratio of a standard normal variable and the square root of a normalized χ^2 r.v., *i.e.*, $\sqrt{\chi^2(n)/n}$, of n degrees of freedom. It was discovered[29] by W. S. Gossett, a chemist working for the Guinness brewery in Dublin, who in his spare time wrote articles on statistics under the pseudonym* "Student". The number of degrees of freedom, n, is not required to be an integer. The t-distribution with non-integral $n > 0$ is useful in certain applications, which is, however, beyond the scope of this course.



Student's t distribution is symmetric about $t = 0$. It approaches the standard normal distribution as the number of degrees of freedom, n, approaches infinity. For $n = 1$ it is identical to the standard Cauchy p.d.f. As $n \to \infty$, it approaches the standard normal distribution. It thus has larger tails and a larger variance than the Gaussian, but not so large as the Cauchy distribution.

We have constructed t from a single observation, x. In a similar way, a r.v. t can be constructed for the mean of a number of r.v.'s each distributed normally with mean μ and standard deviation σ. We know from the reproductive property of the normal p.d.f. that $\bar{x}$ is also normally distributed with mean μ but with a standard deviation of $\sigma/\sqrt{N}$. Thus $z = \frac{\bar{x}-\mu}{\sigma}\sqrt{N}$ is a standard normal r.v. and hence

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}}\sqrt{N} \tag{3.40}$$

*In order to prevent competitors from learning about procedures at Guinness, it was the policy of Guinness that articles by its employees be published under a pseudonym.

is distributed as Student's t with n degrees of freedom. It can be shown[3] that $\bar{x}$ and $\hat{\sigma}^2$ are independent.

3.14 The *F*-distribution

Consider two random variables, χ_1^2 and χ_2^2, distributed as χ^2 with ν_1 and ν_2 degrees of freedom, respectively. We define a new r.v., F, as the ratio of the two reduced χ^2:

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \tag{3.41}$$

The p.d.f. of F may be derived by a method similar to that used for Student's t distribution: Start with the joint p.d.f. of the independent variables χ_1^2, χ_2^2; make a change of variables to F, $v = \chi_2^2$; and integrate out the v dependence. The result is[2]

$$f(F;\nu_1,\nu_2) = \sqrt{\nu_1^{\nu_1}\nu_2^{\nu_2}}\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\frac{F^{\frac{\nu_1}{2}-1}}{(\nu_2+\nu_1 F)^{\frac{\nu_1+\nu_2}{2}}} \tag{3.42}$$

This distribution is known by many names: Fisher-Snedecor distribution, Fisher distribution, Snedecor distribution, variance ratio distribution, and F-distribution.

We could, of course, have written equation 3.41 with the ratio the other way around. By convention, one usually puts the larger value on top so that $F \geq 1$.

Properties:

mean	$\mu = E[F]$	$=$	$\frac{\nu_2}{\nu_2-\nu_1}$	, $\nu_2 > 2$
variance	$V[F]$	$=$	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$	, $\nu_2 > 4$

The distribution is positively skew and tends to mormality as $\nu_1, \nu_2 \longrightarrow \infty$, but only slowly ($\nu_1, \nu_2 > 50$).

The p.d.f. for $Z = \frac{1}{2}\ln F$ has a much faster approach to a Gaussian with a mean of $\frac{1}{2}(\frac{1}{\nu_2} - \frac{1}{\nu_1})$ and variance $\frac{1}{2}(\frac{1}{\nu_2} + \frac{1}{\nu_1})$.

The F-distribution is useful in various hypothesis tests (*cf.* sections 10.4.3 and 10.7.4). However, for the tests it may be more convenient to use

$$U = \frac{\nu_1 F}{\nu_2 + \nu_1 F} \tag{3.43}$$

which is a monotonic function of F and has a beta distribution (*cf.* section 3.15).

3.15 Beta distribution

This is a basic distribution for random variables bounded on both sides. Without loss of generality the bounds are here taken as $0 \leq x \leq 1$. It has two parameters (not necessarily integers): $n, m > 0$. The p.d.f. is

$$f(x;n,m) = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)}x^{m-1}(1-x)^{n-1} \qquad , \qquad 0 \leq x \leq 1 \tag{3.44}$$

$$= 0 \qquad , \qquad \text{otherwise}$$

Properties:

mean	$\mu = E[x]$	$=$	$\frac{m}{m+n}$
variance	$V[x]$	$=$	$\frac{mn}{(m+n)^2(m+n+1)}$

For $m = n = 1$ this becomes the uniform p.d.f.

Do not confuse the beta distribution with the beta function,

$$\beta(y,z) = \frac{\Gamma(y)\Gamma(z)}{\Gamma(y+z)} = \int_0^1 x^{y-1}(1-x)^{z-1}\,\mathrm{d}x \quad , \text{ real } y, z > 0$$

to which it is related, and from which the normalization of the p.d.f. is easily derived.

3.16 Double exponential (Laplace) distribution

This distribution is symmetric about the mean. Its tails fall off less sharply than the Gaussian, but faster than the Cauchy distribution. Note that its first derivative is discontinuous at $x = \mu$.

$$f(x;\mu,\lambda) = \frac{\lambda}{2}\exp\left(-\lambda|x-\mu|\right) \tag{3.45}$$

Properties:

mean	$\mu = E[x]$	$=$	μ
variance	$V[x]$	$=$	$2/\lambda^2$
skewness	γ_1	$=$	0
kurtosis	γ_2	$=$	3
characteristic function	$\phi(t)$	$=$	$\imath t\mu + \frac{\lambda^2}{\lambda^2+t^2}$

It can also be written

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2}}\exp\left(-\sqrt{2}\frac{|x-\mu|}{\sigma}\right) \tag{3.46}$$

3.17 Weibull distribution

Originally invented to describe failure rates in ageing lightbulbs, it describes a wide variety of complex phenomena.

$$f(t;\alpha,\lambda) = \alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^\alpha} \qquad \text{real } t \geq 0 \text{ and } \alpha, \lambda > 0 \tag{3.47}$$

Properties:

mean	$\mu = E[x]$	$=$	$\frac{1}{\lambda}\Gamma\left(\frac{1}{\alpha}+1\right)$
variance	$V[x]$	$=$	$\frac{1}{\lambda^2}\left\{\Gamma\left(\frac{2}{\alpha}+1\right) - \left[\Gamma\left(\frac{1}{\alpha}+1\right)\right]^2\right\}$

The exponential distribution (equation 3.10) is a special case ($\alpha = 1$), when the probability of failure at time t is independent of t.

Chapter 4

Real p.d.f.'s

There are, of course, many other distributions which we have not discussed in the previous section. We may introduce a few more later when needed. Now let us turn to some complicatons which we will encounter in trying to use these distributions.

4.1 Complications in real life

So far we have treated probability and handled some ideal p.d.f.'s. Given the p.d.f. for the physical process we want to study, we can, in principle, calculate the probability of a given experimental result. There are, however, some complications:

- In real life the p.d.f. is quite likely not one of the ideal distributions we have studied. It may be difficult to calculate. Or it may not even be known.

- The range of variables is never the $-\infty$ to $+\infty$ we have so blithely assumed. Either it is limited by physics, *e.g.*, conservation of energy, or by our apparatus, *e.g.*, a given radio telescope only works in a certain range of frequencies, in which case we must use the conditional p.d.f., $f(x|x_{\text{min}} \leq x \leq x_{\text{max}})$. While truncation is usually a complication, making the p.d.f. more difficult to calculate (*e.g.*, we must renormalize, which frequently can only be done by numerical integration), occasionally it is welcome, *e.g.*, the Cauchy p.d.f. becomes well-behaved if truncated at $\mu \pm a$:

$$\begin{aligned} C(x;\mu=0,\alpha=1) &= \frac{1}{\pi}\frac{1}{1+x^2} \\ &\longrightarrow \frac{C(x;0,1)}{\int_{-a}^{+a} C(x;0,1)\,\mathrm{d}x} = \frac{1}{2\arctan a}\cdot\frac{1}{1+x^2} \end{aligned}$$

  which has a finite variance (recall that the Cauchy p.d.f. did not):

$$V[x] = \frac{1}{\arctan a}\int_{-a}^{+a}\frac{x^2}{1+x^2}\,\mathrm{d}x = \frac{a}{\arctan a} - 1$$

- The physical p.d.f. may be modified by the response of the detector. This response must then be convoluted with the physical p.d.f. to obtain the p.d.f. which is actually sampled.

"Now we see in a mirror dimly ...
Now I know in part ..."
—1 Corinthians **13***:12*

4.2 Convolution

Experimentally we often measure the sum of two (or more) r.v.'s. For example, in the decay $\mathrm{n} \to \mathrm{p}\mathrm{e}^-\overline{\nu}_\mathrm{e}$ we want to measure the energy of the electron, which is distributed according to a p.d.f. given by the theory of weak interactions, $f_1(E_1)$. But we measure this energy with some apparatus, which has a certain resolution. Thus we do not record the actual energy E_1 of the electron but $E_1 + \delta$, where δ is distributed according to the resolution function (p.d.f.) of the apparatus, $f_2(\delta)$. What is then the p.d.f., $f(E)$, of the quantity we record, *i.e.*, $E = E_1 + \delta$? This $f(E)$ is called the (Fourier) convolution of f_1 and f_2.

Assume E_1 and δ to be independent. This may seem reasonable since E_1 is from the physical process (n decay) and δ is something extra added by the apparatus, which has nothing at all to do with the decay itself. Then the joint p.d.f. is

$$f_{12}(E_1, \delta) = f_1(E_1)\, f_2(\delta)$$

The c.d.f. of $E = E_1 + \delta$ is then

$$\begin{aligned} F(E) &= \iint\limits_{E_1 + \delta \le E} f_1(E_1) f_2(\delta)\, \mathrm{d}E_1\, \mathrm{d}\delta \\ &= \int_{-\infty}^{+\infty} \mathrm{d}E_1\, f_1(E_1) \int_{-\infty}^{E-E_1} \mathrm{d}\delta\, f_2(\delta) \\ &= \int_{-\infty}^{+\infty} \mathrm{d}E_1\, f_1(E_1)\, F_2(E - E_1) \\ \text{or} \quad &= \int_{-\infty}^{+\infty} \mathrm{d}\delta\, f_2(\delta)\, F_1(E - \delta) \end{aligned}$$

The p.d.f. can then be calculated from the c.d.f.:

$$\begin{aligned} f(E) = \frac{\mathrm{d}F(E)}{\mathrm{d}E} &= \int_{-\infty}^{+\infty} f_1(E_1) f_2(E - E_1)\, \mathrm{d}E_1 \\ \text{or} \qquad &= \int_{-\infty}^{+\infty} f_2(\delta) f_1(E - \delta)\, \mathrm{d}\delta \end{aligned}$$

The characteristic function is particularly useful in evaluating convolutions:

$$\begin{aligned}
\phi_f(t) &= \int e^{\imath tE} f(E)\,\mathrm{d}E \\
&= \int e^{\imath tE} \int f_1(E_1) f_2(E-E_1)\,\mathrm{d}E_1\,\mathrm{d}E \\
&= \int\int e^{\imath tE_1} f_1(E_1) e^{\imath t(E-E_1)} f_2(E-E_1)\,\mathrm{d}E_1\,\mathrm{d}E \\
&\qquad\qquad\qquad\qquad \text{since } E = E_1 + (E-E_1) \\
&= \phi_{f_1}(t)\,\phi_{f_2}(t) \qquad (4.1)
\end{aligned}$$

Thus, assuming that the r.v.'s are independent, the characteristic function of a convolution is just the product of the individual characteristic functions. (This probably looks rather familiar. We have already seen it in connection with the reproductive property of distributions; in that case f_1 and f_2 were the same p.d.f.) Recall that the characteristic function is a Fourier transform. Hence, a convolution, $E = E_1 + \delta$, where δ is independent of E, is known as a Fourier convolution.

Another type of convolution, called the Mellin convolution, involves the product of two random variables, *e.g.*, $E = E_1 R_1$. As we shall see, the Fourier convolution is easily evaluated using the characteristic function, which is essentially a Fourier transform of the p.d.f. Similarly, the Mellin convolution can be solved using the Mellin transformation, but we shall not cover that here.

In the above example we have assumed a detector response independent of what is being measured. In practice, the distortion of the input signal usually depends on the signal itself. This can occur in two ways:

1. Detection efficiency. The chance of detecting an event with our apparatus may depend on the properties of the event itself. For example, we want to measure the frequency distribution of electromagnetic radiation incident on the earth. But some of this radiation is absorbed by the atmosphere. Let $f(x)$ be the p.d.f. for the frequency, x, of incident radiation and let $e(x)$ be the probability that we will detect a photon of frequency x incident on the earth. Both f and e may depend on other parameters, y, *e.g.*, the direction in space in which we look. The p.d.f. of the frequency of the photons which we detect is
$$g(x) = \frac{\int f(x,y)e(x,y)\,\mathrm{d}y}{\int\int f(x,y)e(x,y)\,\mathrm{d}x\,\mathrm{d}y}$$

2. Resolution. To continue with the above example, suppose the detector records frequency x' when a photon of frequency x is incident. Let $r(x',x)$ be the p.d.f. that this will occur. Then
$$g(x') = \int r(x',x) f(x)\,\mathrm{d}x$$
In the case that r is just a function of $x - x'$ we get the simple convolution handled above. Note that resolution effects can lead to values of x' which lie

outside the physical range of x, *e.g.*, an energy of a particle which is larger than the maximum energy allowed by energy conservation. The Central Limit Theorem (chapter 5) will tell us that the detector response, or resolution function, is usually normally distributed for a given input to the detector:

$$\begin{aligned} r(x', x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(x'-x)^2}{\sigma^2}\right] \\ &= N(x'; x, \sigma^2) \qquad\qquad \text{if } \sigma \text{ is constant} \end{aligned}$$

However in practice σ often depends on x, in which case $r(x', x)$ may still have the above form, but is not really a Gaussian.

If the resolution function is Gaussian and if the physical p.d.f., $f(x)$, is also Gaussian, $f(x) = N(x; \mu, \tau^2)$, then you can show, by using the reproductive property of the Gaussian (exercise 19) or by evaluating the convolution using the characteristic function (equation 4.1), that the p.d.f. for x' is also normal:

$$g(x') = \int_{-\infty}^{+\infty} f(x)\, r(x', x)\, \mathrm{d}x = N\left(x'; \mu, \sigma^2 + \tau^2\right)$$

Chapter 5

Central Limit Theorem

5.1 The Theorem

This is a very important theorem; you could call it the 'central' theorem of statistics. It states:

Given n independent variables, x_i, distributed according to p.d.f.'s, f_i, having mean μ_i and variance $V_i = \sigma_i^2$, then the p.d.f. for the sum of the x_i, $S \equiv \sum x_i$, has expectation (mean) $E[S] = \sum \mu_i$ and variance $V[S] = \sum V_i = \sum \sigma_i^2$ and approaches the normal p.d.f. $N\left(S; \sum \mu_i, \sum \sigma_i^2\right)$ as $n \to \infty$:

$$\lim_{n\to\infty} f(S) \to N\left(S; \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \quad , \qquad S = \sum_{i=1}^{n} x_i \tag{5.1}$$

It must be emphasized that the mean and variance must exist.

It is left as an exercise to show that

$$\mu_S = \sum \mu_i \tag{5.2}$$

$$\text{and} \quad \sigma_S^2 = V[S] = \sum V_i = \sum \sigma_i^2 \tag{5.3}$$

Proving the C.L.T. in the general case is a bit too difficult for us. We will only demonstrate it for the restricted case where all the p.d.f's are the same, $f_i = f$. Without loss of generality we can let $\mu = 0$. Then $\sigma^2 = E[x^2]$. We also assume not only that the mean and variance of f are finite, but also that the expectations of higher powers of x are finite such that we can expand the characteristic function of f (equation 2.31):

$$\begin{aligned}\phi_x(t) = E\left[e^{\imath tx}\right] &= 1 + \frac{(\imath t)^2}{2}\sigma^2 + \frac{(\imath t)^3}{3!}E\left[x^3\right] + \ldots \\ &= 1 - \frac{\sigma^2 t^2}{2} + \ldots\end{aligned}$$

Let $u = \frac{x}{\sigma\sqrt{n}}$. The p.d.f. for u has variance $1/n$. Then

$$\phi_u(t) = E\left[e^{\imath tu}\right] = 1 - \frac{t^2}{2n} + \ldots$$

Now recall that the characteristic function of a sum of independent r.v.'s is the product of the individual characteristic functions. Therefore, the characteristic function of $S_u = \sum u_i$ is

$$\phi_{S_u}(t) = [\phi_u(t)]^n = \left[1 - \frac{t^2}{2n} + \ldots\right]^n$$

which in the limit $n \to \infty$ is just an exponential:

$$\phi_{S_u}(t) = \exp\left(-\frac{t^2}{2}\right)$$

But this is just the characteristic function of the standard normal $N(S_u; 0, 1)$. Since $S_u = \sum u_i = \frac{1}{\sigma\sqrt{n}} S$, the p.d.f. for $\sum x_i$ is the normal p.d.f. $N(S; n\mu, n\sigma^2)$.

A corallary of the C.L.T.: Under the conditions of the C.L.T., the p.d.f. of S/n approaches the normal p.d.f. as $n \to \infty$:

$$\lim_{n\to\infty} f\left(\frac{S}{n}\right) = N\left(\frac{S}{n}; \frac{\sum \mu_i}{n}, \frac{\sum \sigma_i^2}{n^2}\right) \quad , \qquad S = \sum_{i=1}^{n} x_i \tag{5.4}$$

or in the case that all the f_i are the same:

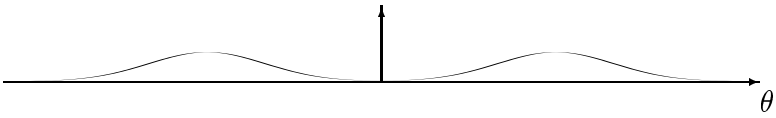
$$\lim_{n\to\infty} f\left(\frac{S}{n}\right) = N\left(\frac{S}{n}; \mu, \frac{\sigma^2}{n}\right) \quad , \qquad S = \sum_{i=1}^{n} x_i \tag{5.5}$$

5.2 Implications for measurements

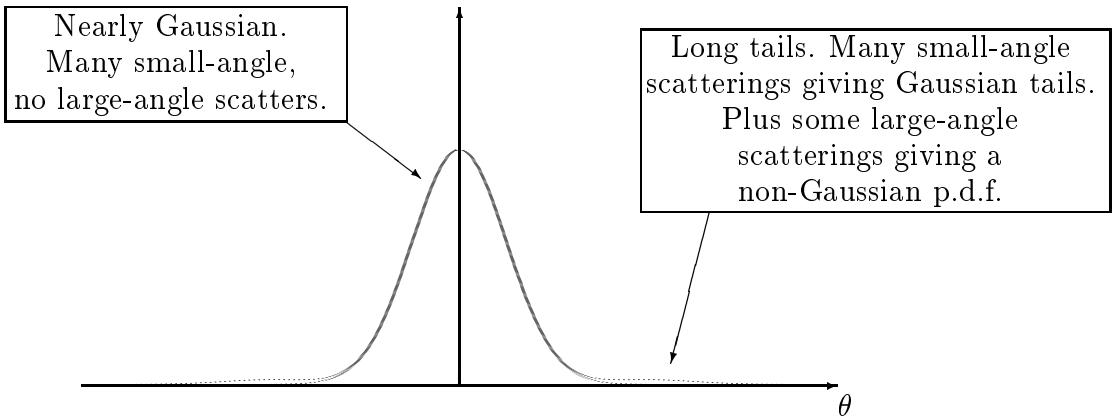
The C.L.T. shows why the Gaussian p.d.f. is so important. Most of what we measure is in fact the sum of many r.v.'s. For example, you measure the length of a table with a ruler. The length you measure depends on a lot of small effects: optical parallax, calibration of the ruler, temperature, your shaking hand, *etc.* A digital meter has electronic noise at various places in its circuitry. Thus, what you measure is not only what you want to measure, but added to it a large number of (hopefully) small contributions. If this number of small contributions is large the C.L.T. tells us that their total sum is Gaussian distributed. This is often the case and is the reason resolution functions are usually Gaussian. But if there are only a few contributions, or if a few of the contributions are much larger than the rest, the C.L.T. is not applicable, and the sum is not necessarily Gaussian.

Consider the passage of particles, *e.g.*, an α particle, through matter. Usually the α undergoes a large number of small-angle scatters producing a small net deflection. This net deflection is Gaussian distributed since it results from a large number of individual scatters. However occasionally there is a large-angle scattering; usually not, but sometimes 1 and very rarely 2. The distribution of the scattering angle θ when there has been one or more large-angle scatters will not be Gaussian, since

1 or 2 is not a large number. Instead, the p.d.f. for θ will be the convolution of the Gaussian for the net deflection from many small-angle scatters with the actual p.d.f. for the large-angle scatters. It will look something like:



Adding this to the Gaussian p.d.f. for the much more likely case of no large-angle scatters will give a p.d.f. which looks almost like a Gaussian, but with larger tails:



This illustrates that the further you go from the mean, the worse the Gaussian approximation is likely to be.

The C.L.T. also shows the effect of repeated measurements of a quantity. For example, we measure the length of a table with a ruler. The variance of the p.d.f. for 1 measurement is σ^2; the variance of the p.d.f. for an average of n measurements is $\frac{\sigma^2}{n}$. Thus σ is reduced by $\sqrt{n}$.

If a r.v. is the product of many factors, then its logarithm is a sum of equally many terms. Assuming that the CLT holds for these terms, then the r.v. is asymptotically distributed as the log-normal distribution.

"You can ... never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant."
—Arthur Conan Doyle: Sherlock Holmes in "The Sign of Four"

Part II

Monte Carlo

Chapter 6

Monte Carlo

The term Monte Carlo is used for calculational techniques which make use of random numbers. These techniques represent the solution of a problem as a parameter of a hypothetical population, and use a random sequence of numbers to construct a sample of the population, from which statistical estimates of the parameter are obtained.

The Monte Carlo solution of a problem thus consists of three parts:

1. choice of the p.d.f. which describes the hypothetical population;
2. generation of a random sample of the hypothetical population using a random sequence of numbers; and
3. statistical estimation of the parameter in question from the random sample.

It is no accident that these three steps correspond to the three parts of these lectures. P.d.f.'s have been covered in part I; this part will cover the generation of a Monte Carlo sample according to a given p.d.f.; and part III will treat statistical estimation, which is done in the same way for Monte Carlo as for real samples.

If the solution of a problem is the number F, the Monte Carlo estimate of F will depend on, among other things, the random numbers used in the calculation. The introduction of randomness into an otherwise well-defined problem may seem rather strange, but we shall see that the results can be very good.

After a short treatment of random numbers (section 6.1) we will treat a common application of the Monte Carlo method, namely integration (section 6.2) for which the statistical estimation is particularly simple. Then, in section 6.3 we will treat methods to generate a Monte Carlo sample which can then be used with any of the statistical methods of part III.

6.1 Random number generators

Random number generators may be classified as true random number generators or as pseudo-random number generators.

6.1.1 True random number generators

True random number generators must be based on random physical processes, *e.g.*,

- the potential across a resistor, which arises from thermal noise.
- the time between the arrival of two cosmic rays.
- the number of radioactive decays in a fixed time interval.

An example of how we could use this last possibility is to turn on a counter for a fixed time interval, long enough that the average number of decays is large. If the number of detected decays is odd, we record a 1; if it is even, we record a 0. We repeat this the number of times necessary to make up the fraction part of our computer's word (assuming a binary computer). We then have a random number between 0 and 1.

Unfortunately, this procedure does not produce a uniform distribution if the probability of an odd number of decays is not equal to that of an even number. To remove this bias we could take pairs of bits: If both bits are the same, we reject both bits; if they are different, we accept the second one. The probability that we end up with a 1 is then the probability that we first get a zero and second a one; the probability that we end up with a zero is the probability that we first get a one and second a zero. Assuming no correlation between successive trials, these probabilities are equal and we have achieved our goal.

The main problem with such generators is that they are very slow. Not wanting to have too dangerous a source, *i.e.*, not too much more than the natural background (cosmic rays are about 200 $\mathrm{m}^{-2}\mathrm{s}^{-1}$), nor too large a detector, it is clear that we will have counting times of the order of milliseconds. For a 24-bit fraction, that means 24 counting intervals per real random number, or 96 intervals if we remove the bias. Thus we can easily spend of the order of 1 second to generate 1 random number!

They are also not, by their very nature, reproducible, which can be a problem when debugging a program.

6.1.2 Pseudo-random number generators

A pseudo-random number generator produces a sequence of numbers calculated by a strictly mathematical procedure, which nonetheless *appears* random by some statistical tests. Since the sequence is not really random, there will certainly exist some other statistical test for which it will fail to appear random.

Several algorithms have been used to produce pseudo-random generators,[30] descriptions of which are beyond the scope of this course. In `FORTRAN77`, generators have usually been introduced as functions with names such as `RAN`. The statement `X = RAN(0)` assigns the next number in the random number sequence to the variable `X`. The argument of the function is a dummy argument which is not used. The generation proceeds from a 'seed', each number in the sequence acting as the seed

for the next. Usually there is a provision allowing the user to set the seed at the start of his program and to find out what the seed is at the end. This feature allows a new run to be made starting where the previous run left off. In `FORTRAN90` this is standardized by providing an intrinsic subroutine, `random_number(h)`, which fills the real variable (or array) `h` with pseudo-random numbers in the interval $[0, 1)$. A subroutine `random_seed` is also provided to input a seed or to inquire what the seed is at any point in the calculation. However, no requirements are made on the quality of the generated sequence, which will therefore depend on the compiler used. In critical applications one may therefore prefer to use some other generator.

Recently, new methods have been developed resulting in pseudo-random number generators far better than the old ones.[31] In particular the short periods, *i.e.*, that the sequence repeats itself, of the old generators has been greatly lengthened. For example the generator `RANMAR` has a period of the order of 10^{43}. The new generators are generally written as subroutines returning an array of random numbers rather than as a function, since the time to call a subroutine or invoke a function is of the same order as the time to generate one number, *e.g.*, `CALL RANMAR (RVEC,90)` to obtain the next 90 numbers in the sequence in the array `RVEC`, which of course must have a dimension of at least 90.

Some pseudo-random number generators generate numbers in the closed interval $[0, 1]$ rather than the open interval. Although it occurs very infrequently (once in 2^{24} on a 32-bit computer), the occurence of an exact 0 can be particularly annoying if you happen to divide by it. The open interval is therefore recommended.

Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin.
— John von Neumann

6.2 Monte Carlo integration

Much of this section has been taken from James[30] and Lyons[8].

We want to evaluate the integral

$$I = \int_a^b y(x)\,\mathrm{d}x \tag{6.1}$$

We will discuss several Monte Carlo methods to do so.

6.2.1 Crude Monte Carlo

A trivial (certainly not the best) numerical method is to divide the interval (a, b) into n subintervals and add up the areas of each subinterval using the value of y at the middle of the interval:

$$I = \frac{b-a}{n}\sum_{i=1}^{n} y(x_i)\ , \qquad x_i = a + \left(i - \frac{1}{2}\right)\frac{b-a}{n}$$

An obvious Monte Carlo method, called crude Monte Carlo, is to do the same sum, but with

$$x_i = a + r_i(b - a)$$

where the r_i are random numbers uniformly distributed on the interval $(0, 1)$.

More formally, the expectation of the function $y(x)$ given a p.d.f. $f(x)$ which is non-zero in (a, b) is given by

$$\mu_y = E[y] = \int_a^b y(x) f(x)\,\mathrm{d}x$$

Since the available pseudorandom number generators sample a uniform distribution, we take $f(x)$ to be the uniform p.d.f. $f(x) = 1/(b - a)$, $a \le x \le b$. Then

$$\mu_y = E[y] = \frac{1}{b-a}\int_a^b y(x)\,\mathrm{d}x = \frac{I}{b-a}$$
$$\sigma_y^2 = V[y] = \frac{1}{b-a}\int_a^b (y-\mu_y)^2\,\mathrm{d}x = \frac{1}{b-a}\int_a^b y^2\,\mathrm{d}x - \mu_y^2$$

Let us emphasize that μ_y and σ_y^2 are the expectation and variance of the function $y(x)$ for a uniform p.d.f. Do not confuse them with the mean and variance of a p.d.f.—$y(x)$ is *not* a p.d.f.

Let $y_i = y(x_i)$ where the x_i are distributed according to $f(x)$, *i.e.*, uniformly. Then, by the C.L.T., the average of the n values y_i approaches the normal distribution for large n:

$$N\left(\frac{\sum y_i}{n}; \mu_y, \frac{\sigma_y^2}{n}\right) = N\left(\frac{\sum y_i}{n}; \frac{I}{b-a}, \frac{\sigma_y^2}{n}\right)$$

We shall see in statistics (sect. 8.3) that an expectation value, *e.g.*, $E[y]$, can be estimated by the sample mean of the quantity, $\bar{y} = \sum y_i/n$.

Thus by generating n values x_i distributed uniformly in (a, b) and calculating the sample mean, we determine the value of $I/(b - a)$ to an uncertainty $\sigma_y/\sqrt{n}$:

$$I = \frac{b-a}{n}\sum_{i=1}^{n} y(x_i) \quad \pm \quad (b-a)\frac{\sigma_y}{\sqrt{n}} \tag{6.2}$$

In practice, if we do not know $\int y\,\mathrm{d}x$, it is unlikely that we know $\int y^2\,\mathrm{d}x$, which is necessary to calculate σ_y. However, we shall see that this too can be estimated from the Monte Carlo points (eq. 8.7):

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Since n is large, we can replace $n - 1$ by n. Multiplying out the sum we then get

$$\widehat{\sigma^2} = (\overline{y^2} - \bar{y}^2)$$

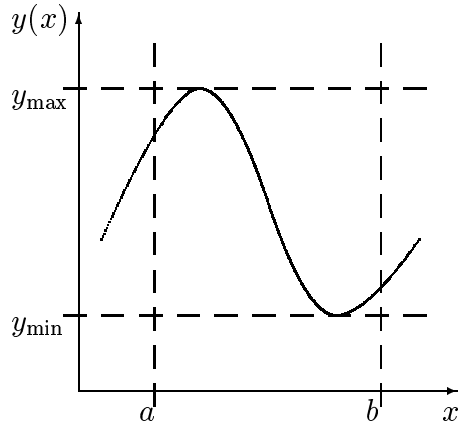
Hence the integral is estimated by

$$I = \int_a^b y(x)\,\mathrm{d}x = (b-a)\left(\bar{y} \quad \pm \quad \frac{1}{\sqrt{n}}\sqrt{\overline{y^2} - \bar{y}^2}\right) \tag{6.3}$$

Generalizing to more than one dimension is straightforward: Points are generated uniformly in the region of integration. The Monte Carlo estimate of the integral is still given by equation 6.3 if the length of the interval, $(b-a)$, is replaced by the volume of the region of integration.

6.2.2 Hit or Miss Monte Carlo

Another method to evaluate the integral (6.1) is by hit or miss Monte Carlo. In this method two random numbers are required per evaluation of $y(x)$. Let $R[x,y]$ be a random number uniformly distributed on (x,y). Then generate a point in the rectangle defined by the minimum and maximum values of y and the limits of integration, a and b:



$$x_i = R[a,b]$$
$$y_i = R[y_{\min}, y_{\max}]$$

If you do not know $y_{\min}$ and $y_{\max}$, you must guess 'safe' values. The generated point is called a

$$\begin{array}{lll} \text{'hit'} & \text{if} & y_i < y(x_i) \\ \text{'miss'} & \text{if} & y_i > y(x_i) \end{array}$$

Then an estimate of I is given by the fraction of points which are hits:

$$I = \frac{n_{\text{hits}}}{n}(b-a)(y_{\max} - y_{\min}) + y_{\min}(b-a)$$

Since hit or miss is a binomial situation, the number of hits follows the binomial p.d.f. with expectation $E[n_{\text{hits}}] = np$ and variance $V[n_{\text{hits}}] = np(1-p)$, where p is the probability of a hit. $V[I]$ is trivially related to $V[n_{\text{hits}}]$:

$$\begin{aligned} V[I] &= \frac{1}{n^2}V(n_{\text{hits}})(b-a)^2(y_{\max} - y_{\min})^2 \\ &= \frac{1}{n}p(1-p)(b-a)^2(y_{\max} - y_{\min})^2 \end{aligned}$$

The probability p, of a hit can be estimated from the result: $\hat{p} = n_{\text{hits}}/n$. Thus

$$\begin{aligned} I = {} & \frac{n_{\text{hits}}}{n}(b-a)(y_{\max} - y_{\min}) + y_{\min}(b-a) \\ & \pm \quad \frac{\sqrt{n_{\text{hits}}}}{n}\sqrt{\left(1 - \frac{n_{\text{hits}}}{n}\right)}\,(b-a)\,(y_{\max} - y_{\min}) \end{aligned} \tag{6.4}$$

Here too, the generalization to more than one dimension is straightforward: Points are generated uniformly in the region of integration and the function value is tested for a hit. The integral is then given by equation 6.4 with $(b-a)$ replaced by the volume of the region in which points were generated.

6.2.3 Buffon's needle, a hit or miss example

An early (1777) application of the Monte Carlo technique was to estimate the value of π. This calculation, known as Buffon's needle,[32] proceeds as follows: Parallel lines separated by distance d are drawn on the floor. A needle of length d is dropped on the floor such that its position (distance of the center of the needle to the nearest line) and its orientation (angle, θ, between the needle and a perpendicular to the lines) are both distributed uniformly. If the needle lies across a line we have a hit, otherwise a miss.

For a given θ, the chance of a hit is given by the conditional p.d.f.

$$f(\text{hit}|\theta) = \frac{\text{projected length of needle on a perpendicular}}{\text{distance between lines}} = \frac{d\cos\theta}{d} = \cos\theta$$

The chance of a hit irrespective of θ is then

$$p = \int_0^{\pi/2} f(\text{hit}|\theta) f(\theta)\,\mathrm{d}\theta = \int_0^{\pi/2} \cos\theta \underbrace{\frac{1}{\frac{\pi}{2}-0}}_{f(\theta)}\,\mathrm{d}\theta = \frac{2}{\pi}$$

Thus an estimate of $2/\pi$ is given by the estimator of p, namely $\hat{p} = n_{\text{hits}}/n$ and an estimate of π by $2n/n_{\text{hits}}$.

6.2.4 Accuracy of Monte Carlo integration

The uncertainty of the Monte Carlo integration decreases, for both crude and hit or miss Monte Carlo, with the number of points, n, as $n^{-1/2}$. However crude Monte Carlo is usually more accurate than the hit or miss method. For example, take the integral involved in Buffon's needle. In crude Monte Carlo,

$$\begin{aligned}
\mu_y = E[y] &= \frac{I}{b-a} = \frac{2}{\pi}\int_0^{\pi/2} \cos\theta\,\mathrm{d}\theta = \frac{2}{\pi} \\
V[y] &= \frac{1}{b-a}\int_0^{\pi/2} \cos^2\theta\,\mathrm{d}\theta - \mu_y^2 \\
&= \frac{1}{2} - \left(\frac{2}{\pi}\right)^2 = 0.0947
\end{aligned}$$

The uncertainty of the estimation of I is then $\sqrt{0.0947}/\sqrt{n} = 0.308/\sqrt{n}$.

On the other hand, hit or miss yields, using $p = 2/\pi$:

$$V[I] = \frac{1}{n}p(1-p)\left(\frac{\pi}{2}\right)^2 = \frac{0.571}{n}$$

The uncertainty of the estimation of I is then $\sqrt{0.571}/\sqrt{n} = 0.756/\sqrt{n}$, which is considerably larger (more than a factor 2) than for crude Monte Carlo.

The uncertainty of Monte Carlo integration is compared with that of numerical methods in the following table:

uncertainty in I calculated from n points		
method	1 dimension	d dimensions
Monte Carlo	$n^{-1/2}$	$n^{-1/2}$
trapezoidal rule	n^{-2}	$n^{-2/d}$
Simpson's rule	n^{-4}	$n^{-4/d}$
m-point Gauss rule	$n^{-(2m-1)}$	$n^{-(2m-1)/d}$

Thus we see that Monte Carlo integration converges much more slowly than other methods, particularly for low numbers of dimensions. Only for greater than 8 dimensions does Monte Carlo converge faster than Simpson's rule, and there is always a Gauss rule which converges faster than Monte Carlo.

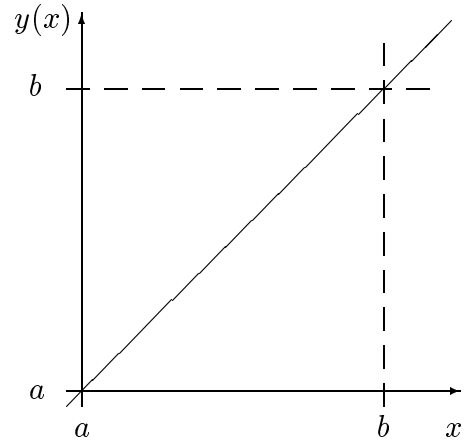
However, there are other considerations besides rate of convergence: The first is the question of feasibility. For example, in 38 dimensions a 10-point Gauss method converges at the same rate as Monte Carlo. However, in the Gauss method, the number of points is fixed, $n = m^d$, which in our example is 10^{38}. The evaluation of even a very simple function requires on the order of microseconds on a fast computer. So 10^{38} is clearly not feasible. (10^{32} sec. $\approx \pi \times 10^{24}$ years, while the age of the universe is only of order 12 Gyr.)

Another problem with numerical methods is the boundary of integration. If the boundary is complicated, numerical methods become very difficult. This is, however, easily handled in Monte Carlo. One simply takes the smallest hyperrectangle that will surround the region of integration and integrates over the hyperrectangle, throwing away the points that fall outside the region of integration. This leads to some inefficiency, but is straightforward. This is one of the chief advantages of the Monte Carlo technique. An example is given by phase space integrals in particle physics. Consider the decay $\mathrm{n} \to \mathrm{pe}^-\overline{\nu}_\mathrm{e}$, the neutron at rest. Calculations for this decay involve 9 variables, p_x, p_y, p_z for each of the 3 final-state particles. However these variables are not independent, being constrained by energy and momentum conservation, $\sum p_x = \sum p_y = \sum p_z = 0$, and $\sum E = m_\mathrm{n} c^2$, where the energy of a particle is given by, $E = \sqrt{m^2c^4 + p_x^2c^2 + p_y^2c^2 + p_z^2c^2}$. This complicated boundary makes an integration by numerical methods difficult; it becomes practically impossible for more than a few particles. However Monte Carlo integration is quite simple: one generates points uniformly in the integration variables, calculates the energy and momentum components for each particle and tests whether momentum and energy are conserved. If not, the point is simply rejected.

Another practical issue might be termed the growth rate. Suppose you have performed an integration and then decide that it is not accurate enough. With

Monte Carlo you just have to generate some more points (starting your random number generator where you left off the previous time). However, with the Gauss rule, you have to go to a higher order m. All the previous points are then useless and you have to start over.

6.2.5 A crude example in 2 dimensions



One of the advantages of Monte Carlo is the ease with which irregular integration regions can be handled. Consider a two-dimensional integral over a triangular region:

$$I = \int_a^b \mathrm{d}x \int_a^x \mathrm{d}y\, g(x,y)$$

We give five ways of estimating this integral using crude Monte Carlo:

1. The obvious way:
   (a) Choose $x_i = R[a,b]$.
   (b) Choose $y_i = R[a,x_i]$.
   (c) Sum the $g(x_i,y_i)$: $I = \frac{(b-a)^2}{2n} \sum_{i=1}^n g(x_i,y_i)$

   This method, although obvious, is incorrect. This is because the points (x_i,y_i) are not uniformly distributed over the region of integration. There are (approximately) the same number of points for $a < x < (a+b)/2$ as for $(a+b)/2 < x < b$, while the areas differ by a factor 3.

2. Rejection method:
   (a) Choose $x_i = R[a,b]$ and $y_i = R[a,b]$.
   (b) Define a new function $z(x,y)$ which is defined on the entire region for which points are generated, but which has the same integral as g:

   $$z_i = \begin{cases} 0, & \text{if } y_i > x_i, \\ g(x_i,y_i), & \text{if } y_i < x_i. \end{cases}$$

   Or, equivalently, reject the point if it does not lie in the region of integration, *i.e.*, if $y_i > x_i$.
   (c) Then sum the z_i:

   $$I = \frac{(b-a)^2}{n} \sum_{i=1}^n z_i$$

3. Rejection method (area of region of integration known): The above rejection method results in a perhaps needlessly large error since we are using Monte Carlo to estimate the integral of z, even where we know that $z = 0$. Another way of looking at it is that we are using Monte Carlo to estimate what fraction, f_a, of the area of point generation is taken up by the area of integration. Hence, if we know this fraction we can remove this contribution to the error by simply rejecting the points not in the area of integration. We proceed as follows:

   (a) Choose $x_i = R[a, b]$ and $y_i = R[a, b]$.

   (b) Reject the point if it does not lie in the region of integration, *i.e.*, if $y_i > x_i$.

   (c) Then sum the $g(x_i, y_i)$ replacing the area of point generation by the area of the region of integration, $f_a(b-a)^2$. In this example we know that $f_a = \frac{1}{2}$. The result is then

   $$I = \frac{1}{2}\frac{(b-a)^2}{n'}\sum_{i=1}^{n'} g(x_i, y_i)$$

   where n' is the number of generated points lying in the region of integration.

   Both rejection methods are correct, but inefficient—both use only half of the points. Sometimes this inefficiency can be overcome by a trick:

4. Folding method (a trick):

   (a) Choose $u_i = R[a, b]$ and $v_i = R[a, b]$.

   (b) Let $\quad x_i = \max(u_i, v_i) \quad$ and $\quad y_i = \min(u_i, v_i)$.

   (c) Then sum the g_i:

   $$I = \frac{(b-a)^2}{2n}\sum_{i=1}^{n} g(x_i, y_i)$$

   This is equivalent to generating points uniformly over the whole square and then folding the square about the diagonal so that all the points fall in the triangular region of integration. The density of points remains uniform.

5. Weighting method. We generate points as in the "obvious", but wrong, method:

   (a) Choose $x_i = R[a, b]$.

   (b) Choose $y_i = R[a, x_i]$.

   (c) But we make a weighted sum, the weight correcting for the unequal

density of points (density $\sim \frac{1}{(x-a)}$):

$$I = \frac{b-a}{n} \sum_{i=1}^{n} (x_i - a)\, g(x_i, y_i) \tag{6.5}$$

The derivation of this formula is left as an exercise (27).

This method is, in fact, an application of the technique of importance sampling (*cf.* section 6.2.6) It may, or may not, be more efficient than folding, depending on the function g. In particular, it will be more efficient when the variance of $(x-a)g$ is smaller than that of g.

6.2.6 Variance reducing techniques

As we have seen, Monte Carlo integration converges rather slowly with n compared to the better numerical techniques. There are, however, several methods of reducing the variance of the Monte Carlo estimation:

Stratification

In this approach we split the region of integration into two or more subregions. Then the integral is just the sum of the integrals over the subregions, *e.g.*, for two subregions,

$$I = \int_a^b y(x)\,\mathrm{d}x = \int_a^c y(x)\,\mathrm{d}x + \int_c^b y(x)\,\mathrm{d}x$$

The variance of I is just the sum of the variances of the subregions. A good choice of subregions and number of points in each region can result in a dramatic decrease in $V[I]$. However, to make a good choice requires knowledge of the function. A poor choice can increase the variance.

Some improvement can always be achieved by simply splitting the region into subregions of equal size and generating the same number of points for each subregion. We illustrate this, using crude Monte Carlo, for the case of two subregions:

For the entire region the variance is (from equation 6.2)

$$V_1(I) = \frac{(b-a)^2}{n}\sigma_y^2 = \frac{(b-a)^2}{n}\left[\frac{1}{b-a}\int_a^b y^2\,\mathrm{d}x - \left(\frac{1}{b-a}\int_a^b y\,\mathrm{d}x\right)^2\right]$$

For two equal regions, the variance is the sum of the variances of the two regions:

$$\begin{aligned} V_2(I) &= \frac{[(b-a)/2]^2}{n/2}\left\{\left[\frac{2}{b-a}\int_a^c y^2\,\mathrm{d}x - \left(\frac{2}{b-a}\int_a^c y\,\mathrm{d}x\right)^2\right]\right. \\ &\qquad \left. + \left[\frac{2}{b-a}\int_c^b y^2\,\mathrm{d}x - \left(\frac{2}{b-a}\int_c^b y\,\mathrm{d}x\right)^2\right]\right\} \\ &= \frac{(b-a)^2}{2n}\left\{\frac{2}{b-a}\int_a^b y^2\,\mathrm{d}x - \frac{4}{(b-a)^2}\left[\left(\int_a^c y\,\mathrm{d}x\right)^2 + \left(\int_c^b y\,\mathrm{d}x\right)^2\right]\right\} \end{aligned}$$

The improvement in variance is given by

$$V_1(I) - V_2(I) = -\frac{1}{n}\left(\int_a^b y\,\mathrm{d}x\right)^2 + \frac{2}{n}\left[\left(\int_a^c y\,\mathrm{d}x\right)^2 + \left(\int_c^b y\,\mathrm{d}x\right)^2\right]$$

Substituting $\quad A = \int_a^c y\,\mathrm{d}x \quad \text{and} \quad B = \int_c^b y\,\mathrm{d}x$

$$\begin{aligned} V_1(I) - V_2(I) &= \frac{1}{n}\left[-(A+B)^2 + 2\left(A^2+B^2\right)\right] \\ &= \frac{1}{n}(A-B)^2 \\ &\geq 0 \end{aligned}$$

Thus some improvement in the variance is attained, although it may be arbitrarily small. This improvement can be qualitatively understood as due to an increased uniformity of the distribution of points.

Importance Sampling

We have seen that (in crude Monte Carlo) the variance of the estimate of the integral is proportional to the variance of the function being integrated (eq. 6.2). Thus the less variation in y, *i.e.*, the more constant $y(x)$ is, the more accurate the integral. We can effectively achieve this by generating more points in regions of large y and compensating for the higher density of points by reducing the value of y (*i.e.*, giving a smaller weight) accordingly. This was also the motivation for stratification.

In importance sampling we change variable in order to have an integral with smaller variance:

$$I = \int_a^b y(x)\,\mathrm{d}x = \int_a^b \frac{y(x)}{g(x)}\,g(x)\,\mathrm{d}x = \int_{G(a)}^{G(b)} \frac{y(x)}{g(x)}\,\mathrm{d}G(x)$$

$$\text{where} \quad G(x) = \int_a^x g(x)\,\mathrm{d}x$$

Thus we must find a function $g(x)$ such that

- $g(x)$ is a p.d.f., *i.e.*, everywhere positive and normalized such that $G(b) = 1$.
- $G(x)$ is known analytically.
- Either $G(x)$ can be inverted (solved for x) or a random number generator is available which generates points (x) according to $g(x)$.
- The ratio $y(x)/g(x)$ is as nearly constant as possible and in any case more constant than $y(x)$, *i.e.*, $\sigma_{y/g} < \sigma_y$.

We then choose values of G randomly between 0 and 1; for each G solve for x; and sum $y(x)/g(x)$. The weighting method of section 6.2.5 was really an application of importance sampling.

Although importance sampling is a useful technique, it suffers in practice from a number of drawbacks:

- The class of functions g which are integrable and for which the integral can be inverted analytically is small—essentially only the trigonometric functions, exponentials, and polynomials. The inversion could in principle be done numerically, but this introduces inaccuracies which may be larger that the gain made in reducing the variance.
- It is very difficult in more than one dimension. In practice one usually uses a g which is a product of one-dimensional functions.
- It can be unstable. If g becomes small in a region, y/g becomes very big and hence the variance also. It is therefore dangerous to use a function g which is 0 in some region or which approaches 0 rapidly.
- Clearly $y(x)$ must be rather well known in order to choose a good function g.

On the other hand, an advantage of this method is that singularities in $y(x)$ can be removed by choosing a $g(x)$ having the same singularities.

Control Variates

This is similar to importance sampling except that instead of dividing by $g(x)$, we subtract it:

$$I = \int y(x)\,\mathrm{d}x = \int [y(x) - g(x)]\,\mathrm{d}x + \int g(x)\,\mathrm{d}x$$

Here, $\int g(x)\,\mathrm{d}x$ must be known, and g is chosen such that $y - g$ has a smaller variance than y. This method does not risk the instability of importance sampling. Nor is it necessary to invert the integral of $g(x)$.

Antithetic Variates

So far, we have always used Monte Carlo points which are independent. Here we deliberately introduce a correlation. Recall that the variance of the sum of two functions is

$$V\left[y_1(x) + y_2(x)\right] = V\left[y_1(x)\right] + V\left[y_2(x)\right] + 2\,\mathrm{cov}[y_1(x), y_2(x)]$$

Thus, if we can write

$$I = \int_a^b y\,\mathrm{d}x = \int_a^b (y_1 + y_2)\,\mathrm{d}x$$

such that y_1 and y_2 have a large negative correlation, we can reduce the variance of I. Clearly, we must understand the function $y(x)$ in order to do this. It is difficult to give general methods, but we will illustrate it with an example:

Suppose that we know that $y(x)$ is a monotonically increasing function of x.

Then let $y_1 = \frac{1}{2}y(x)$ and $y_2 = \frac{1}{2}y\,(b - (x - a))$. Clearly the integral of $(y_1 + y_2)$ is just the integral of y. However, since y is monotonically increasing, y_1 and y_2 are negatively correlated; when y_1 is small, y_2 is large and *vice versa*. If this negative correlation is large enough, $V[y_1 + y_2] < V[y]$.

6.3 Monte Carlo simulation

References for this section are James[30] and Lyons.[8] For further details and additional topics consult Rubinstein.[33]

Monte Carlo problems are usually classified as either integration or simulation. We shall be concerned with simulating experiments in physics. This begins with a theory or hypothesis about the physical process, *i.e.*, with the assumption of an underlying p.d.f., $g(\underline{x}')$, which may then be modified by the response function, $r(\underline{x}, \underline{x}')$, of the experimental apparatus. The expected p.d.f. of the observations is then given by

$$f(\underline{x}) = \int g(\underline{x}')\, r(\underline{x}, \underline{x}')\, \mathrm{d}\underline{x}'$$

The purpose of the simulation is to produce a set of n simulated or 'fake' data points distributed according to $f(\underline{x})$. These can be compared with the real data to test the hypothesis. They can also be used in the planning stage of the experiment to help in its design, *e.g.*, to compare the use of different apparatus, and to test software to be used in the analysis of the experiment.

Since these fake points are distributed according to $f(\underline{x})$, they are in fact just the points generated for the Monte Carlo integration of $\int f(\underline{x})\,\mathrm{d}x$. Simulation is thus, formally at least, equivalent to integration. The purpose is, however, usually different. This means that often a different Monte Carlo method will be preferred for simulation than for integration.

Although we will continue to use the term p.d.f. for $f(\underline{x})$, for the purposes of simulation the normalization is unimportant (at least if we are careful). It is, however, essential that the function not be negative.

The p.d.f. that we wish to simulate, $f(\underline{x})$, can be extremely complicated. The underlying physical p.d.f., $g(\underline{x}')$, may itself involve integrals which will be evaluated by Monte Carlo in the course of the simulation, and the detector description may consist of various stages, each depending on the previous one.

Monte Carlo simulation of such complex processes breaks them down into a series of steps. At each step a particular outcome is chosen from a set of possible outcomes according to a given p.d.f., $f(x)$. In other words, the outcome of the step is a (pseudo-)random number generated according to $f(x)$. But random number generators generally produce uniformly distributed numbers. We therefore must transform the uniformly distributed random numbers into random numbers distributed according to the desired p.d.f. There are three basic methods to do this:

6.3.1 Weighted events

This method is analogous to that of crude Monte Carlo for integration. For a p.d.f., $f(x)$, defined on the interval (a, b), points are generated uniformly in x and given a weight, w. An event then consists of the values x_i and $w_i = f(x_i)(b - a)$. The integral of $f(x)$ over any subinterval of (a, b), *e.g.*, (c, d) with $a \leq c < d \leq b$, is then given by the sum of the weights of the events in that interval:

$$\int_c^d f(x)\,\mathrm{d}x = \frac{1}{n} \sum_{c<x<d} w_i$$

In particular, a weighted histogram of the x_i (c and d are then the various bin limits), represents the p.d.f. and can be directly compared with the data.

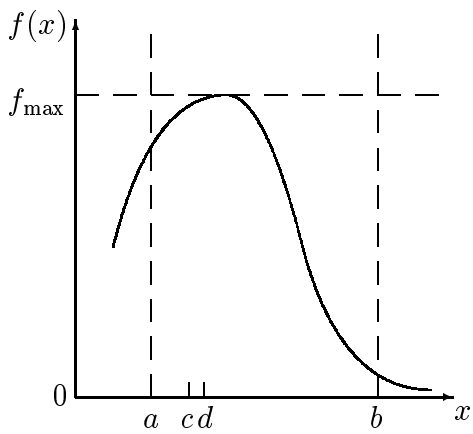
We have seen that integration by crude Monte Carlo gives a smaller variance than the hit-or-miss method, and is therefore generally preferable. However in simulation it is usually deemed preferable not to have weighted events. One prefers to have the Monte Carlo events as much as possible like the real events. In particular, it is usually desirable that the Monte Carlo sample behave statistically like the real event sample, *e.g.*, the variance of the average of n Monte Carlo points should result in the same variance as that of the average of n real points. This is not the case with weighted events. The density of Monte Carlo points where $f(x)$ is small is the same as where $f(x)$ is large, whereas in the real data the density of points is proportional to $f(x)$.

6.3.2 Rejection method

This method is analogous to the hit-or-miss method of Monte Carlo integration. As in hit-or-miss Monte Carlo, we generate points uniformly in x and in $f(x)$

$$x_i = R[a, b]$$
$$r_i = R[0, f_{\max}]$$

where $f_{\max}$ is the maximum value of $f(x)$ in (a, b). Points for which $f(x_i) < r_i$ are then rejected.



The integral over a subinterval (c, d) is then

$$\int_c^d f(x)\,\mathrm{d}x = \frac{n_{c<x<d}}{n}\,(b - a) f_{\max}$$

In hit-or-miss Monte Carlo we also introduced an $f_{\min}$. Since we knew the integral $\int_a^b f_{\min}\,\mathrm{d}x$, it was not necessary to evaluate it by Monte Carlo. It was

therefore better (more efficient) to use all the Monte Carlo points to evaluate $\int_a^b (f - f_{\min})\,\mathrm{d}x$. But here we want to generate all the events for f, not just for $(f - f_{\min})$.

The difficulty with this method lies in knowing $f_{\max}$. If we do not know it, then we must guess a 'safe' value, *i.e.*, a value which we are sure is larger than $f_{\max}$. If we choose $f_{\max}$ too safe, the method becomes inefficient. This method can be made more efficient by choosing different values of $f_{\max}$ in different regions.

This method is the easiest method to use for complicated functions in many dimensions.

6.3.3 Inverse transformation method

Continuous p.d.f.

This is like importance sampling with $g(x) = f(x)$. The resulting integrand is just the uniform distribution. We transform from $f(x)$ to F:

$$f(x)\,\mathrm{d}x = \mathrm{d}F$$

where $F(x)$ is just the c.d.f. of $f(x)$,

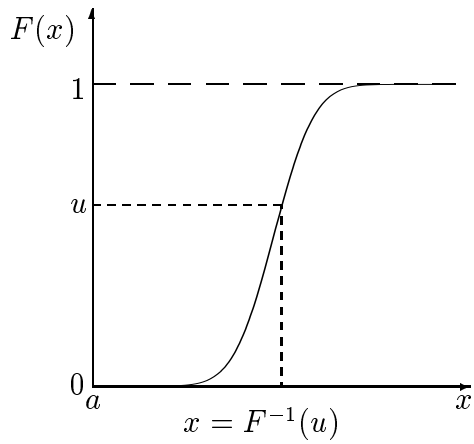
$$F(x) = \int_a^x f(x)\,\mathrm{d}x$$

Instead of generating points uniform in x, we generate points uniformly distributed in F between $F(a)$ and $F(b)$, which are 0 and 1, respectively, if $f(x)$ is a p.d.f. normalized on (a, b):

$$u_i = R\left[F(a), F(b)\right]$$

and calculate the corresponding value of x,

$$x_i = F^{-1}(u_i)$$



The x_i are then distributed as $f(x)$. To see this, recall the results on changing variables (sect. 2.2.6): For a transformation $u \to x = v(u)$ with inverse transformation $u = w(x)$, the p.d.f. for x is given by the p.d.f. for u, $g(u)$, times the Jacobian, *i.e.*,

$$\text{p.d.f. for } x = g(u)\left|\frac{\partial u}{\partial x}\right| = g\left(w(x)\right)\left|\frac{\partial w(x)}{\partial x}\right|$$

Here, $u = F(x)$, $x = F^{-1}(u)$ and u is distributed uniformly, *i.e.*, $g(u) = 1$. The p.d.f. for x is then

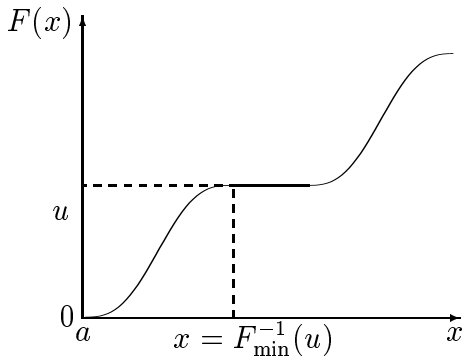
$$\left|\frac{\partial u}{\partial x}\right| = \left|\frac{\partial F(x)}{\partial x}\right| = f(x)$$

Hence, if $g(u)$ is a uniform distribution, the p.d.f. for x is $f(x)$, as desired.

The difficulties with this method are integrating $f(x)$ to obtain $F(x)$ and inverting $F(x)$ to obtain $F^{-1}(u)$. But if this can be done, this is usually the best method.

If F is not one-to-one, we define

$$x = F^{-1}_{\min}(u) = \min(x \text{ for which } F(x) \geq u)$$



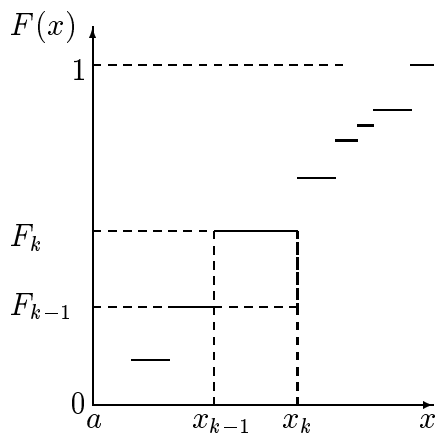
Discrete p.d.f.

For a discrete p.d.f., we can always use this method, since the c.d.f. is always easily calculated. The probability of $X = x_k$ is $P(X = x_k) = f(x_k)$. Then the c.d.f. is

$$F_k = P(X \leq x_k) = \sum_{i=1}^{k} f(x_k)$$

Taking u uniformly distributed between 0 and 1,

$$\begin{aligned} P(F_{k-1} < u \leq F_k) &= \int_{F_{k-1}}^{F_k} \mathrm{d}u \\ &= F_k - F_{k-1} = f(x_k) \end{aligned}$$



Thus, to generate a point, we

1. generate $u_i = R[0, 1]$
2. find the value of k such that

$$F_{k-1} < u_i \leq F_k$$

Then x_k is the desired value of x.

Step 2 of this procedure can involve a lot of steps. You can usually save computer time by starting the comparison somewhere in the middle of the x-range, say at the mean or mode, and then working up or down in x depending on u and F_k.

This is of interest not only for situations with a discrete p.d.f., but also for cases where the p.d.f. is continuous, but not known analytically. The resolution function of an apparatus is often determined experimentally and the resulting p.d.f. expressed as a histogram.

6.3.4 Composite method

It may be advantageous to decompose the desired p.d.f. into a sum of p.d.f.'s which are easier to generate:

$$f(x) = \sum f_k(x) \tag{6.6}$$

Let

$$\alpha_k = \frac{\int_a^b f_k(x)\,\mathrm{d}x}{\sum_j \int_a^b f_j(x)\,\mathrm{d}x} \tag{6.7}$$

Then $\sum \alpha_i = 1$, and α_k is the fraction of the points to be generated according to f_k.

In generating the points, we regard the index k as a discrete r.v. with probability α_k. We first generate $u = R[0,1]$ and use it to select k. Then we generate a value x_i according to $f_k(x)$ using one of the previous methods.

You might ask why not skip the first step and just generate exactly $\alpha_k n$ points according to f_k for each k, where n is the total number of points. This was a method (stratification) to improve the variance in Monte Carlo integration. The answer is that the variance of the Monte Carlo sample would then be different from that of a sample of n real events, while the purpose of simulation is usually to obtain a Monte Carlo sample having the same statistical properties as real events.

6.3.5 Example

As an example of the above methods, we take the p.d.f.

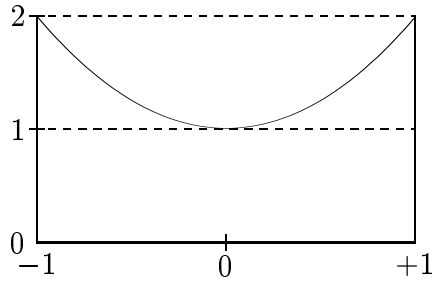
$$f(x) = 1 + x^2$$

in the region $(-1, 1)$. This could be an angular distribution with $x = \cos\theta$. We note that $f(x)$ is not normalized. We could, of course, normalize it, but choose not to do so. For as we shall see, for the purpose of generating events the normalization is not necessary.

Weighted events

This is completely trivial. We generate $x_i = R[-1, 1]$ and assign weight $w_i = 1 + x_i^2$.

Rejection method



We have $f_{\max} = 2$, $a = -1$, $b = +1$. Therefore, we generate

$$\begin{aligned} x_i &= R[-1,+1] &&= 2R[0,1] - 1 \\ r_i &= R[0,2] &&= 2R[0,1] \end{aligned}$$

and reject the point if $r_i > 1 + x_i^2$.

Note that the efficiency of the generation is $\frac{\int_{-1}^{1}(1+x^2)\,\mathrm{d}x}{(b-a)f_{\max}} = \frac{2}{3}$, *i.e.*, $1/3$ of the points are rejected.

Inverse transformation method

We have

$$F(x) = \int_{-1}^{x}(1+x^2)\,\mathrm{d}x = x + \frac{x^3}{3}\Big]_{-1}^{x} = x + \frac{x^3}{3} + \frac{4}{3}$$

Hence $F(-1) = 0$ and $F(1) = 8/3$. Therefore generate $u = R[0,1]$. Then $\frac{8}{3}u$ is uniformly distributed on $[F(-1), F(+1)]$. The corresponding value of x is the solution of

$$\frac{8}{3}u = F(x) = x + \frac{x^3}{3} + \frac{4}{3}$$

The solution is

$$x_i = A + B, \text{ where } A = (4u - 2 + s)^{1/3}$$
$$B = (4u - 2 - s)^{1/3} \quad , \qquad s = \sqrt{1 + 4(1-2u)^2}$$

Note that this requires calculating one square root and two cube roots per point.

Composite method

We write $f(x)$ as the sum of simpler functions. In this case an obvious choice is

$$f(x) = f_a(x) + f_b(x) \qquad \text{with} \quad f_a(x) = 1 \qquad \text{and} \quad f_b(x) = x^2$$

The integrals of these functions are

$$A_a = \int_{-1}^{+1} f_a(x)\,\mathrm{d}x = 2 \qquad \text{and} \qquad A_b = \int_{-1}^{+1} f_b(x)\,\mathrm{d}x = \frac{x^3}{3}\Big]_{-1}^{+1} = \frac{2}{3}$$

Hence we want to generate from f_a with probability $\frac{2}{2+\frac{2}{3}} = \frac{3}{4}$ and from f_b with probability $\frac{1}{4}$.

The first step is therefore to generate $v = R[0,1]$

- If $v \le \frac{3}{4}$ we generate from f_a:

$$u_i = R[0,1]$$
$$x_i = 2u_i - 1$$

- If $v > \frac{3}{4}$ we generate from f_b:

  1. either by the rejection method:

     $$x_i = 2R[0,1] - 1$$
     $$r_i = R[0, f_{b\,\max}] = R[0,1]$$

     repeating until we find a point for which $r_i \leq x_i^2$.

     Note that the efficiency is $\frac{\int_{-1}^{1} x^2\,\mathrm{d}x}{(b-a)f_{b\,\max}} = \frac{1}{3}$ for the points generated here ($1/4$ of the points). But it was 1 for the points distributed according to f_a. The net efficiency is thus $\frac{1}{3}\cdot\frac{1}{4} + 1\cdot\frac{3}{4} = \frac{5}{6}$, a small improvement over the $2/3$ of the simple rejection method.

  2. or by the inverse transformation method:

     $$F_b(x) = \int_{-1}^{x} x^2\,\mathrm{d}x = \left.\frac{x^3}{3}\right]_{-1}^{x} = \frac{x^3}{3} + \frac{1}{3} \qquad F_b(-1) = 0 \qquad F_b(1) = \frac{2}{3}$$

     We generate $u_i = R[0,1]$. Then x_i is the solution of

     $$\frac{2}{3}u_i = \frac{x_i^3}{3} + \frac{1}{3}$$
     $$\text{Hence,} \qquad x_i = (2u_i - 1)^{1/3}$$

     Note that we only have to calculate one cube root; and that only for $1/4$ of the events. This is ~ 12 times faster that the simple inverse transformation method (assuming that square and cube roots take about the same time).

In this example, the composite rejection method turned out to be the fastest with the simple rejection method only slightly slower. The composite inverse transformation method was much faster than the simple inverse transformation method, but still much slower than the rejection method. These results should not be regarded as general. Which method is faster depends on the function f.

6.3.6 Gaussian generator

The Gaussian distribution is one of the most important in physics and statistics. Many methods have been proposed to generate normally distributed points.

Using the Central Limit Theorem

By the C.L.T., the average of a large number of r.v.'s distributed according to almost any p.d.f. will be normally distributed. In particular, for n r.v.'s, u_i, distributed uniformly between 0 and 1, the quantity, g,

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

is approximately distributed as $N(g;0,1)$ for large n. Proof of this is left to the reader (exercise 26).

While simple to program, this generator is not particularly fast and has the feature that the tails are truncated at $\pm n\sigma$, which is usually undesirable. If the absence of long tails is tolerable, this method is usually satisfactory for as few as $n = 12$, where g reduces to

$$g_+ = \sum_{i-1}^{12} u_i - 6$$

Another disadvantage of this method is that it puts severe requirements on the correlations between successive points of the random number generator, in particular on correlations within groups of n successive values of u_i.

A word of caution is perhaps appropriate for clever students who have undoubtedly noticed that instead of summing 12 u_i and subtracting 6, we could have used

$$g_- = \sum_{i=1}^{6} u_i - \sum_{i=7}^{12} u_i$$

So far, so good. But if you try to save computer time by generating both g_+ and g_- with the same 12 values of u_i, you are in trouble: g_+ and g_- are then highly correlated.

A transformation method

Since the Gaussian p.d.f. cannot be integrated in terms of the usually available functions, it is not straightforward to find a transformation from uniformly to Gaussian distributed variables. There is, however, a clever method, which we give without proof, to transform two independent variables, u_1 and u_2, uniformly distributed on (0,1) to two independent variables, g_1 and g_2, which are normally distributed with $\mu = 0$ and $\sigma^2 = 1$:

$$g_1 = \cos(2\pi u_2)\sqrt{-2\ln u_1}$$
$$g_2 = \sin(2\pi u_2)\sqrt{-2\ln u_1}$$

This method is exact, but its speed can be improved upon by effectively generating the sine and cosine by a rejection method:

1. Generate uniform random numbers u_1 and u_2 on (0,1)

2. Calculate $r^2 = (2u_1 - 1)^2 + (2u_2 - 1)^2$.

3. If $r^2 > 1$, then reject u_1 and u_2 and go back to step 1.

4. Otherwise,

$$g_1 = (2u_1 - 1)\sqrt{\frac{-2\ln r^2}{r^2}}$$
$$g_2 = (2u_2 - 1)\sqrt{\frac{-2\ln r^2}{r^2}}$$

This saves the time of evaluating a sine and a cosine at the slight expense of rejecting about 21% of the uniformly generated points.

Part III

Statistics

Chapter 7

Statistics—What is it/are they?

So far, we have considered probability theory. Once we have decided which p.d.f. is appropriate to the problem, we can make direct calculations of the probability of any set of outcomes. Apart from possible uncertainty about which p.d.f. is appropriate, this is a straight-forward and mathematically well defined procedure.

The problem we now address is the inverse of this. We have a set of data which have been sampled from some parent p.d.f. We wish to infer from the data something about the parent p.d.f. Note that here we are assuming that the data are independent, *i.e.*, that the value of a particular datum does not depend on the values of other data, and that all of the data sample the same p.d.f. The statistician speaks of a sample of independent identically distributed *iid* random variables. Usually this will be the case, and some of our methods will depend on this.

The study of calculations using probability is sometimes called direct probability. Statistical inference is sometimes called inverse probability, particularly in the case of Bayesian methods.

We may think we know what the p.d.f. is apart from one or more parameters, *e.g.*, we think it is a Gaussian but want to determine its mean and standard deviation. This is called parameter estimation. It is also called fitting since we want to determine the value of the parameter such that the p.d.f. best 'fits' the data.

On the other hand, we may think we know the p.d.f. and want to know whether we are right. This is called hypothesis testing. Usually both parameter estimation and hypothesis testing are involved, since it makes little sense to try to determine the parameters of an incorrect p.d.f. And frequently an hypothesis to be tested involves some unknown parameter. Nevertheless, we will first treat these as separate topics. A third topic is decision theory or classification.

For all of these topics we shall use statistical methods (or "statistics"), so-called

*It is perhaps interesting to note that the *stat* in *statistics* is the same as in *state*. Statists (advocates of statism, economic control and planning by a highly centralized state), collected data to better enable the state to run the economy. Such data, and quantities calculated from them, came to be called statistics.

because they, statistical methods, make (it, statistics, makes) use of one or more statistics.* A definition: A **statistic** is any function of the observations in a sample, which does not depend on any of the unknown characteristics of the population (parent p.d.f.). An example of a statistic is the sample mean, $\bar{x} = \sum x_i/n$. Each observation, x_i, is, in fact, itself a statistic. In other words, if you can calculate it from the data plus known quantities, it is a statistic. "Statistics" is the branch of applied mathematics which deals with statistics as just defined. Whether the word statistics is singular or plural, thus depends on which meaning you intend.

We have seen in section 2.4 that there are two common interpretations of probability, which we have called frequentist and Bayesian. They give rise to two approaches to statistical inference, usually called classical or frequentist statistics (or inference) and Bayesian inference. The word classical is something of a misnomer, since the Bayesian interpretation is older (Bayes, Laplace). However, in the second half of the 19th century science became more quantitative and objective, even in such fields as biology (Darwin, evolution, heredity, Galton). This gave rise to the frequentist interpretation and the development of frequentist statistics. By about 1935 frequentist statistics, which came to be known as classical statistics, had completely replaced Bayesian thinking. Since around 1960, however, Bayesian inference has been making a comeback.

Probably most physicists would profess to being frequentists, and reflecting this, as well as my own personal bias, the emphasis in the rest of this course will be on classical statistics. However, there are situations where classical statistics is very difficult, or even impossible, to use and where Bayesian statistics is comparatively simple to apply. So, intermixed with classical statistics you will find some Bayesian methods. This is rather unconventional; most books are firmly in one of the two camps, and discussions between frequentists and Bayesians often take on aspects of holy war. It also runs the risk of confusing the student—it is important to know which you are doing.

To understand God's thoughts we must study statistics,
for these are the measure of His purpose.
—Florence Nightingale

Chapter 8

Parameter estimation

8.1 Introduction

In everyday speech, "estimation" means a rough and imprecise procedure leading to a rough and imprecise result. You estimate when you cannot measure exactly. This last sentence is also true in statistics, but only because you can never measure anything exactly; there is always some uncertainty. In statistics, estimation is a precise procedure leading to a result which may be imprecise, but the extent of the precision is, in principle, known. Estimation in statistics has nothing to do with approximation.

The goal of parameter estimation is then to make some sort of statement like $\theta = a \pm b$ where a is, on the basis of the data, the 'best' (in some sense) value of the parameter θ and where it is 'highly probable' that the true value of θ lies somewhere between $a - b$ and $a + b$. We often call b the estimated error on a. If we make a plot, this is represented by a point at $\theta = a$ with a bar running through it from $a - b$ to $a + b$, the 'error bar'. It is usually assumed that the estimate of θ is normally distributed, *i.e.*, that the values of a obtained from many identical experiments would form a normal distribution centered about the true value of θ with standard deviation equal to b. The meaning of $\theta = a \pm b$ is then that a is, in some sense (to be discussed more fully later), the most likely value of θ and that in any case there is, again in some sense, a $\int_{a-b}^{a+b} N(x; a, b^2) = 68.3\%$ chance that the true value of θ lies in the interval $(a - b, a + b)$.[†] This is a special case of a 68.3% 'confidence interval' (*cf.* chapter 9), *i.e.*, an interval within which we are 68.3% confident that the true value lies. We shall see that error bars, or confidence intervals may be difficult to estimate. Just as our estimate of θ has an 'error', so too does our estimate of this 'error'.

Suppose now that we have a set of numbers x_i which are the result of our experiment. This could, *e.g.*, be n measurements of some quantity. Let θ be the

[†]Note that this is different from what an engineer usually means by $a \pm b$, namely that b is the tolerance on a, *i.e.*, that the true value is *guaranteed* to be within $(a - b, a + b)$.

true value of that quantity. The x_i are clustered about θ in some way that depends on the measuring process. We often assume that they are distributed normally about the true value with a width given by the accuracy of the measurement.

It is worth noting the distinction many authors, *e.g.*, Bevington[10], make between the words accuracy and precision, which in normal usage are synonymous. **Accuracy** refers to how close a result is to the true value, whereas **precision** refers to how reproducible the measurements are. Thus, a poorly calibrated apparatus may result in measurements of high precision but poor accuracy. Other authors, *e.g.*, Eadie *et al.*,[4] and James[5] prefer to avoid these terms altogether since neither term is well defined, and to speak only of the variance of the estimates.

Similarly, a distinction is sometimes* made between **error**, the difference between the estimate and the true value, and the **uncertainty**, the square root of the variance of the estimate. Thus accurate means small error and precise means small uncertainty. However, the use of the word 'error' to mean uncertainty is deeply ingrained, and we (like most books) will not make the distinction. Note that with the above distinction, the accuracy and the error are usually unknown, since the true value is usually unknown.

So, we wish to estimate θ. To do this we need an estimator which is a function of the measurements.

As stated in chapter 7, a **statistic** is, by definition, any function of the observations in a sample, $\phi(x_i)$, which does not depend on any of the unknown characteristics of the population (parent p.d.f.). An example of a statistic is the sample mean, $\bar{x} = \sum x_i/n$. In other words, if you can calculate it from the data plus known quantities, it is a statistic.

Since a statistic is calculated from random variables, it is itself a r.v., but a r.v. whose value depends on the particular sample, or set of data. Like all random variables, it is distributed according to some p.d.f. Since the value of the statistic depends on the sample, its p.d.f. is sometimes referred to as the **sampling distribution** or **sampling p.d.f.** in order to distinguish it from the population or parent p.d.f.

An **estimator** is (definition) a statistic, the value of which we will give as our determination of some constant, θ, which is a property of the parent population or parent p.d.f. We will generally denote an estimator of a variable by adding a circumflex (ˆ) to the symbol of the variable. Thus $\hat{\theta}$ is an estimator of θ.

There are in general numerous estimators that one can construct for any θ. Here are several estimators of the mean, μ, of the parent p.d.f., assuming n measurements, x_i:

1. $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — The sample mean. This is probably the most often used estimator of the mean, but it can be sensitive to mismeasured data.

*This is recommended by the International Standards Organization[34].

2. $\hat{\mu} = \frac{1}{10}\sum_{i=1}^{10} x_i$	The sample mean of the first 10 points, ignoring the rest.
3. $\hat{\mu} = \frac{1}{n-1}\sum_{i=1}^{n} x_i$	$n/(n-1)$ times the sample mean.
4. $\hat{\mu} = 5$	Throw away all the data and give the estimate as 5.
5. $\hat{\mu} = \sqrt[n]{\prod_{i=1}^{n} x_i}$	
6.	Make a histogram of the x_i and take $\hat{\mu}$ as the midpoint of the bin containing the most events, *i.e.*, a sort of sample mode. Note that the value will depend on the bin size.
7. $\hat{\mu} = [\min(x_i) + \max(x_i)]/2$	The midrange, *i.e.*, the average of the smallest and the largest x_i. This is very sensitive to the tails of the distribution but may be the best estimator if the p.d.f. is nearly uniform.
8. $\hat{\mu} = \frac{2}{n}\sum_{i=1}^{n/2} x_{2i}$	The sample mean of the even numbered points, ignoring the odd numbered points.
9. $\hat{\mu} = \bar{\mu}_{\text{trimmed}}$	Discard the smallest and largest $y\%$ (*e.g.*, 10%) of the data and then average. This is relatively insensitive to the tails of the distribution, but has a larger variance than the sample mean if there are no problems in the tails.
10. $\hat{\mu} =$ sample median	This is less sensitive to statistical fluctuations in the tails, but it has a larger variance than the sample mean if the p.d.f. is a Gaussian.

Each of these is, by our definition, an estimator. Yet some are certainly better than others. However, which is 'best' depends on the p.d.f. Which is 'best' may also depend on the use we want to make of it. How do we choose which estimator to use? In general we shall prefer an estimator which is 'unbiased', 'consistent', and 'efficient'. We will discuss these and other properties of estimators in the following section. In succeeding sections we will treat three general methods of constructing, or choosing, estimators.

Nothing is easier than to invent
methods of estimation.
—R. A. Fisher

8.2 Properties of estimators

8.2.1 Bias

Since a statistic is a function of r.v.'s, it is itself a r.v. Therefore, it is distributed according to some p.d.f., and we can speak of its expectation value, $E\left[\hat{\theta}\right]$. For an estimator, making use of n observations, the bias b_n is defined as the difference between the expectation of the estimator and the true value of the parameter:

$$b_n(\hat{\theta}) = E\left[\hat{\theta}\right] - \theta = E\left[\hat{\theta} - \theta\right] \tag{8.1}$$

An estimator is unbiased if, for all n and θ, $b_n(\hat{\theta}) = 0$, *i.e.*, if $E\left[\hat{\theta}\right] = \theta$. We include n in this definition since we shall see that some estimators are unbiased only asymptotically, *i.e.*, only for $n \to \infty$.

Mean In general, the sample mean, no. 1 in our list above, is an unbiased estimator of the parent (true) mean:

$$E[\hat{\mu}] = E[\bar{x}] = E\left[\frac{1}{n}\sum x_i\right] = \frac{1}{n}\sum E[x_i] = \frac{1}{n}nE[x] = E[x] = \mu \tag{8.2}$$

On the other hand, the third estimator in our list is biased:

$$E[\hat{\mu}] = E\left[\frac{1}{n-1}\sum x_i\right] = \frac{n}{n-1}\mu$$

although the bias,

$$b_n(\hat{\mu}) = \frac{n}{n-1}\mu - \mu = \frac{\mu}{n-1} \to 0 \text{ , for large } n.$$

This estimator is thus asymptotically unbiased.

If we know the bias, we can construct a new estimator by correcting the old one for its bias. For example, from no. 3 and its bias we construct no. 1 simply by multiplying no. 3 by $(n-1)/n$.

Lack of bias is a reason to prefer no. 1 to no. 3. However, nos. 2 and 8 are also unbiased. The trimmed mean (no. 9) is unbiased if the parent p.d.f. is symmetric about its mean. The sample median (no. 10) is also unbiased if the parent median equals the parent mean. Similarly, nos. 6 and 7 will be unbiased for certain p.d.f.'s.

Variance Now suppose we want to estimate the variance of the parent p.d.f. Assume that we know the true mean, μ. Usually this is not the case, but could be, *e.g.*, if we know that the p.d.f. is symmetric about some value. Then following our above experience with the sample mean, we might expect the sample variance,

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{8.3}$$

to be a good estimator of the parent variance, σ^2. (*N.b.*, do not confuse the standard deviation, σ, of the parent p.d.f. with the 'error' on $\hat{\mu}$.) Assume that the parent variance, σ^2, is finite (exists). Then

$$\begin{aligned}
E\left[s_1^2\right] &= \frac{1}{n} E\left[\sum (x_i - \mu)^2\right] = \frac{1}{n} E\left[\sum \left(x_i^2 - 2x_i\mu + \mu^2\right)\right] \\
&= \frac{1}{n} E\left[\sum x_i^2 - 2\mu \sum x_i + \sum \mu^2\right] = \frac{1}{n}\left[E\left[\sum x_i^2\right] - 2\mu E\left[\sum x_i\right] + n\mu^2\right] \\
&= \frac{1}{n}\left[nE\left[x^2\right] - 2n\mu E[x] + n\mu^2\right] = E\left[x^2\right] - 2\mu^2 + \mu^2 \\
&= \sigma^2 + \mu^2 - 2\mu^2 + \mu^2 \text{ , since } \sigma^2 = E[x^2] - \mu^2 \\
&= \sigma^2
\end{aligned}$$

Thus $\widehat{\sigma^2} = s_1^2$ is an unbiased estimator of the variance of the parent p.d.f., σ^2, if μ is known.

But usually μ is not known. We therefore try using our estimate of μ, $\hat{\mu} = \bar{x}$, instead of μ:

$$s_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2 \tag{8.4}$$

This has the expectation,

$$\begin{aligned}
E\left[s_x^2\right] &= E\left[\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2\right] \\
&= \frac{1}{n}\left(E\left[\sum x_i^2\right] - \frac{1}{n}E\left[\left(\sum x_i\right)^2\right]\right)
\end{aligned} \tag{8.5}$$

The x_i are independent. Hence $E[\sum x_i^2] = nE[x^2]$. Also,

$$\sigma^2 = E\left[x^2\right] - \mu^2$$

$$\text{and} \qquad V\left[\sum x_i\right] = E\left[(\sum x_i)^2\right] - \left(E\left[\sum x_i\right]\right)^2$$

Substituting in (8.5), gives

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

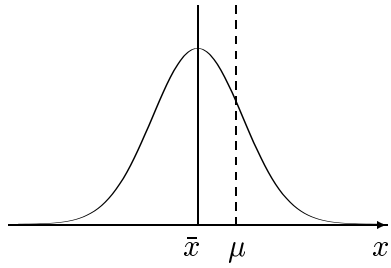
$$\text{Using} \qquad V\left[\sum x_i\right] = \sum V[x_i] = nV[x] = n\sigma^2$$

$$\text{and} \qquad E\left[\sum x_i\right] = nE[x] = n\mu$$

we find

$$\begin{aligned}
E\left[s_x^2\right] &= \frac{1}{n}\left[n\sigma^2 + n\mu^2 - \frac{1}{n}\left(n\sigma^2 + (n\mu)^2\right)\right] \\
&= \frac{1}{n}(n-1)\sigma^2
\end{aligned} \tag{8.6}$$

Thus s_x^2 is a biased estimator of σ^2. The reason is that, not knowing μ, we used our estimate of the mean, $\hat{\mu} = \bar{x}$, the sample mean. The spread of the data about the sample mean is clearly less than its spread about the true mean. Since the variance is the spread about the true mean, s_x^2 underestimates the true variance.



This bias is easily removed. An unbiased estimator for the parent variance when the parent mean is unknown is

$$s^2 = \frac{n}{n-1}\, s_x^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right) = \frac{1}{n-1}\sum (x_i - \bar{x})^2 \tag{8.7}$$

Note that the above calculations did not depend at all on what the parent p.d.f. was, not even on the C.L.T.

If the p.d.f. is Gaussian or if n is large enough that the C.L.T. applies, let

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Then

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

is distributed as χ^2 (section 3.12). There is one relationship among the z_i's:

$$\sum z_i = \frac{1}{\sigma}\sum (x_i - \bar{x}) = \frac{1}{\sigma}\left(\sum x_i - n\bar{x}\right) = 0$$

which follows from the definition of $\bar{x}$. Hence, the p.d.f. for $\sum z_i^2$ is a χ^2 of $n-1$ degrees of freedom. Recall that $E[\chi^2(n-1)] = n-1$. This is another way of seeing that

$$E\left[s^2\right] = E\left[\frac{\sigma^2}{n-1}\sum z_i^2\right] = E\left[\frac{\sigma^2}{n-1}\chi^2\right] = \sigma^2\frac{1}{n-1}E\left[\chi^2\right] = \sigma^2$$

i.e., that $\widehat{\sigma^2} = s^2$ is an unbiased estimator of σ^2 when μ is unknown.

This use of χ^2 is of more than passing interest: In general, if we have n measurements, x_i, of a quantity, with $k \le n$ relationships (constraints) among them, then the χ^2 constructed from the $\sum x_i^2$ will have $n-k$ degrees of freedom.

The $(n-1)$ instead of n in s^2 also makes sense in the limit $n = 1$. With only one measurement of x, you have an estimate $\hat{\mu} = x$ of μ, but no estimate of the width of the distribution. This is consistent with $s^2 = \frac{1}{1-1}(x - \hat{\mu})^2 = \frac{0}{0} =$ indeterminate. However, if μ is known you do not have to use the measurement to estimate μ; you can use it instead to estimate σ^2. Hence s_1 contains n instead of $(n-1)$.

8.2.2 Consistency

If we take more data, we should expect a better (more accurate) estimate of the parameters. An estimator which converges to the true value with increasing n is termed consistent.
Definition: An estimator, $\hat{\theta}$, of θ is consistent if for any $\epsilon > 0$ (no matter how small),

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0 \tag{8.8}$$

This is rather analogous to the definition of convergence of a series except that here it is the probability of the deviation from the true value which approaches 0 rather than the deviation itself. This is therefore sometimes called convergence in probability.

If $\hat{\theta}$ is an average of data which are distributed according to a p.d.f. for which the C.L.T. applies, then $\hat{\theta}$ is a consistent estimator, since the width of the p.d.f., $N(\bar{x}; \mu, \frac{\sigma^2}{n})$ approaches 0 for $n \to \infty$.

In our list of estimators of the mean no. 2 is clearly inconsistent. Nos. 1, 3, and 8 are obviously consistent if the C.L.T. applies. No. 10 is consistent only if the mean and median of the parent p.d.f. are equal. Likewise, the consistency of nos. 6, 7 and 9 depends on the p.d.f.

The usual example of an inconsistent estimator is the sample mean for the Cauchy p.d.f., which, as we have seen, does not have a finite variance. The C.L.T. does not then apply, and in fact $\bar{x}$ is distributed just like x. Thus, $\bar{x}$ does not converge to anything! This illustrates the fact that an unbiased estimator is not necessarily consistent.

8.2.3 Variance of an estimator, efficiency

An estimator is called efficient if it has a small variance, in particular if it has the smallest possible variance (see the following section).

Repetition of an experiment generally results in a different value of our (consistent) estimator. If the variance of the sampling p.d.f. of the estimator, which, for convenience, we will call the variance of the estimator, is small, these values will cluster closely about the true value, or, if the estimator is biased, about the biased (*i.e.*, wrong) value. We will see that in general the variance of an estimator depends on the parent p.d.f., in particular, on the variance (σ^2) of the parent p.d.f.

For example, consider the variance of the sample mean. As we have seen (chapter 5 and exercise 23),

$$V\left[\bar{x} = \frac{1}{n}\sum x_i\right] = \frac{1}{n^2}\sum V[x_i] = \frac{1}{n^2} n V[x] = \frac{\sigma^2}{n} \tag{8.9}$$

Now consider the sample variance, which was defined in equation 8.7. Assuming that the x_i follow a normal p.d.f. (or that n is large and the C.L.T. applies), the

sample variance has variance

$$V\left[s^2\right] = V\left[\frac{1}{n-1}\sigma^2 \sum \frac{(x_i-\bar{x})^2}{\sigma^2}\right] = \left[\frac{\sigma^2}{n-1}\right]^2 V\left[\sum z_i^2\right]$$

where $z_i = \frac{x_i-\bar{x}}{\sigma}$. As we have seen (section 8.2.1), $\sum_{i=1}^n z_i^2$ is distributed as $\chi^2(n-1)$. Thus,

$$V\left[\sum z_i^2\right] = V\left[\chi^2(n-1)\right] = 2(n-1)$$

Hence,

$$V\left[s^2\right] = \frac{2(\sigma^2)^2}{n-1} \tag{8.10}$$

We see that the expressions for the variance of $\bar{x}$ and s^2 both contain σ^2, the variance of the parent p.d.f., which we may not know. (If we do know it we certainly will not be interested in estimating it.) The usual procedure is to use instead our estimate of σ^2, s^2. Then the estimated variances of our estimates are

$$\widehat{V}[\bar{x}] = \frac{s^2}{n} \quad , \quad \widehat{V}\left[s^2\right] = \frac{2(s^2)^2}{n-1} \tag{8.11}$$

Sometimes you do know σ^2. We give two examples: (1) You average many measurements of a quantity, *e.g.*, the length of a table. The p.d.f. is then a convolution of a δ-function about the true length with a resolution function for the measuring apparatus, which is just a Gaussian centered about the true length with σ equal to the resolution. But you have calibrated the measuring apparatus by measuring a standard length a great many times. From this calibration you know σ^2. So you only need to estimate μ. (2) You are designing an experiment and you want to know how many measurements you need to make in order to attain a given accuracy. You then make reasonable assumptions about the p.d.f. and calculate what V will be for the different assumptions about μ, σ^2, and n.

To summarize, assuming that we do not know μ or σ^2, they are estimated by

$$\hat{\mu} = \bar{x} \pm \sqrt{V[\bar{x}]} \qquad \text{and} \qquad \widehat{\sigma^2} = s^2 \pm \sqrt{V[s^2]} \tag{8.12a}$$

$$= \bar{x} \pm \sqrt{\frac{s^2}{n}} \qquad\qquad = s^2 \pm \sqrt{\frac{2}{n-1}}s^2 \tag{8.12b}$$

Note that the 'error' on $\hat{\mu}$ has itself an error. By 'error propagation', which will be covered in section 8.3.6,

$$\begin{aligned} V\left[s^2\right] &= \left(\tfrac{\mathrm{d}s^2}{\mathrm{d}s}\right)^2 V[s] = (2s)^2 V[s] \\ \text{Hence,} \quad V[s] &= \tfrac{1}{4s^2} V\left[s^2\right] = \tfrac{1}{4s^2}\tfrac{2(s^2)^2}{n-1} = \tfrac{s^2}{2(n-1)} \\ \text{and} \quad \sqrt{V[s]} &= \tfrac{\sqrt{s^2}}{\sqrt{2(n-1)}} \end{aligned}$$

The error on the error on $\hat{\mu}$ is then (with δ indicating 'error')

$$\delta(\delta\hat{\mu}) = \sqrt{V[\delta\hat{\mu}]} = \sqrt{V\left[\sqrt{\frac{s^2}{n}}\right]} = \sqrt{\frac{1}{n}V\left[\sqrt{s^2}\right]} = \frac{\sqrt{s^2}}{\sqrt{2n(n-1)}} = \frac{\delta\hat{\mu}}{\sqrt{2(n-1)}}$$

Thus for n not too small, the error on the error on $\hat{\mu}$ is negligible.

8.2.4 Interpretation of the Variance

We usually interpret $V[\hat{q}] = \sigma^2$ as the "square of the expected error" of $\hat{q}$ and we write $q = \hat{q} \pm \delta q$ where $\delta q = \sigma$. If the p.d.f. of $\hat{q}$ is a Gaussian with variance σ^2, then the chance, in some sense, that the true value of q, q_{t}, is within $\hat{q} - \sigma \leq q_{\mathrm{t}} \leq \hat{q} + \sigma$ is

$$P\left(\hat{q} - \sigma \leq q_{\mathrm{t}} \leq \hat{q} + \sigma\right) = \int_{\hat{q}-\sigma}^{\hat{q}+\sigma} N(q; \hat{q}, \sigma^2)\, \mathrm{d}q \approx 0.68$$

In exactly what sense this is so will be discussed in section 9.

We could have used some other quantity to indicate the 'error', *e.g.*, the average of the absolute deviation $\overline{|\hat{q} - q|}$, instead of $\sqrt{\overline{(\hat{q} - q)^2}}$. The variance is conventional for a number of reasons:

- It is low order and hence easy to calculate.
- It is sufficient in the case of a Gaussian, being one of the two parameters of the Gaussian, and the Gaussian is, by the C.L.T., often the asymptotic limit of the p.d.f.
- It is easily converted to a confidence interval in the Gaussian limit (*cf.* chapter 9).

When the p.d.f. of $\hat{q}$ is non-Gaussian one must be careful. If the p.d.f. is skewed, this can be indicated by stating asymmetric errors. But that is not foreseen in the propagation of errors. Also, for a non-Gaussian $P\left(\hat{q} - \sigma \leq q_{\mathrm{t}} \leq \hat{q} + \sigma\right)$ is usually not 68%. Nor is the probability of being within, *e.g.*, 2σ the same in the non-Gaussian case as in the Gaussian case. Nor do the errors even have to be symmetric.

The propagation of errors (*cf.* section 8.3.6) is usually the least trustworthy when there is a dependence on $1/q$. Going to higher orders in the expansion does not necessarily help because the resulting error, though perhaps more accurate, still has the same problems resulting from skewness and the probability content of $\pm 2\sigma$. These questions are often conveniently investigated by Monte Carlo methods. As previously stated, the best cure for these problems is to rewrite the p.d.f. in terms of the parameters you want to estimate.

We shall return to these questions when discussing confidence intervals (chapter 9) and hypothesis testing (chapter 10).

8.2.5 Information and Likelihood

The concepts 'information' and 'likelihood' will be useful in discussing the variance of estimators. We introduce them now:

There are several different definitions of information. They are named after the person who introduced them. We will use that of R. A. Fisher, which is then referred to as the information of R. A. Fisher. However, since we will only treat this one definition of information, we will simply refer to it as information. But bear in mind that the word can have other definitions. We will see that Fisher's definition meets the following requirements, which we find necessary for what we would like the word 'information' to mean:

1. The information should increase if we make more observations.
2. Data, which are irrelevant to the estimation of the parameters we wish to estimate or to the hypothesis we wish to test, should contain no information. Of course the same data may contain information for other parameters or other tests.
3. The precision of the estimation or test should be greater if we have more information.

Present-day, large-scale experiments usually produce a great amount of data of which only a small part is useful for a given measurement or test. The information contained in a datum can be used to decide whether to reject it in order to reduce the amount of data to a manageable size. (It is difficult to work with data on 100 magnetic tapes; working with just one tape, or a small disk file is much easier.) A good criterion for data reduction is to reject the maximum of data with the minimum loss of information. This is usually a compromise, although the rejection of some data may actually result in no loss of information.

Likelihood function: We observe a real random variable, X, sampled from a p.d.f., $f(x;\theta)$, where θ is a parameter. The set of allowed values of X is denoted by Ω_θ, the subscript emphasizing the possible dependence on the parameter. Both X and θ could be sets of values $\underline{X}$ and $\underline{\theta}$, not necessarily of the same dimension.

Consider a set of n independent observations of X, x_i. The joint p.d.f. of the x_i is, since they are independent,

$$\mathcal{L}(\underline{x};\underline{\theta}) = \mathcal{L}(x_1, x_2, \ldots, x_n;\underline{\theta}) = \prod_{i=1}^{n} f(x_i;\underline{\theta}) \tag{8.13}$$

The function $\mathcal{L}$ depends on both the measurements x_i and on the parameters $\underline{\theta}$. However, after having done the experiment, the x_i are fixed. Then $\mathcal{L}$ can be regarded as a function of $\underline{\theta}$ only. $\mathcal{L}$ is called the likelihood function. We also define its logarithm,

$$\ell \equiv \ln \mathcal{L}(x_1, \ldots, x_n;\underline{\theta}) = \sum_{i=1}^{n} \ln f(x_i;\underline{\theta}) \tag{8.14}$$

Information: The information (of R. A. Fisher) given about a parameter θ by an observation of the r.v. $\underline{x}$ is defined as the expectation

$$\begin{aligned} I_{\underline{x}}(\theta) &= E\left[\left(\frac{\partial \ln \mathcal{L}(\underline{x};\theta)}{\partial\theta}\right)^2\right] = E\left[\left(\frac{\partial \ell}{\partial\theta}\right)^2\right] \\ &= \int_{\Omega_\theta} \left(\frac{\partial \ln \mathcal{L}(\underline{x};\theta)}{\partial\theta}\right)^2 \mathcal{L}(\underline{x};\theta)\,\mathrm{d}\underline{x} \end{aligned} \tag{8.15}$$

In the case where there are k parameters, the information is a $k \times k$ matrix:

$$\begin{aligned} \left[I_{\underline{x}}(\underline{\theta})\right]_{ij} &= E\left[\frac{\partial \ln \mathcal{L}(\underline{x};\underline{\theta})}{\partial\theta_i}\frac{\partial \ln \mathcal{L}(\underline{x};\underline{\theta})}{\partial\theta_j}\right] \\ &= \int_{\Omega_\theta} \frac{\partial \ln \mathcal{L}(\underline{x};\underline{\theta})}{\partial\theta_i}\frac{\partial \ln \mathcal{L}(\underline{x};\underline{\theta})}{\partial\theta_j}\mathcal{L}(\underline{x};\underline{\theta})\,\mathrm{d}\underline{x} \end{aligned}$$

This definition of information may seem rather arbitrary, but we shall see that it satisfies the three requirements stated above.

Score: Notation becomes more compact by introducing the score. We define the score of one measurement as

$$S_1 \equiv \frac{\partial}{\partial\theta}\ln f(x;\theta) \tag{8.16}$$

Note that the score, being a function of r.v.'s, is itself a r.v. The score of the entire sample is then defined to be the sum of the scores of each observation:

$$S(\underline{x};\theta) \equiv \sum_{i=1}^{n} S_1(x_i;\theta) \tag{8.17}$$

Then

$$\begin{aligned} S(\underline{x};\theta) &= \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_i;\theta) \\ &= \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln f(x_i;\theta) \\ &= \frac{\partial \ln \mathcal{L}(\underline{x};\theta)}{\partial\theta} \end{aligned}$$

Summarizing,

$$S(\underline{x};\theta) = \frac{\partial \ln \mathcal{L}(\underline{x};\theta)}{\partial\theta} = \sum_{i=1}^{n} S_1(x_i;\theta) = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\ln f(x_i;\theta) \tag{8.18}$$

This result combined with equation 8.15 shows that we can write the information of the sample $\underline{x}$ on the parameter θ as the expectation of the square of the score:

$$I_{\underline{x}}(\theta) = E\left[(S(\underline{x};\theta))^2\right] \tag{8.19}$$

If $\Omega_{\underline{\theta}}$ is independent of $\underline{\theta}$, we can show that the expectation of the score is zero and we can derive another relation between the information and the score. Let us assume that

1. $\Omega_{\underline{\theta}}$ is independent of $\underline{\theta}$, and

2. $\mathcal{L}(\underline{x};\underline{\theta})$ is regular enough that we can interchange the order of $\frac{\partial^2}{\partial\theta_i\partial\theta_j}$ and $\int \mathrm{d}x$.

If condition (1) holds, condition (2) will also generally hold for distributions encountered in physics. Now,

$$\begin{aligned} E[S_1(x;\theta)] &= E\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right] = \int\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right] f(x;\theta)\,\mathrm{d}x \\ &= \int \frac{1}{f(x;\theta)}\left[\frac{\partial}{\partial\theta}f(x;\theta)\right] f(x;\theta)\,\mathrm{d}x \\ &= \int \frac{\partial}{\partial\theta}f(x;\theta)\,\mathrm{d}x \end{aligned}$$

Interchanging the order of integration and differentiation (assumption 2),

$$E[S_1(x;\theta)] = \frac{\partial}{\partial\theta}\int f(x;\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}1 = 0 \tag{8.20}$$

since $f(x;\theta)$ is normalized for all values of θ. Hence,

$$E[S(\underline{x};\theta)] = \sum E[S_1(x_i;\theta)] = 0 \tag{8.21}$$

Using the fact that the variance of a quantity is given by $V[a] = E[a^2] - (E[a])^2$, we see from equations 8.19 and 8.21 that

$$I_{\underline{x}}(\theta) = V[S(\underline{x};\theta)] \tag{8.22}$$

We have shown above (equation 8.19) that in general the information on θ is equal to the expectation of the square of the score. Under the above two assumptions you can show (exercise 31) that the information is also given by

$$I_{\underline{x}}(\theta) = -E\left[\frac{\partial S(\underline{x};\theta)}{\partial\theta}\right] \tag{8.23}$$

These results (equations 8.21 and 8.23) are very useful, but do not forget the assumptions on which they depend.

Does I satisfy the requirements? We can now show that the information increases with the number of independent observations. For n observations,

$$\begin{aligned} I(\theta) &= E\left[\left(\sum_{i=1}^{n} S_1(x_i;\theta)\right)^2\right] \\ &= V\left[\sum_{i}^{n} S_1(x_i;\theta)\right] + \left\{E\left[\sum_{i}^{n} S_1(x_i;\theta)\right]\right\}^2 \end{aligned}$$

where we have used the fact that $V[a] = E[a^2] - (E[a])^2$. The second term is zero under the assumptions that Ω_θ is independent of θ and that the order of differentiation and integration can be interchanged as in the previous paragraph (eq. 8.21). However, let us now relax these assumptions.

Since the x_i are independent, the variance of the sum is just the sum of the variances. And since all the x_i are sampled from the same p.d.f., the variance is the same for all i. A similar argument applies to the second term. Hence,

$$I(\theta) = n\,V\left[S_1(x;\theta)\right] + n^2\,\left\{E\left[S_1(x;\theta\right]\right\}^2 \tag{8.24}$$

Following the same steps for $n = 1$ gives the same expression with $n = 1$. Hence, the information increases with the number of observations, our first requirement for information.

If the assumptions of the previous paragraph apply, the second term in the above equation is zero by equation 8.20. Then,

$$I(\theta) = n\,I_1(\theta) \tag{8.25}$$

and the information of n independent observations is just n times the information of one observation. If the assumptions are not true, the second term may not be zero but will still be positive; hence I will still increase with n.

For data which are irrelevant for the estimation of θ, the p.d.f. will not depend on θ and the score will, from its definition (equations 8.16 and 8.17), be zero. This implies that the information will also be zero, which was our second requirement for information.

We now turn to the third requirement, the connection between the precision of an estimator and the information.

8.2.6 Minimum Variance Bound

It turns out that there is a lower limit to the variance of an estimator under certain general conditions.

Rao-Cramér inequality: Suppose that we have an estimator $\hat{\theta}$ of θ with bias $b_n(\hat{\theta}) = E\left[\hat{\theta}\right] - \theta$, that the variance $V\left[\hat{\theta}\right]$ is finite, and that the range of X does

not depend on θ. Then

$$\begin{aligned}
E\left[\hat{\theta}\, S(\underline{x};\theta)\right] &= \int \ldots \int \hat{\theta} \left[\frac{\partial}{\partial\theta} \ln \mathcal{L}(\underline{x};\theta)\right] \mathcal{L}(\underline{x};\theta)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&= \int \ldots \int \hat{\theta} \left[\frac{1}{\mathcal{L}(\underline{x};\theta)} \frac{\partial}{\partial\theta} \mathcal{L}(\underline{x};\theta)\right] \mathcal{L}(\underline{x};\theta)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&= \int \ldots \int \hat{\theta} \left[\frac{\partial}{\partial\theta} \mathcal{L}(\underline{x};\theta)\right] \mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&= \int \ldots \int \hat{\theta} \frac{\partial}{\partial\theta} \left[\prod_{i=1}^{n} f(x_i;\theta)\, \mathrm{d}x_i\right] \\
&= \int \ldots \int \frac{\partial}{\partial\theta} \left[\hat{\theta} \prod_{i=1}^{n} f(x_i;\theta)\, \mathrm{d}x_i\right]
\end{aligned}$$

The last step follows because $\hat{\theta}$ is a statistic and therefore does not depend on θ. Assuming that we can interchange the order of differentiation and integration, we find

$$\begin{aligned}
E\left[\hat{\theta}\, S(\underline{x};\theta)\right] &= \frac{\partial}{\partial\theta} \int \ldots \int \hat{\theta} \prod_{i=1}^{n} \left[f(x_i;\theta)\, \mathrm{d}x_i\right] \\
&= \frac{\partial}{\partial\theta} E\left[\hat{\theta}\right] = \frac{\partial}{\partial\theta} \left[\theta + b_n(\hat{\theta}))\right] \\
&= 1 + \frac{\partial}{\partial\theta} b_n(\hat{\theta})
\end{aligned}$$

Both $\hat{\theta}$ and $S(\underline{x};\theta)$ are r.v.'s. Their covariance is

$$\begin{aligned}
\operatorname{cov}\left[S(\underline{x};\theta), \hat{\theta}(\underline{x})\right] &= E\left[S(\underline{x};\theta)\, \hat{\theta}(\underline{x})\right] - \underbrace{E\left[S(\underline{x};\theta)\right]}_{=0,\ \text{eq. } 8.21} E\left[\hat{\theta}(\underline{x})\right] \\
&= 1 + \frac{\partial}{\partial\theta} b_n(\hat{\theta})
\end{aligned}$$

Therefore, their correlation coefficient is

$$\rho^2 = \frac{\left\{\operatorname{cov}\left[S, \hat{\theta}\right]\right\}^2}{V[S]\; V\left[\hat{\theta}\right]} = \frac{\left[1 + \frac{\partial}{\partial\theta} b_n(\hat{\theta})\right]^2}{I(\theta)\, V\left[\hat{\theta}\right]}$$

Since $\rho^2 \leq 1$, we have

$$\sigma^2(\hat{\theta}) = V\left[\hat{\theta}\right] \geq \frac{\left[1 + \frac{\partial}{\partial\theta} b_n(\hat{\theta})\right]^2}{I(\theta)} \tag{8.26}$$

Thus, there is a lower bound on the variance of the estimator. For a given set of data and hence a given amount of information, $I(\theta)$, on θ, we can never find an estimator with a lower variance.

The more information we have, the lower this bound is, in accordance with our third requirement for information.

If the estimator is a constant, $\hat{\theta} = c$, then the bias is $b = c - \theta$ and the minimum variance is 0, which is not a very interesting bound since the variance of a constant is always 0.

The inequality (8.26) is usually known as the Rao-Cramér inequality or the Frechet inequality. It was discovered independently by a number of people including Rao,[35] Cramér,[15] and Frechet. The first were Aitken and Silverstone.[36] Although we have assumed that the range of X is independent of θ and that we could interchange the order of differentiation and integration, the result (8.26) can be obtained with somewhat more general assumptions.[11, 13]

In general, we prefer unbiased estimators. In that case the inequality reduces to $\sigma^2(\hat{\theta}) \geq 1/I(\theta)$. This is also the case if the bias of the estimator does not depend on the true value of θ. For more than one parameter this result generalizes to

$$\sigma^2(\hat{\theta}_i) \geq \left[\underline{I}^{-1}(\underline{\theta})\right]_{ii} \tag{8.27}$$

the diagonal element of the inverse of the information matrix.

We define the efficiency of the estimator as

$$\epsilon(\hat{\theta}) = \frac{\sigma^2_{\min}(\hat{\theta})}{\sigma^2(\hat{\theta})} \leq 1 \tag{8.28}$$

which, for unbiased estimators, is just

$$\epsilon(\hat{\theta}) = \frac{1}{\sigma^2(\hat{\theta})\, I(\theta)} \leq 1 \tag{8.29}$$

An estimator whose variance is equal to the minimum variance given by equation 8.26, *i.e.*, has $\epsilon(\hat{\theta}) = 1$, is termed **efficient**. It is not always possible to construct an efficient estimator.

Examples:

Gaussian with known mean. We have seen (section 8.2.1) that $\widehat{\sigma^2} = \sum(x_i - \mu)^2/n$ is an unbiased estimator of the variance of a Gaussian of known mean. It is easy to show (exercise 32) that it is also an efficient estimator.

Exponential. Consider n independent observations from an exponential p.d.f.,

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu} \qquad , \qquad \mu > 0$$

We wish to estimate μ. We note that

$$\ln f(x;\mu) = -\ln \mu - \frac{x}{\mu}$$

The score of one observation is then

$$S_1(x;\mu) = \frac{\partial}{\partial\mu}\left(-\ln\mu - \frac{x}{\mu}\right) = -\frac{1}{\mu} + \frac{x}{\mu^2}$$

The information of one observation is then, using equation 8.19 or 8.23, the latter being applicable since the range of X is independent of μ,

$$\begin{aligned} I_1(\mu) &= E\left[(S_1(x;\mu))^2\right] = -E\left[\frac{\partial S_1(x;\mu)}{\partial\mu}\right] \\ &= -E\left[\frac{1}{\mu^2} - \frac{2x}{\mu^3}\right] = -\frac{1}{\mu^2} + \frac{2}{\mu^2} = \frac{1}{\mu^2} \end{aligned}$$

And the total information of the sample is

$$I(\mu) = nI_1(\mu) = \frac{n}{\mu^2}$$

If $\hat{\mu}$ is unbiased, its minimum variance is then $1/I = \mu^2/n$. We try the sample mean as an estimator: $\hat{\mu} = \bar{x}$. We know (equation 8.2) that the sample mean is always an unbiased estimator of the mean. The variance of the sample mean is

$$\begin{aligned} V[\bar{x}] &= \frac{1}{n}V[x] = \frac{1}{n}\left(E\left[x^2\right] - \mu^2\right) \\ &= \frac{1}{n}\underbrace{\int_0^\infty x^2\frac{1}{\mu}e^{-x/\mu}\,\mathrm{d}x}_{=2\mu^2} - \frac{\mu^2}{n} = \frac{\mu^2}{n} \end{aligned}$$

which is just the minimum variance found above. Thus the sample mean is an efficient estimator of the mean of an exponential p.d.f.

Note that the score is

$$S(x;\mu) = \sum_{i=1}^{n} S_1(x_i;\mu) = -\frac{n}{\mu} + \frac{\sum x_i}{\mu^2} = -I(\mu)\,(\mu - \hat{\mu})$$

Thus the score is a linear function of the estimator. This is not a coincidence, but a general feature of unbiased efficient estimators, as we show in the next section.

8.2.7 Efficient estimators—the Exponential family

In this section we shall show that an efficient estimator can be found if and only if the p.d.f. is a member of a quite general class of functions known as the exponential family.

The minimum variance bound was found using

$$\rho^2 = \frac{\left\{\mathrm{cov}\left[S,\hat{\theta}\right]\right\}^2}{V\left[S\right]\,V\left[\hat{\theta}\right]} \leq 1$$

The equality $\rho = \pm 1$ corresponds to a linear relationship between the variables (exercise 7), *i.e.*, a straight line on a graph of S *vs.* $\hat{\theta}$. Thus, assuming that the conditions of the minimum variance bound hold, an estimator $\hat{\theta}$ can be efficient if and only if it is a linear function of S, with the possible exception of regions where the probability is zero.

Let $A(\theta)$ and $B(\theta)$ be functions of θ, but not of x, and A', B' be their derivatives with respect to θ. Then we can write the linear relationship as

$$\frac{\partial}{\partial\theta} \ln f(x;\theta) \equiv S = A'(\theta)\hat{\theta}(x) + B'(\theta) \tag{8.30}$$

Since $\hat{\theta}$ is a statistic and hence depends only on x, integration over θ gives

$$\ln f(x;\theta) = A(\theta)\,\hat{\theta}(x) + B(\theta) + K(x) \tag{8.31}$$

where the integration constant K may depend on x but not on θ. Then, where the required normalization is included in B and/or K,

$$f(x;\theta) = \exp\left[A(\theta)\,\hat{\theta}(x) + B(\theta) + K(x)\right] \tag{8.32}$$

Any p.d.f. of the above form is said to belong to the exponential family. What we have shown is that an efficient estimator can be found if and only if the p.d.f. is of the exponential family where the estimator enters the exponent in the way shown in equation 8.32.

Note that the efficient estimator is not necessarily unique since the product $A \cdot \hat{\theta}$ can often be factored in more than one way. The estimator $\hat{\theta}$ will be an unbiased estimator for some quantity, although not necessarily for the quantity we want to estimate. It may also not be an estimator which we will be able to use. Let us now calculate the expectation of $\hat{\theta}$ and see for what quantity it is an unbiased estimator: From equation 8.30,

$$\hat{\theta} = \frac{S(x;\theta)}{A'(\theta)} - \frac{B'(\theta)}{A'(\theta)}$$

Since A' and B' do not depend on x, the expectation is then

$$E\left[\hat{\theta}\right] = \frac{1}{A'(\theta)} E\left[S(x;\theta)\right] - \frac{B'(\theta)}{A'(\theta)}$$

Since $E\left[S(x;\theta)\right] = 0$, we have

$$E\left[\hat{\theta}\right] = -\frac{\frac{\partial B(\theta)}{\partial\theta}}{\frac{\partial A(\theta)}{\partial\theta}} \tag{8.33}$$

This is the quantity for which the $\hat{\theta}$ in equation 8.32 is an unbiased, efficient estimator.

If there are k parameters, $\underline{\theta}$, equation 8.32 generalizes to

$$f(x;\underline{\theta}) = \exp\left[\underline{A}(\underline{\theta}) \cdot \hat{\underline{\theta}}(x) + B(\underline{\theta}) + K(x)\right] \tag{8.34}$$

The score for the ith parameter is then

$$S(x;\theta_i) = \frac{\partial}{\partial\theta_i} \ln f(x;\underline{\theta}) = \sum_j \hat{\theta}_j(x) \frac{\partial A_j(\underline{\theta})}{\partial\theta_i} + \frac{\partial B(\underline{\theta})}{\partial\theta_i}$$

Taking the expectation, we arrive at the generalization of equation 8.33, which is a set of k equations:

$$E\left[\hat{\theta}_i\right] = -\frac{\frac{\partial B(\underline{\theta})}{\partial\theta_i} + \sum_{j\neq i} E\left[\hat{\theta}_j\right] \frac{\partial A_j(\underline{\theta})}{\partial\theta_i}}{\frac{\partial A_i(\underline{\theta})}{\partial\theta_i}} \tag{8.35}$$

Examples:

Gaussian. As an example we take the normal p.d.f., $N(x;\mu,\sigma^2)$, which has two parameters $\underline{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$. We write $N(x;\mu,\sigma^2)$ in an exponential form:

$$\begin{aligned} N(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \\ &= \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right] \end{aligned}$$

For n independent observations the p.d.f. becomes

$$\prod_{i=1}^{n} N(x_i;\mu,\sigma^2) = \exp\left[\frac{n\mu}{\sigma^2}\bar{x} - \frac{n}{2\sigma^2}\overline{x^2} - \frac{n}{2}\left(\frac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right]$$

from which we see that we can choose (in equation 8.34)

$$\begin{aligned} &A_1(\theta) = \tfrac{n\mu}{\sigma^2} && \hat{\theta}_1(x) = \bar{x} \\ &A_2(\theta) = -\tfrac{n}{2\sigma^2} && \hat{\theta}_2(x) = \overline{x^2} \\ &B(\theta) = -\tfrac{n}{2}\left(\tfrac{\mu^2}{\sigma^2} + \ln(2\pi\sigma^2)\right) \\ &K(x) = 0 \end{aligned}$$

Then (from equation 8.35)

$\frac{\partial A_1}{\partial\mu} = \frac{n}{\sigma^2}$
$\frac{\partial A_2}{\partial\mu} = 0$
$\frac{\partial B}{\partial\mu} = -n\frac{\mu}{\sigma^2}$

Thus $\hat{\theta}_1 = \bar{x}$ is an efficient and unbiased estimator of

$$-\frac{-n\mu/\sigma^2}{n/\sigma^2} = \mu$$

$\frac{\partial A_1}{\partial\sigma^2} = \frac{n\mu}{\sigma^4}$
$\frac{\partial A_2}{\partial\sigma^2} = \frac{n}{2\sigma^4}$
$\frac{\partial B}{\partial\sigma^2} = \frac{n\mu^2}{2\sigma^4} - \frac{n}{2\sigma^2}$

Thus $\hat{\theta}_2 = \overline{x^2}$ is an efficient and unbiased estimator of $\mu^2 + \sigma^2$. Hence, $\overline{x^2} - \mu^2 = \overline{(x-\mu)^2}$ is an efficient and unbiased estimator of σ^2. However, this is of use to us only if we know μ.

Note the role of the number of observations n. The likelihood function, $\mathcal{L}$, is just the p.d.f. with each term in the exponent replaced by a sum of n terms. Thus $\mathcal{L}$ can be obtained from f by the replacements: $x \to \bar{x}$, $x^2 \to \overline{x^2}$, *etc.* and $A \to nA$, $B \to nB$, and $K \to nK$. But $-\frac{\partial B/\partial\theta}{\partial A/\partial\theta}$ is unchanged by these substitutions. Thus we can work with f instead of $\mathcal{L}$, just replacing any function of x by its average in the expression for $\hat{\theta}$.

Binomial. Discrete p.d.f.'s can also belong to the exponential family. As an example we take the binomial p.d.f.,

$$f(k;n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

which can be written

$$f(k;n,\theta) = \exp\left[k\ln\left(\frac{\theta}{1-\theta}\right) + n\ln(1-\theta) + \ln\binom{n}{k}\right]$$

With n fixed, there is just one parameter to estimate, θ.

$$\begin{aligned} A(\theta) &= \ln\left(\tfrac{\theta}{1-\theta}\right) & \hat{\theta}(k) &= k \\ B(\theta) &= n\ln(1-\theta) & K(k) &= \ln\tbinom{n}{k} \end{aligned}$$

The expectation of the estimator is

$$E\left[\hat{\theta}\right] = -\frac{\partial B/\partial\theta}{\partial A/\partial\theta} = n\theta$$

Thus k is an efficient, unbiased estimator of $n\theta$, or k/n is an efficient, unbiased estimator of θ.

Which estimator is the best? Returning to the list of 10 estimators for the mean at the start of the section, we can ask which of the 10 is the best. Unfortunately, there is no unique answer. In general we prefer unbiased, consistent and efficient estimators. We can clearly reject nos. 2, 3, 4, 5 and 8. Nor is no. 6, the sample mode, a good choice, even when the parent mode equals the parent mean, since it uses so little of the information. However, which of the others is 'best' depends on the parent p.d.f.

The sample mean is efficient for a normal p.d.f. However, for a uniform p.d.f. ($f(x;a,b) = \frac{1}{b-a}$) where the limits (a,b) are unknown, estimator no. 7, $\frac{1}{2}x_{\min} + \frac{1}{2}x_{\max}$, has a smaller variance that $\bar{x}$.

No. 10, the sample median, has a larger variance that the sample mean for a Gaussian p.d.f., but for a 'large-tailed Gaussian' it can be smaller. No. 9, the trimmed sample mean, throws away information but may still be best, in particular if we think that points in the tails are largely due to mismeasurement.

8.2.8 Sufficient statistics

A statistic $T(x)$ is said to be sufficient for the parameter θ if the conditional p.d.f. of x, given T, $f(x|T)$, is independent of θ. (T and θ may of course be multidimensional and of different dimensions.) In other words, T is sufficient if T contains all the information on θ.

Clearly, $\underline{T} = \underline{x}$ is a sufficient statistic since that is all the information we have—on θ or on anything else. But this doesn't help us very much. The importance of sufficiency is in data reduction. If we have a sufficient statistic, $\underline{T}$, of a smaller dimension than the data, $\underline{x}$, we can reduce the amount of data. This can be of enormous practical advantage.

From n independent observations x_i, one can construct $m \leq n$ independent statistics $t, t_1, t_2, \ldots, t_{m-1}$ (in an infinite number of ways). From the definition of marginal and conditional p.d.f.'s we can write the p.d.f. of these statistics as (*cf.* equation 2.28)

$$f(t, t_1, t_2, \ldots, t_{m-1}; \theta) = g(t; \theta)\, h(t_1, t_2, \ldots, t_{m-1}; \theta | t) \tag{8.36}$$

where $g(t; \theta)$ is the marginal p.d.f. of t and h is the conditional p.d.f. Now if h is independent of θ, then clearly the $t_1, t_2, \ldots, t_{m-1}$ contribute nothing to our knowledge of θ. If this is true for any set of t_i and any $m < n$ then t clearly contains all the information on θ. We therefore define a sufficient statistic t as: t is a sufficient statistic for θ if for any choice of $t_1, t_2, \ldots, t_{m-1}$ (which are independent of t),

$$f(t, t_1, t_2, \ldots, t_{m-1}; \theta) = g(t; \theta)\, h(t_1, t_2, \ldots, t_{m-1} | t) \tag{8.37}$$

Now, what does this mean in terms of the likelihood? The likelihood function is the p.d.f. for $\underline{x}$ and is thus related to the f of equation 8.37 by a coordinate transformation. Starting from equation 8.37, let $t_i = x_i$ for $i = 1, 2, \ldots, n-1$. Then

$$f(t, x_1, x_2, \ldots, x_{n-1}; \theta) = g(t; \theta)\, h(x_1, x_2, \ldots, x_{n-1} | t)$$

The p.d.f. in terms of $\underline{x}$ is then

$$\mathcal{L}(\underline{x}; \theta) = g(t; \theta)\, h(x_1, x_2, \ldots, x_{n-1} | t) \left| J\left(\frac{x_1, \ldots, x_n}{x_1, \ldots, x_{n-1}, t}\right)\right|$$

which is, since the Jacobian does not involve θ, of the form

$$\mathcal{L}(\underline{x}; \theta) = g(t; \theta)\, k(\underline{x}) \tag{8.38}$$

Conversely, starting from equation 8.38, we make the transformation

$$\begin{aligned} t &= t(x_1, \ldots, x_n) \\ t_i &= t_i(x_1, \ldots, x_n) \quad , i < m \\ t_i &= x_i \quad , i = m, \ldots, n-1 \end{aligned}$$

$\mathcal{L}(\underline{x};\theta)\,\mathrm{d}\underline{x}$ then transforms to

$$g(t;\theta)\,k(\underline{x})\left|J\left(\frac{t,t_1,\ldots,t_{n-1}}{x_1,\ldots,x_n}\right)\right|\,\mathrm{d}t\,\prod_{i=1}^{n}\mathrm{d}t_i$$

which we integrate over $\mathrm{d}t_m\ldots\mathrm{d}t_{n-1}$ to obtain the p.d.f. $f(t,t_1,\ldots,t_{m-1})$. Neither k nor J depend on θ. However, the integration limits for $t_m,\ldots,t_{n-1}$ $(x_m,\ldots,x_{n-1})$ could depend on θ. If not, it is clear that we obtain the form of equation 8.37. It turns out[11,13] that this is also true even when the integration limits do depend on θ.

Thus equations 8.37 and 8.38 are equivalent. If we can find a statistic t such that the likelihood function can be written in the form of equation 8.38, t is a sufficient statistic for θ.

The sufficient statistics for θ having the smallest dimension are called minimal sufficient statistics for θ. One usually prefers a minimal sufficient statistic since that gives the greatest data reduction.

We have seen that if we can write the p.d.f. in the exponential form of equation 8.34,

$$f(x;\underline{\theta})=\exp\left[\underline{A}(\underline{\theta})\cdot\hat{\underline{\theta}}(x)+B(\underline{\theta})+K(x)\right]$$

then $\hat{\underline{\theta}}$ is an efficient estimator. Such a p.d.f. clearly factorizes like equation 8.38 with

$$\begin{aligned}g(\hat{\underline{\theta}};\underline{\theta})&=\exp\left[\underline{A}(\underline{\theta})\cdot\hat{\underline{\theta}}(x)+B(\underline{\theta})\right]\\ k(x)&=\exp\left[K(x)\right]\end{aligned}$$

Thus, if the range of x does not depend on $\underline{\theta}$, $\hat{\underline{\theta}}(x)$ is not only an efficient estimator of $\underline{\theta}$, but also a sufficient statistic for $\underline{\theta}$. If the range of x depends on $\underline{\theta}$, the situation is more complicated. The reader is referred to Kendall and Stuart[11,13] for the conditions of sufficiency.

8.3 Substitution methods

Now that we know something about the properties of estimators, let us turn to the problem of constructing, or choosing, an estimator. There are three general methods of estimation, which we will examine in turn. We begin with substitution methods.

8.3.1 Frequency substitution

This is the simplest method. It is useful when the parameter to be estimated is a frequency or the function of a frequency. It consists of simply estimating the population (parent) frequency by the experimentally observed (sample) frequency.

Such estimators are also known as **plug-in estimators**, since the data are simply "plugged into" the parameter definition.

For example, if the underlying p.d.f. is a binomial, $B(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$, we would estimate p by $\hat{p} = x/n$. This is unbiased since $E[x] = np$. It is also efficient since B is a member of the exponential family of p.d.f.'s, as we saw in section 8.2.7. And we would estimate a function of p, $g(p)$, by $g(\hat{p}) = g(x/n)$. This method works well for large samples where the C.L.T. assures us that the difference between $E[x]$ and np is a small fraction of np.

Advantages of this method are simplicity and the fact that the estimator is usually consistent. Disadvantages are that the estimator may be biased and that it may not have minimum variance. However, if it is biased, we may be able to reduce the bias, or at least estimate its size by a series expansion:

Suppose that $\hat{\theta}$ is an unbiased estimator of θ. We wish to estimate some function of θ, $g(\theta)$. Following the above prescription, we use $\hat{g} = g(\hat{\theta})$. Then, expanding $\hat{g}$ about the true value of θ, θ_{t}, assuming that the necessary derivatives exist,

$$\hat{g} = g(\hat{\theta}) = g(\theta_{\mathrm{t}}) + (\hat{\theta} - \theta_{\mathrm{t}}) \left.\frac{\partial g(\theta)}{\partial \theta}\right|_{\theta=\theta_{\mathrm{t}}} + \frac{1}{2}(\hat{\theta} - \theta_{\mathrm{t}})^2 \left.\frac{\partial^2 g(\theta)}{\partial \theta^2}\right|_{\theta=\theta_{\mathrm{t}}} + \dots$$

Now we take the expectation. Since $\hat{\theta}$ is assumed unbiased, this gives simply,

$$E[\hat{g}] = g(\theta_{\mathrm{t}}) + \frac{1}{2} E\left[(\hat{\theta} - \theta_{\mathrm{t}})^2\right] \left.\frac{\partial^2 g(\theta)}{\partial \theta^2}\right|_{\theta=\theta_{\mathrm{t}}} + \dots$$

Not knowing the true value θ, we can not calculate $E\left[(\hat{\theta} - \theta_{\mathrm{t}})^2\right]$. But we can estimate it by $V\left[\hat{\theta}\right]$. In the same spirit, we evaluate the derivative at $\theta = \hat{\theta}$ instead of at $\theta = \theta_{\mathrm{t}}$. Thus, to lowest order, there is a bias of approximately $\frac{1}{2} V\left[\hat{\theta}\right] \left.\frac{\partial^2 g(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}}$.

In the case of more than one parameter, $\underline{\theta}$, this becomes

$$\hat{g} = g(\underline{\hat{\theta}}) = g(\underline{\theta}_{\mathrm{t}}) + \sum_i (\hat{\theta}_i - \theta_{\mathrm{t}i}) \left.\frac{\partial g}{\partial \theta_i}\right|_{\underline{\theta}=\underline{\theta}_{\mathrm{t}}} + \frac{1}{2} \sum_i \sum_j (\hat{\theta}_i - \theta_{\mathrm{t}i})(\hat{\theta}_j - \theta_{\mathrm{t}j}) \left.\frac{\partial^2 g}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}=\underline{\theta}_{\mathrm{t}}} + \dots$$

$$E[\hat{g}] = g(\underline{\theta}_{\mathrm{t}}) + \frac{1}{2} \sum_i \sum_j V_{ij}(\underline{\hat{\theta}}) \left.\frac{\partial^2 g}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}=\underline{\hat{\theta}}} + \dots$$

from which we deduce that

$$\hat{g}_1 = \hat{g} - \frac{1}{2} \sum_{i,j} V_{ij} \left.\frac{\partial^2 g}{\partial \theta_i \partial \theta_j}\right|_{\underline{\theta}=\underline{\hat{\theta}}} \tag{8.39}$$

has reduced bias, provided that the correction term is not large or rapidly varying. If that is not true, it is not obvious that going to higher order terms in the expansion would help, since the problem may come from using $\hat{\theta}$ instead of the true value in the expansion. In that case more detailed investigation is needed, perhaps employing Monte Carlo techniques to test the behavior of the estimators under different assumptions for θ.

8.3.2 Method of Moments

The method

This is another substitution method. To estimate a function q of the parameter θ, we write $q(\theta)$ as a function of the moments of the p.d.f.:

$$q(\theta) = g(m_1, m_2, \ldots)$$

where $m_j = E[x^j]$. This can, of course, only be done if all the necessary moments exist. We then estimate $q(\theta)$ by replacing all the parent (population) moments, m_j, in g by the corresponding sample (experimental) moments. Thus,

$$\hat{q} = g(\widehat{m}_1, \widehat{m}_2, \ldots) \quad , \qquad \widehat{m}_j = \overline{x^j} = \frac{1}{n}\sum_i x_i^j \tag{8.40}$$

In this notation $m_1 = \mu$, the parent mean, and $\widehat{m}_1 = \bar{x}$, the sample mean.

For example, to estimate the parent variance, $V[x]$, we write the variance in terms of the moments: $V[x] = \sigma^2 = m_2 - m_1^2$. We then estimate the moments by the corresponding sample moments:

$$\widehat{\sigma^2} = \widehat{m}_2 - \widehat{m}_1^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

As we have previously seen (equation 8.6), this estimator, which we have called s_x^2 (equation 8.4), is biased. Thus the method of moments does not necessarily give unbiased estimators.

As a second example, take the Poisson p.d.f. For this p.d.f., the population mean and the population variance are equal, $\mu = V[x]$. Therefore, we could estimate the mean and the variance either

$$\begin{array}{rcccl} \text{by} & \hat{\theta} & = & \widehat{m}_1 & = \bar{x} \\ \text{or by} & \hat{\theta} & = & \widehat{m}_2 - \widehat{m}_1^2 & = \frac{1}{n}\sum (x_i - \bar{x})^2 \end{array}$$

Thus the method of moments does not necessarily provide a unique estimator.

Variance of sample moments

Of course, a moment estimator, like any estimator, is rather useless unless we also estimate its uncertainty. It can be easily shown (exercise 35) that in general, assuming that the moments exist,

$$V[\widehat{m}_k] = V\left[\frac{1}{n}\sum x_i^k\right] = \frac{1}{n}\left(m_{2k} - m_k^2\right) \tag{8.41}$$

$$\operatorname{cov}[\widehat{m}_j, \widehat{m}_k] = \frac{1}{n}\left(m_{j+k} - m_j m_k\right) \tag{8.42}$$

We can estimate these variances and covariances by replacing the moments by their estimators and $1/n$ by $1/(n-1)$ to remove the bias.

By the C.L.T. the average tends to its expectation under the assumption that the variance is finite. Moments estimators, being averages, are therefore consistent.

A word of caution is in order: If it is necessary to use higher order moments, you should be cautious. They are very sensitive to the tails of the distribution, which is the part of the distribution which is usually the most affected by experimental difficulties.

8.3.3 Descriptive statistics

Moments provide a simple way to describe the data without making any assumption about the parent p.d.f. Since the amount of data in present-day experiments is usually far too large to publish, it is necessary to reduce it to a reasonable volume, but in such a way that it remains useful.

In some cases we have a theory which is in agreement with the data and it is enough that the experimental data agree with the expectation. In other cases we have no theory and the purpose of the experiment is to provide data which can point the way to a theory. The experimental moments of a distribution up to a certain (not too high) order provide a set of numbers with which some future theory can easily be compared.

8.3.4 Generalized method of moments

Instead of the moments $m_i = E[x^i]$, which are moments of the functions x^i, we can use moments of some other set of functions, $u_j(x)$. These moments, $E[u_j]$, are given by

$$E[u_j] = \int u_j(x) f(x; \underline{\theta})\, \mathrm{d}x$$

Thus we have a number of equations for $E[u_j]$ in terms of $\underline{\theta}$. We solve them for the $\underline{\theta}$ in terms of the $E[u_j]$ and substitute the sample moments, $\bar{u}_j$, for the expectations to obtain our estimate of $\underline{\theta}$. We will always need at least as many equations, and hence at least as many functions u_j, as there are parameters to be estimated.

We take as an example the angular distribution of the decay of a vector meson into two pseudo-scalar mesons. The angles θ and ϕ of the decay products in the rest system of the vector meson are distributed as

$$f(\cos\theta, \phi) = \frac{3}{4\pi}\Big[\frac{1}{2}(1-\rho_{00}) + \frac{1}{2}(3\rho_{00}-1)\cos^2\theta - \rho_{1,-1}\sin^2\theta\cos 2\phi \\ -\sqrt{2}\mathrm{Re}\rho_{10}\sin 2\theta\cos\phi\Big]$$

where the ρ's are parameters to be estimated. The data consist of measurements of the angles, θ_i and ϕ_i, for n decays. From inspection of the above expression for f, we choose three functions to estimate the three parameters. The choice is not unique, but an obvious choice is as follows. We then compute the expectation of

each of the functions:

function	expectation
$u_1 = \cos^2\theta$	$E[u_1] = \frac{1}{5}(1 + 2\rho_{00})$
$u_2 = \sin^2\theta\cos 2\phi$	$E[u_2] = -\frac{4}{5}\rho_{1,-1}$
$u_3 = \sin 2\theta\cos\phi$	$E[u_3] = -\frac{4}{5}\sqrt{2}\mathrm{Re}\rho_{10}$

Replacing $E[u_j]$ by the sample mean $\bar{u}_j = \frac{1}{n}\sum u_j(\cos\theta_i, \phi_i)$ gives, *e.g.*,

$$-\frac{4}{5}\sqrt{2}\mathrm{Re}\hat{\rho}_{10} = \bar{u}_3 = \frac{1}{n}\sum_{i=1}^{n}\sin 2\theta_i\cos\phi_i$$

which we solve for $\mathrm{Re}\hat{\rho}_{10}$.

This method is most elegant when the functions u_j form an orthonormal set. Then

$$f(x) = \sum_{i=0}^{\infty} a_i u_i(x) \qquad \text{and} \qquad \int u_j^*(x)u_k(x)\,\mathrm{d}x = \delta_{jk}$$

The expectations are then

$$E[u_k^*(x)] = \int u_k^*(x)f(x)\,\mathrm{d}x = a_k$$

Thus the estimate of the coefficient of the kth term is just the sample mean of the (complex conjugate of the) k^{th} function,

$$\hat{a}_k = \overline{u_k^*}$$

This estimator is unbiased and, by the C.L.T., asymptotically normally distributed about a_k.

8.3.5 Variance of moments

The variance of the kth sample moment, generalized or not, is

$$\begin{aligned} V_{kk} \equiv V[\bar{u}_k] &= \frac{1}{n^2}V\left[\sum_{i=1}^{n}u_k(x_i)\right] = \frac{1}{n}V[u_k(x)] \\ &= \frac{1}{n}E\left[\Big(u_k(x) - E[u_k(x)]\Big)^2\right] \end{aligned} \tag{8.43}$$

which reduces to equation 8.41 for ordinary moments, $u_k(x) = m_k = x^k$. This is estimated by replacing the expectations by the sample means to give

$$\widehat{V}_{kk} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\Big(u_k(x_i) - \bar{u}_k(x)\Big)^2 = \frac{1}{n-1}\left(\overline{u_k^2} - \bar{u}_k^2\right) \tag{8.44}$$

where we have used $\frac{1}{n-1}$ instead of $\frac{1}{n}$ in order to have an unbiased estimate. The general element of the covariance matrix is estimated by

$$\widehat{V}_{jk}\left[\bar{u}\right] = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\Big(u_j(x_i) - \bar{u}_j(x)\Big)\Big(u_k(x_i) - \bar{u}_k(x)\Big) \tag{8.45}$$

$$= \frac{1}{n-1}\left(\overline{u_j\, u_k} - \bar{u}_j\, \bar{u}_k\right) \tag{8.46}$$

8.3.6 Transformation of the covariance matrix under a change of parameters

Frequently it is not one of the moments that we want to estimate, but rather some function of the moments, *e.g.*, $\hat{\rho}_{00} = (5\bar{u}_1 - 1)/2$. We now examine how the covariance matrix for the $\bar{u}_k$ transforms under such a change of parameter. This topic is usually known as **propagation of errors**. This is, of course, applicable to functions of any estimator, not just to moments.

We want to estimate θ which we write as a function of q, $\theta(q)$. We first find an estimate of q, $\hat{q}$, and an estimate of its variance, $\widehat{V}\left[\hat{q}\right]$. To avoid possible misunderstanding, we denote the true (unknown) value of q by q_{t}. The true value of θ is then $\theta(q_{\mathrm{t}})$. Our estimate of q, $\hat{q}$, being a r.v., is of course distributed about q_{t} according to some p.d.f. We wish to (approximately) evaluate the variance of $\hat{\theta}$ from the variance of $\hat{q}$. We assume that $\hat{q}$ is an unbiased estimator of q, which is true, at least asymptotically (C.L.T.), if $\hat{q}$ is a moment.

We expand $\hat{\theta}$ about the true value of q. Then

$$\hat{\theta} = \theta(\hat{q}) = \theta(q_{\mathrm{t}}) + \left.\frac{\partial\theta}{\partial q}\right|_{q=q_{\mathrm{t}}}(\hat{q} - q_{\mathrm{t}}) + \ldots$$

$$\text{and}\quad E\left[\hat{\theta}\right] = \theta(q_{\mathrm{t}}) + \left.\frac{\partial\theta}{\partial q}\right|_{q=q_{\mathrm{t}}} E\left[(\hat{q} - q_{\mathrm{t}})\right] + \ldots$$

Since $\hat{q}$ is unbiased, $E\left[(\hat{q} - q_{\mathrm{t}})\right] = 0$. Thus, to first order, $E\left[\hat{\theta}\right] = \theta(q_{\mathrm{t}})$. Subtracting the second equation from the first gives, to first order,

$$\hat{\theta} - E\left[\hat{\theta}\right] = \left.\frac{\partial\theta}{\partial q}\right|_{q=q_{\mathrm{t}}}(\hat{q} - q_{\mathrm{t}})$$

Hence,

$$V\left[\hat{\theta}\right] \equiv E\left[\left(\hat{\theta} - E\left[\hat{\theta}\right]\right)^2\right] = \left(\frac{\partial\theta}{\partial q}\right)^2_{q=q_{\mathrm{t}}} E\left[(\hat{q} - q_{\mathrm{t}})^2\right]$$

$$= \left(\frac{\partial\theta}{\partial q}\right)^2_{q=q_{\mathrm{t}}} V\left[\hat{q}\right] \tag{8.47}$$

This can be estimated by substituting $\hat{q}$ for q_{t} and our estimate $\widehat{V}[\hat{q}]$ for $V[\hat{q}]$:

$$\widehat{V}\left[\hat{\theta}\right] = \left(\frac{\partial\theta}{\partial q}\right)^2_{q=\hat{q}} \widehat{V}[\hat{q}] \tag{8.48}$$

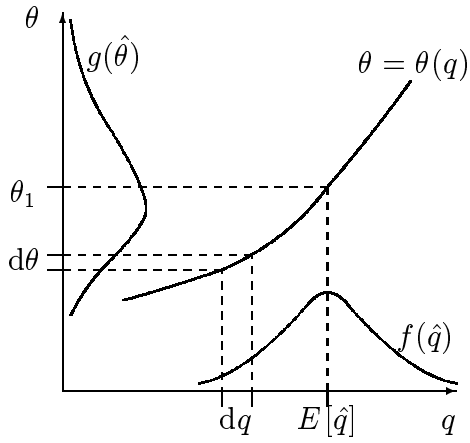
This technique works well only when second and higher order terms are small and when $\hat{q}$ is unbiased.

We give a simple example, a function linear in q. The result is then, in fact, exact since the second and higher order derivatives are zero.

$$\begin{aligned} \theta(q) &= A + Bq \\ \frac{\partial\theta}{\partial q} &= B \\ V\left[\hat{\theta}\right] &= B^2\, V[\hat{q}] \end{aligned} \tag{8.49}$$

The general case is similar to our treatment of change of variables (section 2.2.6). Indeed, it is in principle better to transform the p.d.f. to a new p.d.f. in terms of the parameter we want to estimate, *e.g.*, $f(x;q) \to g(x;\theta)$. In particular it is nice if we can transform to a p.d.f. having θ as its mean (or other low order moment), since sample moments are unbiased estimators. However, in practice such a transformation may be difficult and it may be easier to estimate q than to estimate θ directly.

Consider now the p.d.f.'s for the estimators $\hat{q}$ and $\hat{\theta}$. If the transformation $\theta = \theta(q)$ is non-linear, the shape of the p.d.f. $g(\hat{\theta})$ is changed from that of $f(\hat{q})$ by the Jacobian ($|\partial q/\partial\theta|$ in one dimension), as illustrated in the figure. In regions where $\mathrm{d}\theta < \mathrm{d}q$, the probability piles up faster for θ than for q. Thus in the example the peak in $g(\hat{\theta})$ occurs below $\theta_1 = g\left(E[\hat{q}]\right)$.



In particular, if $f(\hat{q})$ is normal, $g(\hat{\theta})$ is not normal, except for a linear transformation. This is a source of bias, which in the figure manifests itself as a long tail for $g(\hat{\theta})$ resulting in $E\left[\hat{\theta}\right] > \theta_1$.

Now let us treat the multidimensional case, where $\underline{q}$ is of dimension n and $\underline{\theta}$ is of dimension m. Note that $m \leq n$; otherwise not all θ_i will be independent and there will be no unique solution. An example would be a p.d.f. for (x, y) for which we want only to estimate some parameter of the (marginal) distribution for r. In this case, $n = 2$ and $m = 1$.

We can then expand each $\hat{\theta}_i$ about its true value in the same manner as for the one-dimensional case, except that we now must introduce a sum over all parameters:

$$\hat{\theta}_i \equiv \theta_i(\underline{\hat{q}}) = \theta_i(\underline{q}_{\mathrm{t}}) + \sum_{k=1}^{n} \left.\frac{\partial\theta_i}{\partial q_k}\right|_{\underline{q}=\underline{q}_{\mathrm{t}}} (\hat{q}_k - q_{\mathrm{t}\,k}) + \ldots$$

Assuming that $\hat{q}_i$ is unbiased, its expectation is equal to the true value so that to first order,

$$\left(\hat{\theta}_i - E\left[\hat{\theta}_i\right]\right)\left(\hat{\theta}_j - E\left[\hat{\theta}_j\right]\right) = \sum_{k=1}^{n}\sum_{l=1}^{n} \left.\frac{\partial \theta_i}{\partial q_k}\right|_{\underline{q}=\underline{q}_{\mathrm{t}}} \left.\frac{\partial \theta_j}{\partial q_l}\right|_{\underline{q}=\underline{q}_{\mathrm{t}}} (\hat{q}_k - q_{\mathrm{t}\,k})(\hat{q}_l - q_{\mathrm{t}\,l})$$

Taking expectations, and writing in matrix notation, we arrive at the generalization of equation 8.47:

$$V\left[\hat{\underline{\theta}}\right] = \underline{D}^{\mathrm{T}}(\underline{\theta})\, V\left[\hat{\underline{q}}\right]\, \underline{D}(\underline{\theta}) \tag{8.50}$$

where,

$$\underline{D}(\underline{\theta}) = \left.\begin{pmatrix} \frac{\partial\theta_1}{\partial q_1} & \frac{\partial\theta_2}{\partial q_1} & \cdots & \frac{\partial\theta_m}{\partial q_1} \\ \frac{\partial\theta_1}{\partial q_2} & \frac{\partial\theta_2}{\partial q_2} & \cdots & \frac{\partial\theta_m}{\partial q_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\theta_1}{\partial q_n} & \frac{\partial\theta_2}{\partial q_n} & \cdots & \frac{\partial\theta_m}{\partial q_n} \end{pmatrix}\right._{\underline{q}=\underline{q}_{\mathrm{t}}} \tag{8.51}$$

As in the one-dimensional case we estimate this variance by replacing true values by their estimates to arrive at the generalization of equation 8.48:

$$\widehat{V}\left[\hat{\underline{\theta}}\right] = \widehat{\underline{D}}^{\mathrm{T}}(\underline{\theta})\, \widehat{V}\left[\hat{\underline{q}}\right]\, \widehat{\underline{D}}(\underline{\theta}) \tag{8.52}$$

where,

$$\widehat{\underline{D}}(\underline{\theta}) = \left.\begin{pmatrix} \frac{\partial\theta_1}{\partial q_1} & \frac{\partial\theta_2}{\partial q_1} & \cdots & \frac{\partial\theta_m}{\partial q_1} \\ \frac{\partial\theta_1}{\partial q_2} & \frac{\partial\theta_2}{\partial q_2} & \cdots & \frac{\partial\theta_m}{\partial q_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\theta_1}{\partial q_n} & \frac{\partial\theta_2}{\partial q_n} & \cdots & \frac{\partial\theta_m}{\partial q_n} \end{pmatrix}\right._{\underline{q}=\hat{\underline{q}}} \tag{8.53}$$

Warning: $\underline{D}$ is not symmetric.

8.4 Maximum Likelihood method

This method of parameter estimation is very general. It is often the simplest method to use, particularly in complex cases, and maximum likelihood estimators have certain desirable properties.

8.4.1 Principle of Maximum Likelihood

We have already met the likelihood function in section 8.2.5. We repeat its definition here: The likelihood function is the joint p.d.f. for n measurements $\underline{x}$ given parameters $\underline{\theta}$:

$$\mathcal{L}(\underline{x};\underline{\theta}) = \mathcal{L}(x_1, x_2, \ldots, x_n; \underline{\theta}) \tag{8.54}$$

If the x_i are independent, this is just the product of the p.d.f.'s for the individual x_i:

$$\mathcal{L}(\underline{x};\underline{\theta}) = \prod_{i=1}^{n} f_i(x_i;\underline{\theta}) \tag{8.55}$$

where we have included a subscript i on f since it is not necessary that all the x_i have the same p.d.f.

In probability theory this p.d.f. expresses the probability that an experiment identical to ours would result in the n observations $\underline{x}$ which we observed. In probability theory we know $\underline{\theta}$ and the functions f_i, and we calculate the probability of certain results. In statistics this is turned around. We have done the experiment; so we know a set of results, $\underline{x}$. We (think we) know the p.d.f.'s, $f_i(x,\underline{\theta})$. We want to estimate $\underline{\theta}$.

We emphasize that $\mathcal{L}$ is not a p.d.f. for $\underline{\theta}$; if it were we would use the expectation value of $\underline{\theta}$ for $\hat{\underline{\theta}}$. Instead we take eq. 8.54, replace $\underline{\theta}$ by $\hat{\underline{\theta}}$ and solve for $\hat{\underline{\theta}}$ under the condition that $\mathcal{L}$ is a maximum. In other words, our estimate, $\hat{\underline{\theta}}$, of $\underline{\theta}$ is that value of $\underline{\theta}$ which would make our experimental results the most likely of all possible results.

This is the **Principle of Maximum Likelihood**: The best estimate of a parameter θ is that value which maximizes the likelihood function. This can not be proved without defining 'best'. It can be shown that maximum likelihood (ML) estimators have desirable properties. However, they are often biased. Whether the ML estimator really is the 'best' estimator depends on the situation.

It is usually more convenient to work with

$$\ell = \ln \mathcal{L} \tag{8.56}$$

since the product in eq. 8.55 becomes a sum in eq. 8.56. For independent x_i this is

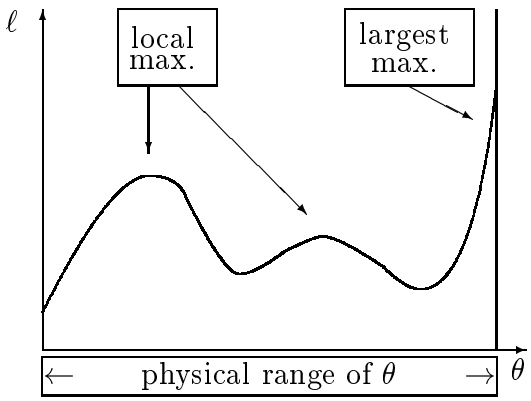
$$\ell = \sum_{i=1}^{n} \ell_i \quad \text{, where} \qquad \ell_i = \ln f_i(x_i;\underline{\theta}) \tag{8.57}$$

Since $\mathcal{L} > 0$, both $\mathcal{L}$ and ℓ have the same extrema, which are found from

$$S_i \equiv \frac{\partial \ell}{\partial \theta_i} = \frac{1}{\mathcal{L}}\frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \tag{8.58}$$

where S_i is the score function (section 8.2.5)

The maximum likelihood condition (8.58) finds an extremum which may be a minimum; so it is important to check. There may also be more than one maximum, in which case one usually takes the highest maximum. The maximum may also be at a physical boundary, in which case



eq. (8.58) may not find it. Usually such problems do not occur for sufficiently large samples. However, this is not always the case.

Note that for the purpose of finding the maximum of $\mathcal{L}$, it is not necessary that $\mathcal{L}$ be normalized. Any factors not depending on θ can be thrown away. This includes factors which depend on $\underline{x}$ but not on θ.

Example: n independent x_i, each distributed normally.

$$\mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$
$$\ell = \sum_{i=1}^{n}\left[-\frac{1}{2}\ln(2\pi) - \ln\sigma_i - \frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right]$$

Suppose that all the μ_i are the same, $\mu_i = \mu$, but that the σ_i are different, but known. This is the case if we make n measurements of the same quantity, each with a different precision, *e.g.*, using different apparatus. The maximum likelihood condition (8.58) is then

$$\frac{\partial \ell}{\partial \mu} = \sum \frac{x_i - \mu}{\sigma_i^2} = \sum \frac{x_i}{\sigma_i^2} - \sum \frac{\mu}{\sigma_i^2} = 0$$

The solution of this equation is the ML estimate of μ:

$$\hat{\mu} = \frac{\sum(x_i/\sigma_i^2)}{\sum(1/\sigma_i^2)} \tag{8.59}$$

which is a weighted average, each x_i weighted by $1/\sigma_i^2$.

The expectation of $\hat{\mu}$ is

$$E[\hat{\mu}] = E\left[\frac{\sum(x_i/\sigma_i^2)}{\sum(1/\sigma_i^2)}\right] = \frac{\sum(E[x_i]/\sigma_i^2)}{\sum(1/\sigma_i^2)} = \frac{\sum(\mu/\sigma_i^2)}{\sum(1/\sigma_i^2)} = \frac{\mu\sum(1/\sigma_i^2)}{\sum(1/\sigma_i^2)} = \mu$$

from which we conclude that this estimate is unbiased. The variance of $\hat{\mu}$ is

$$V[\hat{\mu}] = E\left[\hat{\mu}^2\right] - \left(E[\hat{\mu}]\right)^2 = E\left[\hat{\mu}^2\right] - \mu^2 = \frac{E\left[\left(\sum \frac{x_i}{\sigma_i^2}\right)^2\right]}{\left[\sum\left(\frac{1}{\sigma_i^2}\right)\right]^2} - \mu^2 = \frac{E\left[\sum_i \sum_j \frac{x_i x_j}{\sigma_i^2 \sigma_j^2}\right]}{\left[\sum\left(\frac{1}{\sigma_i^2}\right)\right]^2} - \mu^2$$

Since the x_i are independent,

$$E[x_i x_j] = \begin{cases} E[x_i]\,E[x_j] = \mu_i\mu_j = \mu^2 & \text{if } i \neq j \\ E[x_i^2] = \sigma_i^2 + \mu^2 & \text{if } i = j \end{cases}$$

Therefore, having written the expectation of sums as the sum of expectations and having split the double sum into two parts,

$$\begin{aligned}
V[\hat{\mu}] &= \left(\frac{1}{\sum(1/\sigma_i^2)}\right)^2 \left(\sum_i \frac{\sigma_i^2+\mu^2}{\sigma_i^4} + \sum_i \sum_{j\neq i} \frac{\mu^2}{\sigma_i^2\sigma_j^2}\right) - \mu^2 \\
&= \left(\frac{1}{\sum(1/\sigma_i^2)}\right)^2 \left(\sum_i \frac{1}{\sigma_i^2} + \mu^2 \sum_i \frac{1}{\sigma_i^4} + \mu^2 \sum_i \sum_{j\neq i} \frac{1}{\sigma_i^2\sigma_j^2}\right) - \mu^2 \\
&= \frac{1}{\sum(1/\sigma_i^2)} + \mu^2 \underbrace{\left(\frac{\sum_i \frac{1}{\sigma_i^4} + \sum_i \sum_{j\neq i} \frac{1}{\sigma_i^2\sigma_j^2}}{\left(\sum(1/\sigma_i^2)\right)^2}\right)}_{=1} - \mu^2 \\
&= \frac{1}{\sum(1/\sigma_i^2)}
\end{aligned} \tag{8.60}$$

It is curious that in this example $V[\hat{\mu}]$ does not depend on the x_i, but only on the σ_i. This is not true in general.

We have seen (section 8.2.6) that the Rao-Cramér inequality sets a lower limit on the variance of an estimator. For an unbiased estimator the bound is $1/I$, where I is the information. For μ,

$$\begin{aligned}
I(\mu) &= -E\left[\frac{\partial S(\mu)}{\partial \mu}\right] = -E\left[\frac{\partial^2 \ell}{\partial \mu^2}\right] \\
&= -E\left[\frac{\partial}{\partial \mu}\left(\sum \frac{x_i}{\sigma_i^2} - \sum \frac{\mu}{\sigma_i^2}\right)\right] \\
&= -E\left[-\sum \frac{1}{\sigma_i^2}\right] = \sum \frac{1}{\sigma_i^2}
\end{aligned}$$

Thus $V[\hat{\mu}] = I^{-1}(\mu)$; the variance of $\hat{\mu}$ is the smallest possible. The ML estimator is efficient. This is in fact a general property of ML estimators: The ML estimator is efficient if an efficient estimator exists. We will now demonstrate this.

Properties of maximum likelihood estimators

We have seen in section 8.2.7 that an efficient, unbiased estimator is linearly related to the score function. Assume that such an estimator of θ exists; call it $T(\underline{x})$. Then

$$S(\underline{x}, \theta) = C(\theta)T(\underline{x}) + D(\theta) \tag{8.61}$$

From the maximum likelihood condition, $S(\underline{x}, \hat{\theta}) = 0$, where $\hat{\theta}$ is the ML estimator of θ. Hence the unbiased, efficient estimator $T(\underline{x})$ is related to the ML estimator $\hat{\theta}$ by

$$T(\underline{x}) = -\frac{D(\hat{\theta})}{C(\hat{\theta})} \tag{8.62}$$

We have also seen in section 8.2.6, equation 8.21, that $E[S(\underline{x},\theta)] = 0$ under quite general conditions on f. Therefore, taking the expectation of equation 8.61,

$$E[S(\underline{x},\theta] = C(\theta)E[T(\underline{x})] + D(\theta) = 0$$

Hence,

$$E[T(\underline{x})] = -\frac{D(\theta)}{C(\theta)} \tag{8.63}$$

This is true for any value of θ; in particular it is true for $\theta = \hat{\theta}$, *i.e.*, if the true value of θ is equal to the ML estimate of θ:

$$E\left[T(\underline{x}|\hat{\theta})\right] = -\frac{D(\hat{\theta})}{C(\hat{\theta})} = T(\underline{x}) \tag{8.64}$$

It may seem strange to write $E\left[T(\underline{x}|\hat{\theta})\right]$ since $T(\underline{x})$ does not depend on the value of θ. However, the expectation operator does depend on the value of θ. In fact, since $T(\underline{x})$ is an unbiased estimator of θ,

$$E[T(\underline{x})] = \int T(\underline{x})\, f(\underline{x},\theta)\, \mathrm{d}x = \theta \tag{8.65}$$

Hence,

$$E\left[T(\underline{x}|\hat{\theta})\right] = \hat{\theta}$$

Combining this with equation 8.64 gives

$$T(\underline{x}) = \hat{\theta} \tag{8.66}$$

Thus we have demonstrated that the ML estimator is efficient and unbiased if an efficient, unbiased estimator exists.

If an unbiased, efficient estimator exists, we can derive the following properties:

1. From equations 8.63 and 8.65,

   $$D(\theta) = -\theta\, C(\theta)$$

   Substituting this and equation 8.66 in equation 8.61 yields

   $$S(\underline{x},\theta) = C(\theta)\left[\hat{\theta} - \theta\right] \tag{8.67}$$

2. Assuming that the estimator is efficient means that the Rao-Cramér inequality, equation 8.26, becomes an equality. Collecting equations 8.19, 8.23, and 8.26, results in the variance of an unbiased, efficient estimator $\hat{\theta}$ given by

   $$V\left[\hat{\theta}\right] = \frac{1}{I(\theta)} = \frac{1}{E[S^2]} = -\frac{1}{E\left[\frac{\partial S}{\partial \theta}\right]} = -\frac{1}{E\left[\frac{\partial^2 \ell}{\partial \theta^2}\right]}$$

From (8.67),

$$\frac{\partial S}{\partial \theta} = C'(\theta)\left[\hat{\theta} - \theta\right] - C(\theta) \tag{8.68}$$

Since $\hat{\theta}$ is unbiased, $E\left[\hat{\theta}\right] = \theta_{\mathrm{t}}$, the true value of the parameter. Hence,

$$E\left[\frac{\partial S}{\partial \theta}\right] = -C(\theta_{\mathrm{t}})$$

$$\text{and} \qquad V\left[\hat{\theta}\right] = \frac{1}{C(\theta_{\mathrm{t}})} \tag{8.69}$$

Hence, $C(\theta_{\mathrm{t}}) > 0$.

3. From equation 8.68, we also see that

$$\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = \left.\frac{\partial S}{\partial \theta}\right|_{\theta=\hat{\theta}} = -C(\hat{\theta})$$

Since $C(\theta) > 0$ in the region of the true value, this confirms that the extremum of ℓ, which we have used to determine $\hat{\theta}$, is in fact a maximum.

4. From equation 8.67 and the maximum likelihood condition (equation 8.58), we see that the ML estimator is the solution of

$$0 = S(\underline{x}, \theta) = C(\theta)\left(\hat{\theta} - \theta\right)$$

Since $C(\theta) > 0$ in the region of the true value, this equation can have only one solution, namely $\hat{\theta}$. Hence, the maximum likelihood estimator $\hat{\theta}$ is unique.

Let us return to the Gaussian example. But now assume not only that all $\mu_i = \mu$ but also all $\sigma_i = \sigma$. Unlike the previous example, we now assume that σ is unknown. The likelihood condition gives

$$\left.\frac{\partial \ell}{\partial \mu}\right)_{\hat{\mu},\hat{\sigma}} = \sum\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0$$

$$\left.\frac{\partial \ell}{\partial \sigma}\right)_{\hat{\mu},\hat{\sigma}} = \sum\left(-\frac{1}{\hat{\sigma}} + \frac{(x_i - \hat{\mu})^2}{\hat{\sigma}^3}\right) = 0$$

The first equation gives

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x}$$

Using this in the second equation gives

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

which, as we have previously seen (eq. 8.6), is a biased estimator of σ^2. This illustrates an important, though often forgotten, feature of ML estimators: They are often biased.

To summarize this section: The ML estimator is efficient and unbiased if such an estimator exists. Unfortunately, that is not always the case.

8.4.2 Asymptotic properties

Although, as we have seen in the previous section, the maximum likelihood estimator is efficient and unbiased if an efficient, unbiased estimator exists, in general the ML estimator is neither unbiased nor efficient. However, asymptotically, *i.e.*, for a large number of independent measurements, it (usually) is both unbiased and efficient. To see this we expand the score about $\hat{\theta}$:

$$S(\underline{x},\theta) = \frac{\partial}{\partial\theta}\sum \ln f(x_i,\theta) \approx S(\underline{x},\hat{\theta}) + \left.\frac{\partial S}{\partial\theta}\right|_{\hat{\theta}}\left(\theta-\hat{\theta}\right) + \ldots$$

By the maximum likelihood principle, $S(\underline{x},\hat{\theta}) = 0$. We assume that as $n \to \infty$ higher order terms can be neglected. We are then left with

$$\begin{aligned} S(\underline{x},\theta) &\approx \left.\frac{\partial S}{\partial\theta}\right|_{\hat{\theta}}\left(\theta-\hat{\theta}\right) = \frac{\partial}{\partial\theta}\sum S_1(x_i,\theta)\Big|_{\hat{\theta}}\left(\theta-\hat{\theta}\right) \\ &= \sum \left.\frac{\partial S_1(x_i,\theta)}{\partial\theta}\right|_{\hat{\theta}}\left(\theta-\hat{\theta}\right) \end{aligned}$$

Replacing the sum by n times the sample mean,

$$S(\underline{x},\theta) \approx n\,\overline{\left.\frac{\partial S_1}{\partial\theta}\right|_{\hat{\theta}}}\left(\theta-\hat{\theta}\right) = n\,\overline{\left.\frac{\partial^2}{\partial\theta^2}\ln f(x_i,\theta)\right|_{\hat{\theta}}}\left(\theta-\hat{\theta}\right)$$

Since the sample mean approaches the expectation as $n \to \infty$ provided only that the variance is finite (C.L.T.), asymptotically

$$\begin{aligned} S(\underline{x},\theta) &\approx nE\left[\left.\frac{\partial S_1}{\partial\theta}\right|_{\hat{\theta}}\right]\left(\theta-\hat{\theta}\right) = nE\left[\left.\frac{\partial^2}{\partial\theta^2}\ln f(x_i,\theta)\right|_{\hat{\theta}}\right]\left(\theta-\hat{\theta}\right) \\ &= E\left[\left.\frac{\partial}{\partial\theta}\sum S_1\right|_{\hat{\theta}}\right]\left(\theta-\hat{\theta}\right) = E\left[\left.\frac{\partial^2}{\partial\theta^2}\sum \ln f(x_i,\theta)\right|_{\hat{\theta}}\right]\left(\theta-\hat{\theta}\right) \\ &= E\left[\left.\frac{\partial S}{\partial\theta}\right|_{\hat{\theta}}\right]\left(\theta-\hat{\theta}\right) = E\left[\left.\frac{\partial^2\ell}{\partial\theta^2}\right|_{\hat{\theta}}\right]\left(\theta-\hat{\theta}\right) \\ &= -I(\hat{\theta})\left(\theta-\hat{\theta}\right) \end{aligned} \tag{8.70}$$

the last step following from equation 8.23.

There are several consequences of equation 8.70:

- First we note that asymptotically, $I(\theta) = I(\hat{\theta})$:

  $$I(\theta) = -E\left[\frac{\partial S}{\partial\theta}\right] = E\left[I(\hat{\theta})\right] = I(\hat{\theta})$$

  where the second step follows from equation 8.70 and the last step follows since $I(\hat{\theta})$ is itself an expectation and the expectation of an expectation is just the expectation itself.

- The result, equation 8.70, that $\hat{\theta}$ is linearly related to the score function, implies (section 8.2.7) that $\hat{\theta}$ is unbiased and efficient. This is an important asymptotic property of ML estimators.

- Further, we can integrate equation 8.70,

$$\frac{\partial}{\partial\theta}\ln\mathcal{L} = S(\underline{x},\theta) \approx -I(\hat{\theta})(\theta-\hat{\theta})$$

over θ to find

$$\ell = \ln\mathcal{L} \approx -\frac{I(\hat{\theta})}{2}\left(\hat{\theta}-\theta\right)^2 + \ln k \tag{8.71}$$

where the integration constant, k, is just $k = \mathcal{L}(\hat{\theta}) = \mathcal{L}_{\max}$. Exponentiating,

$$\mathcal{L}(\theta) \approx \mathcal{L}_{\max}\exp\left[-\frac{1}{2}I(\hat{\theta})(\hat{\theta}-\theta)^2\right] \propto N\left(\theta;\hat{\theta},I^{-1}(\hat{\theta})\right) \tag{8.72}$$

Thus, asymptotically, $\mathcal{L}$ is proportional to a Gaussian function of θ with mean $\hat{\theta}$ and variance $1/I(\hat{\theta})$.

Instead of starting with equation 8.70, we could use equation 8.67, which expresses the linear dependence of S on $\hat{\theta}$ for any efficient, unbiased estimator. Integrating equation 8.67 leads to

$$\mathcal{L}(\theta) = \mathcal{L}_{\max}\,\exp\left[-\frac{1}{2}C(\theta)(\hat{\theta}-\theta)^2\right]$$

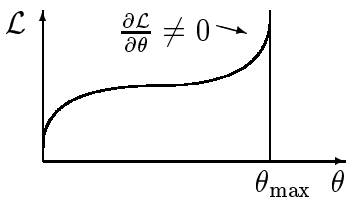
which looks formally similar to equation 8.72 but is not, in fact, a Gaussian function since C depends on θ. Only asymptotically must $C(\theta)$ approach a constant, $C(\theta) \to I(\hat{\theta})$. Nevertheless, $C(\theta)$ *may* be constant for finite n, as we have seen in the example of using $\bar{x}$ to estimate μ of a Gaussian (*cf.* section 8.2.7).

We emphasize again that, despite the form of equation 8.72, $\mathcal{L}$ is not a p.d.f. for θ. It is an experimentally observed function. Nevertheless, the principle of maximum likelihood tells us to take the maximum of $\mathcal{L}$ to determine $\hat{\theta}$, *i.e.*, to take $\hat{\theta}$ equal to the mode of $\mathcal{L}$. In this approximation the mode of $\mathcal{L}$ is equal to the mean, which is just $\hat{\theta}$. In other words the ML estimate is the same as what we would find if we were to regard $\mathcal{L}$ as a p.d.f. for θ and use the expectation (mean) of $\mathcal{L}$ to estimate θ.

Since asymptotically the ML estimator is unbiased and efficient, the Rao-Cramér bound is attained and $V\left[\hat{\theta}\right] = I^{-1}(\theta)$. Thus the variance is also that which we would have found treating $\mathcal{L}$ as a p.d.f. for θ.

We have shown that the ML estimator is, under suitable conditions, asymptotically efficient and unbiased. Let us now specify these conditions (without proof) more precisely:

1. The true value of θ must not be at the boundary of its allowed interval such that the maximum likelihood condition would not be satisfied, *i.e.*, $\frac{\partial \mathcal{L}}{\partial \theta}$ must be zero at the maximum.



2. The p.d.f.'s defined by different values of θ must be distinct, *i.e.*, two values of θ must not give p.d.f.'s whose ratio is not a function of θ. Otherwise there would be no way to decide between them.

3. The first three derivatives of $\ell = \ln \mathcal{L}$ must exist in the neighborhood of $\hat{\theta}$.

4. The information, $I(\theta)$ must be finite and positive definite.

8.4.3 Change of parameters

It is important to understand the difference between a change of parameters and a change of variable. $\mathcal{L}(x;\theta)$ is a p.d.f. for the random variable x. Under a change of variable, $x \longrightarrow y(x)$ and $\mathcal{L}(x;\theta) \longrightarrow \mathcal{L}'(y;\theta)$, the probability must be conserved. Hence, $\mathcal{L}(x;\theta)\,\mathrm{d}x = \mathcal{L}'(y;\theta)\,\mathrm{d}y$. This requirement results (*cf.* section 2.2.6) in

$$\mathcal{L}'(y;\theta) = \mathcal{L}(w(y);\theta)\,|J|$$

where w is the inverse of the transformation $x \longrightarrow y$ and J is the Jacobian of the transformation.

However, for a change of parameters, $\theta \longrightarrow g(\theta)$, the requirement that probability be conserved means that $\mathcal{L}(x;\theta)\,\mathrm{d}x = \mathcal{L}'(x;g)\,\mathrm{d}x$ and consequently that $\mathcal{L}(x;\theta) = \mathcal{L}'(x;g)$. Thus the value of $\mathcal{L}$ is unchanged by the transformation from θ to $g(\theta)$ and $\mathcal{L}'$ is obtained from $\mathcal{L}$ simply be replacing θ by $h(g)$ where h is the inverse of the transformation $\theta \to g$. There is no Jacobian involved.

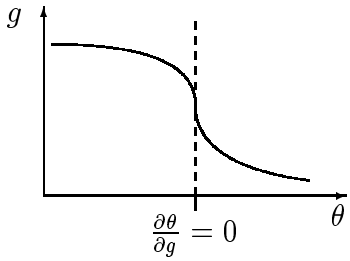
As in frequency substitution, the ML estimator of a function, g, of the parameter θ is just that function for the ML estimator, *i.e.*,

$$\hat{g}(\theta) = g(\hat{\theta})$$

This occurs because, assuming $\frac{\partial \theta}{\partial g}$ exists,

$$\frac{\partial \mathcal{L}}{\partial g} = \frac{\partial \mathcal{L}}{\partial \theta}\frac{\partial \theta}{\partial g}$$

Then the maximum likelihood condition for θ, $\frac{\partial \mathcal{L}}{\partial \theta} = 0$, implies that $\frac{\partial \mathcal{L}}{\partial g} = 0$, which is just the maximum likelihood condition for g.



If $\frac{\partial\theta}{\partial g}$ is zero at some value of θ, this can introduce additional solutions to the likelihood condition for g. This will not usually happen if g is a single-valued function of θ unless there are points of inflection.

Note that $\hat{\theta}$ unbiased does not imply that $\hat{g} = g(\hat{\theta})$ is unbiased and *vice versa*. Asymptotically, both $\hat{\theta}$ and $\hat{g}$ become unbiased and efficient (previous section), but they usually approach this at different rates.

In the case of more than one parameter, $\underline{g}(\underline{\theta})$, the above generalizes to

$$\frac{\partial\mathcal{L}}{\partial g_k} = \sum_i \frac{\partial\mathcal{L}}{\partial\theta_i}\frac{\partial\theta_i}{\partial g_k} = \left(\frac{\partial\mathcal{L}}{\partial\underline{\theta}}\right)^{\mathrm{T}}\left(\frac{\partial\underline{\theta}}{\partial g_k}\right) \tag{8.73}$$

and the information matrix transforms as

$$I_{jk}(\underline{g}) = \left(\frac{\partial g_j}{\partial\underline{\theta}}\right)^{\mathrm{T}} I(\underline{\theta})\left(\frac{\partial g_k}{\partial\underline{\theta}}\right) \tag{8.74}$$

It is not necessary that $\underline{\theta}$ and $\underline{g}$ have the same dimensions.

8.4.4 Maximum Likelihood *vs.* Bayesian inference

Recall Bayes' theorem (section 2.3). Assume that the parameter θ, which we wish to estimate, can have only discrete values, $\theta_1, \theta_2, \ldots, \theta_k$. Applied to the estimation of θ, Bayes' theorem can be stated (*cf.* section 2.4.3)

$$P_{\text{posterior}}(\theta_i \mid \underline{x}) = \frac{P(\underline{x} \mid \theta_i)}{P(\underline{x})} P_{\text{prior}}(\theta_i) \tag{8.75}$$

and it would seem reasonable to choose as our estimate of $\hat{\theta}$ that value θ_i having the largest $P_{\text{posterior}}$, *i.e.*, the mode of the posterior probability.* Since $P_{\text{posterior}}$ is normalized, *i.e.*, $\sum_i P_{\text{posterior}}(\theta_i|\underline{x}) = 1$, we see that $P(\underline{x}) = \sum_i P(\underline{x}|\theta_i)P_{\text{prior}}(\theta_i)$ is the constant which serves to normalize $P_{\text{posterior}}$. We also see that $P(\underline{x}|\theta_i)$ is just the likelihood, $\mathcal{L}(\underline{x};\theta_i)$, apart from normalization.

In the absence of prior knowledge (belief) of θ, Bayes' postulate tells us to assume all values equally likely, *i.e.*, $P_{\text{prior}}(\theta_i) = \frac{1}{k}$. Then the right-hand side of equation 8.75 is exactly $\mathcal{L}(\underline{x};\theta_i)$ (apart from normalization) and maximizing $P_{\text{posterior}}$ is the same as maximizing $\mathcal{L}$. Thus, Bayesian statistics leads to the same estimator as maximum likelihood.

In the more usual case of a continuous parameter, equation 8.75 must be rewrit-

*The mode is not the only choice. A Bayesian could also choose the mean or the median, or some other property of the posterior probability distribution. Asymptotically, of course, $P_{\text{posterior}}$ will be Gaussian, in which case the mode, mean, and median are the same.

ten in terms of probability densities:

$$f_{\text{posterior}}(\theta \mid \underline{x}) = \frac{f(\underline{x} \mid \theta)}{\int f(\underline{x} \mid \theta)\, f_{\text{prior}}(\theta)\, \mathrm{d}\theta}\, f_{\text{prior}}(\theta) \tag{8.76}$$

Assuming Bayes' postulate, $f_{\text{prior}} = \text{constant}$, and again Bayesian statistics is equivalent to maximum likelihood.

But now what happens if we want to estimate the parameter $g = g(\theta)$ rather than θ? Assume that the transformation $g(\theta)$ is one-to-one. Then in the discrete case we just replace θ_i by $g_i = g(\theta_i)$ in equation 8.75. Bayes' postulate again tells us that $P_{\text{prior}} = \frac{1}{k}$ and the same maximum is found resulting in $\hat{g} = g(\hat{\theta})$. However in the continuous case, the change of parameter (*cf.* sections 2.2.6, 8.4.3) involves a Jacobian, since in Bayesian statistics f is a p.d.f. for θ, or in other words, the ML parameter is regarded as the variable of the p.d.f. Hence,

$$f_{\text{posterior}}(g \mid \underline{x}) = f_{\text{posterior}}(\theta \mid \underline{x})\, |J|$$

where J is the Jacobian of the transformation $\theta \to g$. But since the likelihood function is a p.d.f. for $\underline{x}$, not for θ, there is no Jacobian involved in rewriting $\mathcal{L}$ using g instead of θ, *i.e.*, $\mathcal{L}(\underline{x}; g(\theta)) = \mathcal{L}(\underline{x}; \theta)$. Thus, assuming Bayes' postulate for g, $f_{\text{prior}}(g) = \text{constant}$, the value of g which maximizes $f_{\text{posterior}}(g \mid \underline{x})$ is that which maximizes $\mathcal{L}(\underline{x}; \theta)|J|$ rather than $\mathcal{L}(\underline{x}; \theta)$. Bayesian statistics and maximum likelihood thus give different estimates of g. To obtain the same result in ML, the Bayesian would have to use $f_{\text{prior}}(g) = f_{\text{prior}}(\theta)|J|$ rather than the uniform $f_{\text{prior}}(g)$ suggested by Bayes' postulate. In other words, Bayes' postulate can only be applied to θ or g, not simultaneously to both (except when θ and g are linearly related). But how does one choose which?* This is one of the grounds which would cause most physicists to prefer maximum likelihood to Bayesian parameter estimation.

8.4.5 Variance of maximum likelihood estimators

We have seen that the variance of an efficient estimator is given by the Rao-Cramér bound (equation 8.26). Assuming that $\hat{\theta}$ is efficient, substituting $\hat{\theta}$ for θ in this equation gives an estimate of $V\left[\hat{\theta}\right]$. If, in addition, $\hat{\theta}$ is unbiased (or at least that the bias does not depend on θ), this just becomes $V\left[\hat{\theta}\right] = 1/I(\hat{\theta})$. We recall that the ML estimator is efficient if an efficient estimator exists, but that this is not always the case. Nor is the ML estimator always unbiased.

If the estimator is unbiased and efficient

However, asymptotically the ML estimator is both unbiased and efficient. Assuming this to be the case, and also assuming that the range of x does not depend on θ, we can estimate the variance as follows:

*There are arguments for the choice of non-uniform priors (see, *e.g.*, Jeffreys[37]) in certain

1. Using equation 8.19, $I(\theta) = E[S^2]$,

$$V^{-1}\left[\hat{\theta}\right] = I(\theta) = E\left[S^2\right] = E\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$$

which, for more than one parameter generalizes to

$$V_{jk}^{-1}\left[\underline{\hat{\theta}}\right] = E\left[\frac{\partial \ell}{\partial \theta_j}\frac{\partial \ell}{\partial \theta_k}\right] \tag{8.77}$$

If the sample consists of n independent events distributed according to the p.d.f.'s $f_i(x_i;\theta)$, the score is just the sum of the scores for the individual events and

$$V^{-1}\left[\hat{\theta}\right] = E\left[\left(\sum_{i=1}^{n} S_1(x_i;\theta)\right)^2\right]$$

Performing the square and using the fact that the expectation of a sum is the sum of the expectations, we get

$$V^{-1}\left[\hat{\theta}\right] = \sum_{i=1}^{n} E\left[S_1^2(x_i;\theta)\right] + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} E\left[S_1(x_i;\theta)S_1(x_j;\theta)\right]$$

However, the cross terms are zero, which follows from the fact that for independent x_i the expectation of the product equals the product of the expectations and from $E[S_1(x;\theta)] = 0$ (equation 8.20). Therefore, generalizing to more than one parameter,

$$V_{jk}^{-1}\left[\underline{\hat{\theta}}\right] = \sum_{i=1}^{n} E\left[\frac{\partial \ln f_i(x_i;\underline{\theta})}{\partial \theta_j}\frac{\partial \ln f_i(x_i;\underline{\theta})}{\partial \theta_k}\right] \tag{8.78}$$

Not knowing the true value of θ, we estimate this by evaluating it at $\underline{\theta} = \underline{\hat{\theta}}$. If all the f_i are the same, equation 8.78 reduces to

$$V_{jk}^{-1}\left[\underline{\hat{\theta}}\right] = nE\left[\frac{\partial \ln f(x;\underline{\theta})}{\partial \theta_j}\frac{\partial \ln f(x;\underline{\theta})}{\partial \theta_k}\right] \tag{8.79}$$

Rather than calculating the expectation and evaluating it at $\underline{\hat{\theta}}$, we can estimate the expectation value by the sample mean evaluated at $\underline{\hat{\theta}}$:

$$\widehat{V_{jk}^{-1}}\left[\underline{\hat{\theta}}\right] = \sum_{i=1}^{n} \left.\frac{\partial \ln f(x_i;\underline{\theta})}{\partial \theta_j}\right|_{\underline{\hat{\theta}}} \left.\frac{\partial \ln f(x_i;\underline{\theta})}{\partial \theta_k}\right|_{\underline{\hat{\theta}}} \tag{8.80}$$

circumstances. However, they are not completely convincing and remain controversial.

2. I is also given by equation 8.23:

$$I(\theta) = -E\left[\frac{\partial S}{\partial \theta}\right] = -\left.\frac{\partial S}{\partial \hat{\theta}}\right|_{\hat{\theta}=\theta} = -\left.\frac{\partial^2 \ell}{\partial \hat{\theta}^2}\right|_{\hat{\theta}=\theta} \tag{8.81}$$

where the second step follows from the linear dependence of S on $\hat{\theta}$ (equation 8.30) for an unbiased, efficient estimator. The variance is then estimated by evaluating the derivative at $\theta = \hat{\theta}$:

$$\widehat{V^{-1}}\left[\hat{\theta}\right] = -\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\hat{\theta}} \tag{8.82}$$

In the case of more than one parameter, this becomes

$$V_{jk}^{-1}\left[\underline{\hat{\theta}}\right] = I_{jk}(\underline{\theta}) = -E\left[\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right] \tag{8.83}$$

which is estimated by

$$\widehat{V_{jk}^{-1}}\left[\underline{\hat{\theta}}\right] = I_{jk}(\hat{\theta}) = -\left.\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right|_{\underline{\hat{\theta}}} \tag{8.84}$$

which is the Hessian matrix* of $-\ell$. For n independent events, all distributed as $f(x;\underline{\theta})$, the expectations in equations 8.81 and 8.83 can be estimated by a sample mean evaluated at $\hat{\theta}$. Thus

$$\widehat{V_{jk}^{-1}}\left[\underline{\hat{\theta}}\right] = -\sum_{i=1}^{n} \left.\frac{\partial^2 \ln f(x_i;\theta)}{\partial \theta_j \partial \theta_k}\right|_{\underline{\hat{\theta}}} = -n\,\overline{\left.\frac{\partial^2 \ln f(x;\theta)}{\partial \theta_j \partial \theta_k}\right|_{\underline{\hat{\theta}}}} \tag{8.85}$$

The expectation forms (8.77, 8.78, 8.79 and 8.84) are useful for estimating the error we expect before doing the experiment, *e.g.*, to decide how many events we need to have in order to achieve a certain precision under various assumptions for θ. Both the expectation and the sample mean forms (8.80 and 8.85) may be used after the experiment has been done. It is difficult to give general guidelines on which method is most reliable.

Example: Let us apply the two methods to the example of n independent x_i distributed normally with the same μ but different σ_i. Assume that the σ_i are known. Recall that in this case

$$\ell = \sum_{i=1}^{n}\left[-\frac{1}{2}\ln(2\pi) - \ln\sigma_i - \frac{1}{2}\left(\frac{(x_i-\mu)}{\sigma_i}\right)^2\right]$$

*Mathematically it is conditions on the first derivative vector, $\partial\ell/\partial\underline{\hat{\theta}}$, and on the Hessian matrix

1. From equation 8.78,

$$V^{-1}[\hat{\mu}] = \sum_{i=1}^{n} E\left[\left(\frac{\partial \ln f_i(x_i;\mu,\sigma_i)}{\partial \mu}\right)^2\right] = \sum_{i=1}^{n} E\left[\left(-\frac{1}{2}\frac{\partial}{\partial \mu}\left(\frac{x_i-\mu}{\sigma_i}\right)^2\right)^2\right]$$

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

since σ_i is just a parameter of f_i, hence a constant

$$= \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$ since this expectation is 1 for the normal p.d.f.

2. Since $\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma_i^2}$, equation 8.84 yields

$$V^{-1}[\hat{\mu}] = -\frac{\partial^2 \ell}{\partial \mu^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

Thus both methods give $V[\hat{\mu}] = 1/\sum\left(\frac{1}{\sigma_i^2}\right)$. This is the same result we found in section 8.4.1, equation 8.60, where we calculated the variance explicitly from the definition. This was, of course, to be expected since in this example $\hat{\mu}$ is unbiased and efficient and the range of x is independent of μ.

Variance using Bayesian inference

We have emphasized that $\mathcal{L}$ is the p.d.f. for $\underline{x}$ given $\underline{\theta}$ and *not* the p.d.f. for $\underline{\theta}$ given $\underline{x}$. However, using the Bayesian interpretation of probability (sections 2.4.3 and 8.4.4), these two conditional p.d.f.'s are related: By Bayes' theorem,

$$f_{\text{posterior}}(\underline{\theta}|\underline{x}) \propto f(\underline{x}|\underline{\theta})\, f_{\text{prior}}(\underline{\theta})$$

and $f(\underline{x}|\underline{\theta})$ is just the likelihood function $\mathcal{L}(\underline{x};\underline{\theta})$. If we are willing to accept Bayes' postulate (for which there is no mathematical justification) and take the prior p.d.f. for $\underline{\theta}$, $f_{\text{prior}}(\underline{\theta})$, as uniform in $\underline{\theta}$ (within possible physical limits), we have

$$f_{\text{posterior}}(\underline{\theta}|\underline{x}) = \frac{\mathcal{L}(\underline{x};\underline{\theta})}{\int \mathcal{L}(\underline{x};\underline{\theta})\, \mathrm{d}\underline{\theta}} \tag{8.86}$$

where the explicit normalization in the denominator is needed to normalize $f_{\text{posterior}}$, since $\mathcal{L}$ is normalized by $\int \mathcal{L}\, \mathrm{d}x = 1$. Since, in Bayesian inference $\mathcal{L}$ is regarded as a p.d.f. for θ, the covariance matrix of $\hat{\underline{\theta}}$,

$$V_{jk}\left[\hat{\underline{\theta}}\right] = E\left[\left(\hat{\theta}_j - \theta_j\right)\left(\hat{\theta}_k - \theta_k\right)\right] \tag{8.87}$$

that define the maximum of ℓ or the minimum of $-\ell$. The Hessian matrix is positive (negative) definite at a minimum (maximum) of the function and indefinite at a saddle point.

is given by

$$V_{jk}\left[\underline{\hat{\theta}}\right] = \frac{\int \left(\hat{\theta}_j - \theta_j\right)\left(\hat{\theta}_k - \theta_k\right) \mathcal{L}\ \mathrm{d}\underline{\theta}}{\int \mathcal{L}\ \mathrm{d}\underline{\theta}} \tag{8.88}$$

If the integrals in equation 8.88 can not be easily performed analytically, we could use Monte Carlo integration. Alternatively, we can estimate the expectation (8.87) from the data. This is similar to Monte Carlo integration, but instead of Monte Carlo points $\underline{\theta}$ we use the data themselves. Assuming n independent observations x_i, we estimate each parameter for each observation separately, keeping all other parameters fixed at $\underline{\hat{\theta}}$. Thus, $\hat{\theta}_{j(i)}$ is the value of $\hat{\theta}_j$ that would be obtained from using only the ith event. In other words, $\hat{\theta}_{j(i)}$ is the solution of

$$\left.\frac{\partial f_i(x_i;\underline{\theta})}{\partial \theta_j}\right|_{\theta_k=\hat{\theta}_k, k\neq j} = 0$$

With $\mathcal{L}$ regarded as a p.d.f. for $\underline{\theta}$, the $\hat{\theta}_{j(i)}$ are r.v.'s distributed according to $\mathcal{L}$. Their variance about $\underline{\theta}$ thus estimates the variance of $\hat{\theta}$. However, not knowing $\underline{\theta}$ we must use our estimate of it. This leads to the following estimate of the covariance, where in equation 8.87 the expectation has been replaced by an average over the observations, $\hat{\theta}$ by the estimate from one observation $\hat{\theta}_{j(i)}$, and θ_j by our estimate $\hat{\theta}_j$:

$$\widehat{V_{jk}}\left[\underline{\hat{\theta}}\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{j(i)} - \hat{\theta}_j\right)\left(\hat{\theta}_{k(i)} - \hat{\theta}_k\right) \tag{8.89}$$

Equation 8.88 is particularly easy to evaluate when $\mathcal{L}$ is a Gaussian. We have seen that asymptotically $\mathcal{L}$ is a Gaussian function of θ (equation 8.72) and hence that ℓ is parabolic (equation 8.71):

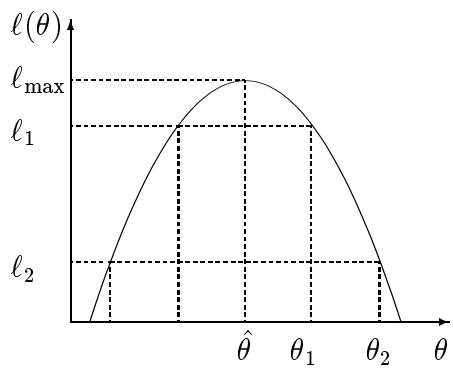
$$\mathcal{L} = \mathcal{L}_{\max}\, e^{-\frac{1}{2}Q^2} \quad , \qquad Q^2 = \frac{(\hat{\theta}-\theta)^2}{\sigma^2} \quad , \qquad \ell = \ln\mathcal{L} = \ln\mathcal{L}_{\max} - \frac{1}{2}Q^2 \tag{8.90}$$

Then, using the Bayesian interpretation, it follows from equation 8.88 that $V\left[\underline{\hat{\theta}}\right] = \sigma^2 = I^{-1}(\underline{\hat{\theta}})$.

However, in the asymptotic limit it is not necessary to invoke the Bayesian interpretation to obtain this result, since we already know from the asymptotic efficiency of the ML estimator that $V\left[\underline{\hat{\theta}}\right] = I^{-1}(\underline{\theta}) = I^{-1}(\underline{\hat{\theta}})$.

A graphical method

In any case, if $\mathcal{L}$ is Gaussian, the values of θ for which $Q^2 = \frac{(\hat{\theta}-\theta)^2}{\sigma^2} = 1$, *i.e.*, the values of θ corresponding to 1 standard deviation "errors", $\hat{\theta} - \theta = \pm\sigma$, are just those values, θ_1, for which ℓ differs from $\ell_{\max}$ by $1/2$. This



provides another way to estimate the uncertainty, $\delta\hat{\theta} = \sqrt{V\left[\hat{\theta}\right]}$, on $\hat{\theta}$: Find the value of θ, θ_1, for which

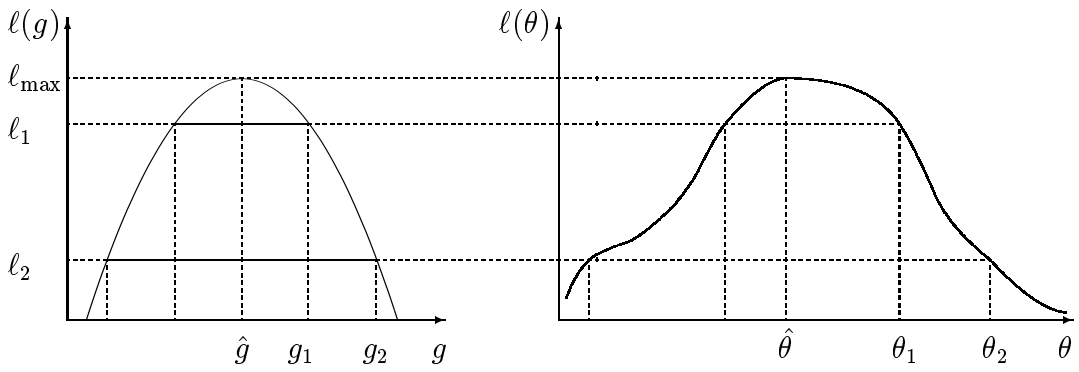
$$\ell_1 = \ell(\theta_1) = \ell_{\max} - \frac{1}{2}$$

The error is then $\delta\hat{\theta} = |\hat{\theta} - \theta_1|$ This could be done graphically from a plot of ℓ *vs.* θ. Similarly, two-standard deviation errors ($Q^2 = 4$) could be found using $\ell_2 = \ell_{\max} - 2$, *etc.* (The change in ℓ corresponding to Q standard deviations is $Q^2/2$.)

But, what do we do if $\mathcal{L}$ is not Gaussian? We can be Bayesian and use equation 8.87 or 8.88. Not wanting to be Bayesian, we can use the following approach. The two approaches will in general give different estimates of the variance, the difference being smallest when $\mathcal{L}$ is nearly of a Gaussian form.

Recall that for efficient, unbiased estimators $\mathcal{L}$ can be Gaussian even for finite n. Imagine a one-to-one transformation $g(\theta)$ from the parameter θ to a new parameter g and suppose that $\hat{g}$ is efficient and unbiased and hence that $\mathcal{L}(g)$ is normal. Such a g may not exist, but for now we assume that it does. We have seen that $\hat{g} = g(\hat{\theta})$. Let h be the inverse transformation, *i.e.*, $\theta = h[g(\theta)]$. Since, by assumption, $\mathcal{L}(g)$ is Gaussian, δg is given by a change of $1/2$ in $\ell(g)$.

But, as we have seen in section 8.4.3, $\mathcal{L}(\theta|\underline{x}) = \mathcal{L}(g(\theta)|\underline{x})$ for all θ; there is no Jacobian involved in going from $\mathcal{L}(\theta)$ to $\mathcal{L}(g)$. This means that since we can find δg from a change of $1/2$ in $\ell(g)$, $\delta\theta$ will be given by the same change.



$\mathcal{L}(\theta)$ need not be a symmetric function of θ, in which case the errors on $\hat{\theta}$ are asymmetric.

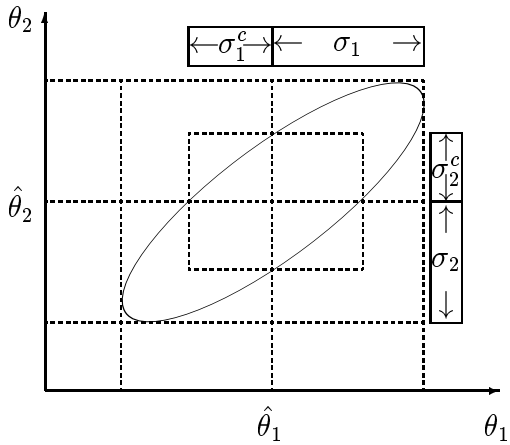
Note that we do not actually need to use the parameter g. We can find $\delta\theta$ directly.

A problem is that such a g may not exist. Asymptotically both $\mathcal{L}(g)$ and $\mathcal{L}(\theta)$ are Gaussian. However, in general, $\mathcal{L}(g)$ and $\mathcal{L}(\theta)$ will approach normality at different rates. It is therefore plausible that there exists some g which is at least nearly

normally distributed for finite n. Since we never actually have to use g, we can only adopt it as an assumption, realizing that the further away $\mathcal{L}$ for the 'best' g is from normality, the less accurate will be our estimation of $\delta\theta$.

This method of error estimation is easily extended to the case of more than one parameter. If all estimators are efficient, $\mathcal{L}$ will be a multivariate normal. We show the example of two parameters, θ_1 and θ_2. The condition of a change of $1/2$ in ℓ, *i.e.*, $Q^2 = 1$, gives an ellipse of constant $\mathcal{L}$ in θ_2 *vs.* θ_1. A distinction must be made, however, between the 'error' and the 'reduced' or 'conditional error', which is the error if the values of the other parameters are all assumed to be equal to their estimated values.

If, for example, θ_2 is held fixed at $\hat{\theta}_2$ and ℓ varied by $1/2$, the conditional error, σ_1^c is found rather than the error σ_1, which is the error that enters the multivariate normal distribution. In practice, the maximum of ℓ, as well as the variation of ℓ by $1/2$, are usually found on a computer using a search technique. However, since it is easier (faster), the program may compute σ^c rather than σ. If the parameters are uncorrelated, $\sigma^c = \sigma$. If parameters are correlated, the correlation should be stated along with the errors, or in other words, the complete covariance matrix should be stated, *e.g.*, as σ_1, σ_2, and ρ, the correlation coefficient.



8.4.6 Summary

- If the sample is large, maximum likelihood gives a unique, unbiased, minimum variance estimate under certain general conditions. However 'large' is not well defined. For finite samples the ML estimate may not be unique, unbiased, or of minimum variance. In this case other estimators may be preferable.

- Maximum likelihood estimators are often the easiest to compute, especially for complex problems. In many practical cases maximum likelihood is the only tractable approach.

- Maximum likelihood estimators are sufficient, *i.e.*, they use all the information about the parameter that is contained in the data. In particular, for small samples ML estimators can be much superior to methods which rely on binned data.

- Maximum likelihood estimators are not necessarily robust. If you use the wrong p.d.f., the ML estimate may be worse than that from some other method.

- The maximum likelihood method gives no way of testing the validity of the underlying theory, *i.e.*, whether or not the assumed p.d.f. is the correct one. In practice this is not so bad: You can always follow the maximum likelihood estimation by a goodness-of-fit test. Such tests will be discussed in section 10.6.

And finally, a practical point: In complex situations, the likelihood condition $\frac{\partial \ell}{\partial \theta_i} = 0$ can not be solved analytically. You then must code the likelihood function and use computer routines to find its maximum. Very clever programs exist as pre-packaged routines for finding the minimum or maximum of a function. Do not be tempted to write your own; take one from a good software library, *e.g.*, that of the Numerical Algorithms Group (NAG) or the MINUIT[38] program from CERN. Note that such programs usually search for a minimum instead of a maximum, so put a minus sign before your ℓ. One usually writes a subroutine which calculates the function for values of $\underline{\theta}$ chosen by the program. The program needs a starting value for $\underline{\theta}$. It evaluates the function at numerous points in $\underline{\theta}$ space, determines the most likely direction in this space to find the minimum (or maximum), and proceeds to search until the minimum is found. The search can usually be speeded up by also supplying a subroutine to calculate the derivatives of ℓ with respect to the θ_i; otherwise the program must do this numerically.

8.4.7 Extended Maximum Likelihood

Applied to n independent events from the same p.d.f., the likelihood method, as discussed so far, is a method to determine parameters governing the shape of the p.d.f. The number of events in the sample is not regarded as a variable.

Fermi proposed to extend the maximum likelihood method by including the number of events as a parameter to be estimated. His motivation was the grand canonical ensemble of statistical mechanics. In the canonical ensemble the number of atoms or molecules is regarded as fixed while in the grand canonical ensemble the number is free to vary.

To incorporate a variable number of events, the ordinary likelihood function is multiplied by the Poisson p.d.f. expressing the probability of obtaining n events when the expected number of events is ν. This expected number of events is then another parameter to be estimated from the data. The likelihood becomes

$$\mathcal{L}(\underline{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta) \longrightarrow \mathcal{L}_{\mathrm{E}}(\underline{x};\theta,\nu) = \frac{e^{-\nu}\nu^n}{n!}\prod_{i=1}^{n} f(x_i;\theta) \tag{8.91}$$

$$\ell(\underline{x};\theta) = \sum_{i=1}^{n} \ln\, f(x_i;\theta) \longrightarrow \ell_{\mathrm{E}}(\underline{x};\theta,\nu) = \sum_{i=1}^{n} \ln\, f(x_i;\theta) - \nu + n\ln\nu - \ln n!$$

$$\ell_{\mathrm{E}}(\underline{x};\theta,\nu) = \sum_{i=1}^{n} \ln\, \nu f(x_i;\theta) - \nu - \ln n!$$

Or,
$$\ell_{\mathrm{E}}(\underline{x};\theta,\nu) = \sum_{i=1}^{n} \ln\, g(x_i;\theta) - \nu \tag{8.92}$$

where $g = \nu f$ is the p.d.f. normalized to ν rather than to 1 and where we have dropped the constant term $\ln n!$ since constant terms are irrelevant in finding the maximum and the variance of estimators.

Just as the grand canonical ensemble can be used even for situations where the number of molecules is in fact constant (non-permeable walls), so also the extended maximum likelihood method. In particular, if there is no functional relationship between ν and θ, the likelihood condition $\partial\ell_{\mathrm{E}}/\partial\nu = 0$ will lead to $\hat{\nu} = n$. Also, $\partial\ell_{\mathrm{E}}/\partial\theta_j = \partial\ell/\partial\theta_j$, which leads to identical estimators $\hat{\theta}_j$ as in the ordinary maximum likelihood method. Nevertheless, we may still prefer to use the extended maximum likelihood method. It can happen that the p.d.f., f, is very difficult to normalize, *e.g.*, involving a lengthy numerical integration. Then, even though the number of events is fixed, we can use the extended maximum likelihood method, allowing the maximum likelihood principle to find the normalization. In this case, the resulting estimate of ν should turn out to be the actual number of events n times the normalization of f and the estimate of the other parameters to be the same as would have been found using the ordinary maximum likelihood method. However, the errors on the parameters will be overestimated since the method assumes that ν can have fluctuations. This overestimation can be removed (*cf.* section 3.9) by

1. inverting the covariance matrix,

2. removing the row and column corresponding to ν,

3. inverting the resulting matrix.

This corresponds to fixing ν at the best value, $\hat{\nu}$. Thus we could also fix $\nu = \hat{\nu}$ and find the errors on $\hat{\theta}$ by the usual ML procedure.

An example: Suppose we have an angular distribution containing N events, F in the forward hemisphere and $B = N - F$ in the backward hemisphere. In the ordinary maximum likelihood method N is regarded as fixed. The p.d.f. for the division of N events into F forward and B backward is the binomial p.d.f.:

$$\begin{aligned}\mathcal{L}(F;p) &= B(F;p,N) = \frac{N!}{F!B!}\,p^F(1-p)^B\\ \ell(F;p) &= F\ln p + B\ln(1-p) + \ln N! - \ln F!B!\end{aligned}$$

The likelihood condition is then

$$\frac{\partial\ell}{\partial p} = \frac{F}{p} - \frac{B}{1-p} = 0 \quad\longrightarrow\quad \hat{p} = \frac{F}{F+B} = \frac{F}{N}$$

Its variance is given by

$$V[\hat{p}] = -\left[\frac{\partial^2 \ell}{\partial p^2}\right]^{-1} = -\left[\frac{F}{p^2} + \frac{B}{(1-p)^2}\right]^{-1}$$

which we estimate by replacing p by $\hat{p}$:

$$\widehat{V}[\hat{p}] = -\left[\frac{F}{\hat{p}^2} + \frac{B}{(1-\hat{p})^2}\right]^{-1} = -\left[\frac{N}{\hat{p}} + \frac{N}{(1-\hat{p})}\right]^{-1} = -\left[\frac{N}{\hat{p}(1-\hat{p})}\right]^{-1} = \frac{\hat{p}(1-\hat{p})}{N}$$

The estimated numbers of forward and backward events, *i.e.*, the estimate of the expectation of the numbers of forward and backward events if the experiment were repeated, are then

$$\widehat{F} = N\hat{p} = F \qquad \text{and} \qquad \widehat{B} = N(1-\hat{p}) = B$$

with variance

$$V\left[\hat{F}\right] = V[N\hat{p}] = N^2 V[\hat{p}]$$

which is estimated by replacing V by $\widehat{V}$:

$$\widehat{V}\left[\hat{F}\right] = N^2 \widehat{V}[\hat{p}] = N\hat{p}(1-\hat{p}) = N\frac{F}{N}\frac{B}{N} = \frac{FB}{N}$$

Similarly,

$$V\left[\hat{B}\right] = V[N(1-\hat{p})] = V[N\hat{p}]$$
$$\widehat{V}\left[\hat{B}\right] = \frac{FB}{N}$$

Further, $\hat{F}$, $\hat{B}$ are completely anticorrelated.

In extended maximum likelihood N is not constant, but Poisson distributed. Hence,

$$\mathcal{L}_{\mathrm{E}} = \frac{e^{-\nu}\nu^N}{N!}\mathcal{L} = \frac{e^{-\nu}\nu^N}{N!}\frac{N!}{F!B!}p^F(1-p)^B$$
$$\ell_{\mathrm{E}} = -\nu + N\ln\nu - \ln N! + F\ln p + B\ln(1-p) + \ln N! - \ln F!B!$$
$$\frac{\partial \ell_{\mathrm{E}}}{\partial \nu} = 1 + \frac{N}{\nu} = 0 \qquad \longrightarrow \qquad \hat{\nu} = N$$

The likelihood condition for p, $\frac{\partial \ell_{\mathrm{E}}}{\partial p} = 0$ gives $\hat{p} = \frac{F}{N}$, the same as in ordinary likelihood. The variance of $\hat{p}$ is also the same. For $\hat{\nu}$, the variance is found as follows:

$$\frac{\partial^2 \ell_{\mathrm{E}}}{\partial \nu^2} = -\frac{N}{\nu^2} \qquad \longrightarrow \qquad V[\hat{\nu}] = \frac{\nu^2}{N}$$

Estimating the variance by replacing ν with $\hat{\nu}$ gives $\widehat{V}(\hat{\nu}) = N$. Further,

$$\frac{\partial^2 \ell_{\mathrm{E}}}{\partial \nu \partial p} = 0 \qquad \longrightarrow \qquad \hat{p} \text{ and } \hat{\nu} \text{ are uncorrelated.}$$

The estimate of the number of forward events is $\hat{F} = \hat{p}\hat{\nu} = F$, with the variance found by error propagation:

$$V\left[\hat{F}\right] = \hat{p}^2 V[\hat{\nu}] + \hat{\nu}^2 V[\hat{p}] = \frac{F^2}{N^2} N + N^2 \frac{\hat{p}(1-\hat{p})}{N} = \frac{F^2}{N} + N \frac{F}{N} \frac{B}{N} = F$$

The result for $\hat{B}$ is similar. Thus,

$$\hat{F} = F \pm \sqrt{F} \qquad \text{and} \qquad \hat{B} = B \pm \sqrt{B}$$

Alternatively, we can write the p.d.f. as a product of Poisson p.d.f.'s, one for forward events and one for backward events (see exercise 13). Again, N is not fixed. The parameters are now the expected numbers of forward, ϕ, and backward, β, events. Then

$$\mathcal{L}_{\mathrm{E}} = \frac{e^{-\phi}\phi^F}{F!} \frac{e^{-\beta}\beta^B}{B!}$$

which leads to the same result:

$$\hat{F} = \hat{\phi} = F \pm \sqrt{F} \qquad \text{and} \qquad \hat{B} = \hat{\beta} = B \pm \sqrt{B}$$

again with uncorrelated errors.

The constraint of fixed N leads to smaller, but correlated, errors in the ordinary maximum likelihood method. The estimates of the numbers of forward and backward events are, however, the same. Which method is correct depends on the question we are asking. To find the fraction of backward events we should use ordinary maximum likelihood. To find the number of backward events that we should expect in repetitions of the experiment where the number of events can vary, we should use extended maximum likelihood.

8.4.8 Constrained parameters

It often happens that the parameters to be estimated are constrained, for instance by a physical law. The imposition of constraints always implies adding some information, and therefore the errors of the parameters are in general reduced. One should therefore be careful not to add incorrect information. One should always test that the data are indeed compatible with the constraints. For example, before fixing a parameter at its theoretical value one should perform the fit with the parameter free and check that the resulting estimate is compatible with the theoretical value. Even if the theory is true, the data may turn out to give an incompatible value because of some experimental bias. Testing the compatibility is usually a good way to discover such experimental problems. How to do this will be discussed in sections 10.4 and 10.6.

The constraints may take the form of a set of equations

$$\underline{g}(\hat{\underline{\theta}}) = 0 \tag{8.93}$$

The most efficient method to deal with such constraints is to change parameters such that these equations become trivial. For example, if the constraint is

$$g(\underline{\theta}) = \theta_1 + \theta_2 - 1 = 0$$

we simply replace θ_2 by $1-\theta_1$ in the likelihood function and maximize with respect to θ_1.

Similarly, boundaries on a parameter, *e.g.*, $\theta_\mathrm{l} < \theta < \theta_\mathrm{h}$, can be imposed by the transformation

$$\theta = \theta_\mathrm{l} + \frac{1}{2}(\sin\psi + 1)(\theta_\mathrm{h} - \theta_\mathrm{l})$$

and maximizing $\mathcal{L}$ with respect to ψ.

When the θ_i are fractional contributions, subject to the constraints

$$0 \le \theta_i \le 1 \qquad ; \qquad \sum_{i=1}^{k} \theta_i = 1$$

one can use the following transformation:

$$\begin{aligned}
\theta_1 &= \xi_1 \\
\theta_2 &= (1-\xi_1)\xi_2 \\
\theta_3 &= (1-\xi_1)(1-\xi_2)\xi_3 \\
\vdots &= \vdots \\
\theta_{k-1} &= (1-\xi_1)(1-\xi_2)(1-\xi_3)\cdots(1-\xi_{k-2})\xi_{k-1} \\
\theta_k &= (1-\xi_1)(1-\xi_2)(1-\xi_3)\cdots(1-\xi_{k-2})(1-\xi_{k-1})
\end{aligned}$$

where the ξ_i are bounded by 0 and 1 using the method given above:

$$\xi_i = \frac{1}{2}(\sin\psi_i + 1)$$

$\mathcal{L}$ is then maximized with respect to the $k-1$ parameters ψ_i. A drawback of this method is that the symmetry of the problem with respect to the parameters is lost.

In general, the above simple methods may be difficult to apply. One then turns to the method of Lagrangian multipliers. Given the likelihood function $\mathcal{L}(\underline{x};\underline{\theta})$ and the constraints $\underline{g}(\underline{\theta}) = 0$, one finds the extremum of

$$F(\underline{x};\underline{\theta},\underline{\alpha}) = \ln\mathcal{L}(\underline{x};\underline{\theta}) + \underline{\alpha}^\mathrm{T}\underline{g}(\underline{\theta}) \tag{8.94}$$

with respect to $\underline{\theta}$ and $\underline{\alpha}$. The likelihood condition (equation 8.58) becomes

$$\left.\frac{\partial F}{\partial\theta_i}\right|_{\substack{\underline{\theta}=\hat{\underline{\theta}}\\ \underline{\alpha}=\hat{\underline{\alpha}}}} = \left.\frac{\partial\ell}{\partial\theta_i}\right|_{\underline{\theta}=\hat{\underline{\theta}}} + \hat{\underline{\alpha}}^\mathrm{T}\left.\frac{\partial\underline{g}(\underline{\theta})}{\partial\theta_i}\right|_{\underline{\theta}=\hat{\underline{\theta}}} = 0 \tag{8.95}$$

$$\left.\frac{\partial F}{\partial \alpha_j}\right|_{\substack{\underline{\theta}=\hat{\underline{\theta}} \\ \underline{\alpha}=\hat{\underline{\alpha}}}} = \underline{g}(\hat{\underline{\theta}}) = 0 \tag{8.96}$$

The estimators of $\underline{\theta}$ found in this way clearly satisfy the constraints (equation 8.93). They also have all the usual properties of maximum likelihood estimators.

To find the variances, we construct the matrix of the negative of the second derivatives:

$$\underline{I} \equiv -E \begin{pmatrix} \dfrac{\partial^2 F}{\partial \underline{\theta} \partial \underline{\theta}'} & \dfrac{\partial^2 F}{\partial \underline{\theta} \partial \underline{\alpha}} \\ \left(\dfrac{\partial^2 F}{\partial \underline{\theta} \partial \underline{\alpha}}\right)^{\mathrm{T}} & \dfrac{\partial^2 F}{\partial \underline{\alpha}^2} \end{pmatrix} = -E \begin{pmatrix} \dfrac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'} & \dfrac{\partial \underline{g}}{\partial \underline{\theta}} \\ \left(\dfrac{\partial \underline{g}}{\partial \underline{\theta}}\right)^{\mathrm{T}} & \underline{0} \end{pmatrix} \equiv \begin{pmatrix} \underline{A} & \underline{B} \\ \underline{B}^{\mathrm{T}} & \underline{0} \end{pmatrix} \tag{8.97}$$

It can be shown[4,5] that the covariance matrix of the estimators is then given by

$$\underline{V}\left[\hat{\underline{\theta}}\right] = \underline{A}^{-1} - \underline{A}^{-1} \underline{B}\, \underline{V}[\hat{\underline{\alpha}}]\, \underline{B}^{\mathrm{T}} \underline{A}^{-1} \tag{8.98}$$

$$\underline{V}[\hat{\underline{\alpha}}] = \left(\underline{B}^{\mathrm{T}} \underline{A}^{-1} \underline{B}\right)^{-1} \tag{8.99}$$

The first term of $\underline{V}\left[\hat{\underline{\theta}}\right]$ is the ordinary unconstrained covariance matrix; the second term is the reduction in variance due to the additional information provided by the constraints. We have implicitly assumed that $\underline{I}$ is not singular. This may not be the case, *e.g.*, when the constraint is necessary to define the parameters unambiguously. One then adds another term to F,

$$F' = F - \underline{g}^2(\underline{\theta})$$

and proceeds as above. The resulting inverse covariance matrix is usually non-singular.[4,5]

Computer programs which search for a maximum will generally perform better if the constraints are handled correctly, rather than by some trick such as setting the likelihood very small when the constraint is not satisfied, since this will adversely affect the program's estimation of derivatives. Also, use of Lagrangian multipliers may not work with some programs, since the extremum can be a saddle point rather than a maximum: a maximum with respect to θ, but a minimum with respect to α. In such a case, "hill-climbing" methods will not be capable of finding the extremum.

8.5 Least Squares method

8.5.1 Introduction

We begin this subject by starting from maximum likelihood and treating the example of n independent x_i, each distributed normally with the same mean but different

σ_i. To estimate μ when all the σ_i are known we have seen that the likelihood function is

$$\mathcal{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma_i}\right)^2\right]$$
$$\ell = -\frac{n}{2}\ln(2\pi) + \sum_{i=1}^{n}\left[-\ln\sigma_i - \frac{(x_i-\mu)^2}{2\sigma_i^2}\right]$$

To maximize $\mathcal{L}$, or ℓ, is equivalent to minimizing $\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma_i^2}$. If μ were known, this quantity would be, assuming each point independent, a $\chi^2(n)$. Since μ is unknown we replace it by an estimate of μ, $\hat{\mu}$. There is then one relationship between the terms of the χ^2 and therefore

$$\chi^2 = \sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\sigma_i^2} \tag{8.100}$$

is a χ^2 not of n, but of $n-1$ degrees of freedom.

The method of least squares takes as the estimator of a parameter that value which minimizes χ^2. The least squares estimator is thus given by

$$\left.\frac{\partial\chi^2}{\partial\mu}\right|_{\mu=\hat{\mu}} = -2\sum_{i=1}^{n}\frac{x_i-\hat{\mu}}{\sigma_i^2} = 0$$

which gives the same estimator as did maximum likelihood (equation 8.59):

$$\hat{\mu} = \frac{\sum\frac{x_i}{\sigma_i^2}}{\sum\frac{1}{\sigma_i^2}} \tag{8.101}$$

Although in this example the least squares and maximum likelihood methods result in the same estimator, this is not true in general, in particular if the p.d.f. is not normal. We will see that although we arrived at the least squares method starting from maximum likelihood, least squares is much more solidly based than maximum likelihood. It is, perhaps as a consequence, also less widely applicable.

The method of least squares is a special case of a more general class of methods whereby one uses some measure of distance, $d_i(x_i,\theta)$, of a data point from its expected value and minimizes the sum of the distances to obtain the estimate of θ. Examples of d, in the context of our example, are

1. $d_i(x_i,\theta) = |x_i - \hat{\mu}|^\alpha$
2. $d_i(x_i,\theta) = \left(\frac{|x_i-\hat{\mu}|}{\sigma_i}\right)^\alpha$

The difference between these two is that in the second case the distance is scaled by the square root of the expected variance of the distance. If all these variances, σ_i^2,

are the same, the two definitions are equivalent. It can be shown[11,13] that the first distance measure with $\alpha = 1$ leads to $\hat{\mu}$ given by the sample median. The second distance measure with $\alpha = 2$ is just χ^2.

The first publication in which least squares was used is by Legendre. In an 1805 paper entitled "Nouvelles méthodes pour la determination des orbites des comètes" he writes:

> Il faut ensuite, lorsque toutes les conditions du problême sont exprimées convenablement, determiner les coëfficiens de manière à rendre les erreurs les plus petites qu'il est possible. Pour cet effet, la méthode qui me paraît la plus simple et la plus générale, consiste à rendre minimum la somme des quarrés des erreurs.

Least squares was not the only method in use in those days (or now). In 1792 Laplace minimized the sum of absolute errors, although he later switched to least squares. Bessel and Encke also used least squares. In 1831, Cauchy suggested, "que la plus grande de toutes les erreurs, abstraction faite du signe, devienne un minimum", *i.e.*, to minimize the maximum of the absolute values of the deviations, $\max|x_i - \hat{\mu}|$. This 'minimax' principle gives a very robust estimation but is not very efficient.[4,5]

We have noted that the χ^2 of equation 8.100 is a χ^2 of $n-1$ degrees of freedom. Thus, if we were to repeat the identical experiment many times, the values of χ^2 obtained would be distributed as $\chi^2(n-1)$, provided that the assumed p.d.f. of the x_i is correct. We would not expect then to get a value of χ^2 which would be very improbable if the assumed p.d.f. were correct. This could provide a reason for deciding that the assumed p.d.f. is incorrect. This built-in test of the validity of the assumed p.d.f. is a feature which was missing in the maximum likelihood method. We will return to this and other hypothesis tests in sections 10.4 and 10.6.

In the example we assumed that the x_i were normally distributed about μ with standard deviation σ_i. If this were not the case, the distribution of the quantity χ^2 for repetitions of the experiment would not follow the expected $\chi^2(n-1)$ distribution. Consequently, the chance of getting a particular value of χ^2 would not be that given by the χ^2 distribution. In other words, the quantity that we have called χ^2 is a χ^2 r.v. only if our assumption that the x_i are distributed normally is correct.

Assuming that we have not rejected the p.d.f., we need to estimate the variance of $\hat{\mu}$. First we can use error propagation (section 8.3.6) to calculate the variance of $\hat{\mu}$, given by equation 8.101, from the variances of the x_i. In our example $\hat{\mu}$ is linear in the x_i; hence the method is exact (equation 8.49):

$$V[\hat{\mu}] = \sum \left(\frac{\partial \hat{\mu}}{\partial x_i}\right)^2 V[x_i] = \left(\frac{1}{\sum(1/\sigma_i)^2}\right)^2 \sum \frac{V[x_i]}{\sigma_i^4} = \frac{1}{\sum \frac{1}{\sigma_i^2}}$$

which agrees with the variance found in the maximum likelihood method (equation 8.60).

We see that in this example (although not in general true) the variance of the estimator does not depend on the value of χ^2. However, it does depend on the shape of $\chi^2(\mu)$:

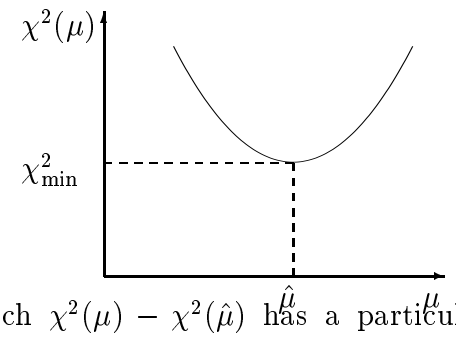
$$\chi^2(\mu) = \sum \left(\frac{x_i - \mu}{\sigma_i}\right)^2$$

$$\left.\frac{\partial \chi^2}{\partial \mu}\right|_{\hat{\mu}} = -\sum \frac{2(x_i - \hat{\mu})}{\sigma_i^2} = 0$$

$$\left.\frac{\partial^2 \chi^2}{\partial \mu^2}\right|_{\hat{\mu}} = 2\sum \frac{1}{\sigma_i^2} = \frac{2}{V[\hat{\mu}]}$$

All higher order derivatives are zero, a consequence of the efficiency of the estimator and the linear relationship between χ^2 and ℓ. Thus the χ^2 is a parabola:

$$\chi^2(\mu) = \chi^2(\hat{\mu}) + \frac{(\hat{\mu} - \mu)^2}{V[\hat{\mu}]}$$



Corresponding to what we did in the maximum likelihood method, we construct the error on $\hat{\mu}$ by finding that value of μ for which $\chi^2(\mu) - \chi^2(\hat{\mu})$ has a particular value. From the above equation we see that a χ^2-difference of 1 occurs when $(\hat{\mu} - \mu)^2 = V[\hat{\mu}]$, *i.e.*, for those values of μ which are one standard deviation from $\hat{\mu}$, or more generally a value of μ for which $\chi^2(\mu) = \chi^2(\hat{\mu}) = n^2$ corresponds to an n standard deviation difference from $\hat{\mu}$.

8.5.2 The Linear Model

In the preceding example we had a number of measurements of a fixed quantity. Now let us suppose that we have a number of measurements y_i of a quantity y which depends on some other quantity x. Assume, for now, that the values x_i are known exactly, *i.e.*, without error. For each x_i, y is measured to be y_i with expected error σ_i. We assume that σ_i does not depend on y_i.

One of the reasons for doing a fit to a curve is to enable us to predict the most likely value of future measurements at a specified x. For example, we wish to calibrate an instrument. Then the predictor variable x would be the value that the instrument reads. The response variable y would be the true value. A fit averages out the fluctuations in the individual readings as much as possible. This only works, of course, if the form used for the curve in the fit is at least approximately correct. Although we will use a one-dimensional predictor variable x, the generalization to more dimensions is straightforward: $x \to \underline{x}$.

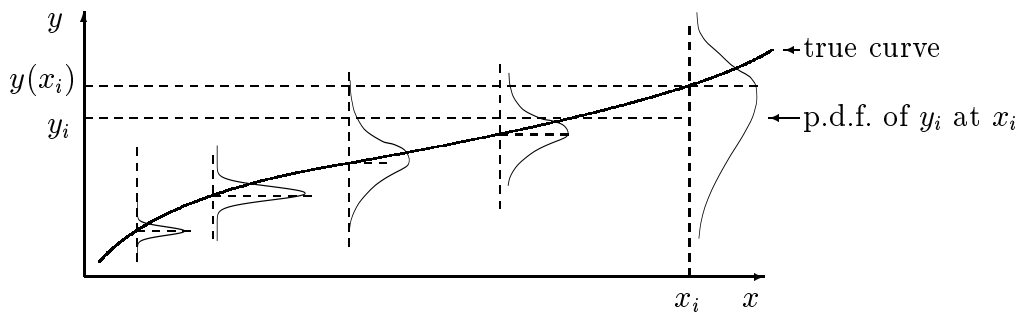
Assume now that we have a model for y *vs.* x in terms of certain parameters θ which are coefficients of known functions of x:

$$y(x) = \theta_1 h_1(x) + \theta_2 h_2(x) + \theta_3 h_3(x) + \ldots + \theta_j h_j(x) \tag{8.102}$$

This is the curve which we fit to the data. There are k parameters, θ_j, to be estimated. The important features of this model are that the h_j are known, distinguishable, functions of x, single-valued over the range of x, and that y is linear in the θ_j. The word 'linear' in the term 'linear model' thus refers to the parameters θ_j and not to the variable x. In some cases the linear model is just an approximation arrived at by retaining only the first few terms of a Taylor series. The functions h_j must be distinguishable, *i.e.*, no h_j may be expressible as a linear combination of the other h_j; otherwise the corresponding θ_j will be indeterminate.

We want to determine the values of the θ_j for which the model (eq. 8.102) best fits the measurements. We assume that any deviation of a point y_i from this curve is due to measurement error or some other unbiased effects beyond our control, but whose distribution is known from previous study of the measuring process to have variance σ_i^2. It need not be a Gaussian. We take as our measure of the distance of the point y_i from the hypothesized curve the squared distance in units of σ_i, as in our example above.

The general term for this fitting procedure is 'Regression Analysis'. This term is of historical origin and like many such terms it is not particularly appropriate; nothing regresses. The term is not much used in physics, where we prefer to speak of least squares fits, but is still in common use in the social sciences and in statistics books. Some authors make a distinction between regression analysis and least squares, reserving the term regression for the case where the y_i (and perhaps the x_i) are means (or other descriptive statistics) of some random variable, *e.g.*, y the average height and x the average weight of Dutch male university students. The mathematics is, however, the same.



We assume that the actual measurements are described by

$$y_i = y(x_i) + \epsilon_i = \sum_{j=1}^{k} \theta_j h_j(x_i) + \epsilon_i \tag{8.103}$$

where the unknown error on y_i has the properties: $E[\epsilon_i] = 0$, $V[\epsilon_i] = \sigma_i^2$, and σ_i^2 is known. The ϵ_i do not have to be normally distributed for most of what we shall do; where a Gaussian assumption is needed, we will say so. Note that if at each x_i the y_i does not have a normal p.d.f., we may be able to transform to a set of variables which does.

Further, we assume for simplicity that each y_i is an independent measurement, although correlations can easily be taken into account by making the error matrix non-diagonal, as will be discussed. The x_i may be chosen any way we wish, including several x_i which are equal. However, we shall see that we need at least k distinct values of x to determine k parameters θ_j.

Estimator

The problem is now to determine the 'best' values of k parameters, θ_j, from n measurements, (x_i, y_i). The deviations from the true curve are ϵ_i. Therefore the "χ^2" is

$$Q^2 = \sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_i^2} \tag{8.104}$$

$$= \sum_{i=1}^{n} \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2$$

$$= \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(y_i - \sum_{j=1}^{k} \theta_j h_j(x_i) \right)^2 \tag{8.105}$$

This is a true χ^2, *i.e.*, distributed as a χ^2 p.d.f., only if the ϵ_i are normally distributed. To emphasize this we use the symbol Q^2 instead of χ^2.

We do not know the actual value of Q^2, since we do not know the true values of the parameters θ_j. The least squares method estimates $\underline{\theta}$ by that value $\hat{\underline{\theta}}$ which minimizes Q^2. This is found from the k equations $(l = 1, \ldots, k)$

$$\frac{\partial Q^2}{\partial \theta_l} = 2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(y_i - \sum_{j=1}^{k} \theta_j h_j(x_i) \right) (-h_l(x_i)) = 0$$

which we rewrite as

$$\sum_{i=1}^{n} \frac{h_l(x_i)}{\sigma_i^2} \sum_{j=1}^{k} \hat{\theta}_j h_j(x_i) = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} h_l(x_i) \tag{8.106}$$

This is a set of k linear equations in k unknowns. They are called the normal equations. It is easier to work in matrix notation. We write

$$\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad ; \quad \underline{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix} \quad ; \quad \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\underline{H} = \begin{pmatrix} h_1(x_1) & h_2(x_1) & \ldots & h_k(x_1) \\ h_1(x_2) & h_2(x_2) & \ldots & h_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x_n) & h_2(x_n) & \ldots & h_k(x_n) \end{pmatrix}$$

Then

$$\underline{H}\,\underline{\theta} = \begin{pmatrix} \sum_{j=1}^{k} \theta_j h_j(x_1) \\ \sum_{j=1}^{k} \theta_j h_j(x_2) \\ \vdots \\ \sum_{j=1}^{k} \theta_j h_j(x_2) \end{pmatrix}$$

and the model (eq. 8.103) can be rewritten

$$\underline{y} = \underline{H}\,\underline{\theta} + \underline{\epsilon} \tag{8.107}$$

Since $E[\underline{\epsilon}] = 0$, we obtain $E\left[\underline{y}\right] = \underline{H}\,\underline{\theta}$. In other words, the expectation value of each measurement is exactly the value given by the model.

The errors σ_i^2 can also be incorporated in a matrix, which is diagonal given our assumption of independent measurements,

$$\underline{V}[\underline{y}] = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{pmatrix}$$

If the measurements are not independent, we incorporate that by setting the off-diagonal elements to the covariances of the measurements. In this matrix notation, the equations for Q^2 (equations 8.104 and 8.105) become

$$Q^2 = \underline{\epsilon}^{\mathrm{T}} \underline{V}^{-1} \underline{\epsilon} \tag{8.108}$$

$$= \left(\underline{y} - \underline{H}\,\underline{\theta}\right)^{\mathrm{T}} \underline{V}^{-1} \left(\underline{y} - \underline{H}\,\underline{\theta}\right) \tag{8.109}$$

To find the estimates of $\underline{\theta}$ we solve

$$\frac{\partial Q^2}{\partial \underline{\theta}} = -2\,\underline{H}^{\mathrm{T}} \underline{V}^{-1} \left(\underline{y} - \underline{H}\,\underline{\theta}\right) = 0 \tag{8.110}$$

which gives the normal equations corresponding to equations 8.106, but now in matrix form:

$$\underset{(k\times n)}{\underline{H}^{\mathrm{T}}} \; \underset{(n\times n)}{\underline{V}^{-1}} \; \underset{(n\times k)}{\underline{H}} \; \underset{(k\times 1)}{\hat{\underline{\theta}}} = \underset{(k\times n)}{\underline{H}^{\mathrm{T}}} \; \underset{(n\times n)}{\underline{V}^{-1}} \; \underset{(n\times 1)}{\underline{y}} \tag{8.111}$$

where we have indicated the dimension of the matrices. The normal equations are solved by inverting the square matrix $\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}$, which is a symmetric matrix since $\underline{V}$ is symmetric. The solution is then

$$\hat{\underline{\theta}} = \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1} \underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{y} \tag{8.112}$$

It is useful to note that the actual sizes of the errors σ_i^2 do not have to be known to find $\hat{\underline{\theta}}$; only their relative sizes. To see this, write $\underline{V} = \sigma^2\underline{W}$, where σ^2 is an arbitrary scale factor and insert this in equation 8.112. The factors σ^2 cancel; thus σ^2 need not be known in order to determine $\hat{\underline{\theta}}$.

Now let us evaluate the expectation of $\underline{\hat{\theta}}$:

$$E\left[\underline{\hat{\theta}}\right] = E\left[\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{y}\right] = \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}E\left[\underline{y}\right]$$
$$= \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)\underline{\theta} = \underline{\theta}$$

Thus $\underline{\hat{\theta}}$ is unbiased, assuming that the model is correct. This is true even for small n. (Recall that maximum likelihood estimators are often biased for finite n.)

Procedures exist for solving the normal equations without the intermediate step of matrix inversion. Such methods are usually preferable in that they usually suffer less from round-off problems.

In some cases, it is more convenient to solve these equations by numerical approximation methods. As discussed at the end of section 8.4.6, programs exist to find the minimum of a function. For simple cases like the linear problem we have considered, use of such programs is not very wasteful of computer time, and its simplicity decreases the probability of an experimenter's error and probably saves his time as well. If the problem is not linear, a case which we shall shortly discuss, such an approach is usually best.

We have stated that there must be no linear relationship between the h_j. If there is, then the columns of $\underline{H}$ are not all independent, and since $\underline{V}$ is symmetric, $\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}$ will be singular. The best approach is then to eliminate some of the h's until the linear relationships no longer exist. Also, there must be at least k distinct x_i; otherwise the same matrix will be singular.

Note that if the number of parameters k is equal to the number of distinct values of x, *i.e.*, $n = k$ assuming all x_i are distinct, then

$$\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1} = \underline{H}^{-1}\underline{V}\left(\underline{H}^{\mathrm{T}}\right)^{-1}$$

Substituting in equation 8.112 yields $\underline{\hat{\theta}} = \underline{H}^{-1}\underline{y}$, assuming that $\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}$ is not singular. Thus $\hat{\theta}$ is independent of the errors. The curve will pass through all the points, if that is possible. It may not be possible; the assumed model may not be correct.

Variance

The covariance matrix of the estimators is given by

$$\underset{(k\times k)}{\underline{V}\left[\underline{\hat{\theta}}\right]} = \underbrace{\left[\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}\right]}_{(k\times n)}\ \underset{(n\times n)}{\underline{V}\left[\underline{y}\right]}\ \underbrace{\left[\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}\right]^{\mathrm{T}}}_{(n\times k)} \tag{8.113}$$

This can be demonstrated by working out a simple example. Alternatively, it follows from propagation of errors (section 8.3.6): Since we are converting from errors on

$\underline{y}$ to errors on $\hat{\underline{\theta}}$, the matrix $\underline{D}$ (equation 8.53) is

$$\underline{D}(\hat{\underline{\theta}}) = \begin{pmatrix} \frac{\partial\hat{\theta}_1}{\partial y_1} & \frac{\partial\hat{\theta}_2}{\partial y_1} & \cdots & \frac{\partial\hat{\theta}_k}{\partial y_1} \\ \frac{\partial\hat{\theta}_1}{\partial y_2} & \frac{\partial\hat{\theta}_2}{\partial y_2} & \cdots & \frac{\partial\hat{\theta}_k}{\partial y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\hat{\theta}_1}{\partial y_n} & \frac{\partial\hat{\theta}_2}{\partial y_n} & \cdots & \frac{\partial\hat{\theta}_k}{\partial y_n} \end{pmatrix}$$

The elements of $\underline{D}$ are found by differentiating equation 8.112, which gives

$$D_{ij}^{\mathrm{T}} = D_{ji} = \frac{\partial\hat{\theta}_i}{\partial y_j} = \left[\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}\right]_{ij} \tag{8.114}$$

or

$$\underline{D} = \left[\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\,\underline{V}^{-1}\right]^{\mathrm{T}} \tag{8.115}$$

The covariance (equation 8.113) then follows from equation 8.52, $V\left[\hat{\underline{\theta}}\right] = \underline{D}^{\mathrm{T}}V\left[\underline{y}\right]\underline{D}$.

What we here call $\underline{V}\left[\underline{y}\right]$ is what we previously just called $\underline{V}$. It is a square, symmetric matrix. Hence $\underline{V}^{-1}$ is also square and symmetric and therefore $(\underline{V}^{-1})^{\mathrm{T}} = \underline{V}^{-1}$. For the same reason $\left[\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\right]^{\mathrm{T}} = \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}$. Therefore, equation 8.113 can be rewritten:

$$\begin{aligned} \underline{V}\left[\hat{\underline{\theta}}\right] &= \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{V}\underline{V}^{-1}\underline{H}\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1} \\ &= \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1} \\ \underline{V}\left[\hat{\underline{\theta}}\right] &= \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1} \end{aligned} \tag{8.116}$$

Equation 8.112 for the estimator $\hat{\underline{\theta}}$ and equation 8.116 for its variance constitute the complete method of linear least squares.

σ^2 unknown

If $\underline{V}(\underline{y})$ is only known up to an overall constant, *i.e.*, $\underline{V} = \sigma^2\underline{W}$ with σ^2 unknown, it can be estimated from the minimum value of Q^2: Defining Q^2 in terms of $\underline{W}$, its minimum value is given by equation 8.108 with $\theta = \hat{\theta}$:

$$Q^2_{\min} = \left(\underline{y} - \underline{H}\,\hat{\underline{\theta}}\right)^{\mathrm{T}}\underline{W}^{-1}\left(\underline{y} - \underline{H}\,\hat{\underline{\theta}}\right) \tag{8.117}$$

If the ϵ_i are normally distributed, $Q^2 = \sigma^2\chi^2$ where the χ^2 has $n-k$ degrees of freedom. The expectation of Q^2 is then

$$E\left[Q^2\right] = E\left[\sigma^2\chi^2\right] = \sigma^2(n-k)$$

Therefore,

$$\widehat{\sigma^2} = \frac{Q^2_{\min}}{n-k} \tag{8.118}$$

is an unbiased estimate of σ^2. It can be shown* that this result is true even when the ϵ_i are not normally distributed.

Interpolation

Having found $\underline{\hat{\theta}}$, we may wish to calculate the value of y for some particular value of x. In fact, the reason for doing the fit is often to be able to interpolate or extrapolate the data points to other values of x. This is done by substituting the estimators in the model. The variance is found by error propagation, reversing the procedure used above to find the variance of $\underline{\hat{\theta}}$. The estimate $\hat{y}_0$ of y at $x = x_0$ and its variance are therefore given by

$$\hat{y}_0 = \underline{H}_0 \underline{\hat{\theta}} \tag{8.119}$$

$$V[\hat{y}_0] = \underline{H}_0\, \underline{V}(\underline{\hat{\theta}})\, \underline{H}_0^{\mathrm{T}} = \underline{H}_0 \left(\underline{H}^{\mathrm{T}}\, \underline{V}^{-1} \left[\underline{y}\right] \underline{H} \right)^{-1} \underline{H}_0^{\mathrm{T}} \tag{8.120}$$

where $\underline{H}_0 = (\, h_1(x_0) \quad h_2(x_0) \quad \ldots \quad h_k(x_0)\,)$, *i.e.*, the H-matrix for the single point x_0.

8.5.3 Derivative formulation

We can derive the above results in another way. The covariance matrix can be found from the derivatives of Q^2: Starting from equation 8.109,

$$\left.\frac{\partial Q^2}{\partial \underline{\theta}}\right|_{\underline{\theta}=\underline{\hat{\theta}}} = -2\, \underline{H}^{\mathrm{T}} \underline{V}^{-1} \left(\underline{y} - \underline{H}\, \underline{\hat{\theta}} \right) \tag{8.121}$$

$$\left.\frac{\partial^2 Q^2}{\partial \underline{\theta}^2}\right|_{\underline{\theta}=\underline{\hat{\theta}}} = +2\, \underline{H}^{\mathrm{T}} \underline{V}^{-1} \underline{H} = 2\, \underline{V}^{-1} \left[\underline{\hat{\theta}}\right] \tag{8.122}$$

This is a very useful way to calculate the covariance, which we have already seen in our simple example of repeated measurements of a fixed quantity in the introduction (section 8.5.1).

In fact, the solution $\underline{\hat{\theta}}$ can be written in terms of the derivatives of Q^2 making it unnecessary to construct $\underline{H}$, $\underline{V}$, and the associated matrix products. To see this we substitute the second derivative, equation 8.122, in equation 8.112. Since we are trying to find $\underline{\hat{\theta}}$, we do not yet know it, and we can not evaluate the derivative at $\underline{\theta} = \underline{\hat{\theta}}$. We therefore evaluate it at some guessed value, $\underline{\theta}_0$. Thus,

$$\underline{\hat{\theta}} = 2 \left(\left.\frac{\partial^2 Q^2}{\partial \underline{\theta}^2}\right|_{\underline{\theta}=\underline{\theta}_0} \right)^{-1} \underline{H}^{\mathrm{T}} \underline{V}^{-1} \underline{y}$$

*See Kendall & Stuart[11], vol. II, section 19.9 and exercise 19.5.

$$= \left(\frac{\partial^2 Q^2}{\partial \underline{\theta}^2} \bigg|_{\underline{\theta}=\underline{\theta}_0} \right)^{-1} \left[\frac{\partial^2 Q^2}{\partial \underline{\theta}^2} \bigg|_{\underline{\theta}=\underline{\theta}_0} \cdot \underline{\theta}_0 - \frac{\partial Q^2}{\partial \underline{\theta}} \bigg|_{\underline{\theta}=\underline{\theta}_0} \right]$$

$$= \underline{\theta}_0 - \left(\frac{\partial^2 Q^2}{\partial \underline{\theta}^2} \bigg|_{\underline{\theta}=\underline{\theta}_0} \right)^{-1} \cdot \frac{\partial Q^2}{\partial \underline{\theta}} \bigg|_{\underline{\theta}=\underline{\theta}_0} \tag{8.123}$$

This is the Newton-Raphson method of solving the equations $\frac{\partial Q^2}{\partial \theta} = 0$. It is exact, *i.e.*, independent of the choice of $\underline{\theta}_0$ for the linear model where the form of Q^2 is a parabola. In the non-linear case, the method can still be used, but iteratively; its success will depend on how close $\underline{\theta}_0$ is to $\hat{\underline{\theta}}$ and on how non-linear the problem is.

The derivative formulation for the least squares solution is frequently the most convenient technique in practical problems. The derivatives we need are

$$\frac{\partial Q^2}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \sum_m \frac{\epsilon_m^2}{\sigma_m^2} = 2 \sum_m \frac{\epsilon_m}{\sigma_m^2} \frac{\partial \epsilon_m}{\partial \theta_i}$$

and

$$\frac{\partial^2 Q^2}{\partial \theta_i \partial \theta_j} = 2 \sum_m \frac{1}{\sigma_m^2} \frac{\partial \epsilon_m}{\partial \theta_i} \frac{\partial \epsilon_m}{\partial \theta_j} + 2 \sum_m \frac{\epsilon_m}{\sigma_m^2} \frac{\partial^2 \epsilon_m}{\partial \theta_i \partial \theta_j}$$

In the linear case, $\frac{\partial^2 \epsilon_m}{\partial \theta_i \partial \theta_j} = 0$, and $\frac{\partial \epsilon_m}{\partial \theta_i} = -h_i(x_m)$. Thus, the necessary derivatives are easy to compute.

Finally, we note that the minimum value of Q^2 is given by

$$Q^2(\hat{\underline{\theta}}) = Q^2(\underline{\theta}_0) + \frac{\partial Q^2}{\partial \underline{\theta}} \bigg|_{\underline{\theta}=\underline{\theta}_0} \cdot (\hat{\underline{\theta}} - \underline{\theta}_0) + \frac{1}{2} (\hat{\underline{\theta}} - \underline{\theta}_0)^{\mathrm{T}} \frac{\partial^2 Q^2}{\partial \underline{\theta}^2} \bigg|_{\underline{\theta}=\underline{\theta}_0} (\hat{\underline{\theta}} - \underline{\theta}_0) \tag{8.124}$$

where we have expanded $Q^2(\hat{\theta})$ about $\underline{\theta}_0$. Third and higher order terms are zero for the linear model.

Just as in the example in the introduction to least squares, we can show, by expanding Q^2 about $\hat{\underline{\theta}}$ that the set of values of $\underline{\theta}$ given by $Q^2(\underline{\theta}) = Q^2_{\min} + 1$ define the one standard deviation errors on $\hat{\underline{\theta}}$. This is the same as the geometrical method to find the errors in maximum likelihood analysis (section 8.4.5), except that here the difference in Q^2 is 1 whereas the difference in ℓ was $1/2$. This is because the covariance matrix here is given by twice the inverse of the second derivative matrix, whereas it was equal to the inverse of the second derivative matrix in the maximum likelihood case.

So far we have made no use of the assumption that the ϵ_i are Gaussian distributed. We have only used the conditions $E(\epsilon_i) = 0$ and $V[\epsilon_i] = \sigma_i^2$ known and the linearity of the model.

8.5.4 Gauss-Markov Theorem

This is the theorem which provides the method of least squares with its firm mathematical foundation. In 1812 Laplace showed that the method of least squares gives

unbiased estimates, irrespective of the parent distribution. Nine years later Gauss proved that among the class of estimators which are both linear combinations of the data and unbiased estimators of the parameters, the method of least squares gives estimates having the least possible variance. This was treated more generally by Markov in 1912. It was extended in 1934 by Aitken to the case where the observations are correlated and have different variances.

We will simply state the theorem without proof:* If $E[\epsilon_i] = 0$ and the covariance matrix of the ϵ_i, $\underline{V}[\underline{\epsilon}]$ is finite and fixed, *i.e.*, independent of $\underline{\theta}$ and $\underline{y}$, (it does not have to be diagonal), then the least squares estimate, $\hat{\underline{\theta}}$ is unbiased and has the smallest variance of all *linear* (in $\underline{y}$), unbiased estimates, regardless of the p.d.f. for the ϵ_i.

Note that

- This theorem concerns only linear unbiased estimators. It may be possible, particularly if $\underline{\epsilon}$ is not normally distributed, to find a non-linear unbiased estimator with a smaller variance. Biased estimators with a smaller variance may also exist.

- Least squares does not in general give the same result as maximum likelihood (unless the ϵ_i are Gaussian) even for linear models. In this case, linear least squares is often to be preferred to linear maximum likelihood where applicable and convenient, since linear least squares is unbiased and has smallest variance. An exception may occur in small samples where the data must be binned in order to do a least squares analysis, causing a loss of information.

- The assumptions are important: The measurement errors must have zero mean and they must be homoscedastic (the technical name for constant variance). Non-zero means or heteroscedastic variances may reveal themselves in the residuals, $y_i - f(x_i)$, *cf.* section 10.6.8.

8.5.5 Examples

A Straight-Line Fit

As an example of linear least squares we do a least squares fit of independent measurements y_i at points x_i assuming the model $y = a + bx$. Thus,

$$\underline{\theta} = \begin{pmatrix} a \\ b \end{pmatrix} \quad ; \quad \underline{h} = \begin{pmatrix} 1 \\ x \end{pmatrix} \quad ; \quad \underline{H} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

and

$$\underline{y} = \underline{H}\,\underline{\theta} + \underline{\epsilon}$$

*For a proof, see for example, Kendall & Stuart[11], chapter 19 (Stuart *et al.*[13], chapter 29), or Eadie *et al.*[4] (or James[5]).

Since the measurements are independent, the covariance matrix is diagonal with

$$V_{ii}(\underline{y}) = V_{ii}(\underline{\epsilon}) = \sigma_i^2 \qquad \text{and} \qquad Q^2 = \underline{\epsilon}^{\mathrm{T}} \underline{V}^{-1} \underline{\epsilon} = \sum_{i=1}^{n} \frac{\epsilon_i^2}{\sigma_i^2} = \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

Hence, using the derivative method,

$$\frac{\partial Q^2}{\partial a} = 0 \qquad \longrightarrow \qquad \hat{a} = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \sum_{i=1}^{n} \frac{y_i - \hat{b}x_i}{\sigma_i^2}$$

$$\frac{\partial Q^2}{\partial b} = 0 \qquad \longrightarrow \qquad \hat{b} = \frac{1}{\sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}} \sum_{i=1}^{n} \frac{x_i y_i - \hat{a}x_i}{\sigma_i^2}$$

Solving, we find

$$\hat{b} = \frac{\left(\sum \frac{x_i y_i}{\sigma_i^2}\right)\left(\sum \frac{1}{\sigma_i^2}\right) - \left(\sum \frac{y_i}{\sigma_i^2}\right)\left(\sum \frac{x_i}{\sigma_i^2}\right)}{\left(\sum \frac{x_i^2}{\sigma_i^2}\right)\left(\sum \frac{1}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

which can in turn be substituted in the expression for $\hat{a}$.

Alternatively, we can solve the matrix equation,

$$\underline{\hat{\theta}} = \left(\underline{H}^{\mathrm{T}} \underline{V}^{-1} \underline{H}\right)^{-1} \underline{H}^{\mathrm{T}} \underline{V}^{-1} \underline{y}$$

which, of course, gives the same result.

Note that if all σ_i are the same, $\sigma_i = \sigma$, then

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \text{and} \qquad \hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \tag{8.125}$$

These are the formulae which are programmed into many pocket calculators. As such, they should only be used when the σ_i are all the same. These formulae are, however, also applicable to the case where not all σ_i are the same if the sample average indicated by the bar is interpreted as meaning a weighted sample average with weights given by $1/\sigma_i^2$, *e.g.*, $\bar{y} = \frac{\sum y_i/\sigma_i^2}{\sum 1/\sigma_i^2}$. The proof is left as an exercise (ex. 40).

Note that at least two of the x_i must be different. Otherwise, the denominator in the expression for $\hat{b}$ is zero. This illustrates the general requirement that there must be at least as many distinct values of x_i as there are parameters in the model; otherwise the matrix $\underline{H}^{\mathrm{T}} \underline{V}^{-1} \underline{H}$ will be singular.

The errors on the least squares estimates of the parameters are given by equation 8.122 or 8.116. With all σ_i the same, equation 8.116 gives

$$\begin{aligned}
\underline{V}\left[\underline{\hat{\theta}}\right] &= \left(\underline{H}^{\mathrm{T}} \underline{V}^{-1} \underline{H}\right)^{-1} = \left(\underline{H}^{\mathrm{T}} \underline{H}\right)^{-1} \sigma^2 \\
&= \sigma^2 \left[\begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \right]^{-1} = \sigma^2 \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \\
&= \frac{\sigma^2}{n \sum x_i^2 - \left(\sum x_i\right)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} = \frac{\sigma^2}{n \sum \left(x_i - \bar{x}\right)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}
\end{aligned}$$

Thus,

$$\begin{pmatrix} V[\hat{a}] & \operatorname{cov}(\hat{a},\hat{b}) \\ \operatorname{cov}(\hat{a},\hat{b}) & V\left[\hat{b}\right] \end{pmatrix} = \frac{\sigma^2}{n\left(\overline{x^2}-\bar{x}^2\right)} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \tag{8.126}$$

Note that by translating the x-axis such that $\bar{x}$ becomes zero, the estimates of the parameters become uncorrelated.

Here too, it is possible to use this formula for the case where not all σ_i are the same. Besides taking the bar as a weighted average, one must also replace σ^2 by its weighted average,

$$\overline{\sigma^2} = \frac{\sum \sigma_i^2/\sigma_i^2}{\sum 1/\sigma_i^2} = \frac{n}{\sum 1/\sigma_i^2} \tag{8.127}$$

Note that the errors are smallest for the largest spread in the x_i. Thus we will attain the best estimates of the parameters by making measurements only at the extreme values of x. This procedure is, however, seldom advisable since it makes it impossible to test the validity of the model, as we shall see.

Having found $\hat{a}$ and $\hat{b}$, we can calculate the value of y for any value of x by simply substituting the estimators in the model. The estimate $\hat{y}_0$ of y at $x = x_0$ is therefore given by

$$\hat{y}_0 = \hat{a} + \hat{b}x_0 \tag{8.128}$$

We note in passing that this gives $\hat{y}_0 = \bar{y}$ for $x_0 = \bar{x}$. The variance of $\hat{y}_0$ is found by error propagation:

$$V[\hat{y}_0] = V[\hat{a}] + x_0^2 V\left[\hat{b}\right] + 2x_0 \operatorname{cov}(\hat{a},\hat{b})$$

Substituting from equation 8.126 gives

$$V[\hat{y}_0] = \frac{\sigma^2}{n}\left[1 + \frac{(x_0-\bar{x})^2}{\left(\overline{x^2}-\bar{x}^2\right)}\right] \tag{8.129}$$

Thus, the closer x_0 is to $\bar{x}$, the smaller the error in $\hat{y}_0$.

A Polynomial Fit

To fit a parabola

$$y = a_0 + a_1 x + a_2 x^2$$

the matrix $\underline{H}$ is

$$\underline{H} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

Assuming that all the σ_i are equal, equation 8.112 becomes

$$\underline{\hat{\theta}} = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{pmatrix} \sum_i 1 & \sum_i x_i & \sum_i x_i^2 \\ \sum_i x_i & \sum_i x_i^2 & \sum_i x_i^3 \\ \sum_i x_i^2 & \sum_i x_i^3 & \sum_i x_i^4 \end{pmatrix}^{-1} \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \\ \sum_i x_i^2 y_i \end{pmatrix}$$

The extension to higher order polynomials is obvious. Unfortunately, there is no simple method to invert such matrices, even though the form of the matrix appears very regular and symmetric. Numerical inversion suffers from rounding errors when the order of the polynomial is greater than six or seven.

One can hope to mitigate these problems by choosing a set of orthogonal polynomials, *e.g.*, Legendre or Tchebycheff (Chebyshev) polynomials, instead of powers of x. The off-diagonal terms then involve products of orthogonal functions summed over the events. The expectation of such products is zero, and hence the sum of their products over a large number of events should be nearly zero. The matrix is then nearly diagonal and less prone to numerical problems.

Even better is to find functions which are exactly orthogonal over the measured data points, *i.e.*, functions, ξ, for which

$$\sum_{i=1}^{n} \xi_j(x_i)\xi_k(x_i)(V^{-1})_{jk} = \delta_{jk}$$

The matrix which has to be inverted, $\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}$, is then simply the unit matrix. An additional feature of such a parametrization is that the estimates of the parameters are independent; the covariance matrix for the parameters is diagonal. Such a set of functions can always be found, *e.g.*, using Schmidt's orthogonalization method* or, more simply, using Forsythe's method.[40] Its usefulness is limited to cases where we are merely seeking a parametrization of the data (for the purpose of interpolation or extrapolation) rather than seeking to estimate the parameters of a theoretical model.

8.5.6 Constraints in the linear model

If the parameters to be estimated are constrained, we can, as in the maximum likelihood case (section 8.4.8), try to write the model in terms of new parameters which are unconstrained. Alternatively, we can use the more general method of Lagrangian multipliers, which we will now discuss for least squares fits.

Suppose the model is $\underline{y} = \underline{H}\,\underline{\theta} + \underline{\epsilon}$ for n observations y_i and k parameters θ_j. The H_{ij} may take any form, *e.g.*, $H_{ij} = x_i^{j-1}$ for a polynomial fit to the observations y_i taken at points x_i as in the previous section.

Suppose that the deviations ϵ_i have covariance matrix $\underline{V}$ and that the parameters $\underline{\theta}$ are subject to m linear constraints,

$$\sum_{j=1}^{k} \ell_{ij}\theta_j = R_i \qquad , \quad i = 1, \ldots, m \tag{8.130}$$

or, in matrix notation,

$$\underline{L}\,\underline{\theta} = \underline{R} \tag{8.131}$$

*See, *e.g.*, Margenau & Murphy[39].

The least squares estimate of $\underline{\theta}$ is then found using a k-component vector of Lagrangian multipliers, $2\underline{\lambda}$, by finding the extremum of

$$Q^2 = \left(\underline{y} - \underline{H}\,\underline{\theta}\right)^{\mathrm{T}} \underline{V}^{-1} \left(\underline{y} - \underline{H}\,\underline{\theta}\right) + 2\underline{\lambda}^{\mathrm{T}} \left(\underline{L}\,\underline{\theta} - \underline{R}\right) \tag{8.132}$$

where the first term is the usual Q^2 and the second term represents the constraints. Differentiating with respect to $\underline{\theta}$ and with respect to $\underline{\lambda}$, respectively, yields the normal equations

$$\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\,\hat{\underline{\theta}} + \underline{L}^{\mathrm{T}}\hat{\underline{\lambda}} = \underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{y} \tag{8.133a}$$

$$\underline{L}\,\hat{\underline{\theta}} = \underline{R} \tag{8.133b}$$

which can be combined to give

$$\begin{pmatrix} \underline{C} & \underline{L}^{\mathrm{T}} \\ \underline{L} & \underline{0} \end{pmatrix} \begin{pmatrix} \hat{\underline{\theta}} \\ \hat{\underline{\lambda}} \end{pmatrix} = \begin{pmatrix} \underline{S} \\ \underline{R} \end{pmatrix} \tag{8.134}$$

where

$$\underline{C} = \underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H} \tag{8.135}$$

$$\underline{S} = \underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{y} \tag{8.136}$$

Assuming that both $\underline{C}$ and $\underline{L}\,\underline{C}^{-1}\underline{L}^{\mathrm{T}}$ can be inverted, the normal equations can be solved for $\hat{\underline{\theta}}$ and $\hat{\underline{\lambda}}$ giving[3–5]

$$\begin{pmatrix} \hat{\underline{\theta}} \\ \hat{\underline{\lambda}} \end{pmatrix} = \begin{pmatrix} \underline{F} & \underline{G}^{\mathrm{T}} \\ \underline{G} & \underline{E} \end{pmatrix} \begin{pmatrix} \underline{S} \\ \underline{R} \end{pmatrix} \tag{8.137}$$

where*

$$\underline{W} = \left(\underline{L}\,\underline{C}^{-1}\underline{L}^{\mathrm{T}}\right)^{-1} \tag{8.138}$$

$$\begin{aligned} \underline{F} &= \underline{C}^{-1} - \underline{C}^{-1}\underline{L}^{\mathrm{T}}\underline{W}\,\underline{L}\,\underline{C}^{-1} \\ &= \left(\underline{1} - \underline{C}^{-1}\underline{L}^{\mathrm{T}}\underline{W}\,\underline{L}\right)\underline{C}^{-1} \end{aligned} \tag{8.139}$$

$$\underline{G} = \underline{W}\,\underline{L}\,\underline{C}^{-1} \tag{8.140}$$

$$\underline{E} = -\underline{W} \tag{8.141}$$

The solutions can then be written

$$\hat{\underline{\theta}} = \underline{F}\,\underline{S} + \underline{G}^{\mathrm{T}}\underline{R} = \underline{F}\,\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{y} + \underline{G}^{\mathrm{T}}\underline{R} \tag{8.142}$$

$$\hat{\underline{\lambda}} = \underline{G}\,\underline{S} + \underline{E}\,\underline{R} = \underline{G}\,\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{\epsilon} \tag{8.143}$$

The covariance matrix can be shown[3–5] to be given by

$$\underline{V}\left[\hat{\underline{\theta}}\right] = \underline{F} \tag{8.144}$$

$$\underline{V}\left[\hat{\underline{\lambda}}\right] = \underline{W} \tag{8.145}$$

$$\mathrm{cov}\left(\hat{\underline{\theta}}, \hat{\underline{\lambda}}\right) = 0 \tag{8.146}$$

*Note that Eadie *et al.*[4] contains a misprint in these equations.

In the unconstrained case the solution was $\underline{\hat{\theta}} = \underline{C}^{-1}\underline{S}$ with covariance matrix $\underline{V}\left[\underline{\hat{\theta}}\right] = \underline{C}^{-1}$. These results are recovered from the above equations by setting terms involving $\underline{L}$ or $\underline{R}$ to zero. From equations 8.139 and 8.144 we see that the constraints reduce the variance of the estimators, as should be expected since introducing constraints adds information. We also see that the constraints introduce (additional) correlations between the $\hat{\theta}_i$.

It can be shown[4,5] that the $\underline{\hat{\theta}}$ are unbiased, and that $E\left[\hat{\lambda}\right] = 0$ as expected.

8.5.7 Improved measurements through constraints

An important use of constraints in the linear model is to improve measurements. As an example, suppose that one measures the three angles of a triangle. We know, of course, that the sum of the three angles must be 180°. However, because of the resolution of the measuring apparatus, it probably will not be. In particle physics one often applies the constraints of energy and momentum conservation to the measurements of the energies and momenta of particles produced in an interaction.* By using this knowledge we can obtain improved values of the measurements.

To do this, we make use of the linear model with constraints as developed in the previous section. We assume that there is just one measurement of each quantity. If there is more than one, they can be averaged and the average used in the fit. The model is here the simplest imaginable, $\underline{y} = \underline{\theta}$, *i.e.*, what we want to estimate is the response variable itself. The measurements are then described by (equation 8.103)

$$y_i = \sum_{j=1}^{n} \theta_i \delta_{ij} + \epsilon_i = \theta_i + \epsilon_i$$

Thus the matrix $\underline{H}$ is just the unit matrix, and, in the absence of constraints, the normal equations have the trivial (and obvious) solution (equation 8.112)

$$\underline{\hat{\theta}} = \left(\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{H}\right)^{-1}\underline{H}^{\mathrm{T}}\underline{V}^{-1}\underline{y} = \underline{y}$$

The best value of a measurement ($\hat{\theta}_i$) is just the measurement itself (y_i).

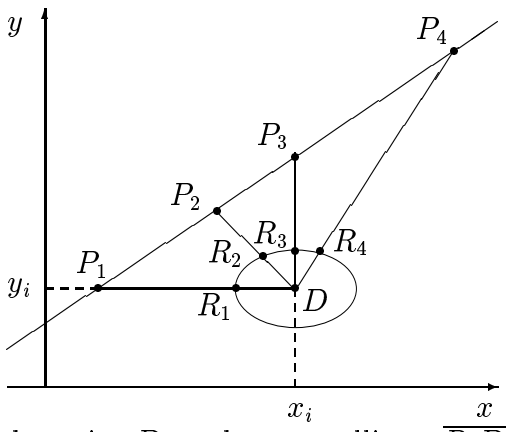
With m linear constraints (equation 8.130 or 8.131) the solution follows immediately from the previous section by setting $\underline{H} = \underline{1}$. The improved values of the measurements are then the $\hat{\theta}_i$. Note that the constraints introduce a correlation between the measurements.

8.5.8 Linear Model with errors in both x and y

So far we have considered the x_i to be known exactly. Now let us drop this restriction and allow the x_i as well as the y_i to have errors: $\sigma_{x\,i}$ and $\sigma_{y\,i}$, respectively.

*In particle physics this procedure is known as kinematical fitting since the constraints usually express the kinematics of energy and momentum conservation.

Straight-line fit



We begin by treating the case of a straight-line fit, $y = a + bx$, from section 8.5.5.

As before, we take Q^2 as the sum of the squares of the distances between the fit line and the measured point scaled by the error on this distance. However, this distance is not unique. This is illustrated in the figure where the ellipse indicates the errors on x_i and y_i. For a point on the line, P_j, the distance to D is $\overline{P_jD}$ and the error is the distance along this line from the point D to the error ellipse, $\overline{R_jD}$:

$$d_j = \frac{\overline{P_jD}}{\overline{R_jD}}$$

Since we want the minimum of Q^2, we also want to take the minimum of the d_j, *i.e.*, the minimum of

$$d_i^2 = \frac{(x - x_i)^2}{\sigma_{xi}^2} + \frac{(y - y_i)^2}{\sigma_{yi}^2} \tag{8.147}$$

where we have assumed that the errors on x_i and y_i are uncorrelated. Substituting $y = a + bx$ and setting $\frac{\mathrm{d}d_i}{\mathrm{d}x} = 0$ results in the minimum distance being given by

$$d_{i\,\mathrm{min}}^2 = \frac{(y_i - a - bx_i)^2}{\sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2}$$

This same result can be found by taking the usual definition of the distance,

$$d_i^2 = \left(\frac{y_i - y(x_i)}{\sigma_i}\right)^2 = \left(\frac{y_i - a - bx_i}{\sigma_i}\right)^2$$

where σ_i is no longer just the error on y_i, $\sigma_{y\,i}$, but is now the error on $y_i - a - bx_i$ and is found by error propagation to be

$$\sigma_i^2 = \sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2$$

Here the error propagation is exact since $y_i - a - bx_i$ is linear in x_i.

We must now find the minimum of

$$Q^2 = \sum_{i=1}^{n} \frac{(y_i - a - bx_i)^2}{\sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2} \tag{8.148}$$

The easiest method is to program it and use a minimization program. However, lets see how far we can get analytically.

Differentiating with respect to a gives

$$\frac{\partial Q^2}{\partial a} = -2\sum_{i=1}^{n} \frac{y_i - a - bx_i}{\sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2}$$

Setting this to zero and solving for a results in

$$\hat{a} = \frac{\sum_{i=1}^{n} \frac{y_i - \hat{b}x_i}{\sigma_{y\,i}^2 + \hat{b}^2\sigma_{x\,i}^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_{y\,i}^2 + \hat{b}^2\sigma_{x\,i}^2}}$$

We note that if all $\sigma_{x\,i} = 0$ this reduces to the expression found in section 8.5.5. Unfortunately, the differentiation with respect to b is more complicated. In practice it is most easily done numerically by choosing a series of values for $\hat{b}$, calculating $\hat{a}$ from the above formula and using these values of $\hat{a}$ and $\hat{b}$ to calculate Q^2, repeating the process until the minimum Q^2 is found.

The errors on $\hat{a}$ and $\hat{b}$ are most easily found from the condition that $Q^2 - Q^2_{\text{min}} = 1$ corresponds to one standard deviation errors.

If all $\sigma_{x\,i}$ are the same and also all $\sigma_{y\,i}$ are the same, the situation simplifies considerably. The above expression for $\hat{a}$ becomes

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{8.149}$$

and differentiation with respect to b leads to

$$\frac{\partial Q^2}{\partial b} = -2\sum_{i=1}^{n} \frac{y_i - \hat{a} - \hat{b}x_i}{\sigma_y^2 + b^2\sigma_x^2} + \frac{\hat{b}\sigma_x^2 \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2}{\sigma_y^2 + b^2\sigma_x^2} = 0$$

Substituting the expression for $\hat{a}$ into this equation then yields

$$\hat{b}^2\sigma_x^2\,\Delta xy - \hat{b}\,(\sigma_x^2\,\Delta y^2 - \sigma_y^2\,\Delta x^2) - \sigma_y^2\,\Delta xy = 0 \tag{8.150}$$

where

$$\begin{aligned} \Delta x^2 &= \overline{x^2} - \bar{x}^2 \\ \Delta y^2 &= \overline{y^2} - \bar{y}^2 \\ \Delta xy &= \overline{xy} - \bar{x}\bar{y} \end{aligned}$$

This is a quadratic equation for $\hat{b}$. Of the two solutions it turns out that the one with a negative sign before the square root gives the minimum Q^2; the one with the plus sign gives the maximum Q^2 of all straight lines passing through the point $(\bar{x}, \bar{y})$. We note that these solutions for $\hat{a}$ and $\hat{b}$ reduce to those found in section 8.5.5 when there is no uncertainty on x ($\sigma_x = 0$).

In general

Now let us consider a more complicated case. Let us represent a data point by the vector $\underline{z}_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$. If the model is a more complicated function than a straight line, or if there is a non-zero correlation between x_i and y_i, the distance measure d_i defined in equation 8.147 becomes

$$d_i^2 = (\underline{z}_i^c - \underline{z}_i)^{\mathrm{T}} \, \underline{V_i}^{-1} \, (\underline{z}_i^c - \underline{z}_i)$$

where $\underline{V_i}$ is the covariance matrix for data point i, $\underline{V_i} = \begin{pmatrix} \sigma_{x\,i}^2 & \mathrm{cov}(x_i, y_i) \\ \mathrm{cov}(x_i, y_i) & \sigma_{y\,i}^2 \end{pmatrix}$ and the point on the curve closest to $\underline{z}_i$ is represented by $\underline{z}_i^c = \begin{pmatrix} x_i^c \\ y_i^c \end{pmatrix}$. The components of $\underline{z}_i^c$ are related by the model: $y_i^c = \underline{H}^{\mathrm{T}}(x_i^c)\,\underline{\theta}$, which can be regarded as constraints for the minimization of Q^2. We then use Lagrangian multipliers and minimize

$$Q^2 = \sum_{i=1}^{n} \left[(\underline{z}_i^c - \underline{z}_i)^{\mathrm{T}} \, \underline{V_i} \, (\underline{z}_i^c - \underline{z}_i) + \lambda_i \left(y_i^c - \underline{H}^{\mathrm{T}}(x_i^c)\underline{\theta} \right) \right] \tag{8.151}$$

with respect to the unknowns:

$$\begin{array}{lll} k & \text{parameters} & \underline{\theta} \\ n & \text{unknowns} & x_i^c \\ n & \text{unknowns} & y_i^c \\ n & \text{unknowns} & \lambda_i \end{array}$$

by setting the derivatives of Q^2 with respect to each of these unknowns equal to zero. The solution of these $3n + k$ equations is usually quite messy and a numerical search for the minimum Q^2 is more practical.

8.5.9 Non-linear Models

For simplicity we again assume that the x_i are exactly known.

If the deviations of the measurements y_i from the true value $y(x_i)$ are normally distributed, the likelihood function is

$$\mathcal{L}(\underline{y}; \underline{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \left(\frac{y_i - y(x_i; \underline{\theta})}{\sigma_i} \right)^2 \right]$$

$$\ell = \ln \mathcal{L} = -\frac{n}{2} \ln(2\pi) + \sum_{i=1}^{n} \left[-\ln \sigma_i - \frac{1}{2} \left(\frac{y_i - y(x_i; \underline{\theta})}{\sigma_i} \right)^2 \right]$$

and $\mathcal{L}$ is maximal when

$$Q^2 = \sum_{i=1}^{n} \left(\frac{y_i - y(x_i; \underline{\theta})}{\sigma_i} \right)^2$$

is minimal. Thus the least squares method yields the same estimates as the maximum likelihood method, and accordingly has the same desirable properties.

When the deviations are not normally distributed, the least squares method may still be used, but it does not have such general optimal properties as to be useful for small n. Even asymptotically, the estimators need not be of minimum variance.[4,5]

In practice, the minimum of Q^2 is usually most easily found numerically using a search program such as MINUIT. However, an iterative solution[3,6] of the normal equations (subject to constraints) may yield considerable savings in computer time.

8.5.10 Summary

The most important properties of the least squares method are

- In the linear model, it follows from the Gauss-Markov theorem that least squares estimators have optimal properties: If the measurement errors have zero expectation and finite, fixed variance, then the least squares estimators are unbiased and have the smallest variance of all linear, unbiased estimators.
- If the errors are Gaussian, least squares estimators are the same as maximum likelihood estimators.
- If the errors are Gaussian, the minimum value of Q^2 provides a test of the validity of the model, at least in the linear model (*cf.* sections 10.4.3 and 10.6.3).
- If the model is non-linear in the parameters and the errors are not Gaussian, the least squares estimators usually do not have any optimal properties.

The least squares method discussed so far does not apply to histograms or other binned data. Fitting to binned data is treated in section 8.6.

8.6 Estimators for binned data

The methods of parameter estimation treated so far were developed and applied either to points (events) sampled from some p.d.f. (moments and maximum likelihood) or to measurements, *i.e.*, the results of some previous analysis (least squares and maximum likelihood). Here we want to apply a least squares method to a sample of events in order to estimate parameters of the underlying p.d.f., much as we did with maximum likelihood.

8.6.1 Minimum Chi-Square

The astute reader will have noticed that the least squares method requires measurements y_i with variance $V[y_i]$ for values x_i of the predictor variable. What do we

do when the data are simply observations of the values of x for a sample of events? This was easily treated in the maximum likelihood method. For a least squares type of estimator we must transform this set of observations into estimates of y at various values of x.

To do this we collect the observations into mutually exclusive and exhaustive classes defined with respect to the variable x. (The extension to more than one variable is straightforward.) An example of such a classification is a histogram and we shall sometimes refer to the classes as bins, but the concept is more general than a histogram. Assume that we have k classes and let π_i be the probability, calculated from the assumed p.d.f., that an observation falls in the ith class. Then

$$\sum_{i=1}^{k} \pi_i = 1$$

and the distribution of observations among the classes is a multinomial p.d.f. Let n be the total number of observations and n_i the number of observations in the ith class. Then $p_i = n_i/n$ is the fraction of observations in the ith class.

The minimum chi-square method consists of minimizing Pearson's[53] χ^2, which we refer to here as Q_1^2,

$$\begin{aligned} Q_1^2 &= n \sum_{i=1}^{k} \frac{(p_i - \pi_i)^2}{\pi_i} = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i} \\ &= n \left(\sum_{i=1}^{k} \frac{p_i^2}{\pi_i} - 1 \right) \end{aligned} \tag{8.152}$$

The estimators $\hat{\theta}_j$ are then the solutions of

$$\frac{\partial Q_1^2}{\partial \theta_j} = n \sum_{i=1}^{k} \frac{\partial Q_1^2}{\partial \pi_i} \frac{\partial \pi_i}{\partial \theta_j} = -n \sum_{i=1}^{k} \left(\frac{p_i}{\pi_i} \right)^2 \frac{\partial \pi_i}{\partial \theta_j} = 0 \tag{8.153}$$

This appears rather similar to the usual least squares method. The 'measurement' is now the observed number of events in a bin, and the model is that there should be $n\pi_i$ events in the bin. Recall (section 3.3) that the multinomial p.d.f. has for the ith bin the expectation $\mu_i = n\pi_i$ and variance $\sigma_i^2 = n\pi_i(1 - \pi_i)$. For a large number of bins, each with small probability π_i, the variance is approximately $\sigma_i^2 = n\pi_i$ and the covariances, $\text{cov}(n_i, n_j) = -n\pi_i\pi_j$, $i \neq j$, are approximately zero. The 'error' used in equation 8.152 is thus that expected from the model and is therefore a function of the parameters. In the least squares method we assumed, as a condition of the Gauss-Markov theorem, that σ_i^2 was fixed. Since that is here not the case, the Gauss-Markov theorem does not apply to minimum χ^2.

This use of the error expected from the model may seem rather surprising, but nevertheless this is the definition of Q_1^2. We note that in least squares the error was actually also an expected error, namely the error expected from the measuring apparatus, not the error estimated from the measurement itself.

In practice, Q_1^2 may be difficult to minimize owing to the dependence of the denominator on the parameters. This consideration led to the modified minimum chi-square method where one minimizes Q_2^2 (Neyman's χ^2), which is defined using an approximation of the observed, *i.e.*, estimated, error, $\sigma_i^2 \approx n_i$, which is valid for large n_i:

$$\begin{aligned} Q_2^2 &= n \sum_{i=1}^{k} \frac{(p_i - \pi_i)^2}{p_i} = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n_i} \\ &= n \left(\sum_{i=1}^{k} \frac{\pi_i^2}{p_i} - 1 \right) \end{aligned} \tag{8.154}$$

The estimators $\hat{\theta}_j$ are then the solutions of

$$\frac{\partial Q_2^2}{\partial \theta_j} = n \sum_{i=1}^{k} \frac{\partial Q_2^2}{\partial \pi_i} \frac{\partial \pi_i}{\partial \theta_j} = 2n \sum_{i=1}^{k} \left(\frac{\pi_i}{p_i} \right) \frac{\partial \pi_i}{\partial \theta_j} = 0 \tag{8.155}$$

From the approximations involved, it is clear that neither Q^2 is a true χ^2 for finite n. However, both become a $\chi^2(k-s)$ asymptotically, where s is the number of parameters which are estimated. Also, it can be shown[11] that the estimators found by both methods are 'best asymptotically normal' (BAN) estimators, *i.e.*, that the estimators are consistent, asymptotically normally distributed, efficient (of minimum variance), and that $\frac{\partial \hat{\theta}}{\partial p_i}$ exists and is continuous for all i. Both Q_1^2 and Q_2^2 thus lead asymptotically to estimators with optimal properties.

8.6.2 Binned maximum likelihood

Alternatively, one can use the maximum likelihood method on the binned data. The multinomial p.d.f. (eq. 3.3) in our present notation is

$$f = \frac{n!}{n_1! n_2! \ldots n_k!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k} = n! \prod_{i=1}^{k} \frac{\pi_i^{n_i}}{n_i!}$$

Dropping factors which are independent of the parameters, the log-likelihood which is to be maximized is given by

$$\ell = \ln \mathcal{L} = \sum_{i=1}^{k} n_i \ln \pi_i \tag{8.156}$$

Note that in the limit of zero bin width this is identical to the usual log-likelihood of equation 8.57. The estimators $\hat{\theta}_j$ are the values of θ for which ℓ is maximum and are given by

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{k} n_i \frac{\partial \ln \pi_i}{\partial \pi_i} \frac{\partial \pi_i}{\partial \theta_j} = n \sum_{i=1}^{k} \left(\frac{p_i}{\pi_i} \right) \frac{\partial \pi_i}{\partial \theta_j} = 0 \tag{8.157}$$

These maximum likelihood estimators are also BAN.

This formulation assumes that the total number of observations, $n = \sum n_i$, is fixed, as did the minimum chi-square methods of the previous section. If this is not the case, the binned maximum likelihood method is easily extended. As in section 8.4.7, the joint p.d.f. is multiplied by a Poisson p.d.f. for the total number of observations. Equivalently (*cf.* exercise 13), we can write the joint p.d.f. as a product of k Poisson p.d.f.'s:

$$f = \prod_{i=1}^{k} \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}$$

where ν_i is the expected number of observations in bin i. This leads to

$$\ell_{\mathrm{E}} = \sum_{i=1}^{k} n_i \ln \nu_i - \sum_{i=1}^{k} \nu_i \tag{8.158}$$

In terms of the present notation, $\nu_i = \nu_{\mathrm{tot}} \pi_i$. But now $\nu_{\mathrm{tot}} = \sum \nu_i$ is not necessarily equal to n.

8.6.3 Comparison of the methods

Asymptotically, all three of these methods are equivalent. How do we decide which one to use? In a particular problem, one method could be easier to compute. However, given the computer power most physicists have available, this is seldom a problem. The question is then which method has the best behavior for finite n.

- Q_1^2 requires a large number of bins with small π_i for each bin in order to neglect the correlations and to approximate the variance by $n\pi_i$. Assuming that the model is correct, this will mean that all n_i must be small.
- In addition, Q_2^2 requires all n_i to be large in order that $\sqrt{n_i}$ be a good estimate of the variance. Thus the n_i must be neither too large nor too small. In particular, an $n_i = 0$ causes Q_2^2 to blow up.
- The binned maximum likelihood method does not suffer from such problems.

In view of the above, it is perhaps not surprising that the maximum likelihood method usually converges faster to efficiency. In this respect the modified minimum chi-square (Q_2^2) is usually the worst of the three methods.[11]

One may still choose to minimize Q_1^2 or Q_2^2, perhaps because the problem is linear so that the equations $\frac{\partial Q^2}{\partial \theta_j} = 0$ can be solved simply by a matrix inversion instead of a numerical minimization. One must then ensure that there are no small n_i, which in practice is usually taken to mean that all n_i must be greater than 5 or 10. Usually one attains this by combining adjacent bins. However, one can just as well combine non-adjacent ones. Nor is there any requirement that all bin widths be equal. One must simply calculate the π_i properly, *i.e.*, as the integral of the

p.d.f. over the bin, which is not always adequately approximated by the bin width times the value of the p.d.f. at the center of the bin.

Since the maximum likelihood method is usually preferred, we can ask why we bin the data at all. Although binning is required in order to use a minimum chi-square method, we can perfectly well do a maximum likelihood fit without binning. Although binning loses information, it may still be desirable in the maximum likelihood method in order to save computing time when the data sample is very large. In choosing the bin sizes one should pay particular attention to the amount of information that is lost. Large bins lose little information in regions where the p.d.f. is nearly constant. Nor is much information lost if the bin size is small compared to the experimental resolution in the measurement of x. It would seem best to try to have the information content of the bins approximately equal. However, even with this criterion the choice of binning is not unique. It is then wise to check that the results do not depend significantly on the binning.

"There are nine and sixty ways of constructing tribal lays.
And – every – single – one – of – them – is – right!"
—Rudyard Kipling

8.7 Practical considerations

In this section we try to give some guidance on which method to use and to treat some complications that arise in real life.

8.7.1 Choice of estimator

Criteria

Faced with different methods which lead to different estimators we must decide which estimator to use. Eadie *et al.*[4] and James[5] give the following order of importance of various criteria for the estimators:

1. Consistency. The estimator should converge to the true value with increasing numbers of observations. If this is not the case, a procedure to remove the bias should be applied.

2. Minimum loss of information. When an estimator summarizes the results of an experiment in a single number, it is of vital interest to subsequent users of the estimate that no other number could contain more information about the parameter of interest.

3. Minimum variance (efficiency). The smaller the variance of the estimator, the more certain we are that it is near the true value of the parameter (assuming it is unbiased).

4. Robustness. If the p.d.f. is not well known, or founded on unsafe assumptions, it is desirable that the estimate be independent of, or insensitive to, departures from the assumed p.d.f. In general, the information content of such estimates is less since one chooses to ignore the information contained in the form of the p.d.f.

5. Simplicity. When a physicist reads the published value of some parameter, he usually presumes that the estimate of the parameter is unbiased, normally distributed, and uncorrelated with other estimates. It is therefore desirable that estimators have these simple properties. If the estimate is not simple, it should be stated how it deviates from simplicity and not given as just a number $\pm$ an error.

6. Minimum computer time. Although not fundamental, this may be of practical concern.

7. Minimum loss of physicist's time. This is also not fundamental; its importance is frequently grossly overestimated.

Compromising between these criteria

The order of the desirable properties above reflects a general order of importance. However, in some situations a somewhat different order would be better. For example, the above list places more importance on minimum loss of information than on minimum variance. These two criteria are related. The minimum variance is bounded by the inverse of the information. However this limit is not always attainable. In such cases it is possible that two estimates t_1 and t_2 of θ are such that $I_2(\theta) > I_1(\theta)$ but $V[t_1] < V[t_2]$. The recommendation here is to choose t_2, the estimate with the greater information. The reason is that, having more information, it will be more useful later when the result of this experiment is combined with results of other experiments. On the other hand, if decisions must be made, or conclusions drawn, on the basis of just this one experiment, then it would be better to choose t_1, the estimate with the smaller variance.

Obtaining simplicity

It may be worth sacrificing some information to obtain simplicity.

Estimates of several parameters can be made uncorrelated by diagonalizing the covariance matrix and finding the corresponding linear combinations of the parameters. But the new parameters may lack physical meaning.

Techniques for bias removal will be discussed below (section 8.7.2).

When sufficient statistics exist, they should be used, since they can be estimated optimally (*cf.* section 8.2.8).

Asymptotically, most usual estimators are unbiased and normally distributed. The question arises how good the asymptotic approximation is in any specific case. The following checks may be helpful:

- Check that the log-likelihood function or χ^2 is a parabolic function of the parameters.
- If one has two asymptotically efficient estimators, check that they give consistent results. An example is the minimum chi-square estimate from two different binnings of the data.
- Study the behavior of the estimator by Monte Carlo techniques, *i.e.*, make a large number of simulations of the experiment and apply the estimator to each Monte Carlo simulation in order to answer questions such as whether the estimate is normally distributed. However, this can be expensive in computer time.

A change of parameters can sometimes make an estimator simpler. For instance the estimate of $\theta_2 = g(\theta_1)$ may be simpler than the estimate of θ_1. However, it is in general impossible to remove both the bias and the non-normality of an estimator in this way[4,5].

Economic considerations

Economy usually implies fast computing. Optimal estimation is frequently iterative, requiring much computer time. The following three approaches seek a compromise between efficiency (minimum variance) and economic cost.

- Linear methods. The fastest computing is offered by linear methods, since they do not require iteration. These methods can be used when the expected values of the observations are linear functions of the parameters. Among linear unbiased estimators, the least squares method is the best, which follows from the Gauss-Markov theorem (section 8.5.4).

  When doing empirical fits, rather than fits to a known (or hypothesized) p.d.f., choose a p.d.f. from the exponential family (section 8.2.7) if possible. This leads to easy computing and has optimal properties.

- Two-step methods. Some computer time can be saved by breaking the problem into two steps:

  1. Estimate the parameters by a simple, fast, inefficient method, *e.g.*, the moments method.
  2. Use these estimates as starting values for an optimal estimation, *e.g.*, maximum likelihood.

Although more physicist's time may be spent in evaluating the results of the first step, this might also lead to a better understanding of the problem.

- Three-step method.

  1. Extract from the data a certain number of statistics which summarize the observations compactly, and if possible in a way which increases insight into the problem. For example, one can make a histogram, which reduces the number of observations to the number of bins in the histogram. Another example is the summary of an angular distribution by the coefficients of the expansion of the distribution in spherical harmonics. These coefficients are rapidly estimated by the moments method (section 8.3.2) and their physical meaning is clear.
  2. Estimate the parameters of interest using this summary data. If the summary data have an intuitive physical meaning this estimation may be greatly simplified.
  3. Use the preliminary estimates from the second step as starting values for an optimal estimation directly from the original data.

  The third step should not be forgotten. It is particularly important when the information in the data is small ('small statistics'). Because of the third step, the second step does not have to be exact, but only approximate.

8.7.2 Bias reduction

We have already given a procedure for bias reduction in section 8.3.1 for the case of an estimator $\hat{g}$ which is calculated from an unbiased estimator $\hat{\theta}$ by a change of variable $\hat{g} = g(\hat{\theta})$. Now let us consider two general methods.

Exact distribution of the estimate known

If the p.d.f. of the estimator is exactly known, the bias $b = E\left[\hat{\theta}\right] - \theta$ can be calculated. If b does not depend on the parameters, we can use the unbiased estimator $\hat{\theta}' = \hat{\theta} - b$ instead of the biased estimator $\hat{\theta}$. The variances of $\hat{\theta}'$ and $\hat{\theta}$, are the same since b is exactly known.

However, b is usually not exactly known since it usually depends on some of the parameters of the p.d.f. It must therefore be estimated. Assuming that we can make an unbiased estimate of the bias, $\hat{b}$, the unbiased estimator of the parameter is $\hat{\theta}' = \hat{\theta} - \hat{b}$, which results in a larger variance for $\hat{\theta}'$ than for $\hat{\theta}$.

Exact distribution of the estimate unknown

There is a straightforward method[4,5] to use in the case that the p.d.f. is not well known or no unbiased estimate of b is possible. Suppose that $\hat{\theta}$ is a biased estimator

which is asymptotically unbiased (as maximum likelihood estimators frequently are). Express $\hat{\theta}$ as a power series in $\frac{1}{N}$, where N is the number of events. The leading term is then θ, independent of N. The N^{-1} term is the leading bias term. Now split the data into two samples, each of $\frac{N}{2}$ events. Let the estimate from the two $\frac{N}{2}$ samples be $\hat{\theta}_1$ and $\hat{\theta}_2$. The expectation of the above expansion will be

$$E\left[\hat{\theta}\right] = \theta + \frac{1}{N}\beta + \mathcal{O}\left(\frac{1}{N^2}\right)$$
$$E\left[\hat{\theta}_1\right] = E\left[\hat{\theta}_2\right] = \theta + \frac{2}{N}\beta + \mathcal{O}\left(\frac{1}{N^2}\right)$$

Thus,

$$E\left[2\hat{\theta} - \frac{1}{2}(\hat{\theta}_1 + \hat{\theta}_2)\right] = \theta + \mathcal{O}\left(\frac{1}{N^2}\right)$$

and we see that we have a method to reduce the bias from $\mathcal{O}\left(\frac{1}{N}\right)$ to $\mathcal{O}\left(\frac{1}{N^2}\right)$. The variance is, however, in general increased by a term of order $\frac{1}{N}$.

A generalization of this method,[11,13] known as the **jackknife**,* estimates θ by

$$\hat{\theta} = N\hat{\theta}_N - (N-1)\overline{\hat{\theta}}_{N-1} \tag{8.159}$$

where $\hat{\theta}_N$ is the estimator using all N events and $\overline{\hat{\theta}}_{N-1}$ is the average of the N estimates possible using $N-1$ events:

$$\overline{\hat{\theta}}_{N-1} = \sum_{i=1}^{N} \hat{\theta}_i / N \tag{8.160}$$

where $\hat{\theta}_i$ is the estimate obtained using all events except event i.

A more general method, of which the jackknife is an approximation, is the **bootstrap** method introduced by Efron.[41–43] Instead of using each subset of $N-1$ observations, it uses samples of size $M \leq N$ randomly drawn, with replacement, from the data sample itself. For details, see, *e.g.*, Reference 43.

8.7.3 Variance of estimators—Jackknife and Bootstrap

The **jackknife** and the **bootstrap** of the previous section provide methods, albeit computer intensive, to evaluate the variance of estimators, which may be used in situations where the usual methods are unreliable, *e.g.*, small statistics where the asymptotic properties of ML estimators are questionable, non-Gaussian errors in least-squares fits,, or non-linear transformations of parameters (*cf.* section 8.3.6).

*Named after a large folding pocket knife, this procedure, like its namesake, serves as a handy tool in a variety of situations where specialized techniques may not be available.

The jackknife estimation of the variance of an estimator $\hat{\theta}$ is given by[43,44]

$$\hat{V}\left[\hat{\theta}\right]_{\mathrm{J}} = \frac{N-1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \overline{\hat{\theta}}_{N-1}\right)^2 \tag{8.161}$$

where the notation is the same as in the previous section.

While the jackknife is often a good method to estimate the variance of an estimator, it can fail miserably when the value of the estimator does not behave smoothly to small changes in the data. An example of such an estimator is the median. Suppose the data consist of 13 points: 5 values smaller than 9; the values 9, 11, and 13; and 5 values larger than 13. The sample median is 11. Removing any one of the smallest 6 values results in a median of 12 (midway between 11 and 13), while removing any one of the largest 6 values results in a median of 10. Removal of the middle value results in 11. There are only 3 different jackknife values; the estimate does not change smoothly with changes in the data, but only in large steps. This failure of the jackknife can be overcome by removing more points to make the jackknife samples. This is known as the **delete-*d* jackknife**; the interested reader is referred, *e.g.*, to Reference 43.

For bootstrap sample size, M, equal to the data sample size, N, there are N^N distinct samples possible, which is very large even for moderate N. Then the bootstrap sampling distribution for an estimator is a good approximation of the true sampling distribution, converging to it as $N \to \infty$ under fairly general conditions. This method is something like Monte Carlo, but uses the data themselves instead of a known (or hypothesized) distribution. The variance of $\hat{\theta}$ is then estimated by the following procedure:

1. Select B independent bootstrap samples, each consisting of N data, drawn with replacement from the real data sample. Usually B in the range 25–200 will suffice,[43] but this can be checked by repeating for large values of B until the improvement is negligible.

2. Evaluate $\hat{\theta}$ for each bootstrap sample, giving B values, $\hat{\theta}_b$.

3. The bootstrap estimate of the variance of $\hat{\theta}$ is then

$$\hat{V}\left[\hat{\theta}\right]_{B} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \overline{\hat{\theta}}_b\right)^2 \tag{8.162}$$

$$\text{where} \qquad \overline{\hat{\theta}}_b = \sum_{b=1}^{B}\hat{\theta}_b/B$$

Note that these two methods are applicable to non-parametric estimators as well as parametric. If the estimators are the result of a parametric fit, *e.g.*, ML, the B bootstrap samples can be generated from the fitted distribution function, *i.e.*, the

parametric estimate of the population, rather than from the data. The estimation of the variance is again given by equation 8.162.

Limitation: It should be clear that the non-parametric bootstrap will not be reliable when the estimator depends strongly on the tail of the distribution, as is the case, *e.g.*, with high-order moments. A bootstrap sample can never contain points larger than the largest point in the data.

8.7.4 Robust estimation

When the form of the p.d.f. is not exactly known, the following questions arise:

1. What kind of parameters can be estimated without any assumption about the form of the p.d.f.? Such estimators are usually called 'distribution-free'. This term may be misleading, for although the estimate itself does not depend on the assumption of a p.d.f., its properties, *e.g.*, the variance, do depend on the actual (unknown) p.d.f.

2. How reliable are the estimates if the assumed form of the p.d.f. is not quite correct?

Center of a symmetric distribution

There is relatively little known about robust estimation. The only case treated extensively in the literature is the estimation of the center of an unknown, symmetric distribution. The center of a distribution may be defined by a 'location parameter' such as the mean, the median, the mode, the midrange, *etc.* Several of these estimators were mentioned in section 8.1. The sample mean is the most obvious and most often used estimator of location because

- By the central limit theorem it is consistent whenever the variance of the p.d.f. is finite.

- It is optimal (unbiased and minimum variance) when the p.d.f. is a Gaussian.

However, if the distribution is not normal, the sample mean may not be the best estimator. For symmetric distributions of finite range, *e.g.*, the uniform p.d.f. or a triangular p.d.f., the location is determined by specifying the end points of the distribution. The midrange is then an excellent estimator. However, for distributions of infinite range, the midrange is a poor estimator.

The following table[4,5] shows asymptotic efficiencies, *i.e.*, the ratio of the minimum variance bound to the variance of the estimator, of location estimators for various p.d.f.'s.

Distribution	Sample median	Sample mean	Sample midrange
Normal	0.64	1.00	0.00
Cauchy	0.82	0.00	0.00
Double exponential	1.00	0.50	0.00
Uniform	0.00	0.00	1.00

None of these three estimators is asymptotically efficient for all four distributions. Nor has any of these estimators a non-zero asymptotic efficiency for all four distributions. As an example take a distribution which is the sum of a normal distribution and a Cauchy distribution having the same mean:

$$f(x) = \beta\, N(x;\mu,\sigma^2) + (1-\beta)\, C(x;\mu,\alpha) \quad , \qquad 0 \le \beta \le 1$$

Because of the Cauchy admixture, the sample mean has infinite variance, as we see in the table, while the sample median has at worst ($\beta = 1$) a variance of $1/0.64 = 1.56$ times the minimum variance bound. This illustrates that the median is generally more robust than the mean.

Other methods to improve robustness involve 'trimming', *i.e.*, throwing away the highest and lowest points before using one of the above estimators. This is particularly useful when there are large tails which come mostly from experimental problems. Such methods are further discussed by Eadie *et al.*[4,5]

Center of an asymmetric distribution

Consider the estimation of the center of a narrow 'signal' distribution superimposed on an unknown but wider 'background' distribution. The asymmetry of the background makes it difficult to use any of the above-mentioned estimators.

A common technique is to parametrize the signal and background in some arbitrary way and to do a maximum likelihood or least squares fit to obtain optimum values of the parameters, including the location parameter of interest. This is a non-robust method because the location estimate depends on the background parametrization and on correlations with other parameters.

A robust technique for this problem is to estimate the mode of the observed distribution. The mode is nearly invariant under variations of a smooth background. An obvious way to estimate the mode is to histogram the data and take the center of the most populated bin. Such a method depends on the binning used. A better method is given by the following procedure: Find the two observations which are separated by the smallest distance, and choose the one which has the closer next nearest neighbor. The estimate of the mode is then taken as the position of this observation. A generalization of this method is that of k nearest neighbors, where the density of observations at a given point is estimated by the reciprocal of the distance between the smallest and largest of the k observations closest to the point.

8.7.5 Detection efficiency and Weights

We are often not able to observe directly the phenomenon we wish to study. The apparatus generally introduces some distortion or bias, the effect of which must be taken into account. Such distortion may take the form of a detection efficiency, *i.e.*, the apparatus may not detect all events and the efficiency of detection may depend on the values of the variables being measured. This problem has already been mentioned in section 4.2.

The method used to account for this distortion depends on the severity of the problem. If the detection efficiency varies greatly over the range of the variables, it will be necessary to treat the problem exactly in order to avoid losing a great deal of information. On the other hand, if the detection efficiency is nearly uniform (say to within 20%), an approximate method will suffice.

Maximum likelihood—ideal method

As already mentioned in section 4.2, the p.d.f. of the observations is the product of the underlying physical p.d.f. and the efficiency function. It often happens that the physical p.d.f. can be written as the product of two p.d.f.'s where the parameters we want to estimate occur in only one of the two. For example, consider the production of particles in an interaction. The energies of the produced particles will not depend on where the interaction took place. The p.d.f. is then a product of a p.d.f. for the place where the interaction takes place and a p.d.f. for the interaction itself. Accordingly we write the physical p.d.f. as

$$f(\underline{x}, \underline{y}; \underline{\theta}, \underline{\psi}) = p(\underline{x}; \underline{\theta})\, q(\underline{y}; \underline{\psi})$$

where the p.d.f.'s p and q are, as usual, normalized: $\int p\, \mathrm{d}\underline{x} = \int q\, \mathrm{d}\underline{y} = 1$. Let $e(\underline{x}, \underline{y})$ be the detector efficiency, *i.e.*, the p.d.f. describing the probability that an event is observed. Then the p.d.f. of the actual observations is

$$g(\underline{x}, \underline{y}; \underline{\theta}, \underline{\psi}) = \frac{p(\underline{x}; \underline{\theta})\, q(\underline{y}; \underline{\psi})\, e(\underline{x}, \underline{y})}{\int p(\underline{x}; \underline{\theta})\, q(\underline{y}; \underline{\psi})\, e(\underline{x}, \underline{y})\, \mathrm{d}\underline{x}\, \mathrm{d}\underline{y}}$$

Note that the efficiency may depend on both $\underline{x}$ and $\underline{y}$. The likelihood of a given set of observations is then

$$\mathcal{L}(\underline{x}_1, \ldots \underline{x}_N, \underline{y}_1 \ldots \underline{y}_N; \underline{\theta}, \underline{\psi}) = \prod_{i=1}^{N} g(\underline{x}_i, \underline{y}_i; \underline{\theta}, \underline{\psi}) = \prod_{i=1}^{N} g_i$$

Hence,
$$\ell = \ln \mathcal{L} = W + \sum_{i=1}^{N} \ln(e_i\, q_i) \tag{8.163}$$

where
$$W = \sum_{i=1}^{N} \ln p_i - N \ln \int pqe\, \mathrm{d}\underline{x}\, \mathrm{d}\underline{y} \tag{8.164}$$

and where
$$p_i = p(\underline{x}_i; \underline{\theta})$$

Suppose now that we are not interested in estimating $\underline{\psi}$, but only $\underline{\theta}$. Then the second term of equation 8.163 does not depend on the parameters and may be ignored. The estimates $\hat{\underline{\theta}}$ and their variances are then found in the usual way treating W as the log-likelihood.

In practice, difficulties arise when pqe is not analytically normalized, but must be normalized numerically by time-consuming Monte Carlo. Moreover, the results depend on the form of q, which may be poorly known and of little physical interest. For these reasons one prefers to find a way of eliminating q from the expressions. Since this will exclude information, it will increase the variances, but at the same time make the estimates more robust.

Troll, to thyself be true—enough.
—Ibsen, "Peer Gynt"

Maximum likelihood—approximate method

We replace W in equation 8.164 by

$$W' = \sum_{i=1}^{N} \left(\frac{1}{e_i} \ln p_i \right) \tag{8.165}$$

Intuitively, the observation of an event with efficiency e_i corresponds, in some sense, to $w_i = 1/e_i$ events having actually occurred. Then the likelihood for all of the events, *i.e.*, the one which is observed *and* the ones which are not, is $p_i^{w_i}$, which results in W'.

Whatever the validity of this argument, it turns out[4,5] that the estimate $\hat{\underline{\theta}}'$ obtained by maximizing W' is, like the usual maximum likelihood estimate, asymptotically normally distributed about the true value. However, care must be taken in evaluating the variance. Using the second derivative matrix of W' is wrong since it assumes that

$$\sum_{i=1}^{N} w_i = N$$

events have been observed. One approach to curing this problem is to renormalize the weights by using $w_i' = N w_i / \sum w_i$ instead of w_i. However, this is only satisfactory if the weights are all nearly equal.

The correct procedure, which we will not derive, results in[4,5]

$$V\left(\hat{\underline{\theta}}'\right) = \underline{H}^{-1} \underline{H}' \underline{H}^{-1} \tag{8.166}$$

where the matrices $\underline{H}$ and $\underline{H'}$ are given by

$$H_{jk} = E\left[\frac{1}{e} \left(\frac{\partial \ln p}{\partial \theta_j} \right) \left(\frac{\partial \ln p}{\partial \theta_k} \right) \right]$$

$$H'_{jk} = E\left[\frac{1}{e^2}\left(\frac{\partial \ln p}{\partial \theta_j}\right)\left(\frac{\partial \ln p}{\partial \theta_k}\right)\right]$$

which may be estimated by the sample mean:

$$\hat{H}_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{e_i p_i^2}\left(\frac{\partial p_i}{\partial \theta_j}\right)\left(\frac{\partial p_i}{\partial \theta_k}\right) \tag{8.167a}$$

$$\hat{H}'_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{e_i^2 p_i^2}\left(\frac{\partial p_i}{\partial \theta_j}\right)\left(\frac{\partial p_i}{\partial \theta_k}\right) \tag{8.167b}$$

evaluated at $\underline{\theta} = \underline{\hat{\theta}}'$. If e is constant, this reduces to the usual estimator of the covariance matrix given in equations 8.78 and 8.80.

Alternatively, one can estimate the matrix elements from the second derivatives:

$$H_{jk} = -\left.\frac{\partial^2 W'}{\partial \theta_j \partial \theta_k}\right|_{\underline{\theta}=\underline{\hat{\theta}}} \quad ; \quad \hat{H}_{jk} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e_i}\frac{\partial^2 \ln p_i}{\partial \theta_j \partial \theta_k}\right]_{\underline{\theta}=\underline{\hat{\theta}}} \tag{8.168a}$$

$$H'_{jk} = -\frac{1}{e}\left.\frac{\partial^2 W'}{\partial \theta_j \partial \theta_k}\right|_{\underline{\theta}=\underline{\hat{\theta}}} \quad ; \quad \hat{H}_{jk} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{e_i^2}\frac{\partial^2 \ln p_i}{\partial \theta_j \partial \theta_k}\right]_{\underline{\theta}=\underline{\hat{\theta}}} \tag{8.168b}$$

If e is constant, this reduces to the usual estimator of the covariance matrix given in equations 8.84 and 8.85.

To summarize: Find the estimates $\underline{\hat{\theta}}'$ by maximizing W' (eq. 8.165). If possible compute $\underline{H}$ and $\underline{H}'$ by equation 8.167 or 8.168; if the derivatives are not known analytically, use equation 8.168, evaluating $\frac{\partial^2 W'}{\partial \theta_j \partial \theta_k}$ numerically. The covariance matrix is then given by equation 8.166.

It is clear from the above formulae that the appearance of one event with a very large weight will ruin the method, since it will cause W' (equation 8.165) to be dominated by one term and will make the variance very large. Accordingly, a better estimate may be obtained by rejecting events with very large weights.

Minimum chi-square—approximate method

Consider a histogram with k bins containing n_i events in the ith bin. Suppose that a model predicts the normalization $n = \sum n_i$ as well as the shape of the distribution. Denote the expected number of events in the ith bin by

$$a_i(\theta) = A(\theta)\frac{\int_i pqe\,\mathrm{d}x}{\int pqe\,\mathrm{d}x} \tag{8.169}$$

where $A(\theta) = \sum a_i$ is the predicted total number of events and $\int_i$ indicates an integral over bin i.

The minimum chi-square and modified minimum chi-square formulae (section

8.6.1) become

$$Q_1^2 = \sum_{i=1}^{k} \frac{(n_i - a_i)^2}{a_i} \quad , \quad \frac{\partial Q_1^2}{\partial \theta} = -\sum_{i=1}^{k} \left[\left(\frac{n_i}{a_i} \right)^2 - 1 \right] \frac{\partial a_i}{\partial \theta}$$

$$Q_2^2 = \sum_{i=1}^{k} \frac{(n_i - a_i)^2}{n_i} \quad , \quad \frac{\partial Q_2^2}{\partial \theta} = -2 \sum_{i=1}^{k} \left[1 - \frac{a_i}{n_i} \right] \frac{\partial a_i}{\partial \theta}$$

So far, this is exact. Now let us introduce the approximate method by removing the dependence on q from equation 8.169. Let b_i be the predicted number of events in bin i when $e = 1$. We want to correct the numbers b_i using the known experimental efficiency to obtain numbers c_i such that

$$E[c_i] = a_i \tag{8.170}$$

From its definition, b_i is given by

$$b_i = B(\theta) \frac{\int_i pq \, \mathrm{d}x}{\int pq \, \mathrm{d}x} = A(\theta) \frac{\int_i pq \, \mathrm{d}x}{\int pqe \, \mathrm{d}x}$$

where $B(\theta)$ is the total number of events predicted when $e = 1$. Combining this equation with equation 8.169, we find

$$a_i = b_i \frac{\int_i pqe \, \mathrm{d}x}{\int_i pq \, \mathrm{d}x}$$

The inverse of this ratio of integrals can be rewritten as

$$\frac{\int_i pqew \, \mathrm{d}x}{\int_i pqe \, \mathrm{d}x} = E_i[w]$$

where $w = \frac{1}{e}$ is the weight. This expectation can be estimated by the sample mean of the weights of the events in the bin:

$$\widehat{E_i[w]} = \frac{\sum_{j=1}^{n_i} w_{ij}}{n_i}$$

where w_{ij} is the weight $(1/e_i)$ of the jth event in the ith bin.

We now define

$$c_i = \frac{b_i n_i}{\sum_{j=1}^{n_i} w_{ij}}$$

From the preceding equations it is clear that this c_i satisfies equation 8.170.

The expressions for Q^2 then use c_i instead of a_i. Writing σ_i^2 for a_i in the case of Q_1^2 and for n_i in the case of Q_2^2, both may be written as

$$Q^2 = \sum_{i=1}^{k} \frac{1}{\sigma_i^2} \left(n_i - b_i \frac{n_i}{\sum_j w_{ij}} \right)^2 = \sum_{i=1}^{k} \frac{1}{\sigma_i'^2} \left(\sum_j w_{ij} - b_i \right)^2$$

where $$\frac{1}{\sigma_i'^2} = \frac{1}{\sigma_i^2}\left(\frac{n_i}{\sum_j w_{ij}}\right)^2$$

The 'error', σ_i', is then given by

$$\sigma_i'^2 = E\left[\left(\sum_{j=1}^{n_i} w_{ij} - b_i\right)^2\right] = E\left[\left(\sum_{j=1}^{n_i} w_{ij}\right)^2\right] - b_i^2$$

since $E\left[\sum_{j=1}^{n_i} w_{ij}\right] = b_i$. Further, one can show that

$$E\left[\left(\sum_{j=1}^{n_i} w_{ij}\right)^2\right] = E\left[\sum_{j=1}^{n_i} w_{ij}^2\right] + E\left[\sum_{j=1}^{n_i}\sum_{\substack{k=1\\k\neq j}}^{n_i} w_{ij}w_{ik}\right] \approx E[n_i]\,E\left[w_i^2\right] + b_i^2$$

$E[w_i^2]$ can be estimated by the sample mean

$$\widehat{E[w_i^2]} = \frac{1}{n_i}\sum_{j=1}^{n_i} w_{ij}^2$$

$E[n_i]$ can be estimated in two ways: from the model, which gives the minimum chi-square method; or from the data, which gives the modified minimum chi-square method. The resulting expressions for Q^2 are

$$\widehat{E[n_i]} = c_i \quad ; \quad Q_1'^2 = \sum_{i=1}^{k}\left[\frac{\left(\sum_{j=1}^{n_i} w_{ij} - b_i\right)^2}{b_i\frac{\sum_{j=1}^{n_i} w_{ij}^2}{\sum_{j=1}^{n_i} w_{ij}}}\right] \tag{8.171}$$

$$\widehat{E[n_i]} = n_i \quad ; \quad Q_2'^2 = \sum_{i=1}^{k}\left[\frac{\left(\sum_{j=1}^{n_i} w_{ij} - b_i\right)^2}{\sum_{j=1}^{n_i} w_{ij}^2}\right] \tag{8.172}$$

Clearly both Q' approach the corresponding Q as the weights all approach 1. As in the unweighted case, the minimum chi-square method (Q_1) is better justified than the modified minimum chi-square method (Q_2). However, if b_i is a linear function of the parameters, the solution of the modified method can in principle be written explicitly, which is much faster than a numerical minimization.

But who can discern his errors?
Clear thou me from hidden faults.
—Psalm **19**.*12*

8.7.6 Systematic errors

If a meter has a random error, then its readings are distributed in some way about the true value. If the error distribution is not specified further, you expect it to

be Gaussian. Thus if it is simply stated that the error is 1%, you expect that this distribution will be a Gaussian distribution with a standard deviation of 1% of the true value. The standard deviation of a single reading will be 1% of that reading. But by making many (N) readings and averaging them, you obtain an estimate of the true value which has a much smaller variance. Usually, the variance is reduced by a factor $1/N$, which follows from the central limit theorem.

If the meter has a systematic error such that it consistently reads 1% too high, the situation is different. The readings are thus correlated. Averaging a large number of readings will not decrease this sort of error, since it affects all the readings in the same way. With more readings, the average will not converge to the true value but to a value 1% higher. It is as though we had a biased estimator.

Systematic errors can be very difficult to detect. For example, we might measure the voltage across a resistor for different values of current. If the systematic error was 1 Volt, all the results would be shifted by 1 Volt in the same direction. If we plotted the voltages against the currents, we would find a straight line, as expected. However, the line would not pass through the origin. Thus, we could in principle discover the systematic effect. On the other hand, with a systematic error of 1% on the voltage, all points would be shifted by 1% in the same direction. The voltages plotted against the currents would lie on a straight line and the line would pass through the origin. The voltages would thus appear to be correctly measured. But the slope of the line would be incorrect. This is the worst kind of systematic error—one which cannot be detected statistically. It is truly a 'hidden fault'.

The size of a systematic error may be known. For example, consider temperature measurements using a thermocouple. You calibrate the thermocouple by measuring its output voltages V_1 and V_2 for two known temperatures, T_1 and T_2, using a voltmeter of known resolution. You then determine some temperatures T by measuring voltages V and using the proportionality of V to T to calculate T:

$$T = \frac{T_2 - T_1}{V_2 - V_1}(V - V_1) + T_1$$

The error on T will include a systematic contribution from the errors on V_1 and V_2 as well as a random error on V. In this example the systematic error is known.

In other cases the size of the systematic error is little more than a guess. Suppose you are studying gases at various pressures and you measure the pressure using a mercury manometer. Actually it only measures the difference in pressure between atmospheric pressure and that in your vessel. For the value of the atmospheric pressure you rely on that given by the nearest meteorological station. But how big is the difference in the atmospheric pressure between the station at the time the atmospheric pressure was measured and your laboratory at the time you did the experiment?

Or, suppose you are measuring a (Gaussian) signal on top of a background. The estimate of the signal (position, width, strength) may depend on the functional form chosen for the background. If you do not know what this form is, you should

try various forms and assign systematic errors based on the resulting variations in the estimates.

Experimental tips

To clear your experiment of 'hidden faults' you should begin in the design of the experiment. Estimate what the systematic errors will be, and, if they are too large, design a better experiment.

Build consistency checks into the experiment, *e.g.*, check the calibration of an instrument at various times during the course of the experiment.

Try to convert a systematic error into a random error. Many systematic effects are a function of time. Examples are electronics drifts, temperature drifts, even psychological changes in the experimenter. If you take data in an orderly sequence, *e.g.*, measuring values of y as a function of x in the order of increasing x, such drifts are systematic. So mix up the order. By making the measurements in a random order, these errors become random.

The correct procedure depends on what you are trying to measure. If there are hysteresis effects in the apparatus, measuring or setting the value of a quantity, *e.g.*, a magnetic field strength, from above generally gives a different result than setting it from below. Thus, if the absolute values are important such adjustments should be done alternatively from above and from below. On the other hand, if only the differences are important, *e.g.*, you are only interested in a slope, then all adjustments should be made from the same side, as the systematic effect will then cancel.

Error propagation with systematic errors

Having eliminated what systematic effects you can, you must evaluate the rest. Different independent systematic errors are, since independent, added in quadrature.* Since random and systematic errors are independent, they too can be added in quadrature to give the total error. Nevertheless, the two types of error are often quoted separately, *e.g.*,

$$R = -1.9 \pm 0.1 \pm 0.4$$

where (conventionally) the first error is statistical and the second systematic. Such a statement is more useful to others, particularly if they want to combine your result with other results which may have the same systematic errors. For this reason, the various contributions to the systematic errors should also be given separately, particularly those which could be common to other experiments. One also sees in this example that more data would not help since the systematic error is much larger than the statistical error.

Error propagation is done using the covariance matrix in the usual way except

*This assumes that the errors are normally distributed. If you know this not to be the case, you should try to combine the errors using the correct p.d.f.'s.

that we keep track of the statistical and systematic contributions to the error. Suppose that we have two 'independent' measurements x_1 and x_2 with statistical errors σ_1 and σ_2 and with a common systematic error s. For pedagogical purposes we can think of the x_i as being composed of two parts, $x_i = x_i^{\mathrm{R}} + x_i^{\mathrm{S}}$, where x_i^{R} has only a random statistical error, σ_i, and x_i^{S} has only a systematic error, s. Then x_1^{R} and x_2^{R} are completely independent and x_1^{S} and x_2^{S} are completely correlated. The variance of x_i is then

$$\begin{aligned} V[x_i] &= E\left[x_i^2\right] - (E[x_i])^2 \\ &= E\left[\left(x_i^{\mathrm{R}} + x_i^{\mathrm{S}}\right)^2\right] - \left(E\left[x_i^{\mathrm{R}} + x_i^{\mathrm{S}}\right]\right)^2 \\ &= \sigma_i^2 + s^2 \end{aligned}$$

The covariance is

$$\begin{aligned} \mathrm{cov}(x_1, x_2) &= E[x_1 x_2] - E[x_1]\,E[x_2] \\ &= E\left[\left(x_1^{\mathrm{R}} + x_1^{\mathrm{S}}\right)\left(x_2^{\mathrm{R}} + x_2^{\mathrm{S}}\right)\right] - E\left[x_1^{\mathrm{R}} + x_1^{\mathrm{S}}\right] E\left[x_2^{\mathrm{R}} + x_2^{\mathrm{S}}\right] \end{aligned}$$

Each term involves four products. Those involving an x_i^{R} cancel leaving

$$\mathrm{cov}(x_1, x_2) = \mathrm{cov}(x_1^{\mathrm{S}}, x_2^{\mathrm{S}}) = s^2$$

Thus the covariance matrix is

$$V = \begin{pmatrix} \sigma_1^2 + s^2 & s^2 \\ s^2 & \sigma_2^2 + s^2 \end{pmatrix}$$

So far we have considered systematic errors which are constants. They also occur as fractions or percentages. The systematic error s is then not a constant but proportional to the measurement (actually to the true value, but for small errors the difference is by definition negligible): $s = \epsilon x$ with, *e.g.*, $\epsilon = 0.01$ for a 1% error. The above analysis is still valid: x_1^{S} and x_2^{S} are still completely correlated. The resulting covariance matrix is

$$V = \begin{pmatrix} \sigma_1^2 + \epsilon^2 x_1^2 & \epsilon^2 x_1 x_2 \\ \epsilon^2 x_1 x_2 & \sigma_2^2 + \epsilon^2 x_2^2 \end{pmatrix}$$

Generalization is rather obvious. If there are several independent sources of systematic error then they are added in quadrature. If there are more variables the matrix is larger. For example, consider three variables with independent statistical errors, a common systematic error s and in addition an independent systematic error t which is shared by x_1 and x_2 but not x_3. The covariance matrix is then

$$V = \begin{pmatrix} \sigma_1^2 + s^2 + t^2 & s^2 + t^2 & s^2 \\ s^2 + t^2 & \sigma_2^2 + s^2 + t^2 & s^2 \\ s^2 & s^2 & \sigma_3^2 + s^2 \end{pmatrix}$$

Least squares fit with systematic errors

Consider a least squares fit where the y-values have not only a statistical error σ, but also a common systematic error s. The covariance matrix for $\underline{y}$ is then

$$V_{ij}\left[\underline{y}\right] = \delta_{ij}\sigma^2 + s^2$$

This is just the covariance matrix previously considered in section 8.5.5 with the addition of s^2 to every element. As an example, consider a fit to a straight line, $y = a + bx$. Using this V and $\underline{\epsilon} = \underline{y} - a - b\underline{x}$, in $Q^2 = \underline{\epsilon}^{\mathrm{T}}\underline{V}\,\underline{\epsilon}$ and solving $\frac{\partial Q^2}{\partial a} = 0$ and $\frac{\partial Q^2}{\partial b} = 0$ leads to the same expressions for the estimators as before (equation 8.125). A common systematic shift of all points up or down clearly has no effect on the slope, and therefore we expect the same variance for $\hat{b}$ as before. However, a systematic shift in y will affect the intercept; consequently, we expect a larger variance for $\hat{a}$.

Chapter 9

Confidence intervals

In the previous chapter we have discussed methods to estimate the values of unknown parameters. As the uncertainty, or "error", $\delta\hat{\theta}$, on the estimate, $\hat{\theta}$, we have been content to state the standard deviations and correlation coefficients of the estimate as found from the covariance matrix or the estimated covariance matrix. This is inadequate in certain cases, particularly when the sampling p.d.f., *i.e.*, the p.d.f. of the estimator is non-Gaussian. In this chapter our interest is to find the range

$$\theta_{\mathrm{a}} \leq \theta \leq \theta_{\mathrm{b}}$$

which contains the true value θ_{t} of θ with "probability" β. We shall see that when the sampling p.d.f. is Gaussian, the interval $[\theta_a, \theta_b]$ for $\beta = 68.3\%$ is the same as the interval of ± 1 standard deviation about the estimated value.

9.1 Introduction

In parameter estimation we found an estimator for a parameter $\hat{\theta}$ and its variance $\sigma^2_{\hat{\theta}} = V\left[\hat{\theta}\right]$ and we wrote the result as $\theta = \hat{\theta} \pm \sigma_{\hat{\theta}}$. Assuming a normal distribution for $\hat{\theta}$, one is then tempted to say, as we did in section 8.2.4, that the probability is 68.3% that

$$\hat{\theta} - \sigma_{\hat{\theta}} \leq \theta_{\mathrm{t}} \leq \hat{\theta} + \sigma_{\hat{\theta}} \tag{9.1}$$

Now, what does this statement mean? If we interpret it as 68.3% probability that the value of θ_{t} is within the stated range, we are using Bayesian probability (*cf.* section 2.4.4) with the assumption of uniform prior probability. This assumption is not always justifiable and often is wrong, as is illustrated in the following example: An empty dish is weighed on a balance. The result is 25.31 ± 0.14 g. A sample of powder is placed on the dish, and the weight is again determined. The result is 25.51 ± 0.14 g. By subtraction and combination of errors, the weight of the powder is found to be 0.20 ± 0.20 g. Our first conclusion is that the scientist should have used a better balance. Next we try to determine some probabilities. From the

normal distribution, there is a probability of about 16% that a value lies lower than $\mu - \sigma$. In this example that means that there is a chance of about 16% that the powder has negative weight (an anti-gravity powder!). The problem here is Bayes' postulate of uniform prior probability. We should have incorporated in the prior knowledge the fact that the weight must be positive, but we didn't.

Let us avoid the problems of Bayesian prior probability and stick to the frequentist interpretation. This will lead us to the concept of **confidence intervals**, developed largely by Neyman,[45] which give a purely frequentist interpretation to equation 9.1. We shall return to the Bayesian interpretation in section 9.9.

Suppose we have a p.d.f. $f(x;\theta)$ which depends on one parameter θ. The probability content β of the interval $[a,b]$ in X-space is

$$\beta = P(a \leq X \leq b) = \int_a^b f(x;\theta)\,\mathrm{d}x \tag{9.2}$$

Common choices for β are 68.3% (1σ), 95.4% (2σ), 99.7% (3σ), 90% (1.64σ), 95% (1.96σ), and 99% (2.58σ), where the correspondence between percent and a number of standard deviations (σ) assumes that f is a Gaussian p.d.f.

If the function f and the parameter θ are known we can calculate β for any a and b. If θ is unknown we try to find another variable $z = z(x,\theta)$ such that its p.d.f., $g(z)$, is independent of θ. If such a z can be found, we can construct an interval $[z_a, z_b]$, where $z_x = z(x,\theta)$, such that

$$\beta = P(z_a \leq Z \leq z_b) = \int_{z_a}^{z_b} g(z)\,\mathrm{d}z \tag{9.3}$$

It may then be possible to use this equation together with equation 9.2 to find an interval $[\theta_-, \theta_+]$ such that

$$P(\theta_- \leq \theta_\mathrm{t} \leq \theta_+) = \beta \tag{9.4}$$

The meaning of this last equation must be made clear. Contrast the following two quite similar statements:

1. The probability that θ_t is in the interval $[\theta_-, \theta_+]$ is β.

2. The probability that the interval $[\theta_-, \theta_+]$ contains θ_t is β.

The first sounds like θ_t is the r.v. and that the interval is fixed. This is incorrect—we are frequentists here, and so θ_t is not a r.v. The second statement sounds like a statement about θ_- and θ_+, which is the correct meaning of equation 9.4. θ_- and θ_+ are the results of the experiment, and hence r.v.'s. To put it slightly differently: Performing the experiment as we have done, we have the probability, β, of finding an interval, $[\theta_-, \theta_+]$, which contains the (unknown) true value of θ, θ_t. If we were to repeat the experiment many times, a fraction β of the experiments would yield an interval containing the true value, *i.e.*, an interval which "covers" the true value.

Turned around, this means that if we assert on the basis of our experiment that the true value of θ lies in the interval $[\theta_-, \theta_+]$, we will be right in a fraction β of

the cases. Thus, β expresses the degree of confidence (or belief) in our assertion; hence the name confidence interval. The quantity β is known by various names: **confidence coefficient**, **coverage probability**, confidence level. However, the last term, "confidence level", is inadvisable, since it is also used for a different concept, which we will encounter in goodness-of-fit tests (*cf.* section 10.6).

The interval $[\theta_-, \theta_+]$ corresponding to a confidence coefficient β is in general not unique; many different intervals exist with the same probability content.

We can, of course, choose to state any one of these intervals. Commonly used criteria to remove this arbitrariness are

1. Symmetric interval: $\hat{\theta} - \theta_- = \theta_+ - \hat{\theta}$.
2. Shortest interval: $\theta_+ - \theta_-$ is the smallest possible, given β.
3. Central interval: the probability content below and above the interval are equal, *i.e.*, $P(\theta < \theta_-) = P(\theta > \theta_+) = (1-\beta)/2$.

For a symmetric distribution having a single maximum these criteria are equivalent. We usually prefer intervals satisfying one (or more) of these criteria. However, non-central intervals will be preferred when there is some reason to be more careful on one side than on the other, *e.g.*, the amount of tritium emitted from a nuclear power station.

Normally distributed estimators. To illustrate the above procedure: Let $t(\underline{x})$ be an estimator of a parameter having true value θ. As we have seen in the previous chapter, many estimators are (at least asymptotically) normally distributed about the true value. Then t is a r.v. distributed as $N(t;\theta,\sigma^2)$. Equation 9.2 is then

$$\beta = P(a \le T \le b) = \int_a^b N(t;\theta,\sigma^2)\,\mathrm{d}t = \mathrm{erf}\left(\frac{b-\theta}{\sigma}\right) - \mathrm{erf}\left(\frac{a-\theta}{\sigma}\right) \tag{9.5}$$

since the c.d.f. of the normal p.d.f. is the error function (*cf.* section 3.7).

If θ is not known, we can not evaluate the integral. Instead, assuming that σ is known, we transform to the r.v. $z = t - \theta$. The interval $[c, d]$ for z corresponds to the interval $[\theta + c, \theta + d]$ for t. Hence, equation 9.3 becomes

$$\beta = P(\theta + c \le T \le \theta + d) = \int_{\theta+c}^{\theta+d} N(t;\theta,\sigma^2)\,\mathrm{d}t = \mathrm{erf}\left(\frac{d}{\sigma}\right) - \mathrm{erf}\left(\frac{c}{\sigma}\right) \tag{9.6}$$

We can, for a given β, now choose an interval $[\theta + c, \theta + d]$ satisfying this equation. Now $t \le \theta + d$ implies that $\theta \ge t - d$ and $t \ge \theta + c$ implies that $t - c \ge \theta$. Hence, the above interval in t-space corresponds to the interval $[t-d, t-c]$ in θ-space, and we have the desired confidence interval for θ:

$$\beta = P(t - d \le \theta \le t - c) \tag{9.7}$$

Again, we emphasize that although this looks like a statement concerning the probability that θ is in this interval, it is not, but instead means that we have a probability β of being right when we assert that θ is in this interval.

If neither θ or σ is known, one chooses the standardized variable $z = \frac{t-\theta}{\sigma}$. The probability statement about Z is

$$\beta = P(c \leq Z \leq d) = \int_c^d N(z;0,1)\,\mathrm{d}z = \mathrm{erf}(d) - \mathrm{erf}(c) \tag{9.8}$$

which can be converted into a probability statement for θ:

$$\beta = P(t - d\sigma \leq \theta \leq t + c\sigma) \tag{9.9}$$

For the normal distribution this conversion is easy, due to the symmetry of the distribution between Z and θ. Note however that equation 9.9 does not help us very much since we do not know σ. We will discuss this further in section 9.4.2

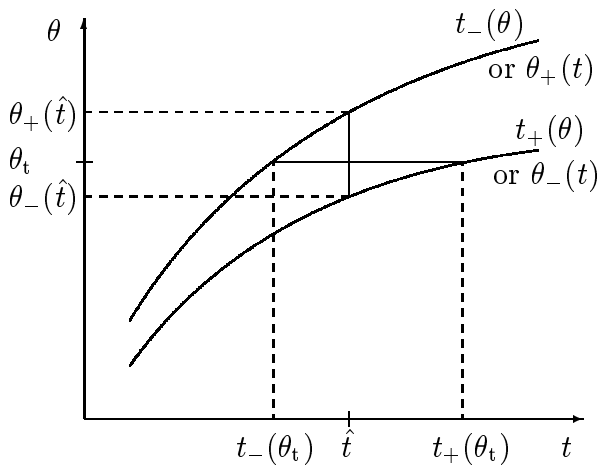
9.2 Confidence belts

Now let us see how we construct confidence intervals for an arbitrary p.d.f.[45] Suppose that $t(\underline{x})$ is an estimator of the parameter θ with p.d.f. $f(t|\theta)$. For a given value of θ, there will be values of t, $t_-(\theta)$ and $t_+(\theta)$ such that

$$\beta = P(t_- \leq T \leq t_+) = \int_{t_-}^{t_+} f(t|\theta)\,\mathrm{d}t \tag{9.10}$$

These values of t then define an interval in t-space, $[t_-, t_+]$, with probability content β. Usually the choice of t_- and t_+ is not unique, but may be fixed by an additional criterion, *e.g.*, by requiring a central interval:

$$\int_{-\infty}^{t_-} f(t|\theta)\,\mathrm{d}t = \frac{1-\beta}{2} = \int_{t_+}^{+\infty} f(t|\theta)\,\mathrm{d}t \tag{9.11}$$



We do not, of course, know the true value of θ, and hence we are unable to solve this equation for t_- and t_+. Nevertheless, we can make a plot of $t_-(\theta)$ and $t_+(\theta)$ *vs.* θ, which can also be viewed as a plot of, respectively, $\theta_+(t)$ and $\theta_-(t)$ *vs.* t. The region between the t_- and t_+ curves is known as a confidence belt.

For an unbiased, normally distributed estimator, as in the previous section, $f(t|\theta) = N(t;\theta,\sigma^2)$ and the lines for $\beta = 0.683$ would be, from equation 9.11, $t_-(\theta) = \theta - \sigma$ and $t_+(\theta) = \theta + \sigma$.

For any value of θ, the chance of finding a value of t in the interval $[t_-(\theta), t_+(\theta)]$ is β, by construction. Conversely, having done an experiment giving a value $t = \hat{t}$, the values of θ_- and θ_+ corresponding to $t_+ = \hat{t}$ and $t_- = \hat{t}$ can be read off of the plot as indicated. The interval $[\theta_-, \theta_+]$ is then a confidence interval of probability content β for θ. This can be seen as follows:

Suppose that θ_t is the true value of θ. A fraction β of experiments will then result in a value of t in the interval $[t_-(\theta_\mathrm{t}), t_+(\theta_\mathrm{t})]$. Any such value of t would yield, by the above-indicated method, an interval $[\theta_-, \theta_+]$ which would include θ_t. On the other hand, the fraction $1-\beta$ of experiments which result in a value of t not in the interval $[t_-(\theta_\mathrm{t}), t_+(\theta_\mathrm{t})]$ would yield an interval $[\theta_-, \theta_+]$ which would not include θ_t. Thus the probability content of the interval $[\theta_-, \theta_+]$ is also β.

To summarize, given a measurement $\hat{t}$, the central β confidence interval ($\theta_- \leq \theta \leq \theta_+$) is the solution of

$$\int_{-\infty}^{\hat{t}} f(t|\theta_+)\ \mathrm{d}t = \frac{1-\beta}{2} = \int_{\hat{t}}^{+\infty} f(t|\theta_-)\ \mathrm{d}t \tag{9.12}$$

If $f(t)$ is a normal p.d.f., which is often (at least asymptotically, as we have seen in chapter 8) the case, this interval is identical for $\beta = 68.3\%$ to $[\hat{\theta}-\sigma_{\hat{\theta}} < \theta < \hat{\theta}+\sigma_{\hat{\theta}}]$. If $f(t)$ is not Gaussian, the interval of $\pm 1\sigma$ (σ^2 the variance of $\hat{\theta}$) does not necessarily correspond to $\beta = 68.3\%$. In this case the uncertainty should be given which *does* correspond to $\beta = 68.3\%$. Such an interval is not necessarily symmetric about $\hat{\theta}$.

In 'pathological' cases, the confidence belt may wiggle in such a way that the resulting confidence interval consists of several disconnected pieces. While mathematically correct, the use of such disconnected intervals may not be very meaningful.

9.3 Confidence bounds

As mentioned above, the choice of confidence interval is usually not unique. In many cases we prefer a central interval. But sometimes an extremely non-central interval is preferable from a physical standpoint. In particular, confidence bounds, *i.e.*, upper or lower limits, are useful when the 'best' value of a parameter is found to be close (or perhaps beyond) a physical boundary.

For an upper limit, $t_+(\theta)$ is chosen infinite (or equal to the maximum allowed value of t). Then, the function $t_-(\theta)$ is defined (equation 9.10) by

$$\beta = P(T > t_-) = \int_{t_-}^{+\infty} f(t|\theta)\ \mathrm{d}t$$

For a measurement $\hat{t}$, θ_+ is read from this $t_-(\theta)$ curve as in the previous section. In other words, the upper limit, θ_+ is the solution of

$$\beta = P(\theta < \theta_+) = \int_{\hat{t}}^{+\infty} f(t|\theta_+)\ \mathrm{d}t \tag{9.13}$$

The statement is then that $\theta < \theta_+$ with confidence β, and such an assertion will be correct in a fraction β of the cases.

Lower limits are defined analogously: The lower limit θ_-, for which $\theta > \theta_-$ with confidence β, is found from

$$\beta = P(\theta > \theta_-) = \int_{-\infty}^{\hat{t}} f(t|\theta_-)\ \mathrm{d}t \tag{9.14}$$

Note that we have defined these limits as $>$ and $<$, whereas we used $\geq$ and $\leq$ for confidence intervals. Some authors also use $\geq$ and $\leq$ for confidence bounds. For continuous estimators, this makes no difference. However, for discrete estimators, *e.g.*, a number of events, the integral over the p.d.f. of the estimator is replaced by a sum, and then this difference is important. This will be discussed further for the Poisson p.d.f. (section 9.6).

9.4 Normal confidence intervals

The example of a normally distributed estimator has already been discussed in the introduction (section 9.1). There we saw that the situation is different depending on whether σ is or is not known.

9.4.1 σ known

If the variance, σ^2, of the estimator is known, the confidence interval is easily calculated, as shown in the introduction. Suppose we have n measurements of an exact quantity, μ, like the mass of a ball, using an apparatus of known resolution, σ_a. The estimate, $\hat{\mu} = \bar{x}$, of the quantity is then normally distributed as $N(\hat{\mu}; \mu, \sigma^2 = \sigma_a^2/n)$, and confidence intervals (equation 9.7) are computed using σ and the error function (equation 9.6). The central confidence belt is defined by straight lines corresponding to $t_\pm = \mu \pm b\sigma$, where b is the number of standard deviations corresponding to probability β.

9.4.2 σ unknown

But suppose that we do not know the resolution of the apparatus. As shown in the introduction, it is still possible to give a confidence interval, but only in terms of σ (equations 9.8 and 9.9). Since σ is not known, this is not particularly useful.

Rather, the approach is to estimate σ from the data. In the simple example of a set of n measurements of the same quantity, x, with an apparatus of constant, but unknown resolution, σ, the mean is estimated by $\hat{\mu} = \bar{x}$. As we have seen (equation 8.7), the resolution is then estimated by

$$\hat{\sigma} = s = \sqrt{\frac{n}{n-1}\overline{(x-\bar{x})^2}}$$

and the variance of the estimator is estimated by

$$V[\hat{\mu}] = \frac{s^2}{n}$$

Although $z = \frac{x-\mu}{\sigma}$ is distributed as a standard normal p.d.f., *i.e.*, z^2 is distributed as χ^2, the corresponding variable for the case of unknown σ,

$$t = \frac{x-\mu}{\hat{\sigma}} = \frac{(x-\mu)/\sigma}{\hat{\sigma}/\sigma} = \frac{z}{\hat{\sigma}/\sigma}$$

is not. Instead, it follows Student's t distribution (section 3.13). It is therefore not correct to determine a confidence interval for μ from the normal p.d.f.

Qualitatively we can understand that the confidence region will be somewhat larger with σ unknown than with σ known, since the region must also take into account fluctuations of s from the true value of σ. It can be shown[6,11,13] that the central β-confidence interval is given by

$$\mu_{\pm} = \hat{\mu} \pm T(\tfrac{1}{2}(1+\beta); n-1)\sqrt{V[\hat{\mu}]} \tag{9.15}$$

The factor T is derived from the c.d.f. of Student's t distribution. It is the value of t for which the c.d.f. is equal to $\frac{1}{2}(1+\beta)$:

$$\int_{-\infty}^{T} t(x; n-1)\,\mathrm{d}x = \frac{1}{2}(1+\beta) \tag{9.16}$$

In the case of a least squares fit to measurements y_i, all having the same (unknown) Gaussian error σ, this generalizes to

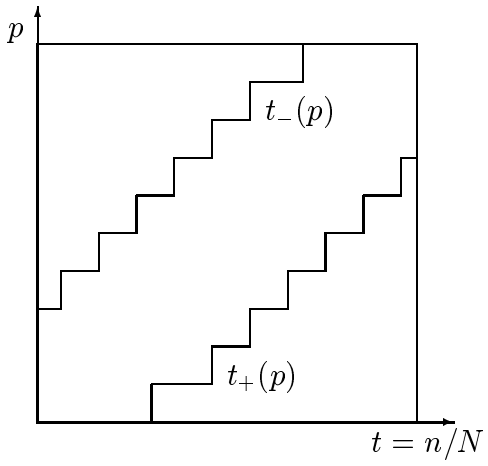
$$\theta_{i\pm} = \hat{\theta}_i \pm T(\tfrac{1}{2}(1+\beta); n-k)\sqrt{V\left[\hat{\theta}_i\right]} \tag{9.17}$$

where n is the number of points and k the number of parameters in the model.

9.5 Binomial confidence intervals

For a binomial p.d.f., $B(n; N, p)$, for which we want to estimate the parameter p, the experimental observation is the number of successes, n, in N trials. The estimator of p is then n/N.

For a given number of trials and various values of p, the confidence-belt diagram can be constructed as before using sums instead of integrals. Since the estimator of p, $t = n/N$ can take on only discrete values, the $t_-(p)$ and $t_+(p)$ curves will have a staircase-like form. Also, it will not usually be possible to find an interval for β exactly equal to say 95%. One normally then takes the next higher possible value, *i.e.*, one takes an interval with probability content slightly larger that 95%.



For example, to find the 95% central confidence interval for p, given that we observe n successes in N trials, we first find the regions $p < p_+$ and $p > p_-$ using the discrete analogues of equations 9.13 and 9.14 to find 97.5% upper and lower limits

$$P(p < p_+) = \sum_{k=n+1}^{N} B(k; N, p_+) \geq 0.975 \tag{9.18a}$$

$$P(p > p_-) = \sum_{k=0}^{n-1} B(k; N, p_-) \geq 0.975 \tag{9.18b}$$

The smallest value of p_+ and the largest value of p_- satisfying these equations give the central 95% confidence interval $[p_-, p_+]$. In other words, we find the upper and lower limits for $1 - \frac{1-\beta}{2}$ and then exclude these regions.

Using the $\geq$ in these equations rather than taking the values of p for which the equality is most nearly satisfied means that if no value gives an equality, we take the next larger value for p_+ and the next smaller value for p_-. This is known as being *conservative*. It implies that for some values of p we have **overcoverage**, which means that for some values of p the coverage probability is actually greater than the 95% that we claim, *i.e.*, that $P(p_- < p < p_+) > 0.95$ instead of $= 0.95$. This is not desirable, but the alternative would be to have undercoverage for other values of p. Since we do not know what the true value of p is—if we did know, we would not be doing the experiment—the lesser of two evils is to accept overcoverage in order to rule undercoverage completely out.

9.6 Poisson confidence intervals

9.6.1 Large N

If the number of observed events is large, the Poisson p.d.f. is well approximated by a Gaussian, and the Gaussian p.d.f. may be used to determine the confidence

interval.

9.6.2 Small N — Confidence bounds

If the number of events is smaller a confidence interval may be determined in the same way as for the binomial p.d.f.

However, for very small numbers of events one frequently prefers to state upper or lower limits. The Poisson p.d.f. is a particularly important case for such limits, since many random processes follow the Poisson p.d.f. (section 3.4).

Some experiments search for rare or 'forbidden' processes and conclude by stating upper limits for their occurrence. For example, we may search for the decay $\mu \to e\gamma$, which is forbidden in the standard theory of weak interactions, but which would be allowed in various proposed generalizations of this theory. Detection of such a decay would show that the standard theory was only an approximate theory, and the rate, *i.e.*, the fraction of μ's which decay through this mode, would help to choose among the various alternative theories. Usually such experiments find a few events which are consistent with the searched-for process, but which are not necessarily evidence for it because of possible background processes. The experimental result is then stated as an upper limit for the process.

On the other hand, a theory may predict that some process must not be zero. Then an experiment will seek to give a lower limit.

When n events have been observed, the β upper limit μ_+ for the parameter μ of the Poisson p.d.f. is, from equation 9.13, the solution of

$$\begin{aligned}\beta = P(\mu < \mu_+) &= \sum_{k=n+1}^{\infty} P(k;\mu_+) = \sum_{k=n+1}^{\infty} e^{-\mu_+}\frac{\mu_+^k}{k!} \\ &= 1 - \sum_{k=0}^{n} P(k;\mu_+) = 1 - \sum_{k=0}^{n} e^{-\mu_+}\frac{\mu_+^k}{k!}\end{aligned} \tag{9.19}$$

The solution is easily found using the fact that the sum in the right-hand side of equation 9.19 is related to the c.d.f. of the χ^2-distribution for $2(n+1)$ degrees of freedom.[4,5,46] Thus,

$$1 - \beta = \sum_{k=0}^{n} P(k;\mu_+) = P\left[\chi^2(2n+2) > 2\mu_+)\right] = \int_{2\mu_+}^{\infty} \chi^2(2n+2)\,\mathrm{d}\chi^2 \tag{9.20}$$

The upper limit μ_+ can thus be found from a table of the c.d.f. of $\chi^2(2n+2)$. Lacking a table, equation 9.19 can be solved by iteration.

Let us emphasize, perhaps unnecessarily, exactly what the upper limit means: If the true value of μ is really μ_+, the probability that a repetition of the experiment will find a number of events which is as small or smaller than n is $1-\beta$; for a true value of μ larger than μ_+, the chance is even smaller. Thus we say that we are 'β confident' that μ is less than μ_+. In making such statements, we will be right in a fraction β of the cases.

Similarly, the β lower limit, μ_-, is the solution of

$$\beta = \sum_{k=0}^{n-1} P(k;\mu_-) = \sum_{k=0}^{n-1} e^{-\mu_-} \frac{\mu_-^k}{k!} \tag{9.21}$$

which can be found from the c.d.f. of the χ^2-distribution for $2n$ degrees of freedom. Thus,

$$\beta = \sum_{k=0}^{n-1} P(k;\mu_-) = P\left[\chi^2(2n) > 2\mu_-)\right] = \int_{2\mu_-}^{\infty} \chi^2(2n)\,\mathrm{d}\chi^2 \tag{9.22}$$

The fact that it is here $2n$ degrees of freedom instead of $2(n+1)$ as for the upper limit is because there are only n terms in the sum of equation 9.22 whereas there were $n+1$ terms in the upper limit case, equation 9.20.

9.6.3 Background

As mentioned above, there is usually background to the signal. The background is also Poisson distributed. The sum of the two Poisson-distributed quantities is also Poisson distributed (section 3.7), with mean equal to the sum of the means of the signal and background, $\mu = \mu_\mathrm{s} + \mu_\mathrm{b}$. Assume that μ_b is known with negligible error. However, we do not know the actual number of background events, n_b, in our experiment. We only know that $n_\mathrm{b} \leq n$. If $\mu_\mathrm{b} + \mu_\mathrm{s}$ is large we may approximate the Poisson p.d.f. by a Gaussian and take the number of background events as $\hat{n}_\mathrm{b} \approx \mu_\mathrm{b}$. Then $\hat{\mu}_\mathrm{s} = n - \hat{n}_\mathrm{b} = n - \mu_\mathrm{b}$, with variance $V[\hat{\mu}_\mathrm{s}] = V[n] + V[\hat{n}_\mathrm{b}] = n + \mu_\mathrm{b}$.

An upper limit may be found by replacing μ_+ in equation 9.19 by $(\mu_+ + \mu_\mathrm{b})$. A lower limit may be found from equation 9.21 by a similar substitution. The results are

$$\mu_+ = \mu_+(\text{nobackground}) - \mu_\mathrm{b} \tag{9.23}$$

$$\mu_- = \mu_-(\text{nobackground}) - \mu_\mathrm{b} \tag{9.24}$$

A difficulty arises when the number of observed events is not large compared to the expected number of background events. The situation is even worse when the expected number of background events is greater than the number of events observed. For small enough n and large enough μ_b, equation 9.23 will lead to a negative upper limit. So, if you follow this procedure, you may end up saying something like "the number of whatever-I-am-trying-to-find is less than -1 with 95% confidence." To anyone not well versed in statistics this sounds like nonsense, and you probably would not want to make such a silly sounding statement. Of course, 95% confidence means that 5% of the time the statement is false. This is simply one of those times, but still it sounds silly. We will return to this point in section 9.12.

9.7 Use of the likelihood function or χ^2

We have seen in section 8.4.5 how to estimate the variance of a maximum likelihood estimator. Using the asymptotic normality of maximum likelihood estimators, we can find confidence intervals as for any normally distributed quantity with known variance (equations 9.6 and 9.7):

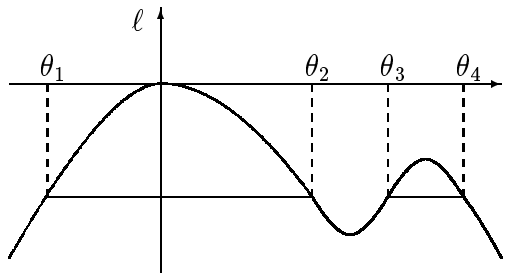
$$\hat{\theta} - d \leq \theta \leq \hat{\theta} + c \quad \text{with confidence} \quad \beta = \operatorname{erf}\left(\frac{d}{\sigma_{\hat{\theta}}}\right) - \operatorname{erf}\left(\frac{c}{\sigma_{\hat{\theta}}}\right)$$

With smaller samples it is usually most convenient to use the likelihood ratio (difference in log likelihood) to estimate the confidence interval. Then, relying on the assumption that a change of parameters would lead to a Gaussian likelihood function (*cf.* section 8.4.5), the region for which $\ell > \ell_{\max} - a^2/2$, or equivalently (*cf.* section 8.5.1) $\chi^2 < {\chi^2}_{\min} + a^2$, corresponds to a probability content

$$\beta = \int_{-a}^{+a} N(z;0,1)\,\mathrm{d}z = \operatorname{erf}(a) - \operatorname{erf}(-a)$$

In 'pathological' cases, *i.e.*, cases where there is more than one maximum, as pictured here, the situation is less clear. Applying the above procedure would lead to disconnected intervals, whereas the interval for the transformed parameter would give a single interval. It is sometimes said that it is nevertheless correct to state a β confidence interval as

$$\theta_1 \leq \theta \leq \theta_2 \quad \text{or} \quad \theta_3 \leq \theta \leq \theta_4$$



However, this statement seems to be the result of confusing confidence intervals with fiducial intervals (section 9.8). Be that as it may, the usefulness of such intervals is rather dubious, and in any case gives an incomplete picture of the situation. One should certainly give more details than just stating these intervals.

The application of other methods of estimating the variance of $\hat{\theta}$ to finding confidence intervals for finite samples is discussed in some detail by Eadie *et al.*[4] and James[5].

9.8 Fiducial intervals

Confidence intervals, as developed by Neyman and discussed in the previous sections, use a fully frequentist approach to probability. R. A. Fisher, a few years earlier, had followed a somewhat different, also frequentist, approach to interval estimation[24]. His intervals are called **fiducial intervals**. A third approach is the much older Bayesian one, which will be presented in the next section.

Fisher's concept of information (section 8.2.5) is intimately related to the likelihood function. So too is his fiducial interval.

In section 8.4.2 we saw that asymptotically the likelihood function $\mathcal{L}(\underline{x};\underline{\theta})$ becomes (under rather general assumptions) a Gaussian function of the parameters $\underline{\theta}$. This does not mean (as we have repeatedly emphasized) that $\mathcal{L}$ is a p.d.f. for $\underline{\theta}$. That only happens in a Bayesian interpretation, which we are not making here. Recall that the principle of maximum likelihood, *i.e.*, that the best estimate of $\underline{\theta}$ is that value of $\underline{\theta}$ for which the likelihood function is a maximum, was not derived, but assumed on intuitive grounds. In the same way we go now a step further and assume, again intuitively, that $\mathcal{L}$ represents our level of credence in a value of $\underline{\theta}$. A fiducial interval for a degree of credence β is defined as an interval $[\theta_1, \theta_2]$ such that

$$\beta = \frac{\int_{\theta_1}^{\theta_2} \mathcal{L} \, \mathrm{d}\theta}{\int_{-\infty}^{+\infty} \mathcal{L} \, \mathrm{d}\theta} \tag{9.25}$$

This procedure is supported by the connection we have seen (section 8.4.2) between the asymptotic Gaussian shape of $\mathcal{L}$ and the variance of the maximum likelihood estimator. And just as with the maximum likelihood method, the attractiveness of fiducial intervals is based on asymptotic properties.

As with confidence intervals, a supplementary criterium, such as a central interval, is needed in addition to equation 9.25 to uniquely define a fiducial interval.

Often the confidence interval and fiducial interval approaches lead to the same interval. However, the approach, and hence the meaning, is different. The confidence approach says that if we assert that the true value is in a 95% interval we will be right 95% of the time. However, in the fiducial approach the same assertion means that we are 95% sure that we are right *this* time. This shift in emphasis is the same as in the meaning of the likelihood function itself: We can regard $\mathcal{L}(x;\theta)$ as an elementary probability in which θ is fixed and x varies, *i.e.*, as the p.d.f. for the r.v. X. On the other hand, we can regard it as a likelihood in which x is fixed and θ varies, as is done in the maximum likelihood method. Similarly, in interval estimation, we can regard θ as a constant and set up containing intervals which are random variables (the confidence interval approach); or we can regard the observations as fixed and set up intervals based on some undefined intensity of belief in the values of the parameter generating the observations (the fiducial interval approach).

Today, fiducial intervals are seldom used, since they lack a firm mathematical basis. If one is a frequentist, one generally prefers confidence intervals.

9.9 Credible (Bayesian) intervals

Confidence intervals are based on the frequentist interpretation of probability and are statements about the probability of experimental results. Fiducial intervals are also based on the frequentist interpretation of probability (the parameters $\underline{\theta}$ have fixed true values) but represent our credence (or belief) about the values of the

parameters. However, we may prefer to use Bayesian probability. In this case we can construct intervals, $[a, b]$, called **credible intervals**, Bayesian confidence intervals, or simply **Bayesian intervals**, such that β is the probability that parameter θ is in the interval:

$$\beta = P(a \leq \theta \leq b) = \int_a^b f(\theta|x)\ \mathrm{d}\theta \tag{9.26}$$

where $f(\theta|x)$ is the Bayesian posterior p.d.f. As with confidence and fiducial intervals, supplementary conditions, such as centrality, are needed to uniquely specify the interval. We have seen in section 8.4.5 that, assuming Bayes' postulate, $f(\theta|x)$ is just the likelihood function $\mathcal{L}(x;\theta)$, apart, perhaps, from normalization.

9.10 Discussion of intervals

We have presented three approaches to interval estimation: confidence intervals, fiducial intervals, and credible (or Bayesian) intervals. In cases where the likelihood function is a Gaussian function of the parameters, as is usually true asymptotically, these approaches (with a suitable choice of prior in the Bayesian case) all lead to the same interval. Though this is comforting, we must realize that in less ideal circumstances the intervals given by the different approaches may be different. This does not mean that any of the approaches is wrong, but rather that they are answering different questions or making different assumptions.

The virtue of the confidence interval approach is its firm grounding in frequentist probability. The Bayesian approach is also firmly grounded, but loses something in objectivity by its subjective Bayesian interpretation of probability as a degree of belief. Further, it suffers from its need for an arbitrary choice of prior probability (Bayes' postulate). The fiducial approach is well-grounded only where its results are identical to the other approaches. Extension to other cases is more a question of intuition.

We thus are inclined to prefer the confidence interval approach even though it is a very complicated procedure compared to the other approaches. However, the confidence interval approach is unable to incorporate prior information, as we will see in the next section.

9.11 Measurement of a bounded quantity

Let us return to the example in the introduction (section 9.1). A dish is weighed, a sample is placed on the dish and the combination is weighed, and then the mass of the sample is estimated by subtracting the mass of the dish from the mass of the dish plus sample. If the mass of the sample is smaller than or comparable to the resolution of the balance, the confidence interval $[-\infty, 0]$ will have a non-negligible probability content. This is clearly ridiculous and comes about because we have not made use of our knowledge that the mass must be positive. Such a situation

can also occur when we must subtract a number of background events from the observed number of events to find the number of events in the signal; a number of events also can not be negative.

The problem is how to incorporate this constraint (or prior knowledge) into the confidence interval. In the confidence interval approach there is no way to do this. The best we can do is to choose an interval which does not contain the forbidden region (< 0 in our example). Consider the figure showing confidence belts in section 9.2. Suppose that we know that $\theta_{\rm t} > \theta_{\rm min}$. We can think of several alternatives to the interval $[\theta_-, \theta_+]$ when $\theta_- < \theta_+$:

1. $[\theta_{\rm min}, \theta_+]$. But this is the same interval we would have found using a confidence belt with t_+ shifted upwards such that the t_+ curve passes through the point $(\theta_{\rm min}, \hat{t}\,)$. This confidence belt clearly has a smaller β. This places us in the position of stating the same confidence for two intervals, the one completely contained in, and smaller than, the other.

2. $[\theta_{\rm min}, \theta_+^{''}]$, where $\theta_+^{''}$ is the solution of $t_{\rm min}(\theta) = t_{\rm min}$, with $t_{\rm min} = t_+(\theta_{\rm min})$. This is the interval we would have stated had we found $\hat{t} = t_+(\theta_{\rm min})$. So, apparently the fact that we found a lower value of $\hat{t}$ does not mean anything—*any* value of $\hat{t}$ smaller than $t_+(\theta_{\rm min})$ leads to the *same* confidence interval! This procedure is clearly unsatisfactory.

3. $[\theta_{\rm min}, \theta_+^{'}]$, where $\theta_+^{'}$ is determined from a new confidence belt constructed such that the t_+ curve passes through the point $(\theta_{\rm min}, \hat{t}\,)$. The t_- curve is taken as that curve which together with this new t_+ curve gives the required β. This approach seems better than the previous two. However, it is still unsatisfactory since it relies on the measurement to define the confidence belt.

The situation is even worse if not only $\theta_-(\hat{t}) < \theta_{\rm min}$ but also $\theta_+(\hat{t}) > \theta_{\rm max}$. Then we find ourselves in the absurd situation of, *e.g.*, stating the conclusion of our experiment as $-0.2 < \theta < 1.2$ with 95% confidence when we *know* that $0 < \theta < 1$—we are only 95% confident that θ is within its physical limits! The best procedure to follow has been the subject of much interest lately among high energy physicists, particularly those trying to measure the mass of the neutrino and those searching for hypothetical new particles. The most reasonable procedure seems to be[47] that of Feldman and Cousins,[48] who rediscovered a prescription previously given by Kendall and Stuart.[11]

On the other hand, in the fiducial approach physical boundaries are easily incorporated. The likelihood function is simply set to zero for unphysical values of the parameters and renormalized. Equation 9.25 is thus replaced by

$$\beta = \frac{\int_{\theta_1}^{\theta_2} \mathcal{L}\ {\rm d}\theta}{\int_{\theta_{\rm min}}^{\theta_{\rm max}} \mathcal{L}\ {\rm d}\theta} \tag{9.27}$$

Also the Bayesian approach has no difficulty in incorporating the physical limits. They are naturally imposed on the prior probability. If the prior probability is uniform within the physical limits, the result is the same interval as in the fiducial approach (equation 9.27).

Note, however, that in order to combine with the results of other experiments, the (nonphysical) estimate and its variance should be stated, as well as the confidence interval. This, in fact, should also be done for quantities which are not bounded.

9.12 Upper limit on the mean of a Poisson p.d.f. with background

In section 9.6.3 we introduced the problem of measuring an upper limit on the number of (Poisson distributed) events for a particular process in the presence of background. This is related to the problems of the previous section. The number of events can not be negative; it is a bounded quantity. Within the classical confidence limit approach, the most reasonable procedure here too is that of Feldman and Cousins[48].

Another approach is to determine an upper limit by an extension of the argument of section 9.6.[46,49] As in that section, let n be the number of events observed, n_b the expected number of background events, and μ_+ the upper limit on μ_s. Then μ_+ is that value of μ_s such that any random repetition of the current experiment would, if μ_s actually equals μ_+, result in more than n events and would also have $n_\mathrm{b} \leq n$, all with probability β. Thus, in equation 9.19 the sum, which is the probability of $\leq n$ events given $\mu = \mu_+$, is replaced by the same probability given $\mu = \mu_\mathrm{b} + \mu_+$ normalized to the probability that $n_\mathrm{b} \leq n$.

$$\begin{aligned}\beta &= 1 - \frac{P(\leq n \text{ events})}{P(\leq n \text{ background events})} \\ &= 1 - \frac{e^{-(\mu_{\mu_+ + \mathrm{b}})} \sum_{k=0}^{n} \frac{(\mu_+ + \mu_\mathrm{b})^k}{k!}}{e^{-\mu_\mathrm{b}} \sum_{k=0}^{n} \frac{\mu_\mathrm{b}^k}{k!}} \end{aligned} \tag{9.28}$$

This equation must be solved for μ_+. In practice this is best done numerically, adjusting μ_+ until the desired β is obtained. However, to incorporate the probability that $n_\mathrm{b} \leq n$, we have been Bayesian. The result is thus a credible upper limit rather than a classical upper limit.

When μ_b is not known to a negligible error, the same approach can be used. However, we must integrate over the p.d.f. for n_b. It is most convenient to use a Monte Carlo technique. We generate a sample of Monte Carlo experiments taking μ_b randomly distributed according to our knowledge of μ_b (usually normally) and with a fixed μ_s. Experiments with $n_\mathrm{b} > n$ are rejected. The sum in equation 9.19 or 9.21 is then estimated by the fraction of remaining Monte Carlo experiments satisfying

the corresponding probability. The process is repeated for different values of μ_s until the desired value of β is found.[50]

"Which way ought I to go to get from here?"
"That depends a good deal on where you want to get to," said the Cat.
"I don't much care where—" said Alice.
"Then it doesn't matter which way you go," said the Cat.
—Lewis Carroll, "Alice in Wonderland"

Chapter 10

Hypothesis testing

10.1 Introduction

In chapter 8 we were concerned with estimating parameters of a p.d.f. using a statistic calculated from observations assumed to be distributed according to that p.d.f. In chapter 9 we sought an interval which we were confident (to some specified degree) contained the true value of the parameter. In this chapter we will be concerned with whether some previously designated value of the parameter is compatible with the observation, or even whether the assumed p.d.f. is compatible. In a sense, this latter question logically precedes the estimation of the value of a parameter, since if the p.d.f. is incompatible with the data there is little sense in trying to estimate its parameters.

When the hypothesis under test concerns the value of a parameter, the problems of hypothesis testing and parameter testing are related and techniques of parameter estimation will lead to analogous testing procedures. If little is known about the value of a parameter, you will want to estimate it. However, if a theory predicts it to have a certain value, you may prefer to test whether the data are compatible with the predicted value. In either case you should be clear which you are doing. That others are often confused about this is no excuse.

10.2 Basic concepts

The question here is thus one of hypothesis testing. We make some hypothesis and want to use experimental observations to test whether it is correct. Not all scientific hypotheses can be tested statistically. For instance, the hypothesis that

every particle in the universe attracts every other particle can not be tested statistically. Statistical hypotheses concern the distributions of observable random variables. Suppose we have N such observations. We denote them by a vector $\underline{x}$ in an N-dimensional space, Ω, called the sample space (section 2.1.2), which is the space of all possible values of $\underline{x}$, *i.e.*, the space of all possible results of an experiment. A statistically testable hypothesis is one which concerns the probability of a particular observation $\underline{X}$, $P(\underline{X} \in \Omega)$.

Suppose that $\underline{x}$ consists of a number of independent measurements of a r.v., x_i. Let us give four examples of statistical hypotheses concerning $\underline{x}$:

1. The x_i are distributed normally with particular values of μ and σ.

2. The x_i are distributed normally with a particular value of μ.

3. The x_i are distributed normally.

4. The results of two experiments, x_{1i} and x_{2i} are distributed identically.

Each of these hypotheses says something about the distribution of probability over the sample space and is hence statistically testable by comparison with observations.

Examples 1 and 2 specify a p.d.f. and certain values for one or both of its parameters. Such hypotheses are called **parametric** hypotheses. Example 3 specifies the form of the p.d.f., but none of its parameters, and example 4 does not even specify the form of the p.d.f. These are examples of **non-parametric hypotheses**, *i.e.*, no parameter is specified in the hypothesis. We shall mainly concentrate on parametric hypotheses, leaving non-parametric hypotheses to section 10.7.

Examples 1 and 2 differ in that 1 specifies all of the parameters of the p.d.f., whereas 2 specifies only a subset of the parameters. When all of the parameters are specified the hypothesis is termed simple; otherwise composite. If the p.d.f. has n parameters, we can define an n-dimensional parameter space. A **simple** hypothesis selects a unique point in this space. A **composite** hypothesis selects a subspace containing more than one point. The number of parameters specified exactly by the hypothesis is called the number of constraints. The number of unspecified parameters is called the number of degrees of freedom of the hypothesis. Note the similarity of terminology with that used in parameter estimation:

	Parameter Estimation	Hypothesis Testing
n = number of	observations	parameters
k = number of	parameters to be estimated	parameters specified by the hypothesis (constraints)
$n - k$ = number of	degrees of freedom	

To test an hypothesis on the basis of a random sample of observations, we must divide the sample space Ω into two subspaces. If the observation $\underline{x}$ lies in one of these subspaces, call it ω, we shall reject the hypothesis; if $\underline{x}$ lies in the complementary region, $\omega^* = \Omega - \omega$, we shall accept the hypothesis. The subspace ω is called the **critical region** of the test, and ω^* is called the **acceptance region**.

A few words are in order regarding this terminology. In science we can never completely reject or accept an hypothesis. Nevertheless, the words "reject" and "accept" are in common usage. They should be understood as meaning "the observations are unfavorable" or "favorable" to the hypothesis. Since acceptance or rejection is never certain, it is clear that we also need to be able to state our degree of confidence in acceptance or rejection, just as when constructing confidence intervals we did so with a specified confidence.

The hypothesis being tested is generally designated H_0 and is called the **null hypothesis**. For the time being, we will assume that H_0 is a simple hypothesis, *i.e.*, it specifies the p.d.f. completely. We can then calculate the probability that a random observation will fall in the critical region, and we can choose this region such that this probability is equal to some pre-chosen value, α,

$$P(\underline{x} \in \omega | H_0) = \alpha \tag{10.1}$$

This value α is thus the probability of rejecting H_0 if H_0 is true. It is called the **size** of the test or the **level of significance**, although this latter term can be misleading. For a discrete p.d.f. the possible values of α will also be discrete, while for a continuous p.d.f. any value of α is possible.

In general, there will be many, often an infinity, of subspaces ω of the same size α. Which of them should we use? In other words, which of all possible observations should we regard as favoring and which as disfavoring H_0?

To decide which subspace to take as ω, we need to know what the alternatives are. It is perfectly possible that an observation is unlikely under H_0 but even more unlikely under an alternative hypothesis. Forced to choose between the two we would not want to reject H_0. Thus whether we accept or reject H_0 depends on what the alternative hypothesis, usually designated H_1, is.

It should now be clear that a critical region (or, synonymously, a test) must be judged by its properties both when H_0 is true and when H_0 is false. We want to accept H_0 if it is true and reject it if it is false. Our decision, *i.e.*, acceptance or rejection, can be wrong in two ways:

1. **Error of the first kind**, or **loss**, or **false negative**: H_0 is true, but we reject it.

2. **Error of the second kind**, or **contamination**, or **false positive**: H_0 is false, but we accept it.

The probability of making an error of the first kind is equal to the size of the critical region, α. The probability of making an error of the second kind depends

on the alternative hypothesis and is denoted* by β:

$$P(\underline{x} \in \omega^* | H_1) = \beta \tag{10.2}$$

The complementary probability,

$$P(\underline{x} \in \omega | H_1) = 1 - \beta \tag{10.3}$$

is called the **power** of the test of H_0 against H_1. The specification of H_1 when giving the power is clearly essential since β depends on H_1.

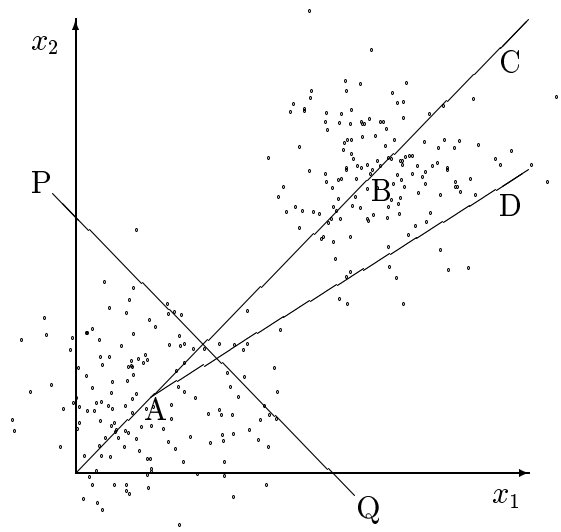
Clearly, we would like a test to have small values of both α and β. However, it is usually a trade-off: decreasing α frequently increases β and *vice versa*. Let us consider two examples where H_0 and H_1 are both simple hypotheses.

Example 1. Consider H_0 and H_1 both of which hypothesize that the r.v. X is normally distributed with standard deviation σ. The difference between the hypotheses lies in the value of μ. For H_0 it is μ_0 and for H_1 it is μ_1. We make two independent observations x_1 and x_2 to test H_0 against H_1.

The two observations can be represented by a point in Ω, which is a plane having x_1 and x_2 as axes. The joint p.d.f. under H_0 is a bivariate normal distribution centered at the point A, *i.e.*, at $x_1 = x_2 = \mu_0$. The density of points about A in the figure is meant to represent this p.d.f. Under H_1 the p.d.f. is the same except that it is centered at the point B, $x_1 = x_2 = \mu_1$.

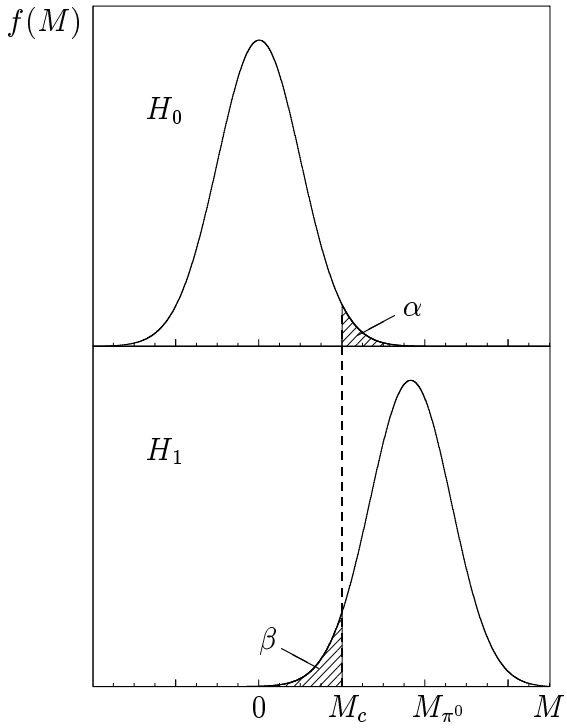
A test of H_0 could be made by defining ω by the line PQ with H_0 to be rejected if the point representing the observations lies above the line PQ. Another possible critical region is that between the lines CA and AD. Both of these regions have the same probability under H_0 and hence the same size, α.

However, the values of β are much different. The first test will almost always reject H_0 when H_1 is true, while the second test will often wrongly accept H_0. Thus β is much larger for the second test, and hence the power of the first test is larger. It should be obvious that the more powerful test is preferable.



Example 2. In the previous example the sample space was only two dimensions. When the dimensionality is larger, it is inconvenient to formulate the test in terms of the complete sample space. Rather, a small number

*The symbols α and β are used by most authors for the probabilities of errors of the first and



(frequently one) of test statistics is defined and the test is formulated in terms of them. In fact, as we shall later see, in some cases a single test statistic provides the best test. Recall that a statistic is a function only of the observations and does not depend on any assumptions about the p.d.f.

Suppose that we want to distinguish K^-p elastic scattering events from inelastic scattering events where a π^0 is produced. The hypotheses are then

H_0: $K^-p \to K^-p$
H_1: $K^-p \to K^-p\pi^0$

If the experiment measures the momenta and energies of charged particles but does not detect neutral particles, a convenient test statistic is the missing mass, the mass of the neutral system in the final state. This is easily calculated from the energies and momenta of the initial- and final-state charged particles. The true value of the missing mass is $M = 0$ under H_0, and $M = 135\ \mathrm{MeV}/c^2$ under H_1. We can choose a critical region $M > M_c$. The corresponding loss and contamination are shown in the figure. The choice of M_c will be governed by balancing our interest in both small loss and small contamination.

Note that the actual contamination in our sample of elastic events depends on the *a priori* abundance of inelastic events produced. If this is small compared to that of elastic events, we can tolerate a large value of β.

10.3 Properties of tests

In parameter estimation we were faced with the problem of choosing the best estimator. Here a similar situation arises: we seek the best test. To aid us, we examine some properties of tests.

10.3.1 Size

In the previous section we defined (equation 10.1) the size, α, of a test as the probability that the test would reject the null hypothesis when it is true. If H_0 is a simple hypothesis, the size of a test can be calculated. In other cases it is not

second kind. However, some authors use $1 - \beta$ where we use β.

always possible. Clearly a test of unknown size is worthless.

10.3.2 Power

We have defined (equation 10.3) the power, $1-\beta$, of a test of one hypothesis H_0 against another hypothesis H_1 as the probability that the test would reject H_0 when H_1 is true. If H_1 is a simple hypothesis, the power of a test can be calculated. If H_1 is composite, the power can still be calculated, but is in general no longer a constant but a function of the parameters.

Suppose that H_0 and H_1 specify the same p.d.f., the difference being the value of the parameter θ:

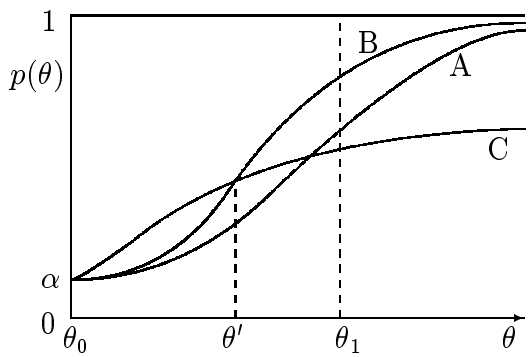
$$\begin{aligned} H_0&: \quad \theta = \theta_0 \\ H_1&: \quad \theta \neq \theta_0 \end{aligned}$$

The contamination, β, is then a function of θ, as is the power:

$$p(\theta) = 1 - \beta(\theta) \tag{10.4}$$

Note that by definition, $p(\theta_0) = 1 - \beta(\theta_0) = \alpha$.

Tests may then be compared on the basis of their power function. If H_0 and H_1 are both simple, the best test of size (at significance level) α is the test with maximum power at $\theta = \theta_1$, the value specified by H_1. In the figure, test B has the largest power for $\theta > \theta'$ and in particular at $\theta = \theta_1$, whereas test C is more powerful for $\theta_0 < \theta < \theta'$.



If for a given value of θ a test is at least as powerful as any other possible test of the same size, it is called a **most powerful (MP) test** at that value of θ, and its critical region is called a **best critical region (BCR)**. A test which is most powerful for all regions of θ under consideration is called a **uniformly most powerful (UMP) test**. Clearly, if a test is MP at θ_1 and the test is independent of θ_1, then it is UMP. It is frequently not possible to find an UMP test, although we will see in section 10.4.1 that if H_0 and H_1 are both simple hypotheses, then an UMP test always exists. Unfortunately, in real life an UMP test does not usually exist. An UMP test which is also unbiased (section 10.3.4) is called **UMPU**.

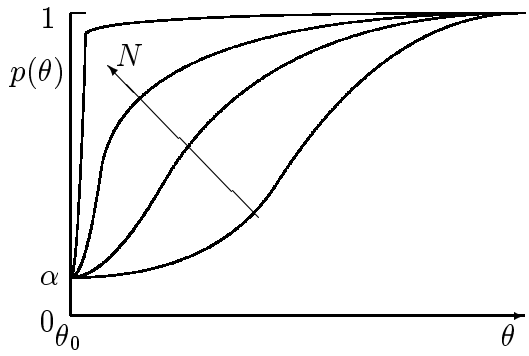
10.3.3 Consistency

A highly desirable property of a test is that, as the number of observations increases, it should distinguish better between the hypotheses. A test is termed

consistent if the power tends to unity as the number of observations increases:

$$\lim_{N\to\infty} P(\underline{x} \in \omega | H_1) = 1$$

where $\underline{x}$ is the set of N observations and ω is the critical region under H_0. The power function thus tends to a step function as $N \to \infty$.



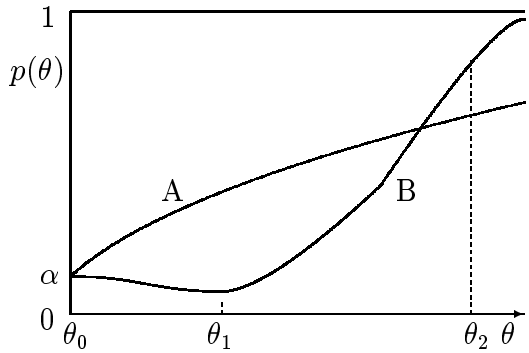
10.3.4 Bias

A test is biased if the power function is smaller at a value of θ corresponding to H_1 than at the value, θ_0, specified by H_0, *i.e.*, when there exists a value θ for which

$$p(\theta) = 1 - \beta(\theta) < \alpha \quad , \quad \theta \neq \theta_0$$

An example is test B at $\theta = \theta_1$ in the figure. In such a case the chance of accepting H_0 is greater when $\theta = \theta_1$ than when $\theta = \theta_0$, which means we are more likely to accept H_0 when it is false than when it is true. Such a test is clearly undesirable in general.

In some situations it may be preferable to use a biased test. For example, test B may be chosen rather than test A if it is particularly important to be able to discriminate against $\theta = \theta_2$, where test B is more powerful than A. However, in so doing all discrimination between H_0 and H_1 in the region of θ_1 is lost.



The definition of a biased test can be formulated in a way which is also applicable for composite hypotheses. Let H_0 specify that θ is in some interval θ_0. Then a test is unbiased if

$$P(\underline{x} \in \omega | \theta) \begin{cases} \leq \alpha, & \text{for all } \theta \in \theta_0 \\ \geq \alpha, & \text{for all } \theta \notin \theta_0 \end{cases}$$

In real life it is usually possible to find an unbiased test.

10.3.5 Distribution-free tests

Most of the time we do not invent our own tests, but instead use some standard test. To be 'standard', the distribution of the test statistic, and hence the size of

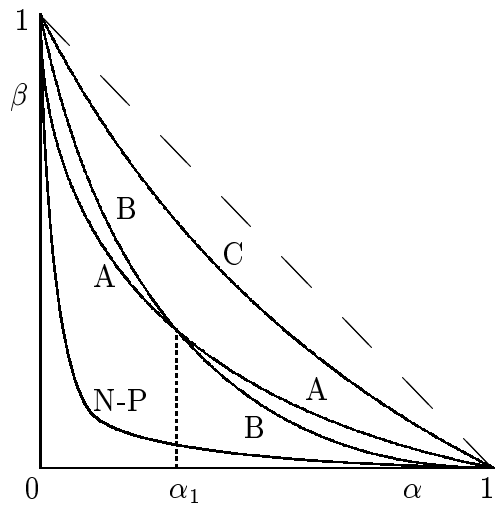
the critical region, must be independent of the p.d.f. specified by H_0. It can only depend on whether H_0 is true. Such a test is called distribution-free. An example is the well-known Pearson's χ^2 test, which we shall meet shortly.

It should be emphasized that it is only the size or level of significance of the test which does not depend on the distributions specified in the hypotheses. Other properties of the test do depend on the p.d.f.'s. In particular, the power will depend on the p.d.f. specified in H_1.

10.3.6 Choice of a test

Traditionally, the choice of a test is done by first specifying the loss α and then choosing the test on the basis of the power. This procedure assumes that the risk of an error of the first kind (loss) is a given constant, and that one only has to minimize the risk of an error of the second kind (contamination).

However, this is frequently not the case. We want to have both kinds of errors as small as possible. It is then advantageous to take both α and β as variables in comparing the tests. Assume, for simplicity, that both H_0 and H_1 are simple, specifying θ_0 and θ_1, respectively. Then, for a given test and a given value of $\alpha = p(\theta_0)$, one can determine $\beta = 1 - p(\theta_1)$. Repeating for different values of α, a curve giving β as a function of α can be constructed, as shown in the figure.



The dashed line in the figure corresponds to $1 - \beta = \alpha$, so that all unbiased test curves will lie entirely below this line, passing through the points (1,0) and (0,1). Since we desire to have both α and β small, test C in the figure is clearly inferior to the others for all values of α and β. If both H_0 and H_1 are simple, there always exists a test (the Neyman-Pearson test, *cf.* section 10.4.1) which is at least as good as any other test for all α and β. If this test is too complicated, or in the case of composite hypotheses, one could be in the position of choosing, for example, between tests A and B. Clearly, test A should be chosen for $\alpha < \alpha_1$ and test B for $\alpha > \alpha_1$.

If the hypotheses are composite, the figure can become a multidimensional diagram with new axes corresponding to θ or to other unspecified parameters. Or each test can be represented by a family of curves in the α-β plane.

A minor difficulty arises when discrete distributions are involved, since only a discrete set of α's are then available, and the α-β curves are discontinuous.

The above techniques allow one to choose the best test. Whether it is good

enough depends on the cost (in terms of such things as time and money) of making an error, *i.e.*, a wrong decision.

10.4 Parametric tests

10.4.1 Simple Hypotheses

The Neyman-Pearson test

When both H_0 and H_1 are simple hypotheses, the problem of finding the best critical region (BCR), or most powerful (MP) test, of size α is particularly straightforward, as was shown by Neyman and Pearson.[51]

We suppose that the r.v. $\underline{x}$ is distributed under H_0 as $f(\underline{x};\theta_0)$ and under H_1 as $g(\underline{x};\theta_1)$. Then equations 10.1 and 10.3 can be written

$$P(\underline{x} \in \omega_\alpha \mid H_0) = \int_{\omega_\alpha} f(\underline{x};\theta_0)\ \mathrm{d}\underline{x} = \alpha \tag{10.5}$$

$$P(\underline{x} \in \omega_\alpha \mid H_1) = \int_{\omega_\alpha} g(\underline{x};\theta_1)\ \mathrm{d}\underline{x} = 1-\beta \tag{10.6}$$

We want to find the critical region ω_α which, for a given value of α, maximizes $1-\beta$. Rewriting equation 10.6, we have

$$\begin{aligned} 1-\beta &= \int_{\omega_\alpha} \frac{g(\underline{x};\theta_1)}{f(\underline{x};\theta_0)} f(\underline{x};\theta_0)\ \mathrm{d}\underline{x} \\ &= E_{\omega_\alpha}\left[\left.\frac{g(\underline{x};\theta_1)}{f(\underline{x};\theta_0)}\right| H_0\right] \end{aligned}$$

which is the expectation of $g(\underline{x};\theta_1)/f(\underline{x};\theta_0)$ in the region ω_α assuming that H_0 is true. This will be maximal if we choose the region ω_α as that region containing the points $\underline{x}$ for which this ratio is the largest. In other words, we order the points $\underline{x}$ according to this ratio and add these points to ω until ω has reached the size α. The BCR thus consists of the points $\underline{x}$ satisfying

$$\frac{f(\underline{x};\theta_0)}{g(\underline{x};\theta_1)} \leq c_\alpha$$

where c_α is chosen such that ω_α is of size α (equation 10.5).

This ratio is, for a given set of data, just the ratio of the likelihood functions, which is known as the likelihood ratio. We therefore use the test statistic

$$\lambda = \frac{\mathcal{L}(\underline{x}|H_0)}{\mathcal{L}(\underline{x}|H_1)} \tag{10.7}$$

and

$$\begin{aligned} \text{reject } H_0 \text{ if } \lambda &\leq c_\alpha \\ \text{accept } H_0 \text{ if } \lambda &> c_\alpha \end{aligned}$$

This is known as the Neyman-Pearson test.

An Example

As an example, consider the normal distribution treated in example 1 of section 10.2. Both H_0 and H_1 hypothesize a normal p.d.f. of the same variance, but different means, μ_0 under H_0 and μ_1 under H_1. The variance is, for both hypotheses, specified as σ^2. The case where the variance is not specified is treated in section 10.4.3. The likelihood function under H_i for n observations is then

$$\begin{aligned}\mathcal{L}(\underline{x}|H_i) &= (2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j-\mu_i)^2}{\sigma^2}\right] \\ &= (2\pi)^{-n/2} \exp\left[-\frac{n}{2\sigma^2}\left\{s^2+(\bar{x}-\mu_i)^2\right\}\right]\end{aligned}$$

where $\bar{x}$ and s^2 are the sample mean and sample variance, respectively. Hence, our test statistic is (equation 10.7)

$$\begin{aligned}\lambda = \frac{\mathcal{L}(\underline{x}|H_0)}{\mathcal{L}(\underline{x}|H_1)} &= \exp\left[\frac{n}{2\sigma^2}\left\{(\bar{x}-\mu_1)^2-(\bar{x}-\mu_0)^2\right\}\right] \\ &= \exp\left[\frac{n}{2\sigma^2}\left\{2\bar{x}(\mu_0-\mu_1)+(\mu_1^2-\mu_0^2)\right\}\right]\end{aligned}$$

and the BCR is defined by $\lambda \le c_\alpha$ or

$$\bar{x}(\mu_0-\mu_1)+\frac{1}{2}(\mu_1^2-\mu_0^2) \le \frac{\sigma^2}{n}\ln c_\alpha$$

which becomes

$$\bar{x} \ge \frac{1}{2}(\mu_1+\mu_0) - \frac{\sigma^2}{n}\frac{\ln c_\alpha}{\mu_1-\mu_0} \qquad \text{if } \mu_1 > \mu_0 \tag{10.8}$$

$$\bar{x} \le \frac{1}{2}(\mu_1+\mu_0) + \frac{\sigma^2}{n}\frac{\ln c_\alpha}{\mu_0-\mu_1} \qquad \text{if } \mu_1 < \mu_0 \tag{10.9}$$

Thus we see that the BCR is determined by the value of the sample mean. This should not surprise us if we recall that $\bar{x}$ was an efficient estimator of μ (section 8.2.7).

In applying the test, we reject H_0 if $\mu_1 > \mu_0$ and $\bar{x}$ is above a certain critical value (equation 10.8), or if $\mu_1 < \mu_0$ and $\bar{x}$ is below a certain critical value (equation 10.9).

To find this critical value, we recall that $\bar{x}$ itself is a normally distributed r.v. with mean μ and variance σ^2/n. (This is the result of the central limit theorem, but when the p.d.f. for x is normal, it is an exact result for all n.) We will treat the case of $\mu_1 > \mu_0$ and leave the other case as an exercise for the reader.

For $\mu_1 > \mu_0$, the right-hand side of equation 10.8 is just $\bar{x}_\alpha$ given by

$$\sqrt{\frac{n}{2\pi\sigma^2}}\int_{\bar{x}_\alpha}^{\infty} \exp\left[-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2\right]\,\mathrm{d}\bar{x} = \alpha$$

Transforming to a standard normal variable,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{10.10}$$

we can rewrite this in terms of the standard normal integral, which is given by the error function (section 3.7):

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{\infty} e^{-z^2/2}\,\mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z_\alpha} e^{-z^2/2}\,\mathrm{d}z = \mathrm{erf}(-z_\alpha) \tag{10.11}$$

For example, for $\alpha = 0.05$ we find in a table that $z_\alpha = 1.645$. For $\mu_0 = 2$, $\sigma = 1$, and $n = 25$, this value of z_α inserted in equation 10.10 yields $\bar{x}_\alpha = 2.33$. Then if $\bar{x} > 2.33$, we reject H_0 with a level of significance of 5%.

The power of the test can also be easily computed in this example. It is

$$\sqrt{\frac{n}{2\pi\sigma^2}} \int_{\bar{x}_\alpha}^{\infty} \exp\left[-\frac{n}{2\sigma^2}(\bar{x} - \mu_1)^2\right]\,\mathrm{d}\bar{x} = 1 - \beta$$

which, in terms of the error function and the z_α defined above, can be written

$$1 - \beta = 1 - \mathrm{erf}\left(\frac{\sqrt{n}}{\sigma}(\mu_0 - \mu_1) + z_\alpha\right) = \mathrm{erf}\left(\frac{\sqrt{n}}{\sigma}(\mu_1 - \mu_0) - z_\alpha\right) \tag{10.12}$$

We see that the power increases monotonically with both n and $\mu_1 - \mu_0$.

10.4.2 Simple H_0 and composite H_1

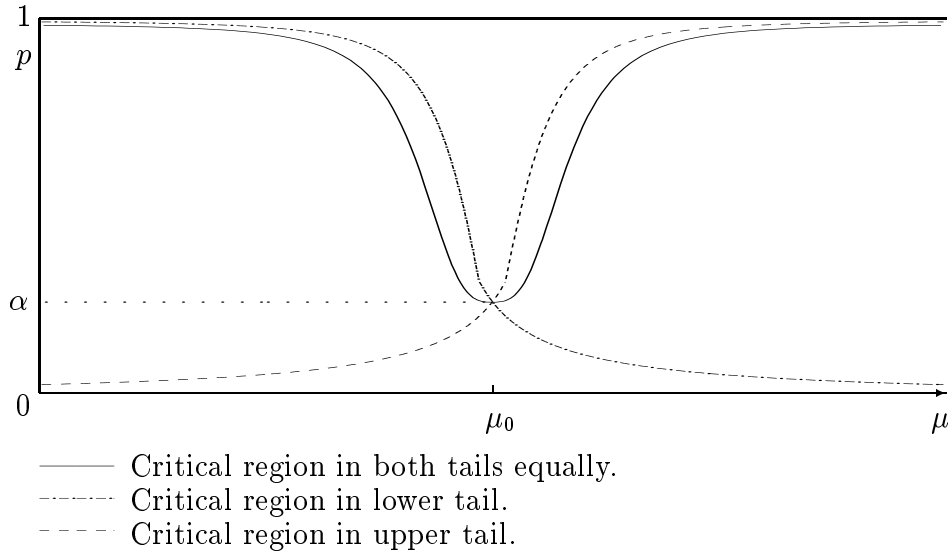
In the previous section we have seen how to construct the best test between two simple hypotheses. Unfortunately, no such generally optimal method exists when H_0 and/or H_1 is not simple.

Suppose that we want to test a simple H_0 against a composite H_1. Let us first treat an H_1 which is just a collection of simple hypotheses, *e.g.*, under H_0 $\theta = \theta_0$, and under H_1 $\theta = \theta_1$ or θ_2 or θ_3 or $\ldots\theta_n$. We could imagine testing H_0 against each of these alternatives separately using a MP test as found in section 10.4.1. However, this would lead in general to a different critical region in each case and most likely to acceptance of H_0 in some cases and rejection in others. We are therefore led to inquire whether there exists one BCR for all the alternative values. A test using such a BCR would be UMP.

UMP test for the exponential family

Unfortunately, an UMP test does not generally exist. One important case where an UMP test does exist is when the p.d.f. of H_0 and H_1 is of the exponential family (section 8.2.7), but then only for 'one-sided' tests.[4,5] We illustrate this for the Gaussian p.d.f. from our results of section 10.4.1.

In that example we saw that for $\mu_1 > \mu_0$ a BCR was given by $\bar{x} \geq b_\alpha$ and for $\mu_1 < \mu_0$ by $\bar{x} \leq a_\alpha$. Thus if H_1 contains only values greater than, or only values less than μ_0, we have a (one-sided) UMP test, but not if H_1 allows values of μ on both sides of μ_0. In such cases we would intuitively expect that a compromise critical region defined by $\bar{x} \leq a_{\alpha/2}$ or $\bar{x} \geq b_{\alpha/2}$ would give a satisfactory 'two-sided' test, and this is what is usually used. It is, of course, less powerful than the one-sided tests in their regions of applicability as is illustrated in the figure.



Maximizing local power

If no UMP test exists, it can be a good idea to look for a test which is most powerful in the neighborhood of the null hypothesis. This is the place where a test will usually be least powerful. Consider the two simple hypotheses, both specifying the same p.d.f.,

$$H_0\colon\ \theta = \theta_0 \qquad\qquad H_1\colon\ \theta = \theta_1 = \theta_0 + \Delta$$

where Δ is small.

The log-likelihood can be expanded about θ_0,

$$\ln \mathcal{L}(\underline{x};\theta_1) = \ln \mathcal{L}(\underline{x};\theta_0) + \Delta \left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0} + \ldots$$

Since we are treating two simple tests, we can use the Neyman-Pearson test (equation 10.7) to reject H_0 if the likelihood ratio is smaller than some critical value:

$$\lambda = \frac{\mathcal{L}(\underline{x}|H_0)}{\mathcal{L}(\underline{x}|H_1)} \leq c_\alpha$$

This is equivalent to

$$\ln \mathcal{L}(\underline{x};\theta_0) - \ln \mathcal{L}(\underline{x};\theta_1) \leq \ln c_\alpha$$

or (assuming $\Delta > 0$)

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0} \geq k_\alpha \quad , \qquad k_\alpha = -\frac{\ln c_\alpha}{\Delta}$$

Now, if the observations are independent and identically distributed, we know from section 8.2.5 that under H_0 the expectation of $\mathcal{L}$ is a maximum and

$$E\left[\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0}\right] = 0$$
$$E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right] = nI$$

for n independent observations, where I is the information on θ for 1 observation. Under suitable conditions $\frac{\partial \ln \mathcal{L}}{\partial \theta}$ is approximately normally distributed with mean 0 and variance nI. The value of k_α corresponding to a particular choice of size α can then be found as in section 10.4.1 (equation 10.11):

$$\alpha = \mathrm{erf}(-z_\alpha) \quad , \qquad \text{where} \quad z_\alpha = \frac{k_\alpha}{\sqrt{nI}}$$

In this way, a locally most powerful test is approximately given by rejecting H_0 if

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0} \geq z_\alpha \sqrt{nI} \quad , \qquad \alpha = \mathrm{erf}(-z_\alpha) \tag{10.13}$$

10.4.3 Composite hypotheses—same parametric family

We now turn to the more general case where both H_0 and H_1 are composite hypotheses. We make a distinction between the case where the p.d.f.'s specified in the hypotheses belong to one continuous family from the case where they belong to distinct families. In the first case the only difference between the hypotheses is the specification of the parameters, *e.g.*,

$$H_0\colon\ f(x;\theta) \quad , \qquad \text{with} \ \ \theta < \theta_0$$
$$H_1\colon\ f(x;\theta) \quad , \qquad \text{with} \ \ \theta > \theta_0$$

However, in the second case the p.d.f.'s are different and may even involve different numbers of parameters. In this section we will treat the first case.

Likelihood ratio test

We have seen (section 8.4) that the maximum likelihood method gave estimators which, under certain conditions, had desirable properties. A method of test construction closely related to it is the likelihood ratio method proposed by Neyman

and Pearson[52] in 1928. It has played a similar role in the theory of tests to that of the maximum likelihood method in the theory of estimation. As we have seen (sect. 10.4.1), this led to a MP test for simple hypotheses.

Assume that the N observations, $\underline{x}$, are independent and that both hypotheses specify the p.d.f. $f(\underline{x};\underline{\theta})$. Then the likelihood function is

$$\mathcal{L}(\underline{x};\underline{\theta}) = \prod_{i=1}^{N} f(x_i;\underline{\theta})$$

We denote the total parameter space by Θ and a subspace of it by ν. Then the hypotheses can be specified by

$$\begin{aligned} H_0&: \ \underline{\theta} \in \nu \\ H_1&: \ \underline{\theta} \in \Theta - \nu \end{aligned}$$

Examples, where for simplicity we assume that there are only two parameters $\underline{\theta} = (\theta_1, \theta_2)$, are

Example	1	2	3
H_0	$\theta_1 = a$ and $\theta_2 = b$	$\theta_1 = c$, θ_2 unspecified	$\theta_1 + \theta_2 = d$
H_1	$\theta_1 \neq a$ and/or $\theta_2 \neq b$	$\theta_1 \neq c$, θ_2 unspecified	$\theta_1 + \theta_2 \neq d$

In the first example H_0 is in fact a simple hypothesis.

We use the term conditional maximum likelihood for the maximum of the likelihood function for $\underline{\theta}$ in the region specified by H_0. Similarily, the unconditional maximum likelihood is the maximum of the likelihood in the entire parameter space. We define as the test statistic the **maximum likelihood ratio**, λ, as the ratio of the conditional maximum likelihood to the unconditional maximum likelihood:

$$\lambda = \frac{\mathcal{L}_{\nu\,\max}(\underline{x};\underline{\theta})}{\mathcal{L}_{\Theta\,\max}(\underline{x};\underline{\theta})} \tag{10.14}$$

Clearly, $0 \leq \lambda \leq 1$. Given what we know about the maximum likelihood method for parameter estimation, it certainly seems reasonable that this statistic would provide a reasonable test. In the limit of H_0 and H_1 both being simple, it is equivalent to the Neyman-Pearson test (equation 10.7, section 10.4.1). The success of the maximum likelihood ratio as a test statistic is due to the fact that it is always a function of a sufficient statistic for the problem. Its main justification is its past success. It has been found very frequently to result in a workable test with good properties, at least for large sets of observations.

The hypotheses to be tested can usually be written in the form

H_0:	$\theta_i = \theta_{i0}$	for $i = 1, 2, \ldots, r$	(denote this by $\underline{\theta}_r = \underline{\theta}_{r0}$)
	θ_j unspecified	for $j = 1, 2, \ldots, s$	(denote this by $\underline{\theta}_s$)
H_1:	$\theta_i \neq \theta_{i0}$	for $i = 1, 2, \ldots, r$	(denote this by $\underline{\theta}_r \neq \underline{\theta}_{r0}$)
	θ_j unspecified	for $j = 1, 2, \ldots, s$	

Hypotheses which do not specify exact values for parameters, but rather relationships between parameters, *e.g.*, $\theta_1 = \theta_2$, can usually be reformulated in terms of other parameters, *e.g.*, $\theta_1' = \theta_1 - \theta_2 = 0$ and $\theta_2' = \theta_1 + \theta_2$ unspecified. We can introduce the more compact notation of $\mathcal{L}(\underline{x}; \underline{\theta}_r, \underline{\theta}_s)$, *i.e.*, we write two vectors of parameters, first those which are specified under H_0 and second those which are not. The unspecified parameters $\underline{\theta}_s$ are sometimes referred to as 'nuisance' parameters.

In this compact notation, the test statistic can be rewritten as

$$\lambda = \frac{\mathcal{L}\left(\underline{x}; \underline{\theta}_{r0}, \hat{\hat{\underline{\theta}}}_s\right)}{\mathcal{L}\left(\underline{x}; \hat{\underline{\theta}}_r, \hat{\underline{\theta}}_s\right)} \tag{10.15}$$

where $\hat{\hat{\underline{\theta}}}_s$ is the value of $\underline{\theta}_s$ at the maximum of $\mathcal{L}$ in the restricted region ν and $\hat{\underline{\theta}}_r$ and $\hat{\underline{\theta}}_s$ are the values of $\underline{\theta}_r$ and $\underline{\theta}_s$ at the maximum of $\mathcal{L}$ in the full region Θ.

If H_0 is true, we expect λ to be near to 1. The critical region will therefore be

$$\lambda \leq c_\alpha \tag{10.16}$$

where c_α must be determined from the p.d.f. of λ, $g(\lambda)$, under H_0. Thus, for a test of size α, c_α is found from

$$\alpha = \int_0^{c_\alpha} g(\lambda)\, \mathrm{d}\lambda \tag{10.17}$$

It is thus necessary to know how λ is distributed. Furthermore, to perform this integration, $g(\lambda)$ must not depend on any of the unspecified (nuisance) parameters. Luckily, this is so for most statistical problems.

Example: As an example, let us again take a normal p.d.f. with H_0 specifying the mean as $\mu = \mu_0$ and H_1 specifying $\mu \neq \mu_0$. Both hypotheses leave σ unspecified; thus σ is a nuisance parameter. Then

$$\mathcal{L}(\underline{x}; \mu, \sigma) = (2\pi\sigma^2)^{-N/2} \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

We have seen (section 8.4.1) that the unconditional maximum likelihood estimators are

$$\hat{\mu} = \bar{x}$$
$$\hat{\sigma}^2 = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Thus, the unconditional likelihood is

$$\mathcal{L}(x; \hat{\mu}, \hat{\sigma}) = (2\pi s^2)^{-N/2} \exp\left[-\frac{1}{2}N\right]$$

Under H_0, the maximum likelihood estimator is

$$\hat{\hat{\sigma}}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_0)^2 = s^2 + (\bar{x} - \mu_0)^2$$

Therefore the conditional maximum likelihood is given by

$$\mathcal{L}(x; \mu_0, \hat{\hat{\sigma}}) = \left\{2\pi\left[s^2 + (\bar{x} - \mu_0)^2\right]\right\}^{-N/2} \exp\left[-\frac{1}{2}N\right]$$

The likelihood ratio is then

$$\lambda = \left\{\frac{s^2}{s^2 + (\bar{x} - \mu_0)^2}\right\}^{\frac{1}{2}N} \tag{10.18}$$

Consequently,

$$\lambda^{2/N} = \frac{1}{1 + \frac{t^2}{N-1}} \qquad , \qquad t^2 = \frac{N(\bar{x} - \mu_0)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{10.19}$$

This t is a Student's t-statistic with $N - 1$ degrees of freedom (equation 3.40). We see that λ is a monotonically decreasing function of t^2. Recall that the t-distribution is symmetric about zero. The critical region, $\lambda < \lambda_\alpha$, therefore corresponds to the two regions $t < t_{-\alpha/2}$ and $t > t_{\alpha/2}$. The values of $t_{\pm\alpha/2}$ corresponding to a particular test size α can be found from the Student's t-distribution, and from that value the corresponding value of λ_α follows using the above equation. It can be shown that this test is UMPU.[11,13]

Asymptotic distribution of the likelihood ratio

In order to determine the critical region of the likelihood ratio, λ, it is necessary to know how it is distributed under H_0. Sometimes we can find this distribution quite easily, as in the example of the previous section. But often it is difficult, since the distribution is unknown or since it is awkward to handle. One can sometimes use Monte Carlo, but this is not always satisfactory. The usual procedure is to consider the asymptotic distribution of the likelihood ratio, and use it as an approximation to the true distribution.

We know that asymptotically the maximum likelihood estimator $\hat{\theta}$ attains the minimum variance bound and that $\hat{\theta}$ becomes normally distributed according to the likelihood function. Suppressing the normalization factor, the likelihood function is of the form

$$\mathcal{L}(\underline{x}; \underline{\theta}) = \mathcal{L}(\underline{x}; \underline{\theta}_r, \underline{\theta}_s) \propto \exp\left[-\frac{1}{2}(\hat{\underline{\theta}} - \underline{\theta})^{\mathrm{T}}\underline{I}(\hat{\underline{\theta}} - \underline{\theta})\right] \tag{10.20}$$

where $\underline{I}$ is the information matrix for $\underline{\theta}$,

$$\underline{I} = \begin{bmatrix} \underline{I}_r & \vdots & \underline{I}_{rs} \\ \cdots & \vdots & \cdots \\ \underline{I}_{rs}^{\mathrm{T}} & \vdots & \underline{I}_s \end{bmatrix}$$

Thus, equation 10.20 can be written

$$\mathcal{L}(\underline{x};\underline{\theta}_r,\underline{\theta}_s) \propto \exp\left\{-\tfrac{1}{2}\left[\begin{array}{l}(\hat{\underline{\theta}}_r-\underline{\theta}_r)^{\mathrm{T}}\underline{I}_r(\hat{\underline{\theta}}_r-\underline{\theta}_r)\\ +2\,(\hat{\underline{\theta}}_r-\underline{\theta}_r)^{\mathrm{T}}\underline{I}_{rs}(\hat{\underline{\theta}}_s-\underline{\theta}_s)\\ +\ (\hat{\underline{\theta}}_s-\underline{\theta}_s)^{\mathrm{T}}\underline{I}_s(\hat{\underline{\theta}}_s-\underline{\theta}_s)\end{array}\right]\right\} \tag{10.21}$$

At the maximum of $\mathcal{L}$ under H_1, $\hat{\underline{\theta}}_r = \underline{\theta}_r$ and $\hat{\underline{\theta}}_s = \underline{\theta}_s$. Thus the exponent of equation 10.21 is zero and equation 10.21 becomes $\mathcal{L} \propto 1$. Under H_0, we must replace $\hat{\underline{\theta}}_s$ in equation 10.21 by $\hat{\hat{\underline{\theta}}}_s$ At the maximum of $\mathcal{L}$, we have $\hat{\hat{\underline{\theta}}}_s = \underline{\theta}_s$. Thus, under H_0 equation 10.21 becomes

$$\mathcal{L} \propto \exp\left[-\frac{1}{2}(\hat{\underline{\theta}}_r-\underline{\theta}_{0r})^{\mathrm{T}}\underline{I}_r(\hat{\underline{\theta}}_r-\underline{\theta}_{0r})\right]$$

Taking the ratio, we find

$$\lambda = \exp\left[-\frac{1}{2}(\hat{\underline{\theta}}_r-\underline{\theta}_{0r})^{\mathrm{T}}\underline{I}_r(\hat{\underline{\theta}}_r-\underline{\theta}_{0r})\right]$$

or

$$-2\ln\lambda = (\hat{\underline{\theta}}_r-\underline{\theta}_{0r})^{\mathrm{T}}\underline{I}_r(\hat{\underline{\theta}}_r-\underline{\theta}_{0r})$$

From the property that $\mathcal{L}$ is normally distributed, it follows that $-2\ln\lambda$ is a distributed as χ^2 with r degrees of freedom under H_0, where r is the number of parameters specified under H_0. For a test of size α, we therefore reject H_0 if

$$-2\ln\lambda > \chi^2_\alpha \qquad \text{where} \quad \int_{\chi^2_\alpha}^{\infty} \chi^2(r)\,\mathrm{d}\chi^2 = \alpha$$

Under H_1, it turns out that $-2\ln\lambda$ is distributed as a non-central χ^2 with r degrees of freedom and non-centrality parameter

$$K = (\hat{\underline{\theta}}_r-\underline{\theta}_{0r})^{\mathrm{T}}\underline{I}_r(\hat{\underline{\theta}}_r-\underline{\theta}_{0r})$$

The non-central χ^2 distribution, $\chi'^2(r,K)$, is the distribution of a sum of variables distributed normally with a non-zero mean and unit variance. It can be used to calculate the power of the test:[4,5,11,13]

$$p = 1-\beta = \int_{\chi^2_\alpha}^{\infty} \mathrm{d}F_1$$

where F_1 is the c.d.f. of χ'^2.

The asymptotic properties of the likelihood ratio test which have been found in this section depend on the asymptotic properties of the likelihood function, which in turn rest on regularity assumptions about the likelihood function. In particular, we have assumed that the range of the p.d.f. does not depend on the value of a parameter. Nevertheless, it turns out that under certain conditions $-2\ln\lambda$ is even then distributed as χ^2, but with $2r$ instead of r degrees of freedom.[11,13]

Small sample behavior of the likelihood ratio

Although the asymptotic properties of the likelihood ratio for hypotheses having p.d.f.'s of the same family are quite simple, the small sample behavior is not so easy. The usual approach is to find a correction factor, f, such that $-(2\ln\lambda)/f$ is distributed as $\chi^2(r)$ even for small N.[4,5,11,13] Only the case of the linear model will be treated here.

Linear model: A particular case is the linear model (section 8.5.2) in which the N observations y_i are assumed to be related to other observations x_i, within random errors ϵ_i, by a function linear in the k parameters θ_j,

$$y_i = y(x_i) + \epsilon_i = \sum_{j=1}^{k} \theta_j h_j(x_i) + \epsilon_i$$

We assume that the ϵ_i are normally distributed with mean 0 and variance σ^2. We wish to test whether the θ_j have the specified values θ_{0j}, or more generally, whether they satisfy some set of r linear constraints,

$$\underline{A}\,\underline{\theta} = \underline{b} \tag{10.22}$$

where $\underline{A}$ and $\underline{b}$ are specified under H_0. Under H_1, the $\underline{\theta}$ may take on any set of values not satisfying the constraints of equation 10.22.

The likelihood for both H_0 and H_1 is given by

$$\begin{aligned}\mathcal{L}(\underline{x};\underline{\theta}) &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{k}\theta_j h_j(x_i)\right)^2\right] \\ &= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2}Q^2\right]\end{aligned}$$

We now distinguish two cases:

Variance known. We first treat the case of known variance σ^2. The estimates of the parameters are given by the least squares solutions (section 8.5), with constraints for H_0 yielding $\hat{\theta}_{0j}$ and without constraints for H_1 yielding $\hat{\theta}_{1j}$. The likelihood ratio, λ, is then given by

$$-2\ln\lambda = \frac{1}{\sigma^2}\sum_{i=1}^{N}\left[y_i - \sum_{j=1}^{k}\hat{\theta}_{0j}h_j(x_i)\right]^2 - \frac{1}{\sigma^2}\sum_{i=1}^{N}\left[y_i - \sum_{j=1}^{k}\hat{\theta}_{1j}h_j(x_i)\right]^2 = Q_0^2 - Q_1^2 \tag{10.23}$$

It has been shown[4,5] that the second term can be expressed as the first term plus a quadratic form in the ϵ_i, and hence that $-2\ln\lambda$ is distributed as a χ^2 of r degrees of freedom. This result is true exactly for all N, not just asymptotically. It also holds if the errors are not independent but have a known covariance matrix.

The test thus consists of performing two least squares fits, one with and one without the constraints of H_0. Each fit results in a value of Q^2, the difference of which, $Q_0^2 - Q_1^2$, is a $\chi^2(r)$. H_0 is then rejected if $Q_0^2 - Q_1^2 > \chi_\alpha^2$ where $\int_{\chi_\alpha^2}^\infty \chi^2(r)\,\mathrm{d}\chi^2 = \alpha$.

We can qualitatively understand this result in the following way: Asymptotically, Q_0^2 is a $\chi^2(N-k+r)$ and Q_1^2 is a $\chi^2(N-k)$. From the reproductive property of the χ^2 distribution (section 3.12), the difference of these χ^2 is also a χ^2 with a number of degrees of freedom equal to the difference of degrees of freedom of Q_0^2 and Q_1^2, namely r. Thus the above result follows.

Variance unknown. If the variance σ^2 is unknown, it must be estimated from the data. Under H_0 the estimate of σ^2 is

$$s_0^2 = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \sum_{j=1}^{k}\hat{\theta}_{0j}h(x_i)\right]^2$$

and the maximum likelihood becomes

$$\mathcal{L}(\underline{x}; H_0) = \frac{1}{(2\pi)^{N/2}(s_0^2)^{N/2}}\exp\left[-\frac{N}{2}\right]$$

The expressions for H_1 are similar. The likelihood ratio is then

$$\lambda = \left(\frac{s_1^2}{s_0^2}\right)^{N/2} \tag{10.24}$$

or

$$\lambda^{-2/N} = 1 + \frac{s_0^2 - s_1^2}{s_1^2} \tag{10.25}$$

It can be shown that $(s_0^2 - s_1^2)/\sigma^2$ and s_1^2/σ^2 are independently distributed as χ^2 with r and $N-k$ degrees of freedom, respectively. The ratio,

$$F = \frac{N-k}{r}\,\frac{s_0^2 - s_1^2}{s_1^2} \tag{10.26}$$

is therefore distributed as the F-distribution (section 3.14). H_0 is then rejected if $F > F_\alpha$, where $\int_{F_\alpha}^\infty F(r, N-k)\,\mathrm{d}F = \alpha$.

However, under H_1, $(s_0^2 - s_1^2)/\sigma^2$ is distributed as a non-central χ^2. This leads to a non-central F-distribution from which the power of the test can be calculated.[11,13]

10.4.4 Composite hypotheses —different parametric families

When the p.d.f. specified by H_1 can not be attained by varying the parameters of the p.d.f. of H_0, we speak of different parametric families of functions. The

distribution of the likelihood ratio then usually turns out to depend on N as well as on which hypothesis is true. The likelihood ratio can still be used as a test, but these dependences must be properly taken into account.[4,5] The tests are therefore more complicated.

The easiest method to treat this situation is to construct a comprehensive family of functions

$$h(\underline{x};\theta,\underline{\phi},\underline{\psi}) = (1-\theta)f(\underline{x};\underline{\phi}) + \theta g(\underline{x};\underline{\psi})$$

by introducing an additional parameter θ.

What we really want to test is H_0 against H_1,

$$\begin{array}{lll} H_0: & f(\underline{x};\underline{\phi}) & , \quad \underline{\phi} \text{ unspecified} \\ H_1: & g(\underline{x};\underline{\psi}) & , \quad \underline{\psi} \text{ unspecified} \end{array}$$

Instead, we can use the composite function to test H_0 against H_1':

$$\begin{array}{llll} H_0: & h(\underline{x};\theta,\underline{\phi},\underline{\psi}) & , \quad \theta = 0, & \underline{\phi},\underline{\psi} \text{ unspecified} \\ H_1': & h(\underline{x};\theta,\underline{\phi},\underline{\psi}) & , \quad \theta \neq 0, & \underline{\phi},\underline{\psi} \text{ unspecified} \end{array}$$

using the maximum likelihood ratio as in the previous section:

$$\lambda = \frac{\mathcal{L}(\underline{x};\theta=0,\hat{\hat{\underline{\phi}}},\underline{\psi})}{\mathcal{L}(\underline{x};\hat{\theta},\hat{\underline{\phi}},\hat{\underline{\psi}})} = \left(\frac{f(\underline{x};\hat{\hat{\underline{\phi}}})}{(1-\hat{\theta})f(\underline{x};\hat{\phi}) + \hat{\theta}g(\underline{x};\hat{\underline{\psi}})}\right)^N \tag{10.27}$$

Then under H_0, $-2\ln\lambda$ is distributed asymptotically as $\chi^2(1)$ since one constraint ($\theta = 0$) has been imposed on the parameter space.

The power of the test can be found using the fact that, under H_1, $-2\ln\lambda$ is distributed as a non-central χ^2, $\chi'^2(1,K)$ with 1 degree of freedom and non-centrality parameter $K = \theta^2/S$ where

$$S = E\left[\frac{\left[f(\underline{x};\underline{\phi}) - g(\underline{x};\underline{\psi})\right]^2}{\left[(1-\theta)f(\underline{x};\underline{\phi}) + \theta g(\underline{x};\underline{\psi})\right]^2}\right] \tag{10.28}$$

Since this test compares $f(\underline{x};\underline{\phi})$ with a mixture of f and g, it is not expected to be very powerful.

In practice, one would also make a test of H_1 against the mixture, *i.e.*, define a new H_0' corresponding to $\theta = 1$, and test this against the mixture H_1' in the same manner as above, hoping that H_0 or H_0', but not both, would be rejected.

10.5 And if we are Bayesian?

If we are Bayesian, our belief in (the probability of) H_0 or H_1 is simply given by Bayes' theorem. After an experiment giving result $\underline{x}$, the probability of H_i ($i = 0, 1$) is

$$P(H_i|\underline{x}) = \frac{P(\underline{x}|H_i)}{P(\underline{x}|H_0) + P(\underline{x}|H_1)} P_{\mathrm{p}}(H_i) \tag{10.29}$$

where $P_{\rm p}(H_i)$ is the probability of H_i before (prior to) doing the experiment and $P(\underline{x}|H_i)$ is the probability of obtaining the result $\underline{x}$ if H_i is true, which is identical to $\mathcal{L}(\underline{x}|H_i)$. We can compare $P(H_0|\underline{x})$ and $P(H_1|\underline{x})$, *e.g.*, by their ratio. If both H_0 and H_1 are simple hypotheses,

$$\frac{P(H_0|\underline{x})}{P(H_1|\underline{x})} = \frac{P(\underline{x}|H_0)}{P(\underline{x}|H_1)}\frac{P_{\rm p}(H_0)}{P_{\rm p}(H_1)} \tag{10.30}$$

$$= \lambda\,\frac{P_{\rm p}(H_0)}{P_{\rm p}(H_1)} \tag{10.31}$$

where λ is just the likelihood ratio (eq. 10.7). This leads to statements such as "the probability of H_0 is, *e.g.*, 20 times that of H_1". Note, however, that here, as always with Bayesian statistics, it is necessary to assign prior probabilities. In the absence of any prior knowledge, $P_{\rm p}(H_0) = P_{\rm p}(H_1)$. The test statistic is then λ, just as in the Neyman-Pearson test (section 10.4.1). However now the interpretation is a probability rather than a level of significance.

Suppose that H_1 is a composite hypothesis where a parameter θ is unspecified. Equation 10.30 remains valid, but with

$$P(\underline{x}|H_1) = \int f(\underline{x},\theta|H_1)\ {\rm d}\theta \tag{10.32}$$

$$= \int P(\underline{x}|\theta,H_1)\,f(\theta|H_1)\ {\rm d}\theta \tag{10.33}$$

Now, $P(\underline{x}|\theta,H_1)$ is identical to $\mathcal{L}(\underline{x};\theta)$ under H_1 and $f(\theta|H_1)$ is just the prior p.d.f. of θ under H_1. In practice, this may not be so easy to evaluate. Let us therefore make some simplifying assumptions for the purpose of illustration. We know that asymptotically $\mathcal{L}(\underline{x};\theta)$ is proportional to a Gaussian function of θ (eq. 8.72). Let us take a prior probability uniform between $\theta_{\rm min}$ and $\theta_{\rm max}$ and zero otherwise. Then, with $\sigma^2_{\hat\theta}$ the variance of the estimate, $\hat\theta$, of θ, equation 10.33 becomes

$$P(\underline{x}|H_1) = \mathcal{L}_{\rm max}(\underline{x};\theta)\int \exp\left(-\frac{(\theta-\hat\theta)^2}{2\sigma^2_{\hat\theta}}\right)\frac{1}{\theta_{\rm max}-\theta_{\rm min}}\,{\rm d}\theta \tag{10.34}$$

$$= \frac{\mathcal{L}_{\rm max}(\underline{x};\theta)}{\theta_{\rm max}-\theta_{\rm min}}\int_{\theta_{\rm min}}^{\theta_{\rm max}} \exp\left(-\frac{(\theta-\hat\theta)^2}{2\sigma^2_{\hat\theta}}\right)\,{\rm d}\theta \tag{10.35}$$

$$= \mathcal{L}_{\rm max}(\underline{x};\theta)\,\frac{\sigma_{\hat\theta}\sqrt{2\pi}}{\theta_{\rm max}-\theta_{\rm min}} \tag{10.36}$$

where we have assumed that the tails of the Gaussian cut off by the integration limits $\theta_{\rm min}$, $\theta_{\rm max}$ are negligible. Thus equation 10.30 becomes

$$\frac{P(H_0|\underline{x})}{P(H_1|\underline{x})} = \lambda\,\frac{P_{\rm p}(H_0)}{P_{\rm p}(H_1)}\,\frac{\theta_{\rm max}-\theta_{\rm min}}{\sigma_{\hat\theta}\sqrt{2\pi}} \tag{10.37}$$

where λ is now the maximum likelihood ratio $\lambda = \mathcal{L}(\underline{x}|H_0)/\mathcal{L}_{\max}(\underline{x}|H_1)$. Note that there is a dependence not only on the prior probabilities of H_0 and H_1, but also on the prior probability of the parameter θ.

Someone remarked to me once:
"Physicians shouldn't say, 'I have cured this man',
but, 'this man didn't die under my care'."
In physics too, instead of saying,
"I have explained such and such phenomenon",
one might say, "I have determined causes for it
the absurdity of which cannot be conclusively proved."
—Georg Christoph Lichtenberg

10.6 Goodness-of-fit tests

10.6.1 Confidence level or p-value

As in the previous section, we are concerned with testing an hypothesis H_0 at some significance level α. Again, H_0 will be rejected if a test statistic has a value which lies in the critical region ω. The difference with the previous section lies in the alternative hypothesis H_1. Now H_1 is simply *not* H_0, *i.e.*, H_1 is the set of all possible alternatives to H_0. Thus H_1 can not be formulated and consequently, the chance of an error of the second kind can not be known. Nor can most of the tests of the previous section (including the use of Bayesian probability) be applied, involving as they do the likelihood ratio, for if we do not specify H_1, we can not calculate the likelihood under H_1.

Goodness-of-fit tests compare the experimental data with the p.d.f. specified under H_0 and lead to the statement that the data are consistent or inconsistent with H_0. Usually one states a confidence level, *e.g.*, "The data are consistent with H_0 at a confidence level of 80%." The **confidence level*** (CL), also known as ***p*-value**,† is the size that the test would have if the critical region were such that the test statistic were at the boundary between rejection and acceptance of $\boldsymbol{H_0}$. In other words, it is the probability, assuming $\boldsymbol{H_0}$ is true, of obtaining a value of the test statistic as "bad" as or "worse" than that actually obtained. Thus, a high confidence level means that if $\boldsymbol{H_0}$ is true there is a large chance of obtaining data 'similar' to ours. On the contrary, if CL is small there is a small chance, and $\boldsymbol{H_0}$

*Many authors use $1 - \text{CL}$ where we use CL.

†The preferable term is p-value, since it eliminates confusion with the confidence level of confidence intervals (chapter 9), which, although related, is different. Nevertheless, the term confidence level is widely used, especially by physicists.

can be rejected. Despite the suggestive "p", the p-value is not a probability; it is a random variable.

We shall only consider distribution-free tests, for the practical reason that they are widely applicable. To apply a test, one needs to know the p.d.f. of the test statistic in order to calculate the confidence level. For the well-known tests tables and/or computer routines are widely available. For a specific problem it may be possible to construct a better test, but it may not be so much better that it is worth the effort.

10.6.2 Relation between Confidence level and Confidence Intervals

The same integrals are involved in confidence intervals and goodness-of-fit tests. To illustrate this, consider a r.v., x, which is distributed normally, $f(x) = N(x; \mu, \sigma^2)$. For n points, assuming σ^2 known, the estimator of the mean, $t = \bar{x}$, is also normally distributed:

$$f(t) = N(t; \mu, \sigma^2/n)$$

The coverage probability (or confidence coefficient or confidence level) of the confidence interval $[\mu_-, \mu_+]$, *e.g.*, for a central confidence interval from equation 9.12, is given by equation 9.10,

$$\beta = \int_{t_-(\mu)}^{t_+(\mu)} N(t; \mu, \sigma^2/n)\ \mathrm{d}t \tag{10.38}$$

which holds for any value of μ.

If H_0 states that x is distributed normally with mean $\mu = 0$,

$$H_0: \qquad f(x) = N(x; 0, \sigma^2) \qquad \text{or} \qquad f(t) = N(t; 0, \sigma^2/n)$$

and if the data give $t = \bar{x}$, the confidence level or p-value (for a symmetric two-sided test) is

$$\begin{aligned} \mathrm{CL} &= \int_{-\infty}^{-|\bar{x}|} N(t; 0, \sigma^2/n)\,\mathrm{d}t + \int_{+|\bar{x}|}^{+\infty} N(t; 0, \sigma^2/n)\,\mathrm{d}t \\ &= 1 - \int_{-|\bar{x}|}^{+|\bar{x}|} N(t; 0, \sigma^2/n)\,\mathrm{d}t \end{aligned} \tag{10.39}$$

Note the similarity of the integrals in equations 10.38 and 10.39. We see that the coverage probability of the interval $[-|\bar{x}|, +|\bar{x}|]$, β, is related to the p-value by $\mathrm{CL} = 1 - \beta$. However, for the confidence interval, the coverage probability is specified first and the interval, $[\mu_-, \mu_+]$, is the random variable, while for the goodness-of-fit test the hypothesis is specified ($\mu = \mu_0$) and the p-value is the r.v.

Referring to the confidence belt figure of section 9.2, and supposing that θ_t is the hypothesized value of the parameter μ_0, $t_-(\mu_0)$ and $t_+(\mu_0)$ are the values of

$\hat{t}$ which would give CL $= 1 - \beta$. Put another way, if we decide to reject H_0 if CL $< \alpha$, then the regions outside the confidence belt for $\beta = 1 - \alpha$ is the rejection region. Thus the confidence belt defines the acceptance region of the corresponding goodness-of-fit test.

10.6.3 The χ^2 test

Probably the best known and most used goodness-of-fit test is the χ^2 test. We have already frequently alluded to it. We know (section 3.12) that the sum of N normally distributed r.v.'s is itself a r.v. which is distributed as $\chi^2(N)$. Hence, assuming that our measurements, y_i, have a normally distributed error, σ_i, the sum

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - f_i)^2}{\sigma_i^2} \tag{10.40}$$

where f_i is the value that y_i is predicted to have under H_0, will be distributed as $\chi^2(N)$. The CL is easily calculable from the χ^2 distribution:

$$\text{CL} = \int_{X^2}^{\infty} \chi^2(z; N)\ \mathrm{d}z \tag{10.41}$$

This X^2 is just the quantity that is minimized in a least squares fit (where we denoted it by Q^2). In the linear model, assuming Gaussian errors, $X^2 = Q^2_{\min}$ is still distributed as χ^2 even though parameters have been estimated by the method. However the number of degrees of freedom is reduced to $N - k$, where k is the number of parameters estimated by the fit. If constraints have been used in the fit (*cf.* section 8.5.6), the number of degrees of freedom is increased by the number of constraints, since each constraint among the parameters reduces by one the number of free parameters estimated by the fit. If the model is non-linear, $X^2 = Q^2_{\min}$ is only asymptotically distributed as $\chi^2(N - k)$.

It is sometimes argued that the χ^2 test should be two-tailed rather than one-tailed, *i.e.*, that H_0 should be rejected for unlikely small values of X^2 as well as for unlikely large values. Arguments given for this practice are that such small values are likely to have resulted from computational errors, overestimation of the measurement errors σ_i, or biases (unintentional or not) in the data which have not been accounted for in making the prediction. However, while an improbably small value of X^2 might well make one suspicious that one or more of these considerations had occurred (and indeed several instances of scientific fraud have been discovered this way), such a low X^2 can not be regarded as a reason for rejecting H_0.

10.6.4 Use of the likelihood function

It is often felt that since the likelihood function is so useful in parameter estimation and in the formulation of tests of hypotheses, it should also be useful as a goodness-of-fit test. Frequently the statement is made that it can not be used for this purpose. In fact, it can be used, but it is usually difficult to do so.

The problem is that in order to use the value of $\mathcal{L}$ as a test, we must know how $\mathcal{L}$ is distributed in order to be able to calculate the confidence level. Suppose that we have N independent observations, x_i, each distributed as $f(x)$. The log likelihood is then just

$$\ell = \sum_{i=1}^{N} \ln f(x_i)$$

If no parameter is estimated from the data, the mean of ℓ is just

$$E[\ell] = \int \sum_{i=1}^{N} \ln f(x_i)\, \mathcal{L}\ \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_N = N \int \ln f(x)\, f(x)\ \mathrm{d}x$$

Similarly higher moments could be calculated, and from these moments (just the first two if N is large and the central limit theorem is applicable) the distribution of ℓ, $g(\ell)$, could be reconstructed. The confidence level would then be given by

$$\mathrm{CL} = \int_{-\infty}^{\ell} g(\ell)\ \mathrm{d}\ell \tag{10.42}$$

If parameters are estimated by maximum likelihood, the calculations become much more complicated. A simple, but expensive, solution is to generate Monte Carlo experiments. From each Monte Carlo experiment one calculates ℓ and thus obtains an approximate distribution for ℓ from which the CL can be determined.

10.6.5 Binned data

We now consider tests of binned data.* Since binning data loses information, we should expect such tests to be inferior to tests on individual data. Further, we must be sure to have a sufficient number of events in each bin, since most of the desirable properties of the tests are only true asymptotically.

However, binning the data removes the difficulty that H_1 is completely unspecified, since the number of events in a bin must be distributed multinomially. Thus both H_0 and H_1 specify the multinomial p.d.f. Some or all of the parameters are specified under H_0; none of them are specified under H_1 further than that they are different from those specified under H_0.

Likelihood ratio test

Suppose that we have k bins with n_i events in bin i and $\sum_{i=1}^{k} n_i = N$. Let H_0 be a simple hypothesis, *i.e.*, all parameters are specified. Let p_i be the probability content of bin i under H_0 and q_i the probability content under the true p.d.f., which we of course do not know. The likelihood under H_0 and under the true p.d.f.

*Although we use the term 'binned', which suggests a histogram, any classification of the observations may be used. See also section 8.6.1.

are then, from the multinomial p.d.f., given by

$$\mathcal{L}_0(\underline{n}|\underline{p}) = N! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!}$$
$$\mathcal{L}(\underline{n}|\underline{q}) = N! \prod_{i=1}^{k} \frac{q_i^{n_i}}{n_i!}$$

An estimate $\hat{q}_i$ of the true probability content can be found by maximizing $\mathcal{L}(\underline{n}|\underline{q})$ subject to the constraint $\sum_{i=1}^{k} q_i = 1$. The result* is

$$\hat{q}_i = \frac{n_i}{N}$$

The test statistic is then the likelihood ratio (*cf.* section 10.4.3)

$$\lambda = \frac{\mathcal{L}_0(\underline{n}|\underline{p})}{\mathcal{L}(\underline{n}|\hat{\underline{q}})} = N^N \prod_{i=1}^{k} \left(\frac{p_i}{n_i}\right)^{n_i} \tag{10.43}$$

The exact distribution of λ is not known. However, we have seen in section 10.4.3 that $-2\ln\lambda$ is asymptotically distributed as $\chi^2(k-1)$ under H_0, where the number of degrees of freedom, $k-1$, is the number of parameters specified. The multinomial p.d.f. has only $k-1$ parameters (p_i) because of the restriction $\sum_{i=1}^{k} p_i = 1$. If H_0 is not simple, *i.e.*, not all p_i are specified, the test can still be used but the number of degrees of freedom must be decreased accordingly.

Pearson's χ^2 test

The classic test for binned data is the χ^2 test proposed by Karl Pearson[53] in 1900. It makes use of the asymptotic normality of a multinomial p.d.f. to find that under H_0 the statistic

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - N\pi_i)^2}{N\pi_i} \tag{10.44}$$

is distributed asymptotically as $\chi^2(k-1)$.

If H_0 is not simple, its free parameters can be estimated, (section 8.6.1) by the minimum chi-square method. In that method, the quantity which is minimized with respect to the parameters (equation 8.152) is just Pearson's X^2. The minimum value thus found therefore serves to test the hypothesis. It can be shown that in this case X^2 is asymptotically distributed as $\chi^2(k-s-1)$ where s is the number of parameters which are estimated. This is also true if the binned maximum likelihood method (section 8.6.2) is used to estimate the parameters.[11,13] Simi-

*This was derived for the binomial p.d.f. in section 8.4.7. It may be trivially extended to the multinomial case by treating each bin separately as binomially distributed between that bin and all the rest.

larly, the quantity which is minimized in the modified minimum chi-square method (equation 8.154) is also asymptotically distributed as $\chi^2(k-s-1)$.

But what if we estimate the parameters by a different method? In particular, as is frequently the case, what if we estimate the parameters by maximum likelihood using the individual data rather than the binned data? It then turns out[11,13] that X^2 is still distributed as χ^2, but with a number of degrees of freedom, d, between that of the binned fit and the fully specified case, *i.e.*, $k-s-1 \leq d \leq k-1$. The exact number of degrees of freedom depends on the p.d.f. The test is then no longer distribution free, although for large k and small s it is nearly so.

Equation 10.44 assumes that H_0 only predicts the shape of the distribution, *i.e.*, the probability, π_i, that an event will be in bin i, with $\sum \pi_i = 1$. If also the total number of events is predicted by H_0, the distribution is no longer multinomial, but rather a multinomial times a Poisson or, equivalently, the product of k Poisson distributions. The test statistic is then

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - \nu_i)^2}{\nu_i} \tag{10.45}$$

where, under H_0, ν_i is the mean (and variance) of the Poisson distribution for bin i. Since each bin is independent, there are now k degrees of freedom, and X^2 is distributed asymptotically as $\chi^2(k-s)$.

Pearson's χ^2 test makes use of the squares of the deviations of the data from that expected under H_0. Tests can be devised which use some other measure of deviation, replacing the square of the absolute value of the deviation by some other power and scaling the deviation or not by the expected variance. Such tests are, however, beyond the scope of this course.

Choosing optimal bin size

If one is going to bin his data, he must define the bins. If the number of bins is small, too much information may be lost. But a large number of bins may mean that there are too few events per bin. Most of the results for binned data are only true asymptotically, *e.g.*, the normal limit of the multinomial p.d.f. or the distribution of $-2\ln\lambda$ or X^2 as χ^2.

There are, in fact, two questions which play a role here. The first is whether the binning may be decided on the basis of the data; the second concerns the minimum number of events per bin. At first glance it would seem that the bin boundaries should not depend on the observations themselves, *i.e.*, that we should decide on the binning before looking at the data. If the bin boundaries depend on the data, then the bin boundaries are random variables, and no provision has been made in our formalism for fluctuations in the position of these boundaries. On the other hand, the asymptotic formalism holds for any set of fixed bins, and so we might expect that it does not matter which of these sets we happen to choose, and this has indeed been shown to be so.[11,13]

Intuitively, we could expect that we should choose bins which are equiprobable under H_0. Pearson's χ^2 test is consistent (asymptotically unbiased) whatever the binning, but for finite N it is not, in general, unbiased. It can be shown[4,5,11,13] that for equiprobable bins it is locally unbiased, *i.e.*, unbiased against alternatives which are very close to H_0, which is certainly a desirable property.

Having decided on equiprobable bins, the next question is how many bins. Clearly, we must not make the number of bins k too large, since the multinormal approximation to the multinomial p.d.f. will no longer be valid. A rough rule which is commonly used is that no expected frequency, Np_i, should be smaller than ~ 5. However, according to Kendall and Stuart,[11] there seems to be no general theoretical basis for this rule. Cochran goes even further and claims[4] that the asymptotic approximation remains good so long as not more than 20% of the bins have an expected number of events between 1 and ~ 5.

This does not necessarily mean that it is best to take $k = N/5$ bins. By maximizing local power, one can try to arrive at an optimal number of bins. The result[4,5] is

$$k = b\left[\frac{\sqrt{2}(N-1)}{\lambda_\alpha + \lambda_{1-p_0}}\right]^{2/5} \tag{10.46}$$

where $\alpha = 1 - \int_{-\lambda_\alpha}^{\lambda_\alpha} N(x;0,1)\,\mathrm{d}x$ is the size of the test for a standard normal distribution and p_0 is the local power. In general, for a simple hypothesis a value for b between 2 and 4 is good, the best value depending on the p.d.f. under H_0. Typical values for k (N/k) using $b = 2$ are given in the following table. We see from the table that there is only a mild sensitivity of the number of bins to α and p_0. For $N = 200$, 25–30 bins would be reasonable.

		p_0	
N	α	0.5	0.8
200	0.01	27 (7.4)	24 (8.3)
	0.05	31 (6.5)	27 (7.4)
500	0.01	39 (13)	35 (14)
	0.05	45 (11)	39 (13)

Thus we are led to the following recommendations for binning:

1. Determine the number of bins, k, using equation 10.46 with $b \sim 2$ to 4.
2. If N/k turns out to be too small, decrease k to make $N/k \geq 5$.
3. Define the bins to have equal probability content, either from the p.d.f. specified by H_0 or from the data.
4. If parameters have to be estimated (H_0 does not specify all parameters), use maximum likelihood on the individual observations, but remember that the test statistic is then only approximately distribution-free.

Note, however, that, regardless of the above prescription, if the p.d.f. under H_0 does not include resolution effects, one should not choose bins much smaller than the resolution.

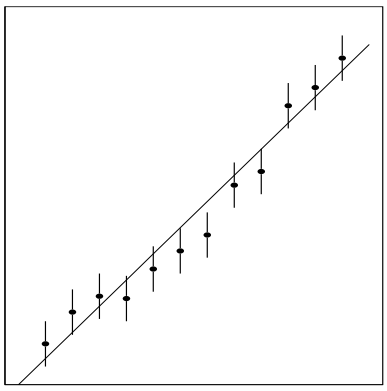
Even with the above prescription, the specification of the bins is still not unique. The usual method in one dimension would be to define a bin as an interval in the variable, $\text{bin}_i = (x_i, x_i + \delta_i)$. However, there is nothing in the above prescription to forbid defining a single bin as consisting of more than one (nonadjacent) interval. This might even be desirable from the point of view H_0. For example, H_0 might specify a p.d.f. that is symmetric about 0, and we might only be interested in testing this hypothesis against alternatives which are also symmetric about 0. Then it would be appropriate to define bins as intervals in $|x|$ rather than in x.

In more than one dimension the situation is more ambiguous. For example, to construct equiprobable bins in two dimensions, the easiest way is to first find equiprobable bins in x and then for each bin in x to find equiprobable bins in y. This is easily generalized to more dimensions. However, one could equally well choose first to construct bins and y and then in x, which in general would yield different bins. One could also choose different numbers of bins in x than in y. The choice depends on the individual situation. One should prefer smaller bins in the variable for which H_0 is most sensitive.

There is, obviously, one taboo: You must *not* try several different choices of binning and choose the one which gives the best (or worst) confidence level.

10.6.6 Run test

χ^2 tests make use of the squares of the deviations of the data from that expected under H_0. Thus they only use the size of the deviations and ignore their signs. However, the signs of the deviations are also important, and systematic deviations of the same sign indicate that the hypothesis is unlikely, as is illustrated in the figure.



A test which uses only the sign of the deviations is the run test. A run is defined as a set of adjacent points all having the same sign of deviation. The data and curve in the figure have deviations $AAABBBBBBAAA$, where A represents a positive and B a negative deviation. There are thus three runs, which seems rather small. We would expect the chance of an A to equal that of a B and to show no correlation between points if the hypothesis were true. This implies that we should expect runs to be short; a long run of 6 points as in the figure should be unlikely. In fact, this expectation is strictly true only if H_0 is a simple hypothesis.

To be more quantitative, let k_A be the number of positive deviations and k_B the number of negative deviations. Let $k = k_A + k_B$. Given k_A and k_B, we can calculate the probability that there will be r runs. If either k_A or k_B is zero, there is, necessarily only one run, and $P(r = 1) = 1$.

Given k_A and k_B, the number of different ways to arrange them is

$$\binom{k}{k_A} = \frac{k!}{k_A!k_B!}$$

Suppose that there are r runs. First, suppose that r is even and that the sequence begins with an A. Then there are k_A A-points and $r/2-1$ divisions between them. For the example of the figure this is $AAA|AAA$. With k_A A's there are $k_A - 1$ places to put the first dividing line, since it can not go at the ends. Then there are $k_A - 2$ places to put the second dividing line, since it can not go at the ends or next to the first dividing line. In total there are $\binom{k_A-1}{r/2-1}$ ways to arrange the dividing lines among the A's. There is a similar factor for arrangement of the B's and a factor 2 because we assumed we started with an A and it could just have well been a B. Thus the probability of r runs, for r even, is

$$P(r) = \frac{2\binom{k_A-1}{r/2-1}\binom{k_B-1}{r/2-1}}{\binom{k}{k_A}} \tag{10.47}$$

Similarly, one finds for r odd

$$P(r) = \frac{\binom{k_A-1}{(r-3)/2}\binom{k_B-1}{(r-1)/2} + \binom{k_A-1}{(r-1)/2}\binom{k_B-1}{(r-3)/2}}{\binom{k}{k_A}} \tag{10.48}$$

From these it can be shown that the expectation and variance of r are

$$E[r] = 1 + \frac{2k_Ak_B}{k} \tag{10.49}$$

$$V[r] = \frac{2k_Ak_B(2k_Ak_B - k)}{k^2(k-1)} \tag{10.50}$$

The critical region of the test is defined as improbably low values of r, $r < r_\alpha$. For k_A and k_B greater than about 10 or 15, one can use the Gaussian approximation for r. For smaller numbers one can compute the probabilities directly using equations 10.47 and 10.48. In our example, $k_A = k_B = 6$. From equations 10.49 and 10.50 we expect $r = 7$ with variance 2.73, or $\sigma = 1.65$. We observe 3 runs, which differs from the expected number by $4/1.65 = 2.4$ standard deviations. Using the Gaussian approximation, this corresponds to a (one-tailed) confidence level of 0.8%. Exact calculation using equations 10.47 and 10.48 yields $P(1) + P(2) + P(3) = 1.3\%$. Whereas the χ^2 is acceptable ($\chi^2 = 12$ for 12 points), the run test suggests that the curve does not fit the data.

The run test is much less powerful than a χ^2 test, using as it does much less information. But the two tests are completely independent and hence they can be combined. An hypothesis may have an acceptable χ^2, but still be wrong and rejectable by the run test. Unfortunately, the run test is applicable only when H_0 is simple. If parameters have been estimated from the data, the distribution of the number of runs is not known and the test can not be applied.

10.6.7 Tests free of binning

Since binning loses information, we should expect tests which do not require binning to be in principle better than tests which do.

The successful bin-free tests are based on the c.d.f., $F(x)$, under H_0 and consist of in some way comparing this c.d.f. with the data. To do so involves the concept of order statistics, which are just the observations, x_i, ordered in some way, *i.e.*, renumbered as $x_{(j)}$. In one dimension this is trivial. For n observations, the order statistics obey

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

In more than one dimension it is rather arbitrary, implying as it were a reduction of the number of dimensions to one. Even in one dimension the ordering is not free of ambiguity since we could equally well have ordered in descending order. We could also make a change of variable which changes the order of the data.

We define the sample c.d.f. for n observations as

$$S_n(x) = \begin{cases} 0 & , \quad x < x_{(1)} \\ \frac{r}{n} & , \quad x_{(r)} \leq x < x_{(r+1)} \\ 1 & , \quad x_{(n)} \leq x \end{cases} \tag{10.51}$$

which is simply the fraction of the observations not exceeding x. Clearly, under H_0, $S_n(x) \to F(x)$ as $n \to \infty$. The tests consist of comparing $S_n(x)$ with $F(x)$. We shall discuss two such tests, the Smirnov-Cramér-von Mises test and the Kolmogorov test. Unfortunately, both are only applicable to simple hypotheses, since the distribution of the test statistic is not distribution-free when parameters have been estimated from the data.

Smirnov-Cramér-von Mises test

As a measure of the difference between $S_n(x)$ and $F(x)$ this test uses the statistic

$$\begin{aligned} W^2 &= \int_0^1 [S_n(x) - F(x)]^2 \, \psi(x) \, \mathrm{d}F \\ &= \int_{-\infty}^{+\infty} [S_n(x) - F(x)]^2 \, \psi(x) f(x) \, \mathrm{d}x \end{aligned}$$

with $\psi(x) = 1$. We see that W^2 is the expectation of $[S_n(x) - F(x)]^2$ under H_0. Inserting S_n (equation 10.51) and performing the integral results in

$$nW^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[F(x_{(i)}) - \frac{2i-1}{2n} \right]^2 \tag{10.52}$$

*A more complete and more accurate table is given by Anderson and Darling,[54] who also consider the test statistic with $\psi(x) = \{F(x)\,[1 - F(x)]\}^{-1}$.

The asymptotic distribution of nW^2 has been found, and from it critical regions have been computed. Those corresponding to frequently used test sizes are given in the following table. The asymptotic distribution is reached remarkably rapidly. To the accuracy of this table,* the asymptotic limit is reached[4,5,11] for $n \geq 3$.

Test size α	Rejection region $nW^2 >$
0.10	0.347
0.05	0.461
0.01	0.743
0.001	1.168

Kolmogorov test

This test also compares S_n and $F(x)$, but only uses the maximum difference: The Kolmogorov (or Smirnov, or Kolmogorov-Smirnov) test statistic is the maximum deviation of the observed distribution $S_n(x)$ from the c.d.f. $F(x)$ under H_0:

$$D_n = \max\{|S_n(x) - F(x)|\} \qquad \text{for all } x \tag{10.53}$$

The asymptotic distribution of D_n yields the critical regions shown in the table. This approximation is considered satisfactory for more than about 80 observations.[4,5,11] Computer routines also exist.†

Test size α	Rejection region $\sqrt{n}D_n >$
0.01	1.63
0.05	1.36
0.10	1.22
0.20	1.07

Alternatively, one can take the maximum positive deviation,

$$D_n^+ = \max\{+[S_n(x) - F(x)]\} \qquad \text{for all } x \tag{10.54}$$

It can be shown that $4n(D_n^+)^2$ is distributed asymptotically as a χ^2 of 2 degrees of freedom. The same holds for D_n^-,

$$D_n^- = \max\{-[S_n(x) - F(x)]\} \qquad \text{for all } x \tag{10.55}$$

Or, as proposed by Kuiper,[56] one can use

$$V = D_n^+ + D_n^- \tag{10.56}$$

Asymptotic critical regions of V can be calculated.[55,57]

The sensitivity of the Kolmogorov test to deviations from the c.d.f. is not independent of x. It is more sensitive around the median value and less sensitive in the tails. This occurs because the difference $|S_n(x) - F(x)|$ does not, under H_0 have a probability distribution that is independent of x. Rather, its variance is proportional to $F(x)\,[1 - F(x)]$, which is largest at $F = 0.5$. Consequently,

†See, *e.g.*, *Numerical Recipes*.[55]

the significance of a large deviation in a tail is underweighted in the test. The Kolmogorov test therefore turns out to be more sensitive to departures of the data from the median of H_0 than to departures from the width. Various modifications of the Kolmogorov test statistic have been proposed[54,58,59] to ameliorate this problem.

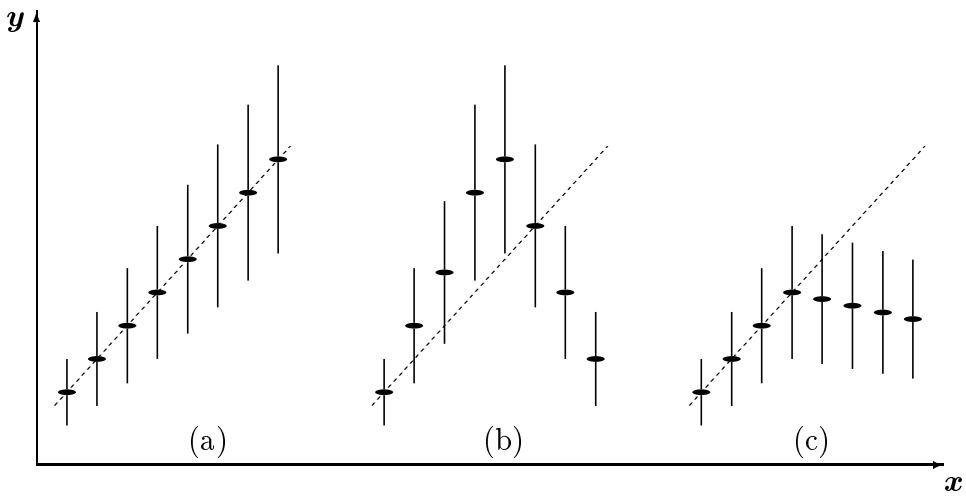
Although the distribution of the test statistic, D_n, is generally unknown if parameters have been estimated from the data, there are cases where the distribution has been calculated, *e.g.*, when H_0 specifies an exponential distribution whose mean is estimated from the data.[60] It also may be possible to determine the distribution of the test statistic yourself, *e.g.*, using Monte Carlo techniques.

10.6.8 But use your eyes!

A few words of caution are appropriate at this point. As illustrated by the figure at the start of the section on the run test (section 10.6.6), one test may give an acceptable value while another does not. Indeed, it is in the nature of statistics that this must sometimes occur.

Also, a fit may be quite good over part of the range of the variable and quite bad over another part. The resulting test value will be some sort of average goodness, which can still have an acceptable value. And so: *Do not rely blindly on a test.* Use your eyes. Make a plot and examine it.

There are several useful plots you can make. One is, as was done to illustrate the run test, simply a plot of the data with the fit distribution superimposed. Of course, the error bars should be indicated. It is then readily apparent if the fit is bad only in some particular region, and frequently you get an idea of how to improve the hypothesis. This is illustrated in the figure where the fit (dashed line) in (a) is perfect, while in (b) higher order terms are clearly needed and in (c) either higher orders or a discontinuity are required.



Since it is easier to see departures from a horizontal straight line, you could instead plot the residuals, $y_i - f(x_i)$, or even better, the residuals divided by their

error, $(y_i - f(x_i))/\delta$, where δ can be either the error on the data, or the expected error from a fit.

It may happen that there is only one or just a few data points which account for almost all the deviation from the fit. These are known as outliers. One is tempted to throw such points away on the assumption that they are due to some catastrophic error in the data taking, *e.g.*, writing down 92 instead of 29. However, one must be careful. Statistics can not really help here. You have to decide on the basis of what you know about your apparatus. Automatic outlier rejection should be avoided. It is said* that the discovery of the hole in the ozone layer above the south pole was delayed several years because computer programs automatically rejected the data which indicated its presence.

It's not right to pick only what you like,
but to take all of the evidence.
—Richard P. Feynman

I don't see the logic of rejecting data
just because they seem incredible.
—Sir Fred Hoyle

*Cited by Barlow[1] from *New Scientist*, 31 March 1988.

10.7 Non-parametric tests

The main classes of non-parametric problems which can be solved by distribution-free methods are

1. The two-sample problem. We wish to test whether two (or more generally k) samples are distributed according to the same p.d.f.

2. Randomness. A series of n observations of a single variable is ordered in some way, *e.g.*, in the time at which the observation was made. We wish to test that all of the observations are distributed according to the same p.d.f., *i.e.*, that there has been no change in the p.d.f. as a function of, *e.g.*, time.

3. Independence of variables. We wish to test that a bivariate (or multivariate) distribution factorizes into two independent marginal distributions, *i.e.*, that the variables are independent (*cf.* section 2.2.4).

These are all hypothesis-testing problems, which are similar to the goodness-of-fit problem in that the alternative hypothesis is simply *not* H_0.

The first two of the above problems are really equivalent to the third, even though the first two involve observations of just one quantity. For problem 1, we can combine the two samples $x_i^{(1)}$ and $x_i^{(2)}$ into one sample by defining a second variable $y_i = 1$ or 2 depending on whether x_i is from the first or the second sample. Independence of x and y is then equivalent to independence of the two samples. For problem 2, suppose that the x_i of problem 3 are just the observations of problem 2 and that the y_i are the order of the observations. Then independence of x_i and y_i is equivalent to no order dependence of the observations of problem 1. Let us begin then with problem 3.

10.7.1 Tests of independence

We have a sample of observations consisting of pairs of real numbers, (x, y) distributed according to some p.d.f., $f(x, y)$, with marginal p.d.f.'s, $g(x)$ and $h(y)$. We wish to test

$$H_0: \qquad f(x, y) = g(x)\, h(y)$$

Sample correlation coefficient

An obvious test statistic is the sample correlation coefficient (*cf.* equation 2.27).

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{s_x s_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{s_x s_y} \tag{10.57}$$

where $\bar{x}$ and $\bar{y}$ are the sample means and s_x and s_y are the sample variances of x and y, respectively, and $\overline{xy}$ is the sample mean of the product xy. Under H_0, x

and y are independent, which leads to the following expectations:

$$E\left[\sum x_i y_i\right] = \sum E\left[x_i y_i\right] = \sum E\left[x_i\right] E\left[y_i\right] = nE\left[x\right] E\left[y\right]$$

Since $E\left[\bar{x}\bar{y}\right] = E\left[x\right] E\left[y\right]$, it follows that

$$E\left[r\right] = 0$$

Higher moments of r can also be easily calculated. It turns out that the variance is $V\left[r\right] = \frac{1}{n-1}$. Thus, the first two moments are exactly equal to the moments of the bivariate normal distribution with zero correlation. Further, the third and fourth moments are asymptotically approximately equal to those of the normal distribution. From this it follows[11] that

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{10.58}$$

is distributed approximately as Student's t-distribution with $(n-2)$ degrees of freedom, the approximation being very accurate even for small n. The confidence level can therefore be calculated from the t-distribution. H_0 is then rejected for large values of $|t|$.

Rank tests

The rank of an observation x_i is simply its position, j, among the order statistics (*cf.* section 10.6.7), *i.e.*, the position of x_i when all the observations are ordered. In other words,

$$\mathbf{rank}(x_i) = j \qquad \mathbf{if} \quad x_{(j)} = x_i \tag{10.59}$$

The relationship between statistics, order statistics and rank is illustrated in the following table:

	i	1	2	3	4	5	6
statistic (measurement)	x_i	7.1	3.4	8.9	1.1	2.0	5.5
order statistic	$x_{(i)}$	1.1	2.0	3.4	5.5	7.1	8.9
rank	$\mathbf{rank}(x_i)$	5	3	6	1	2	4

For each pair of observations (x_i, y_i), the difference in rank

$$D_i = \mathbf{rank}(x_i) - \mathbf{rank}(y_i) \tag{10.60}$$

is calculated. Spearman's rank correlation coefficient is then defined as

$$\rho = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n} D_i^2 \tag{10.61}$$

which can take on values between -1 and 1. If x and y are completely correlated, x_i and y_i will have the same rank and D_i will be zero, leading to $\rho = 1$. It can be shown[1,11] that for large n (≥ 10) ρ has the same distribution as r in the previous section, and Student's t-distribution can be used, substituting ρ for r in equation 10.58.

10.7.2 Tests of randomness

Given n observations, x_i, ordered according to some other variable, *e.g.*, time, called the trend variable, we wish to test whether the x_i are random in, *i.e.*, independent of, the trend variable, t. H_0 is then that all the x_i are distributed according to the same p.d.f.

As already remarked, we can test for randomness in the same way as for independence by making a y-variable equal to the trend variable, $y_i = t_i$.

If the trend is assumed to be monotonic, additional tests are possible. The reader is referred to Kendall and Stuart.[11,13]

10.7.3 Two-sample tests

Given independent samples of n_1 and n_2 observations, we wish to test whether they come from the same p.d.f. The hypothesis to be tested is thus

$$H_0: \qquad f_1(x) = f_2(x)$$

If both samples contain the same number of observations ($n_1 = n_2$), we can group the two samples into one sample of pairs of observations and apply one of the tests for independence. However, we can also adapt (without the restriction $n_1 = n_2$) any of the goodness-of-fit tests (section 10.6) to this problem.

Kolmogorov test

The Kolmogorov test (*cf.* section 10.6.7) adapted to the two-sample problem compares the sample c.d.f.'s of the two samples. Equations 10.53-10.55 become

$$D_{n_1 n_2} = \max\{|S_{n_1}(x) - S_{n_2}(x)|\} \qquad \text{for all } x \tag{10.62}$$

$$D^{\pm}_{n_1 n_2} = \max\{\pm[S_{n_1}(x) - S_{n_2}(x)]\} \qquad \text{for all } x \tag{10.63}$$

However, now the critical values given in section 10.6.7 are in terms of $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1 n_2}$ rather than $\sqrt{n} D_n$ and $4\frac{n_1 n_2}{n_1+n_2}(D^{\pm}_{n_1 n_2})^2$ rather than $4nD^{\pm}_n$, respectively.

Run test

The two samples are combined keeping track of the sample from which each observation comes. Runs in the sample number, rather than in the sign of the deviation,

are then found. In the notation of section 10.6.6, A and B correspond to an observation coming from sample 1 and sample 2, respectively. The test then follows as in section 10.6.6.

χ^2 test

Consider two histograms with identical binning. Let n_{ji} be the number of entries in bin i of histogram j. Each histogram has k bins and a total of N_j entries. The Pearson χ^2 statistic (equation 10.44) becomes a sum over all bins of both histograms,

$$X^2 = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(n_{ji} - N_j p_i)^2}{N_j p_i} \tag{10.64}$$

Under H_0 the probability content p_i of bin i is the same for both histograms and it is estimated from the combined histogram:

$$\hat{p}_i = \frac{n_{1i} + n_{2i}}{N_1 + N_2}$$

Substituting this for p_i in equation 10.64 results, after some work, in

$$X^2 = (N_1 + N_2) \left[\frac{1}{N_1} \sum_{i=1}^{k} \frac{n_{1i}^2}{n_{1i} + n_{2i}} + \frac{1}{N_2} \sum_{i=1}^{k} \frac{n_{2i}^2}{n_{1i} + n_{2i}} - 1 \right] \tag{10.65}$$

In the usual limit of a large number of events in each bin, X^2 is distributed as a $\chi^2(k-1)$. The number of degrees of freedom is $k-1$, since that is the number of parameters specified by H_0. In other words, there are $2(k-1)$ free bins, and $(k-1)$ parameters are estimated from the data, leaving $(k-1)$ degrees of freedom.

This is directly generalizable to more than two histograms. For r histograms,

$$X^2 = \left[\sum_{j=1}^{r} N_r \right] \cdot \left[\sum_{j=1}^{r} \left(\frac{1}{N_r} \sum_{i=1}^{k} \frac{n_{ji}^2}{\sum_{j=1}^{r} n_{ji}} \right) - 1 \right] \tag{10.66}$$

which, for all n_{ji} large, behaves as χ^2 with $(r-1)(k-1)$ degrees of freedom.

Mann-Whitney test

As previously mentioned, the two-sample problem can be viewed as a test of independence for which, as we have seen, rank tests can be used. A rank test appropriate for this problem is the Mann-Whitney test, which is also known as the Wilcoxon* test, the rank sum test, or simply the U-test. Let the observations of the first sample be denoted x_i and those of the second sample y_i. Rank them together.

*Wilcoxon proposed the test before Mann and Whitney, but his name is also used for another test, the Wilcoxon matched pairs test, which is different. The use of Mann-Whitney here eliminates

This results in a series like $xyyxxyx$. For each x value, count the number of y values that follow it and add up these numbers. In the above example, there are 3 y values after the first x, 1 after the second, 1 after the third, and 0 after the fourth. Their sum, which we call U_x is 5. Similarly, $U_y = 3 + 3 + 1 = 7$. In fact, you only have to count for one of the variables, since

$$U_x + U_y = N_x N_y$$

U_x can be computed in another way, which may be more convenient, by finding the total rank, R_x, of the x's, which is the sum of the ranks of the x_i. In the example this is $R_x = 1 + 4 + 5 + 7 = 17$. Then U_x is given by

$$U_x = N_x N_y + \frac{N_x(N_x + 1)}{2} - R_x \tag{10.67}$$

Under H_0, one expects $U_x = U_y = \frac{1}{2} N_x N_y$. Asymptotically, U_x is distributed normally[1,11] with mean $\frac{1}{2} N_x N_y$ and variance $\frac{1}{12} N_x N_y (N_x + N_y + 1)$, from which (two-tailed) critical values may be computed. For small samples, one must resort to tables.

This test can be easily extended[11] to r samples: For each of the $\frac{1}{2} r(r-1)$ pairs of samples, U_x is calculated (call it U_{pq} for the samples p and q) and summed

$$U = \sum_{p=1}^{r} \sum_{q=p+1}^{r} U_{pq} \tag{10.68}$$

Asymptotically U is distributed normally under H_0 with mean and variance:

$$E[U] = \frac{1}{4} \left(N^2 - \sum_{p=1}^{r} N_p^2 \right) \tag{10.69}$$

$$V[U] = \frac{1}{72} \left[N^2(2N + 3) - \sum_{p=1}^{r} N_p^2 (2N_p + 3) \right] \tag{10.70}$$

where $N = \sum_{p=1}^{r} N_p$.

10.7.4 Two-Gaussian-sample tests

The previous two-sample tests make *no* assumptions about the distribution of the samples and are completely general. If we know something about the distribution we can make more powerful tests. Often, thanks to the central limit theorem, the distribution is (at least to a good approximation) Gaussian. If this is not the case, a simple transformation such as $x \to \ln x$, $x \to x^2$, or $x \to 1/x$ may result in a distribution which is nearly Gaussian. If we are testing whether two samples have the same distribution, testing the transformed distribution is equivalent to testing the original distribution. We now consider tests for two samples under the assumption that both are normally distributed.

Test of equal mean

As we have already done several times when dealing with normal distributions, we distinguish between cases where the variance of the distributions is or is not known.

Known σ: Suppose we have two samples, x_i and y_i, both known to have a Gaussian p.d.f. with variance σ_x^2 and σ_y^2, respectively. If $\sigma_x^2 = \sigma_y^2$, the hypothesis that the two Gaussians are the same is equivalent to the hypothesis that their means are the same, or that the difference in their means, $\theta = \mu_x - \mu_y$, is zero. An obvious test that the means are equal, also valid when $\sigma_x^2 \neq \sigma_y^2$ is given by an estimate of this difference, $\hat{\theta} = \hat{\mu}_x - \hat{\mu}_y$, which has variance

$$V\left[\hat{\theta}\right] = V\left[\hat{\mu}_x\right] + V\left[\hat{\mu}_y\right] = \frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}$$

We know that the difference of two normally distributed random variables is also normally distributed. Therefore, $\hat{\theta}$ will be distributed as a Gaussian with variance $V\left[\hat{\theta}\right]$ and mean 0 or non-0 under H_0 and H_1, respectively. H_0 is then rejected for large $|\hat{\theta}|$ and the size of the test follows from the integral of the Gaussian over the critical region as in sections 10.4.1 and 10.4.2. This is, of course, just a question of how many standard deviations $\hat{\theta}$ is from zero, and rejection of H_0 if $\hat{\theta}$ is found to be too many σ from zero.

Unknown σ: If the parent p.d.f. of each sample is known to be normal, but its variance is unknown, we can estimate the variance for each sample:

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{N_x}(x_i - \bar{x})^2}{N_x - 1} \quad ; \quad \hat{\sigma}_y^2 = \frac{\sum_{i=1}^{N_y}(y_i - \bar{y})^2}{N_y - 1} \tag{10.71}$$

A Student's-t variable can then be constructed. Recall that such a r.v. is the ratio of a standard Gaussian r.v. to the square root of a reduced χ^2 r.v. Under H_0, $\mu_x = \mu_y$ and $\hat{\theta} = (\bar{x} - \bar{y})/\sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}$ is normally distributed with mean 0 and variance 1. From equation 10.71 we see that

$$\chi^2 = \frac{(N_x - 1)\hat{\sigma}_x^2}{\sigma_x^2} + \frac{(N_y - 1)\hat{\sigma}_y^2}{\sigma_y^2} \tag{10.72}$$

is distributed as χ^2 with $N_x + N_y - 2$ degrees of freedom, the loss of 2 degrees of freedom coming from the determination of $\bar{x}$ and $\bar{y}$. The ratio, $\hat{\theta}/\sqrt{\chi^2/(N_x + N_y - 2)}$, is then distributed as Student's t. However, we can calculate this only if σ_x and

possible confusion.

σ_y can be eliminated from the expression. This occurs if $\sigma_x = \sigma_y$, resulting in

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{N_x} + \frac{1}{N_y}}} \tag{10.73}$$

$$\textbf{where} \qquad S^2 = \frac{(N_x - 1)\hat{\sigma}_x^2 + (N_y - 1)\hat{\sigma}_y^2}{N_x + N_y - 2} \tag{10.74}$$

Note that S^2 is in fact just the estimate of the variance obtained by combining both samples.

We emphasize that this test rests on two assumptions: (1) that the p.d.f. of both samples is Gaussian and (2) that both Gaussians have the same variance. The latter can be tested (*cf.* section 10.7.4). As regards the former, it turns out that this test is remarkably robust. Even if the parent p.d.f. is not Gaussian, this test is a good approximation.[11] This was also the case for the sample correlation (section 10.7.1).

Correlated samples: In the above we have assumed that the two samples are uncorrelated. A common case where samples are correlated is in testing the effect of some treatment. For example, the light transmission of a set of crystals is measured. The crystals are then treated in some way and the light transmission is measured again. One could compare the means of the sample before and after treatment. However, we can introduce a correlation by using the simple mathematical relation $\sum x_i - \sum y_i = \sum(x_i - y_i)$. A crystal whose light transmission was lower than the average before the treatment is likely also to be below the average after the treatment, *i.e.*, there is a positive correlation between the transmission before and after. This reduces the variance of the before-after difference, θ: $\sigma_\theta^2 = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$. We do not have to know the correlation, or indeed σ_x or σ_y, but can estimate the variance of $\theta = x - y$ directly from the data:

$$\hat{\sigma}_\theta^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\theta_i^2 - \bar{\theta}^2\right) \tag{10.75}$$

Again we find a Student's-t variable: $\hat{\theta} = \bar{\theta}$ is normally distributed with variance σ_θ^2/N. Thus, $\sqrt{N}\bar{\theta}/\sigma_\theta$ is a standard normal r.v. Further, $(N-1)\hat{\sigma}_\theta^2/\sigma_\theta^2$ is a χ^2 r.v. of $N-1$ degrees of freedom. Hence, the ratio

$$t = \frac{\bar{\theta}\sqrt{N}}{\hat{\sigma}_\theta} \tag{10.76}$$

is a Student's-t variable of $N-1$ degrees of freedom, one degree of freedom being lost by the determination of $\bar{\theta}$, a result already known from equation 3.40.

Test of equal variance

One could approach this problem as above for the means, *i.e.*, estimate the variance of each sample and compare their difference with zero. However, this requires knowing the means or, if unknown, estimating them. Further, we must know how this difference is distributed.

A more straightforward approach makes use of the F-distribution (*cf.* section 3.14), which is the p.d.f. for the ratio of two reduced χ^2 variables. For each sample, the estimate of the variance (equation 8.3 or 8.7 depending on whether the mean is known) divided by the true variance is related to a χ^2 (*cf.* equation 10.72). Thus

$$F = \frac{\chi_x^2/(N_x - 1)}{\chi_y^2/(N_y - 1)} = \frac{\hat{\sigma}_x^2/\sigma^2}{\hat{\sigma}_y^2/\sigma^2} \tag{10.77}$$

is distributed as the F-distribution. The σ^2 cancels in this expression, and consequently F can be calculated directly from the data. We could just as well have used $1/F$ instead of F; both have the same p.d.f. By convention F is taken > 1. The parameters of the F-distribution are $\nu_1 = N_x - 1$, $\nu_2 = N_y - 1$ if $\hat{\sigma}_x^2$ is in the numerator of equation 10.77.

"Never trust to general impressions, my boy,
but concentrate yourself upon details."
—Arther Conan Doyle: Sherlock Holmes in
"A Case of Identity"

10.7.5 Analysis of Variance

Analysis of Variance (AV or ANOVA), originally developed by R. A. Fisher in the 1920's, is widely used in the social sciences, and there is much literature—entire books—about it. In the physical sciences it is much less frequently used and so will be only briefly treated here in the context of testing whether the means of k normal samples are equal. The method is much more general. In particular, it can be used for parameters in the linear model. As usual, Kendall and Stuart[11,13] provide a wealth of information.

The basic method: One-way classification

Given k samples, each normally distributed with the same unknown variance, σ^2, we want to test whether the means of all samples are the same. Suppose that sample

i contains N_i measurements and has a sample moment $\bar{y}_i$, which estimates its true mean μ_i. Using all $N = \sum_{i=1}^{k} N_i$ measurements we can calculate the overall sample mean $\bar{y}$ in order to estimate the overall true mean μ. The null hypothesis is that $\mu = \mu_i$ for all i.

If the μ_i differ we can expect the $\bar{y}_i$ to differ more from $\bar{y}$ than would be expected from the variance of the parent Gaussian alone. Unfortunately, we do not know σ, which would enable us to calculate this expectation. We can, however, estimate σ from the data. We can do this in two ways: from the variation of y within the samples and from the variation of $\bar{y}$ between samples. The results of these two determinations can be compared and tested for equality. To do this we will construct an F variable (section 3.14). Recall that F is the ratio of two reduced χ^2 variables.

The expected error on the estimated mean is $\sigma/\sqrt{N}$. Therefore, under H_0

$$\chi^2(k) = \sum_{i=1}^{k} \frac{(\bar{y}_i - \mu)^2}{\sigma^2/N_i}$$

is distributed as $\chi^2(k)$. Since μ is unknown, we replace it by its estimate (obtained from the entire sample) to obtain a χ^2 of $k-1$ degrees of freedom:

$$\chi^2(k-1) = \sum_{i=1}^{k} \frac{N_i\,(\bar{y}_i - \bar{y})^2}{\sigma^2} \tag{10.78}$$

A second χ^2 variable is obtained from the estimate of σ for each sample

$$\hat{\sigma}_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(y_j^{(i)} - \bar{y}_i\right)^2 \tag{10.79}$$

(where $y_j^{(i)}$ is element j of sample i) by a weighted average:

$$\hat{\sigma}^2 = \frac{1}{N-k} \sum_{i=1}^{k} (N_i - 1)\,\hat{\sigma}_i^2 \tag{10.80}$$

which is a generalization of equation 10.74. Then $(N-k)\hat{\sigma}^2/\sigma^2$ is a χ^2 r.v. with $N-k$ degrees of freedom, since k sample means, $\bar{y}_i$, have also been determined.

The ratio of these two χ^2 variables, normalized by dividing by their respective numbers of degree of freedom, is an r.v. distributed as $F(k-1, N-k)$:

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k} N_i(\bar{y}_i - \bar{y})^2}{\frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{N_i}(y_j^{(i)} - \bar{y}_i)^2} \tag{10.81}$$

If the hypothesis of equal means is false, the $\bar{y}_i$ will be different and the numerator of equation 10.81 will be larger than expected under H_0 while the denominator, being an average of the sample variance within samples, will be unaffected (remember that the true variance of all samples is known to be the same). Hence large values of F are used to reject H_0 with a confidence level determined from the one-tailed critical values of the F distribution. If there are only two samples, this analysis is equivalent to the previously described two-sample test using Student's t distribution.

Multiway analysis of variance

Let us examine the situation of the previous section in a slightly different way. An estimate of the variance of the (Gaussian) p.d.f. is given by $\hat{\sigma}^2 = Q/(N-1)$ where the "sum of squares" (SS), denoted here by Q (in contrast to previous sections where Q^2 was used), is given (*cf.* equation 8.118) by

$$Q = (N-1)\,\hat{\sigma}^2 = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{10.82}$$

Under H_0, Q/σ^2 is a χ^2 of $N-1$ degrees of freedom. Equation 10.82 can be rewritten

$$\begin{aligned} Q &= \sum_{i=1}^{k} \sum_{j=1}^{N_i} (y_j^{(i)} - \bar{y})^2 \qquad (10.83) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(y_j^{(i)} - \bar{y}_i + \bar{y}_i - \bar{y}\right)^2 \\ &= \sum_{i=1}^{k} \left\{ \sum_{j=1}^{N_i} \left[\left(y_j^{(i)} - \bar{y}_i\right)^2 + (\bar{y}_i - \bar{y})^2 \right] + 2\,(\bar{y}_i - \bar{y}) \sum_{j=1}^{N_i} \left(y_j^{(i)} - \bar{y}_i\right) \right\} \end{aligned}$$

The second term is zero since both its sums are equal:

$$\sum_{j=1}^{N_i} y_j^{(i)} = \sum_{j=1}^{N_i} \bar{y}_i = N_i \bar{y}_i$$

Hence,

$$Q = (N-1)\,\hat{\sigma}^2 = \sum_{i=1}^{k} (N_i - 1)\,\hat{\sigma}_i^2 + \sum_{i=1}^{k} N_i\,(\bar{y}_i - \bar{y})^2 \tag{10.84}$$

There are thus two contributions to our estimate of the variance of the p.d.f.: The first term is the contribution of the variance of the measurements within the samples; the second is that of the variance between the samples. Also the number of degrees of freedom are partitioned. As we have seen in the previous section, the first and second terms are related to χ^2 variables of $N-k$ and $k-1$ degrees of freedom, respectively, and their sum, $N-1$, is the number of degrees of freedom of the χ^2 variable associated with $\hat{\sigma}^2$.

Now suppose the samples are classified in some way such that each sample has two indices, *e.g.*, the date of measurement and the person performing the measurement. We would like to partition the overall variance between the various sources: the variance due to each factor (the date and the person) and the innate residual variation. In other words, we seek the analog of equation 10.84 with three terms. We then want to test whether the mean of the samples is independent of each factor separately.

Of course, the situation can be more complicated. There can be more than two factors. The classification is called "crossed" if there is a sample for all combinations of factors. More complicated is the case of "nested" classification where this is not the case. Further, the number of observations in each sample can be different. We will only treat the simplest case, namely two-way crossed classification.

We begin with just one observation per sample. As an example, suppose that there are a number of technicians who have among their tasks the weighing of samples. As a check of the procedure, a reference sample is weighed once each day by each technician. One wants to test (a) whether the balance is stable in time, *i.e.*, gives the same weight each day, and (b) that the weight found does not depend on which technician performs the measurement.

In such a case the measurements can be placed in a table with each row corresponding to a different value of the first factor (the date) and each column to a value of the second factor (the technician). Suppose that there are R rows and C columns. The total number of measurements is then $N = RC$. We use subscripts r and c to indicate the row and column, respectively. The sample means of row r and column c are given, respectively, by

$$\bar{y}_{r.} = \frac{1}{C}\sum_{c=1}^{C} y_{rc} \qquad ; \qquad \bar{y}_{.c} = \frac{1}{R}\sum_{r=1}^{R} y_{rc} \tag{10.85}$$

In this notation a dot replaces indices which are averaged over, except that the dots are suppressed if all indices are averaged over ($\bar{y} \equiv \bar{y}_{..}$). We now proceed as in equations 10.82-10.84 to separate the variance (or more accurately, the sum of squares, SS) between rows from the rest:

$$Q = \sum_r \sum_c (y_{rc} - \bar{y})^2 \tag{10.86}$$

$$= \sum_r \sum_c (y_{rc} - \bar{y}_r)^2 + C\sum_r (\bar{y}_{r.} - \bar{y})^2 \tag{10.87}$$

where C is, of course, the same for all rows and hence can be taken out of the sum over r. The second term, to be denoted Q_R, is the contribution to the SS due to variation between rows while the first term contains both the inter-column and the innate, or residual, contributions.

We can, in the same way, separate the SS between rows from the rest. The result can be immediately written down by exchanging columns and rows in equation 10.87:

$$Q = \sum_c \sum_r (y_{rc} - \bar{y}_{.c})^2 + R\sum_c (\bar{y}_{.c} - \bar{y})^2 \tag{10.88}$$

The residual contribution, Q_{W}, to the SS can be obtained by subtracting the inter-row and inter-column contributions from the total:

$$\begin{aligned} Q_{\mathrm{W}} &= Q - Q_{\mathrm{R}} - Q_{\mathrm{C}} \\ &= \sum_r \sum_c (y_{rc} - \bar{y})^2 - C\sum_r (\bar{y}_{r.} - \bar{y})^2 - R\sum_c (\bar{y}_{.c} - \bar{y})^2 \end{aligned}$$

which, using the fact that

$$\sum_r \sum_c y_{rc} = C \sum_r \bar{y}_{r.} = R \sum_c \bar{y}_{.c} = CR\bar{y} \tag{10.89}$$

can be shown to be equal to

$$Q_{\mathrm{W}} = \sum_r \sum_c (y_{rc} - \bar{y}_{r.} - \bar{y}_{.c} + \bar{y})^2$$

We have thus split the variance into three parts. The number of degrees of freedom also partitions:

Two-way Crossed Classification – Single Measurements			
Factor	SS		d.o.f.
Row	Q_{R}	$= C \sum_r (\bar{y}_{r.} - \bar{y})^2$	$R - 1$
Column	Q_{C}	$= R \sum_c (\bar{y}_{.c} - \bar{y})^2$	$C - 1$
Residual	Q_{W}	$= \sum_r \sum_c (y_{rc} - \bar{y}_{r.} - \bar{y}_{.c} + \bar{y})^2$	$RC - R - C + 1$
Total	Q	$= \sum_r \sum_c (y_{rc} - \bar{y})^2$	$RC - 1$

Divided by their respective numbers of degrees of freedom, the SS are, under H_0, all estimators of σ^2. The hypotheses H_0^{R}, that the means of all rows are equal, and H_0^{C}, similarly defined for columns, can be separately tested by the one-tailed F-test using, respectively,

$$F^{\mathrm{R}} = \frac{\frac{1}{R-1} Q_{\mathrm{R}}}{\frac{1}{(R-1)(C-1)} Q_{\mathrm{W}}} \quad , \quad F^{\mathrm{C}} = \frac{\frac{1}{C-1} Q_{\mathrm{C}}}{\frac{1}{(R-1)(C-1)} Q_{\mathrm{W}}} \tag{10.90}$$

Let us now look at this procedure somewhat more formally. What we, in fact, have done is used the following model for our measurements:

$$y_{rc} = \mu + \theta_r + \omega_c \quad , \quad \sum_r \theta_r = \sum_c \omega_c = 0 \tag{10.91}$$

which is a linear model with $R + C + 1$ parameters subject to $R + C$ constraints. The measurements are then equal to $\mu + \theta_r + \omega_c + \epsilon_{rc}$ where the measurement errors, ϵ_{rc}, are assumed to be normally distributed with the same variance. The hypothesis to be tested is that all the θ_r and ω_c are 0. The θ_r and the ω_c can be tested separately. The least squares estimator for θ_r is

$$\hat{\theta}_r = \frac{1}{C} \sum_c y_{rc} - \hat{\mu} = \bar{y}_{r.} - \bar{y} \tag{10.92}$$

If all θ_r are zero, which is the case under H_0, then

$$\chi^2 = \sum_r \frac{\hat{\theta}_r^2}{\sigma^2/C} = \frac{C \sum_r (\bar{y}_{r.} - \bar{y})^2}{\sigma^2} = \frac{Q_{\mathrm{R}}}{\sigma^2} \tag{10.93}$$

is a χ^2 of $R-1$ degrees of freedom. However, since σ^2 is unknown, we can not use this χ^2 directly.

As shown above, a second, independent χ^2 can be found, namely Q_{W}/σ^2, which is from that part of the sum of squares not due to inter-row or inter-column variation. This χ^2 is then combined with that of equation 10.93 to make an F-test. for the hypothesis that all θ_r are zero. Similarly, an F-test can be derived for the hypothesis that all ω_c are zero. The method can be extended to much more complicated linear models.

However, we will go just one step further: two-way crossed classification with several, K, observations per class. We limit ourselves to the same number, K, for all classes. It is now possible to generalize the model by allowing "interaction" between the factors. The model is

$$y_{rck} = \mu + \theta_r + \omega_c + v_{rc} \quad , \quad \sum_r \theta_r = \sum_c \omega_c = \sum_r \sum_c v_{rc} = 0 \qquad (10.94)$$

where k is the index specifying the observation within class rc.

In our example of different technicians and different dates, the variance among technicians can now depend on the date. (On a day a technician does not feel well the measurements might show more variation.)

The null hypothesis that all θ_r, ω_c, and v_{rc} are zero is equivalent to three hypotheses all being true, namely H_0^{R} that all θ_r are zero, a similar H_0^{C} for columns, and H_0^{I} that all v_{rc} are zero. These three hypotheses can all be tested separately.

Here too, the procedure of equations 10.82-10.84 can be followed with the addition of a sum over k. The result is the partition of the sum of squares over four terms:

Two-way Crossed Classification			
Factor	SS		d.o.f.
Row	Q_{R}	$= CK \sum_r (\bar{y}_{r..} - \bar{y})^2$	$R-1$
Column	Q_{C}	$= RK \sum_c (\bar{y}_{.c.} - \bar{y})^2$	$C-1$
Interaction	Q_{I}	$= K \sum_r \sum_c (\bar{y}_{rc.} - \bar{y}_{r..} - \bar{y}_{.c.} + \bar{y})^2$	$RC-R-C+1$
Residual	Q_{W}	$= \sum_r \sum_c \sum_k (y_{rck} - \bar{y}_{rc.})^2$	$RC(K-1)$
Total	Q	$= \sum_r \sum_c \sum_k (y_{rck} - \bar{y})^2$	$RCK-1$

where the averages are, *e.g.*,

$$\bar{y}_{r..} = \frac{1}{CK} \sum_c \sum_k y_{rck} \quad , \quad \bar{y}_{rc.} = \frac{1}{K} \sum_k y_{rck}$$

F-tests can be constructed using Q_{R}, Q_{C}, and Q_{I} together with Q_{W}.

Part IV

Bibliography

1. R. J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences* (Wiley, 1989)

2. Siegmund Brandt, *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*, Third edition (Springer 1999)

3. A. G. Frodesen, O. Skjeggestad, and H. Tøfte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Bergen, 1979)

4. W. T. Eadie *et al.*, *Statistical Methods in Experimental Physics* (North-Holland, 1971)

5. Frederick James, *Statistical Methods in Experimental Physics*, 2nd edition (World Scientific, 2006)

6. G. P. Yost, *Lectures on Probability and Statistics* (Lawrence Berkeley Laboratory report LBL-16993, 1984)

7. Glen Cowan, *Statistical Data Analysis* (Oxford University Press, 1998)

8. Louis Lyons, *Statistics for Nuclear and Particle Physicists* (Cambridge University Press, 1986)

9. Stuart L. Meyer, *Data Analysis for Scientists and Engineers* (Wiley, 1975)

10. Philip R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, 1969)

11. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics* (Griffin, vol. I, $4^{\mathbf{th}}$ ed., 1977; vol. II, $4^{\mathbf{th}}$ ed., 1979; vol. III, $3^{\mathbf{rd}}$ ed., 1976)

12. Alan Stuart and Keith Ord, *Kendall's Advanced Theory of Statistics, vol. 1, Distribution Theory* (Arnold, 1994).

13. Alan Stuart, Keith Ord, and Steven Arnold, *Kendall's Advanced Theory of Statistics, vol. 2A, Classical Inference and the Linear Model* (Arnold, 1999).

14. Anthony O'Hagan, *Kendall's Advanced Theory of Statistics, vol. 2B, Bayesian Inference* (Arnold, 1999).

15. Harald Cramér, *Mathematical Methods of Statistics* (Princeton Univ. Press, 1946)

16. William H. Press, "Understanding Data Better with Bayesian and Global Statistical Methods", preprint astro-ph/9604126 (1996).

17. anon., *Genesis* **13**.*9*

18. T. Bayes, "An essay towards solving a problem in the doctrine of chances", *Phil. Trans. Roy. Soc.*, **liii** (1763) 370; reprinted in *Biometrika* **45** (1958) 293.

19. P. S. de Laplace, "Mémoire sur la probabilité des causes par les évenements", *Mem. Acad. Sci. Paris* **6** (1774) 621; Eng. transl.: "Memoir on the probability of the causes of events" with an introduction by S. M. Stigler, *Statist. Sci.* **1** (1986) 359.

20. P. S. de Laplace, *Théorie analytique des probabilités* (Courcier Imprimeur, Paris, 1812); $3^{\mathbf{rd}}$ edition with supplements (1820).

21. A. N. Kolmogorov, *Foundations of the Theory of Probability* (Chelsea, 1950)

22. T. Bayes, *An introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst* (1736)

23. Richard von Mises, *Wahrscheinlichkeit, Statistik und Wahrheid* (1928); reprinted as *Probability, Statistics, and Truth* (Dover, 1957)

24. R. A. Fisher, *Proc. Cambridge Phil. Soc.* **26** (1930) 528.

25. L. von Bortkiewicz, *Das Gesatz der kleinen Zahlen* (Teubner, Leipzig, 1898)

26. Pierre Simon de Laplace, *Histoire de l'Acadamie* (1783) 423.

27. A. de Moivre, *Approximatio ad summam terminorum binomii* $(a+b)^n$ *in seriem expansi* (1733). An English translation is included in *The Doctrine of Chances* (second edition, 1738, and third edition, 1756); the third edition has been reprinted (Chelsea Publ. Co., New York, 1967)

28. K. F. Gauß, *Theoria motus corporum celestium* (Perthes, Hamburg, 1809); Eng. transl., *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections* (Dover, New York, 1963)

29. "Student", *Biometrika* **6** (1908) 1.

30. F. James, *Rep. on Progress in Physics* **43** (1980) 1145.

31. F. James, *Computer Phys. Comm.* **60** (1990) 329.

32. G. Buffon, "Essai d'arithmétique morale," *Histoire naturelle, générale et particulière, Supplément* **4** (1777) 46.

33. R. Y. Rubinstein, *Simulation and the Monte Carlo Method* (Wiley, 1981)

34. International Organization for Standardization (ISO), *Guide to the expression of uncertainty in measurement* (Geneva, 1993)

35. C. R. Rao, *Bull. Calcutta Math. Soc.* **37** (1945) 81.

36. A. C. Aitken and H. Silverstone, *Proc. Roy. Soc. Edin.* **A 61** (1942) 186.

37. H. Jeffreys, *Theory of Probability* (Oxford Univ. Press, 1961)

38. CERN program library, entry D506;
F. James and M. Roos, *Computer Phys. Comm.* **10** (1975) 343.

39. Henry Margenau and George Mosely Murphy, *The Mathematics of Physics and Chemistry* (vol. I, Van Nostrand, 1956)

40. G. Forsythe, *J. Soc. Indust. Appl. Math.* **5** (1957) 74.

41. B. Efron, *Ann. Statist.* **7** (1979) 1.

42. B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (S.I.A.M., 1982)

43. B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall 1993)

44. J.W. Tukey, *Ann. Math. Stat.* **29** (1958) 614.

45. J. Neyman, *Phil. Trans. Roy. Soc. London*, series A, **236** (1937) 333.

46. Particle Data Group: *'Review of Particle Physics'*, *Phys. Rev. D* **54** (1996) 1.

47. Particle Data Group: *'Review of Particle Physics'*, *Eur. Phys. J. C* **15** (2000) 1.

48. G. J. Feldman and R. D. Cousins, *Phys. Rev. D* **57** (1998) 3873.

49. O. Helene, *Nucl. Instr. Methods* **212** (1983) 319.

50. W. J. Metzger, *'Upper limits'*, Internal report University of Nijmegen HEN-297 (1988)

51. J. Neyman and E. S. Pearson, *Phil. Trans.* **A 231** (1933) 289.

52. J. Neyman and E. S. Pearson, *Biometrika* **A 20** (1928) 175 and 263.

53. K. Pearson, *Phil. Mag.* **5** (1900) 157.

54. T. W. Anderson and D. A. Darling, *Ann. Math. Stat.* **23** (1952) 193.

55. William H. Press *et al.*, *Numerical Recipes in FORTRAN90: The Art of Scientific Computing*, *Numerical Recipes in FORTRAN77: The Art of Scientific Computing*, *Numerical Recipes in C++: The Art of Scientific Computing*, *Numerical Recipes in C: The Art of Scientific Computing*, (Cambridge Univ. Press).

56. N. H. Kuiper, *Proc. Koninklijke Nederlandse Akademie van Wetenschappen*, ser. A **63** (1962) 38.

57. M. A. Stephens, *J. Roy. Stat. Soc.*, ser. B **32** (1970) 115.

58. D. A. Darling, *Ann. Math. Stat.* **28** (1957) 823.

59. J. R. Michael, *Biometrika* **70,1** (1983) 11.

60. J. Durbin, *Biometrika* **62** (1975) 5.

Part V

Exercises

1. In statistics we will see that the moments of the parent distribution can be 'estimated', or 'measured', by calculating the corresponding moment of the data, *e.g.*, $\overline{x} = \frac{1}{n}\sum x_i$ gives an estimate of the mean μ and $\sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$ estimates σ, *etc.*

   (a) Histogram the following data using a suitable bin size.

90	90	79	84	78	91	88	90	85	80
88	75	73	79	78	79	67	83	68	60
73	79	69	74	76	68	72	72	75	60
61	66	66	54	71	67	75	49	51	57
62	64	68	58	56	79	63	68	64	51
58	53	65	57	59	65	48	54	55	40
49	42	36	46	40	37	53	48	44	43
35	39	30	41	41	22	28	36	39	51

   These data will be available in a file, which can be read, *e.g.*, in `FORTRAN` by

```
READ(11,'(10F4.0)') X
```

   where `X` is an array defined by `REAL X(80)`.

   (b) Estimate the mean, standard deviation, skewness, mode, median and FWHM (full width at half maximum) using the data and using the histogram bin contents and the central values of the bins.

   You may find the `FORTRAN` subroutine `FLPSOR` useful: `CALL FLPSOR(X,N)`, where `N` is the dimension, *e.g.*, 80, of the array `X`. After calling this routine, the order of the elements of `X` will be in ascending order.

2. Verify by making a histogram of 1000 random numbers that your random number generator indeed gives an approximately uniform distribution in the interval 0 to 1.

   Make a two-dimensional histogram using successive pairs of random numbers for the x and y coordinates. Does this two-dimensional distribution also appear uniform? Calculate the correlation coefficient between x and y.

3. Let $X_i, i = 1, 2, ..., n$, be n independent r.v.'s uniformly distributed between 0 and 1, *i.e.*, the p.d.f. is $f(x) = 1$ for $0 \leq x \leq 1$ and $f(x) = 0$ otherwise. Let Y be the maximum of the n X_i: $Y = \max(X_1, X_2, ..., X_n)$. Derive the p.d.f. for Y, $g(y)$. Hint: What is the c.d.f. for Y?

4. For two r.v.'s, x and y, show that

$$V[x + y] = V[x] + V[y] + 2\,\mathrm{cov}(x, y)$$

5. Show that the skewness can be written

$$\gamma_1 = \frac{1}{\sigma^3}\left(E\left[x^3\right] - 3E\left[x\right]E\left[x^2\right] + 2E\left[x\right]^3\right)$$

6. The Chebychev Inequality. Assume that the p.d.f. for the r.v. X has mean μ and variance σ^2. Show that for any positive number k, the probability that x will differ from μ by more than k standard deviations is less than or equal to $1/k^2$, *i.e.*, that

$$P\left(|x-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

7. Show that $|\,\mathrm{cov}(x,y)| \leq \sigma_x\sigma_y$, *i.e.*, that the correlation coefficient, $\rho_{x,y} = \mathrm{cov}(x,y)/\sigma_x\sigma_y$, is in the range $-1 \leq \rho \leq 1$ and that $\rho = \pm 1$ if and only if x and y are linearly related.

8. A beam of mesons, composed of 90% pions and 10% kaons, hits a Čerenkov counter. In principle the counter gives a signal for pions but not for kaons, thereby identifying any particular meson. In practice it is 95% efficient at giving a signal for pions, and also has a 6% probability of giving an accidental signal for a kaon. If a meson gives a signal, what is the probability that the particle was a pion? If there is no signal, what is the probability that it was a kaon?

9. Mongolian swamp fever (MSF) is such a rare disease that a doctor only expects to meet it once in 10000 patients. It always produces spots and acute lethargy in a patient; usually (60% of cases) they suffer from a raging thirst, and occasionally (20% of cases) from violent sneezes. These symptoms can arise from other causes: specifically, of patients who do not have MSF, 3% have spots, 10% are lethargic, 2% thirsty, and 5% complain of sneezing. These four probabilities are independent.

   Show that if you go to the doctor with all these symptoms, the probability of your having MSF is 80%. What is the probability if you have all these symptoms except sneezing?

10. Suppose that an antimissile system is 99.5% efficient in intercepting incoming ballistic missiles. What is the probability that it will intercept all of 100 missiles launched against it? How many missiles must an aggressor launch to have a better than even chance of one or more penetrating the defenses? How many missiles would be needed to ensure a better than even chance of more than two missiles evading the defenses?

11. A student is trying to hitch a lift. Cars pass at random intervals, at an average rate of 1 per minute. The probability of a car giving a student a lift is 1%. What is the probability that the student will still be waiting:

(a) after 60 cars have passed?

(b) after 1 hour?

12. Show that the characteristic function of the Poisson p.d.f.,

$$P(r;\mu) = \frac{\mu^r e^{-\mu}}{r!}$$

is

$$\phi(t) = \exp\left[\mu\left(e^{it} - 1\right)\right]$$

Use the characteristic function to prove the reproductive property of the Poisson p.d.f.

13. A single number often used to characterize an angular distribution is the forward-backward ratio, F/B, or the forward-backward asymmetry, $\frac{F}{N}$,where F is the number of events with $\cos\theta > 0$, B is the number of events with $\cos\theta < 0$, and $N = F + B$ is the total number of events. Assume that the events are independent and that the event rate is constant, for both forward and backward events.

Clearly, only two of the three variables, F, B, N, are independent. We can regard this situation in two ways:

(a) The number of events N is Poisson distributed with mean μ and they are split into F and $B = N - F$ following a binomial p.d.f., $B(F; N, p_F)$, *i.e.*, the independent variables are N and F.

(b) The F events and B events are both Poisson distributed (with parameters $\mu_{\mathbf{F}}$ and $\mu_{\mathbf{B}}$), and the total is just their sum, *i.e.*, the independent variables are F and B.

Show that both ways lead to the same p.d.f.

14. Show that the Poisson p.d.f. tends to a Gaussian with mean μ and variance $\sigma^2 = \mu$ for large μ, *i.e.*,

$$P(r;\mu) \longrightarrow N(r;\mu,\mu)$$

For $\mu = 5.3$, what is the probability of 2 or less events? Approximating the discrete Poisson by the continuous Gaussian p.d.f., ≤ 2 should be regarded as < 2.5, half way between 2 and 3. What is the probability in this approximation?

15. For a Gaussian p.d.f.:

(a) What is the probability of a value lying more than 1.23σ from the mean?

(b) What is the probability of a value lying more than $\mathbf{2.43}\boldsymbol{\sigma}$ above the mean?

(c) What is the probability of a value lying less than $\mathbf{1.09}\boldsymbol{\sigma}$ below the mean?

(d) What is the probability of a value lying above a point $\mathbf{0.45}\boldsymbol{\sigma}$ below the mean?

(e) What is the probability that a value lies more than $\mathbf{0.5}\boldsymbol{\sigma}$ but less than $\mathbf{1.5}\boldsymbol{\sigma}$ from the mean?

(f) What is the probability that a value lies above $\mathbf{1.2}\boldsymbol{\sigma}$ on the low side of the mean, and below $\mathbf{2.1}\boldsymbol{\sigma}$ on the high side?

(g) Within how many standard deviations does the probability of a value occurring equal 50%?

(h) How many standard deviations correspond to a one-tailed probability of 99%?

16. During a meteor shower, meteors fall at the rate of 15.7 per hour. What is the probability of observing less than 5 in a given period of 30 minutes? What value do you find if you approximate the Poisson p.d.f. by a Gaussian p.d.f.?

17. Four values (3.9, 4.5, 5.5, 6.1) are drawn from a normal p.d.f. whose mean is known to be 4.9. The variance of the p.d.f. is unknown.

    (a) What is the probability that the next value drawn from the p.d.f. will have a value greater than 7.3?

    (b) What is the probability that the mean of three new values will be between 3.8 and 6.0?

18. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two independent r.v.'s, each distributed uniformly between 0 and 1. Define $\boldsymbol{z_\pm = x \pm y}$.

    (a) How are $\boldsymbol{z_+}$ and $\boldsymbol{z_-}$ distributed?

    (b) What is the correlation between $\boldsymbol{z_+}$ and $\boldsymbol{z_-}$; between $\boldsymbol{z_+}$ and $\boldsymbol{y}$?

    It will probably help your understanding of this situation to use Monte Carlo to generate points uniform in $\boldsymbol{x}$ and $\boldsymbol{y}$ and to make a two-dimensional histogram of $\boldsymbol{z_+}$ *vs.* $\boldsymbol{z_-}$.

19. Derive the reproductive property of the Gaussian p.d.f., *i.e.*, show that if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent r.v.'s distributed normally as $\boldsymbol{N(x;\mu_x,\sigma_x^2)}$ and $\boldsymbol{N(y;\mu_y,\sigma_y^2)}$, respectively, then $\boldsymbol{z = x + y}$ is also normally distributed as $\boldsymbol{N(z;\mu_z,\sigma_z^2)}$. Show that $\boldsymbol{\mu_z = \mu_x + \mu_y}$ and $\boldsymbol{\sigma_z^2 = \sigma_x^2 + \sigma_y^2}$. Derive also the p.d.f. for $\boldsymbol{z = x - y}$, for $\boldsymbol{z = (x+y)/2}$, and for $\boldsymbol{z = \bar{x} = \sum_{i=1}^{n} x_i/n}$ when all the $\boldsymbol{x_i}$ are normally distributed with the same mean and variance.

20. For the bivariate normal p.d.f. for x, y with correlation coefficient ρ, transform to variables u, v such that the covariance matrix is diagonal and show that

$$\sigma_u^2 = \frac{\sigma_x^2 \cos^2\theta - \sigma_y^2 \sin^2\theta}{\cos^2\theta - \sin^2\theta}$$
$$\sigma_v^2 = \frac{\sigma_y^2 \cos^2\theta - \sigma_x^2 \sin^2\theta}{\cos^2\theta - \sin^2\theta}$$
$$\text{where} \quad \tan 2\theta = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$

21. Show that for the bivariate normal p.d.f., the conditional p.d.f., $f(y|x)$, is a normal p.d.f. with mean and variance,

$$E[y|x] = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) \quad \text{and} \quad V[y|x] = \sigma_y^2(1 - \rho^2)$$

22. For a three-dimensional Gaussian p.d.f. the contours of constant probability are ellipsoids defined by constant

$$G = (\underline{x} - \underline{\mu})^{\mathrm{T}}\underline{V}^{-1}(\underline{x} - \underline{\mu})$$

Find the probability that a point is within the ellipsoid defined by $G = 1$.

23. Given n independent variables, x_i, distributed according to f_i having mean, μ_i, and variance, $V_i = \sigma_i^2$, show that $S = \sum x_i$ has mean $\mu_S = E[S] = \sum \mu_i$ and variance $V[S] = \sum V_i = \sum \sigma_i^2$. What are the expected value and variance of the average of the x_i, $\bar{x} = \frac{1}{n}\sum x_i$?

24. Derive the reproductive property of the Cauchy p.d.f. Does the p.d.f. of the sum of n independent, Cauchy-distributed r.v.'s, approach the normal p.d.f. in the limit $n \to \infty$?

25. Let x and y be independent r.v.'s, each distributed normally with mean 0 and variances σ_x^2 and σ_y^2, respectively.

    (a) Derive the p.d.f. of the r.v. $z = x/y$.

    (b) Describe a method to generate random numbers distributed as a standard Cauchy p.d.f. Try it.

26. (a) Show that for n independent r.v.'s, x_i, uniformly distributed between 0 and 1, the p.d.f. for

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

approaches $N(g; 0, 1)$ for $n \to \infty$.

(b) Demonstrate the result (a) by generating by Monte Carlo the distribution of g for $n = 1$, 2, 3, 5, 10, 50 and comparing it to $N(g; 0, 1)$.

(c) If the x_i are uniformly distributed in the intervals $[0.0, 0.2]$ and $[0.8, 1.0]$. *i.e.*,

$$\begin{aligned} f(x) &= \tfrac{1}{0.4}, \quad 0.0 \le x \le 0.2 \ or \ 0.8 \le x \le 1.0 \\ &= 0\ , \quad \text{otherwise,} \end{aligned}$$

what distribution will g approach? Demonstrate this by Monte Carlo as in (b).

27. Show that the weighting method used in the two-dimensional example of crude Monte Carlo integration (sect. 6.2.5, eq. 6.5) is in fact an application of the technique of importance sampling.

28. Perform the integral $I = \int_0^1 x^3\, dx$ by crude Monte Carlo using 100, 200, 400, and 800 points. Estimate not only I, but also the error on I. Does the error decrease as expected with the number of points used?

    Repeat the determination of I 50 times using 200 (different) points each time and histogram the resulting values of I. Does the histogram have the shape that you expect? Also evaluate the integral by the following methods and compare the error on I with that obtained by crude Monte Carlo:

    (a) using hit or miss Monte Carlo and 200 points.

    (b) using crude Monte Carlo and stratification, dividing the integration region in two, (0,0.5) and (0.5,1), and using $f \cdot 200$ points in (0,0.5) and $(1 - f) \cdot 200$ points in (0.5,1), where f = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Plot the error on I *vs.* f.

    (c) as (b) but for f=0.5 with various intervals, (0,c) and (c,1), for c = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. Make a plot of the estimated error on I *vs.* c.

    (d) using crude Monte Carlo and antithetic variables x and $(1 - x)$ and 200 points.

    (e) as (d) but with only 100 points.

    (f) using importance sampling with the function $g(x) = x^2$ and 200 points.

29. Generate 20000 Monte Carlo points with $x > 0$ distributed according to the distribution

$$f(x) = \frac{1}{2}\left(\frac{1}{\tau}e^{-x/\tau} + \frac{1}{\lambda}e^{-x/\lambda}\right)$$

for $\tau = 3$ and $\lambda = 10$. Do this for (a) the weighting, (b) the rejection, and (c) the composite methods using inverse transformations. Which method

is easiest to program? Which is fastest? Make histograms of the resulting distribution in each case and verify that the distribution is correct.

If you can only detect events with $1 < x < 10$, what fraction of the events will you detect? Suppose in addition, that your detector has a detection efficiency given by

$$e = \begin{cases} 0, & \text{if } x < 1 \text{ or } x > 10 \\ (x-1)/9, & \text{if } 1 < x < 10 \end{cases}$$

How can you arrive at a histogram for the x-distribution of the events you detect? There are various methods. Which should be the best?

30. Generate 1000 points, x_i, from the Gaussian p.d.f. $N(x; 10, 5^2)$. Use each of the following estimators to estimate the mean of X: sample mean, sample median, and trimmed sample mean (10%).

    Repeat assuming we only measure values of X in the interval (5,25), *i.e.* if an x_i is outside this range, throw it away and generate a new value.

    Repeat this all 25 times, histogramming each estimation of the mean. From these histograms determine the variance of each of the six estimators.

31. Under the assumptions that the range of the r.v. X is independent of the parameter θ and that the likelihood, $\mathcal{L}(x;\theta)$, is regular enough to allow interchanging $\frac{\partial^2}{\partial\theta^2}$ and $\int \mathbf{d}x$, derive equation 8.23,

$$I_x(\theta) = -E\left[\frac{\partial}{\partial\theta} S(x;\theta)\right]$$

32. Show that the estimator $\widehat{\sigma^2} = \sum(x_i - \mu)^2/n$ is an efficient estimator of the variance of a Gaussian p.d.f. of known mean by showing that its variance is equal to I^{-1}.

33. Using the method of section 8.2.7, find an efficient and unbiased estimator for σ^2 of a normal p.d.f. when μ is known and there is thus only one parameter for the distribution.

34. We count the number of decays in a fixed time interval, T. We do this N times yielding the results n_i, $i = 1, ..., N$. The source is assumed to consist of a large number of atoms having a long half-life. The data, n_i, are therefore assumed to be distributed according to a Poisson p.d.f., the parameter of which can be estimated by $\hat{\mu} = \bar{n}$ (section 8.3.2). Suppose, however, that we want instead to estimate the probability of observing no decays in the time interval, T.

    (a) What is the estimator in the frequency method of estimation?

(b) Derive a less biased estimator.

(c) Derive the variances of both the estimator and the less biased estimator.

35. (a) Derive equations 8.41 and 8.42, *i.e.*, show that the variance of the rth sample moment is given by

$$V\left[\overline{x^r}\right] = \frac{1}{n}\left[E\left[x^{2r}\right] - \left(E\left[x^r\right]\right)^2\right]$$

and that

$$\text{cov}\left[\overline{x^r}, \overline{x^q}\right] = \frac{1}{n}\left[E\left[x^{r+q}\right] - E\left[x^r\right]E\left[x^q\right]\right]$$

(b) Derive an expression in terms of sample moments to estimate the variance, $V\left[\widehat{\sigma^2}\right]$, of the moments estimator of the parent variance, $\widehat{\sigma^2} = \widehat{m}_2 - \widehat{m}_1^2$

36. We estimate the values of x and y by their sample means, $\bar{x}$ and $\bar{y}$, which have variances σ_x^2 and σ_y^2. The covariance is zero. We want to estimate the values of r and θ which are related to x and y by

$$r^2 = x^2 + y^2 \qquad \text{and} \qquad \tan\theta = \frac{y}{x}$$

Following the substitution method, what are $\hat{r}$ and $\hat{\theta}$? Find the variances and covariance of $\hat{r}$ and $\hat{\theta}$.

37. We measure $x = 10.0 \pm 0.5$ and $y = 2.0 \pm 0.5$. What is then our estimate of x/y? Use Monte Carlo to investigate the validity of the error propagation.

38. We measure $\cos\theta$ and $\sin\theta$, both with standard deviation σ. What is the ML estimator for θ? Compare with the results of exercise 36.

39. Decay times of radioactive atoms are described by an exponential p.d.f. (equation 3.10):

$$f(t;\tau) = \frac{1}{\tau}\, e^{-t/\tau}$$

(a) Having measured the times t_i of n decays, how would you estimate τ and the variance $V\left[\hat{\tau}\right]$ (1) using the moments method and (2) using the maximum likelihood method? Which method do you prefer? Why?

(b) Generate 100 Monte Carlo events according to this p.d.f. with $\tau = 10$, (*cf.* exercise 29) and calculate $\hat{\tau}$ and $V\left[\hat{\tau}\right]$ using both the moments and the maximum likelihood methods. Are the results consistent with $\tau = 10$? Which method do you prefer? Why?

(c) Use a minimization program, *e.g.*, `MINUIT`, to find the maximum of the likelihood function for the Monte Carlo events of (39b). Evaluate $\boldsymbol{V}[\hat{\boldsymbol{\tau}}]$ using both the second-derivative matrix and the variation of $\boldsymbol{l}$ by $1/2$. Compare the results for $\hat{\boldsymbol{\tau}}$ and $\boldsymbol{V}[\hat{\boldsymbol{\tau}}]$ with those of (39b).

(d) Repeat (39b) 1000 times making histograms of the value of $\hat{\boldsymbol{\tau}}$ and of the estimate of the error on $\hat{\boldsymbol{\tau}}$ for each method. Do you prefer the moments or the maximum likelihood expression for $\boldsymbol{V}[\hat{\boldsymbol{\tau}}]$? Why?

(e) Suppose that we can only detect times $\boldsymbol{t} < \mathbf{10}$. What is then the likelihood function? Use a minimization program to find the maximum of the likelihood function and thus $\hat{\boldsymbol{\tau}}$ and its variance. Does this value agree with $\boldsymbol{\tau} = \mathbf{10}$?

(f) Repeat (39b) and (39e) with 10000 Monte Carlo events.

40. Verify that a least squares fit of independent measurements to the model $\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{bx}$ results in estimates for $\boldsymbol{a}$ and $\boldsymbol{b}$ given by

$$\hat{\boldsymbol{a}} = \bar{\boldsymbol{y}} - \hat{\boldsymbol{b}}\bar{\boldsymbol{x}} \qquad \text{and} \qquad \hat{\boldsymbol{b}} = \frac{\overline{\boldsymbol{xy}} - \bar{\boldsymbol{x}}\bar{\boldsymbol{y}}}{\overline{\boldsymbol{x}^2} - \bar{\boldsymbol{x}}^2}$$

where the bar indicates a weighted sample average with weights given by $\mathbf{1}/\boldsymbol{\sigma}_i^2$, as stated in section 8.5.5.

41. Use the method of least squares to derive formulae to estimate the value (and its error), $\boldsymbol{y} \pm \boldsymbol{\delta y}$, from a set of $\boldsymbol{n}$ measurements, $\boldsymbol{y}_i \pm \boldsymbol{\delta y}_i$. Assume that the $\boldsymbol{y}_i$ are uncorrelated. Comment on the relationship between these formulae and those derived from ML (equations 8.59 and 8.60).

42. Perform a least squares fit of a parabola

$$\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\boldsymbol{x} + \boldsymbol{\theta}_3\boldsymbol{x}^2$$

for the four independent measurements: $\mathbf{5 \pm 2, 3 \pm 1, 5 \pm 1, 8 \pm 2}$ measured at the points $\boldsymbol{x} = \mathbf{-0.6, -0.2, 0.2, 0.6}$, respectively. Determine not only the $\hat{\boldsymbol{\theta}}_i$ and their covariances, but also calculate the value of $\boldsymbol{y}$ and its uncertainty at $\boldsymbol{x} = \mathbf{1}$.

To invert a matrix you can use the routine `RSINV`:

```
CALL RSINV (N,A,N,IFAIL)
```

where `A` is a symmetric, positive matrix of dimension `(N,N)`. If the matrix inversion is successful, `IFAIL` is returned as 0.

43. The three angles of a triangle are independently measured to be $\mathbf{63°}$, $\mathbf{34°}$, and $\mathbf{85°}$, all with a resolution of $\mathbf{1°}$.

(a) Calculate the least squares estimate of the angles subject to the requirement that their sum be $\mathbf{180°}$.

(b) Calculate the covariance matrix of the estimators.

44. Generate events as in exercise 39b. Histogram the times t_i and use the two minimum chi-square methods and the binned maximum likelihood method to estimate the lifetime τ. Use a minimization program, *e.g.*, `MINUIT`, to find the minima and maximum. Compare the results of these three methods and those of exercise 39b.

45. In section 8.7.4 is a table comparing the efficiencies of various location estimators for various distributions. Generate 10000 random numbers from a standard normal distribution and estimate the mean using each of the estimators in the table. Repeat this 1000 times making histograms of the values of each estimator. The standard deviation of these histograms is an estimate of the standard deviation of the estimator. Are these in the ratio expected from the table?

46. Consider a long-lived radioactive source.

    (a) In our detector it produces 389 counts in the first minute and 423 counts in the second minute. Assuming a 100% efficient detector, what is the best estimation of the activity of the source?

    (b) What can you say about the best value and uncertainty for the activity of the source from the following set of independent measurements?

$$\mathbf{1.08 \pm 0.13} \quad , \quad \mathbf{1.04 \pm 0.07} \quad , \quad \mathbf{1.13 \pm 0.10} \quad \text{Bq.}$$

47. A current is determined by measuring the voltage V across a standard resistor. The voltmeter has a resolution σ_V and a systematic error s_V. We measure two currents using the same resistor and voltmeter. Since the resistance is unchanged between the measurements, we regard its uncertainty as entirely systematic. Find the covariance matrix for the two currents, which are calculated using Ohm's law, $I_i = V_i/R$.

48. We measure a quantity X 25 times using an apparatus of unknown but constant resolution. The average value of the measurements is $\bar{x} = \mathbf{128}$. The estimate of the variance is $s^2 = \frac{1}{24}\sum(x - \bar{x})^2 = \mathbf{225}$. What is the 95% confidence interval on the true value, μ, of the quantity X?

49. You want to determine the probability, p, that a student passes the statistics exam. Since there are only two possible outcomes, pass and fail, the appropriate p.d.f. is binomial, $B(k; N, p)$.

    (a) Construct the confidence belt for a 95% central confidence interval for p for the case that 10 students take the exam and k pass, *i.e.*, draw $k_+(p)$ and $k_-(p)$ curves on a p *vs.* k plot.

(b) Assume that 8 of the 10 pass. Find the 95% central confidence interval from the plot constructed in (a) and by solving equation 9.18.

50. An experiment studying the decay of the proton (an extremely rare process, if it occurs at all) observes 7 events in 1 year for a sample of $\mathbf{10^6}$ kg of hydrogen.

    (a) Assume that there is no background. Give a 90% central confidence interval and a 90% upper limit for the expected number of proton decays and from these calculate the corresponding interval and limit for the mean lifetime of the proton.

    (b) Repeat (a) assuming that background processes are expected to contribute an average of 3 events per year.

    (c) Repeat (a) assuming 8 expected background events per year.

51. Construct a most powerful (MP) test for one observation, x, for the hypothesis that X is distributed as a Cauchy distribution,

$$f(x) = \frac{1}{\pi\,[1 + (x - \theta)^2]}$$

with $\theta = 0$ under H_0 and $\theta = 1$ under H_1. What is the size of the test if you decide to reject H_0 when $x > 0.5$?

52. Ten students each measure the mass of a sample, each with an error of 0.2 g:

10.2 10.4 9.8 10.5 9.9 9.8 10.3 10.1 10.3 9.9 g

    (a) Test the hypothesis that they are all measurements of a sample whose true mass is 10.1 g.

    (b) Test the hypothesis that they are all measurements of the same sample.

53. On Feb. 23, 1987, the Irvine-Michigan-Brookhaven experiment was counting neutrino interactions in their detector. The time that the detector was on was split into ten-second intervals, and the number of neutrino interactions in each interval was recorded. The number of intervals containing i events is shown in the following table. There were no intervals containing more than 9 events.

Number of events	0	1	2	3	4	5	6	7	8	9
Number of intervals	1042	860	307	78	15	3	0	0	0	1

This date was also the date that astronomers first saw the supernova S1987a.

    (a) Test the hypothesis that the data are described by a Poisson distribution.

(b) Test the hypothesis that the data are described by the sum of two Poisson distributions, one for a signal of 9 events within one ten-second interval, and another for the background of ordinary cosmic neutrinos.

54. Marks on an exam are distributed over male and female students as follows (it is left to your own bias to decide which group is male):

Group 1	39	18	3	22	24	29	22	22	27	28	23	48
Group 2	42	23	36	35	38	42	33					

Assume that test scores are normally distributed within each group.

(a) Assume that the variance of the scores of both groups is the same, and test the hypothesis that the mean is also the same for both groups.

(b) Test the assumption that the variance of the scores of both groups is the same.

55. The light transmission of crystals is degraded by ionizing radiation. Folklore, and some qualitative physics arguments, suggest that it can be (partially) restored by annealing. To test this the light transmission of 7 crystals, which have been exposed to radiation, is measured. The crystals are then annealed, and their light transmission again measured. The results:

Crystal	1	2	3	4	5	6	7
Before	29	30	42	34	37	45	32
After	36	26	46	36	40	51	33
difference	7	**−4**	4	2	3	6	1

Assume that the uncertainty in the measurement of the transmission is normally distributed.

(a) Test whether the light transmission has improved using only the mean of the before and after measurements.

(b) Test whether the light transmission has improved making use of the measurements per crystal, *i.e.*, using the differences in transmission.

For the following exercises you will be assigned a file containing the data to be used. It will consist of 3 numbers per event, which may be read, *e.g.*, in `FORTRAN` by

```
READ(11,'(I5)') NEVENTS
READ(11,'(3F10.7)') ((E(I,IEV),I=1,3),IEV=1,NEVENTS)
```

The data may be thought of as being the measurement of the radioactive decay of a neutral particle at rest into two new particles, one positive and one negative, with

`E(1,IEV) =` x, the mass of the decaying particle as determined from the energies of the decay products. The mass values have a certain spread due to the resolution of our apparatus and/or the Heisenberg uncertainty principle (for a very short-lived particle).

`E(2,IEV) =` $\cos\theta$, the cosine of the polar angle of the positive decay particle's direction.

`E(3,IEV) =` ϕ/π, the azimuthal angle, divided by π, of the positive decay particle's direction. Division by π results in a more convenient quantity to histogram.

Assume that the decay is of a vector meson to two pseudo-scalar mesons. The decay angular distribution is then given by

$$f(\cos\theta,\phi) = \frac{3}{4\pi}\left[\frac{1}{2}(1-\rho_{00}) + \frac{1}{2}(3\rho_{00}-1)\cos^2\theta - \rho_{1,-1}\sin^2\theta\cos 2\phi \right.$$
$$\left. -\sqrt{2}\mathrm{Re}\rho_{10}\sin 2\theta\cos\phi\right]$$

A1. Use the moments method to estimate the mass of the particle and the decay parameters ρ_{00}, $\rho_{1,-1}$, and $\mathrm{Re}\rho_{10}$. Also estimate the variance and standard deviation of the p.d.f. for x. Estimate also the errors of all of the estimates.

A2. Use the maximum likelihood method to estimate the decay parameters ρ_{00}, $\rho_{1,-1}$, and $\mathrm{Re}\rho_{10}$ using a program such as `MINUIT` to find the maximum of the likelihood function. Determine the errors on the estimates using the variation of the likelihood.

A3. Assume that x is distributed normally. Determine μ and σ using maximum likelihood. Also determine the covariance matrix of the estimates.

A4. Assume that x is distributed normally. Determine μ and σ using both the minimum χ^2 and the binned maximum likelihood methods. Do this twice, once with narrow and once with wide bins. Compare the estimates and their covariance matrix obtained with these two methods with each other and with that of the previous exercise.

A5. Test the assumption of vector meson decay against the hypothesis of decay of a scalar meson, in which case the angular distribution must be isotropic.

For the following exercises you will be assigned a file containing the data to be used. It is the same situation as in the previous exercises except that it is somewhat more realistic, having some background to the signal.

B1. From an examination of histograms of the data, make some reasonable hypotheses as to the nature of the background, *i.e.*, propose some functional form for the background, $\boldsymbol{f_b(x)}$ and $\boldsymbol{f_b(\cos\theta, \phi)}$.

B2. Modify your likelihood function to include your hypothesis for the background, and use the maximum likelihood method to estimate the decay parameters $\boldsymbol{\rho_{00}}$, $\boldsymbol{\rho_{1,-1}}$, and $\mathbf{Re}\boldsymbol{\rho_{10}}$ as well as the fraction of signal events. Also determine the position of the signal, $\boldsymbol{\mu}$, and its width, $\boldsymbol{\sigma}$, under the assumption that the signal $\boldsymbol{x}$ is normally distributed. Determine the errors on the estimates using the variation of the likelihood.

B3. Develop a way to use the moments method to estimate, taking into account the background, the decay parameters $\boldsymbol{\rho_{00}}$, $\boldsymbol{\rho_{1,-1}}$, and $\mathbf{Re}\boldsymbol{\rho_{10}}$. Estimate also the errors of the estimates.

B4. Determine the goodness-of-fit of the fits in the previous two exercises. There are several goodness-of-fit tests which could be applied. Why did you choose the one you did?