

# **Advanced Statistics for Data Analysis**

Course Notes, A. Y. 2014-15

Edoardo Milotti

*Dipartimento di Fisica, Università di Trieste, Trieste, Italy*  
e-mail: *milotti@units.it*



# Contents

<b>Preface</b>	<b>ix</b>
<b>1 Basic probability concepts</b>	<b>1</b>
1.1 Probability and relative frequency . . . . .	1
1.2 Rudiments of combinatorics . . . . .	6
1.3 Stirling's formula . . . . .	8
1.3.1 Simple, heuristic argument . . . . .	8
1.3.2 A modern proof . . . . .	10
1.4 What's the use of all this? . . . . .	11
<b>2 The algebra of probability</b>	<b>15</b>
2.1 Elementary events. The sample space . . . . .	15
2.1.1 The Addition Law for Probabilities . . . . .	16
2.2 Conditional probability . . . . .	20
2.2.1 Statistical independence . . . . .	22
2.2.2 Bayes' theorem . . . . .	23
2.2.3 Gambler's ruin . . . . .	23
2.2.4 Bertrand's paradox . . . . .	26
2.3 The Borel-Cantelli lemmas . . . . .	30
2.3.1 The first Borel-Cantelli lemma . . . . .	32
2.3.2 The second Borel-Cantelli lemma . . . . .	33
2.4 What's the use of all this? . . . . .	34
<b>3 Random variables</b>	<b>37</b>
3.1 Discrete and Continuous Random Variables. Distribution Functions. . . . .	37
3.1.1 The uniform distribution . . . . .	40
3.1.2 Example: segment length . . . . .	40
3.1.3 Example: Buffon's needle . . . . .	41
3.2 Mathematical Expectation . . . . .	42
3.3 Variance, standard deviation and covariance . . . . .	45
3.3.1 Higher moments . . . . .	46

3.4	Chebyshev's inequality, the laws of large numbers and other important inequalities . . . . .	47
3.4.1	Chebyshev's inequality and the weak law of large numbers . . . . .	47
3.4.2	Other inequalities in probability theory . . . . .	48
3.4.3	The strong law of large numbers . . . . .	50
3.4.4	Chernoff bound . . . . .	53
3.5	What's the use of all this? . . . . .	56
<b>4</b>	<b>Probability distributions</b>	<b>57</b>
4.1	The uniform distribution . . . . .	57
4.2	The Bernoulli distribution . . . . .	58
4.3	The binomial distribution . . . . .	58
4.4	The multinomial distribution . . . . .	60
4.5	The Poisson distribution . . . . .	63
4.5.1	The lottery ticket problem . . . . .	66
4.5.2	Cell plating statistics . . . . .	66
4.6	The exponential distribution . . . . .	66
4.6.1	Detector dead time . . . . .	67
4.7	The De Moivre-Laplace theorem. The normal distribution . . . . .	69
4.7.1	The error function . . . . .	74
4.7.2	High-rate limit of the Poisson distribution . . . . .	74
4.8	Transformations of random variables . . . . .	75
4.8.1	Sum of two random variables . . . . .	75
4.8.2	General considerations on continuous random variable transformations . . . . .	76
4.8.3	Probability density function of a generic function of several random variables . . . . .	77
4.8.4	Error propagation . . . . .	78
4.8.5	Orthogonal transformations of random variables . . . . .	79
4.9	Multivariate Gaussian distribution . . . . .	81
4.10	The log-normal distribution . . . . .	83
4.11	The gamma distribution . . . . .	84
4.12	The Cauchy-Lorentz (Breit-Wigner) distribution . . . . .	85
4.13	The Landau distribution . . . . .	87
4.14	The Rayleigh distribution . . . . .	87
4.15	A short selection of other important probability distributions	88
4.16	Bertrand's paradox, again . . . . .	90
4.16.1	Rotational invariance . . . . .	90
4.16.2	Scale invariance . . . . .	91
4.16.3	Translational invariance . . . . .	92
4.17	Probability in Quantum Mechanics . . . . .	94
4.18	What's the use of all this? . . . . .	98

<b>5 The Central Limit Theorem (CLT)</b>	<b>101</b>
5.1 Generating Functions . . . . .	101
5.1.1 Definition and simple examples . . . . .	101
5.1.2 Probability generating functions . . . . .	104
5.1.3 Photomultiplier noise statistics . . . . .	107
5.2 Characteristic functions . . . . .	112
5.3 The Central Limit Theorem (CLT) . . . . .	118
5.3.1 Convergence rate . . . . .	119
5.3.2 Additive and multiplicative processes . . . . .	120
5.4 What's the use of all this? . . . . .	121
<b>6 Markov processes</b>	<b>125</b>
6.1 Transition Probabilities . . . . .	125
6.1.1 The discrete-time, discrete-space random walk . . . . .	128
6.2 The Ehrenfest urn model . . . . .	129
6.3 Classification of States. Persistent and Transient States . . . . .	130
6.3.1 Example: the simple random walk . . . . .	134
6.3.2 Example: higher dimensional random walks on a lattice	134
6.4 Limiting Probabilities and Stationary Distributions . . . . .	137
6.5 Time reversal. Detailed balance . . . . .	139
6.6 Continuous time Markov processes . . . . .	140
6.6.1 The H-theorem and approach to equilibrium . . . . .	142
6.7 An important application: the Hidden Markov Models (HMM)	143
6.8 What's the use of all this? . . . . .	144
<b>7 Descriptive statistics</b>	<b>149</b>
7.1 Descriptive statistics . . . . .	149
7.1.1 Sample mean . . . . .	149
7.1.2 Sample variance . . . . .	149
7.1.3 Variance of sample mean . . . . .	151
7.1.4 Sample covariance . . . . .	151
7.1.5 The correlation coefficient . . . . .	152
7.2 Probability distributions of estimators in descriptive statistics	153
7.2.1 Distribution of the sample mean of exponentially distributed samples . . . . .	154
7.2.2 Mean of the sample mean distribution . . . . .	155
7.2.3 Example with exponentially distributed samples . . . . .	157
7.2.4 Gaussian approximation . . . . .	157
7.3 The german tank problem . . . . .	158
7.3.1 Order statistics . . . . .	161
7.4 Inadequacies of the standard deviation . . . . .	164
7.4.1 Sample variance of correlated data. The Allan variance.	166
7.5 What's the use of all this? . . . . .	167

<b>8 Monte Carlo Methods</b>	<b>171</b>
8.1 Introduction to Monte Carlo . . . . .	171
8.2 Empirical probability distributions . . . . .	174
8.3 Uniformly distributed pseudo-random numbers . . . . .	175
8.3.1 Randomness vs. chaotic dynamics . . . . .	176
8.3.2 Standard methods . . . . .	177
8.3.3 RANLUX . . . . .	181
8.4 The transformation method . . . . .	182
8.4.1 Gaussian variates with the transformation method (Box-Müller) . . . . .	183
8.5 The acceptance-rejection method . . . . .	184
8.5.1 Examples: angular distributions in $e^+e^-$ collider physics	186
8.6 Cross-sections and the transformation method . . . . .	188
8.7 Monte Carlo simulation of non relativistic electron scattering in matter . . . . .	190
8.8 What's the use of all this? . . . . .	207
<b>9 Random sampling real data: the statistical bootstrap</b>	<b>213</b>
9.1 Introduction . . . . .	213
9.2 Statistical bootstrap . . . . .	214
9.3 Variance of bootstrap estimates . . . . .	218
9.4 Example with real data (from Efron and Gong) . . . . .	220
9.5 Robust statistics . . . . .	222
9.6 What's the use of all this? . . . . .	224
<b>10 Parametric statistics: maximum likelihood and confidence intervals</b>	<b>229</b>
10.1 Bayes' Theorem and the Principle of Maximum Likelihood . .	229
10.2 An example with exponentially distributed data . . . . .	231
10.3 Estimators and their properties . . . . .	233
10.4 Invariance of ML estimates and bias . . . . .	234
10.5 Normally distributed data . . . . .	235
10.6 Consistency of ML estimators . . . . .	235
10.7 Estimator efficiency and the Cramér-Rao-Fisher bound . .	236
10.7.1 Bartlett identities . . . . .	236
10.7.2 The Cramér-Rao-Fisher bound . . . . .	238
10.7.3 Example with exponentially distributed data . . . . .	239
10.8 Optimality and Gaussianity of consistent ML estimators at large $n$ . . . . .	240
10.9 Interlude on information measures . . . . .	242
10.9.1 Boltzmann entropy . . . . .	242
10.9.2 Shannon information . . . . .	243
10.9.3 Kullback-Leibler divergence . . . . .	244
10.9.4 Fisher information . . . . .	246

10.10 Confidence intervals . . . . .	249
10.10.1 Bayesian credible intervals . . . . .	249
10.10.2 Frequentist confidence intervals . . . . .	254
10.10.3 Example: mean of exponentially distributed data . . .	258
10.10.4 Confidence intervals for a more complex estimator: the correlation coefficient . . . . .	259
10.10.5 Graphical method to evaluate estimator variance in the ML method . . . . .	261
10.10.6 The Feldman-Cousins construction . . . . .	261
10.11 Some interesting examples . . . . .	262
10.11.1 Angular estimate . . . . .	262
10.11.2 Cauchy distribution . . . . .	263
10.12 Extended ML . . . . .	263
10.12.1 A simple example: determination of a forward-backward asymmetry . . . . .	263
10.13 ML with binned data . . . . .	268
10.13.1 Signal enhancement . . . . .	268
10.13.2 ML with binned data . . . . .	269
10.13.3 Example: radioactive decay . . . . .	271
10.13.4 Example: estimate of the exponent of a power-law distribution . . . . .	272
10.14 What's the use of all this? . . . . .	276
<b>11 Fitting data</b>	<b>279</b>
11.1 Parametric models and normally distributed data: the method of least squares . . . . .	279
11.1.1 Correlated data . . . . .	279
11.1.2 Expansion about the minimum . . . . .	280
11.1.3 Criticism of the ML approach to least squares . . . . .	280
11.2 Chi-square and confidence intervals . . . . .	281
11.3 The $\chi^2$ distribution . . . . .	281
11.4 Straight line fit . . . . .	285
11.4.1 Numerical example . . . . .	287
11.5 General linear model . . . . .	292
11.5.1 Chi-square distribution . . . . .	293
11.6 Nonlinear least squares . . . . .	293
11.7 Least squares with binned data . . . . .	294
11.8 Polynomial fits . . . . .	294
11.9 What's the use of all this? . . . . .	295
<b>12 Taking decisions</b>	<b>301</b>
12.1 Definitions and basic concepts . . . . .	301
12.2 P-values . . . . .	305
12.3 The Neyman-Pearson lemma . . . . .	305

12.4 Using $\chi^2$ to test goodness-of-fit . . . . .	307
12.5 Significance of a signal . . . . .	309
12.6 A decision problem: compatibility of measurements . . . . .	316
12.6.1 Standard deviations known . . . . .	316
12.6.2 Standard deviations unknown: Student's $t$ . . . . .	316
12.6.3 Another application of Student's $t$ : confidence limits .	318
12.7 What's the use of all this? . . . . .	318
<b>Index</b>	<b>325</b>

# Preface

This text is growing from a collection of notes and handouts for the course on Advanced Statistics for Data Analysis at the Physics Department of the University of Trieste, and has already reached a considerable size. The birth pangs of all books are marked by many errors and omissions, and I shall be grateful to all students who find them and point them out to me.

The main purpose of the course is to introduce students to the basic techniques and methods used to analyze data from physics experiments, and thus the course is very specific and is far from giving a complete overview of the many, beautiful ideas of modern statistics. The standard frequentist approach to statistics for physicists was set many years ago by the notes of Jay Orear (LBL preprint UCRL-8417), that were based on a series of lectures given at the Radiation Laboratory in the summer of 1958<sup>1</sup>. In the preface to these notes, Jay Orear writes that the primary source for the basic material and approach was Enrico Fermi, so the basic statistical toolbox of most physicists is yet another lasting legacy of Enrico Fermi!

A few years later – in 1964 – Frank Solmitz wrote a short, seminal review paper *Analysis of Experiments in Particle Physics*, which was published in the Annual Review of Nuclear Science. And finally the book *Statistical Methods in Experimental Physics* by W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet (North-Holland, 1971) was another deeply influential source of statistical ideas, which derived directly from the statistics book of Kendall and Stuart *The Advanced Theory of Statistics* (first published in 1948, it went through several editions for many years, and it is still in print today). This book championed the traditional frequentist approach, although it also expounded Bayesian ideas, and it is still being sold today (2nd edition, revised by F. James).

The present notes owe much to a more modern text by G. Cowan, *Statistical Data Analysis*, and to similar, related books by R. J. Barlow, *Statistics. A Guide to the Use of Statistical Methods in the Physical Sciences*, and W. J. Metzger, *Statistical Methods in Data Analysis*.

---

<sup>1</sup>The notes have been revised and reprinted again in 1982 as a Cornell preprint, CLNS 82/511

In recent years the computationally-intensive Bayesian methods have been revived and far advanced, thanks to the power of modern computers. Although these methods now fully complement the standard frequentist approach – and sometimes provide more powerful and effective avenues to data analysis – they are not yet included in these notes. However, a “Bayesian extension” is planned and shall be made available in the future.

The notes also include a review of basic probability concepts – which are needed for a sound logical basis of the statistical methods – and applications and examples where statistical ideas often play an important role as modeling concepts (as in the Monte Carlo simulations) as well as analysis tools.

Where appropriate, I introduce short historical annotations as well: indeed, we should never forget the efforts, the hard work, and sometimes the sufferings, of the scientists who first conceived these important ideas.

Edoardo Milotti, Fall 2014

# Chapter 1

## Basic probability concepts

### 1.1 Probability and relative frequency

The study of probability began with gambling (see the appendix to this chapter for a historical introduction, which contains a text excerpted from T. Apostol, “Calculus, vol. 2, 2nd ed.” [3]), and some of the earliest problems were posed by de Méré, a french nobleman. In particular, he noticed that when throwing three dice, a total of 11 turns out more often than a total of 12, even though according to his reasoning they should appear equally often.

He counted the possible outcomes as follows:

- a total of 11 turns out in 6 ways: (6, 4, 1), (6, 3, 2), (5, 5, 1), (5, 4, 2), (5, 3, 3), (4, 4, 3);
- a total of 12 also turns out in 6 ways: (6, 5, 1), (6, 4, 2), (6, 3, 3), (5, 5, 2), (5, 4, 3), (4, 4, 4);

however this is wrong, why?

The answer is simple, *it is just bad counting*. Take the first configuration, (6, 4, 1), there are 6 ways to obtain it (6, 4, 1), (6, 1, 4), (4, 6, 1), (4, 1, 6), (1, 6, 4), (1, 4, 6); this is true of all configurations where all faces turn out different. When there are 2 equal faces there are just 3 different ways, like (5, 5, 1), (5, 1, 5), (1, 5, 5). Finally when all faces are equal, there is only one way to obtain the given configuration. This means that 11 turns out in 27 ways, while 12 turns out in 25 ways. Since there is a total of  $6^3 = 216$  ways to throw three dice,  $P(11) = 27/216 = 0.125$ ,  $P(12) = 25/216 \approx 0.116$ .

Thus we see that combinatorics, the art and science of counting all configurations, plays an important role in determining the probability of dice throws. The same is true in many similar examples, and finding probabilities boils down to some form of counting.

A little practice also shows that there is a hidden order in the randomness. Indeed, each individual throw is somewhat different from other throws and in general we cannot predict its outcome. However, consider the following series of throws: we notice that some values of the sum of the three faces occur more often than others, while some values are quite rare (for instance there is no 3 in the series, although it is a possible value)

```
(2, 1, 5), sum = 8
(1, 1, 4), sum = 6
(2, 1, 2), sum = 5
(5, 4, 4), sum = 13
(1, 6, 4), sum = 11
(3, 3, 1), sum = 7
(3, 3, 1), sum = 7
(6, 6, 3), sum = 15
(1, 5, 3), sum = 9
(2, 1, 3), sum = 6
(3, 2, 3), sum = 8
(6, 5, 6), sum = 17
(4, 6, 4), sum = 14
(1, 3, 6), sum = 10
(6, 2, 3), sum = 11
(6, 5, 2), sum = 13
(5, 2, 4), sum = 11
(2, 1, 2), sum = 5
(5, 2, 4), sum = 11
(1, 5, 1), sum = 7
```

Figure 1.1 shows histograms of the sum of the three faces. We see that as the number of throws  $n$  increases, histograms look more and more regular.

Here we learn an important lesson: the overall behavior of the dice throw is regular for a large number of throws, and becomes more so as the number increases. This is an example of the *law of large numbers*, that we shall prove in the next chapters (see also figure 1.2).

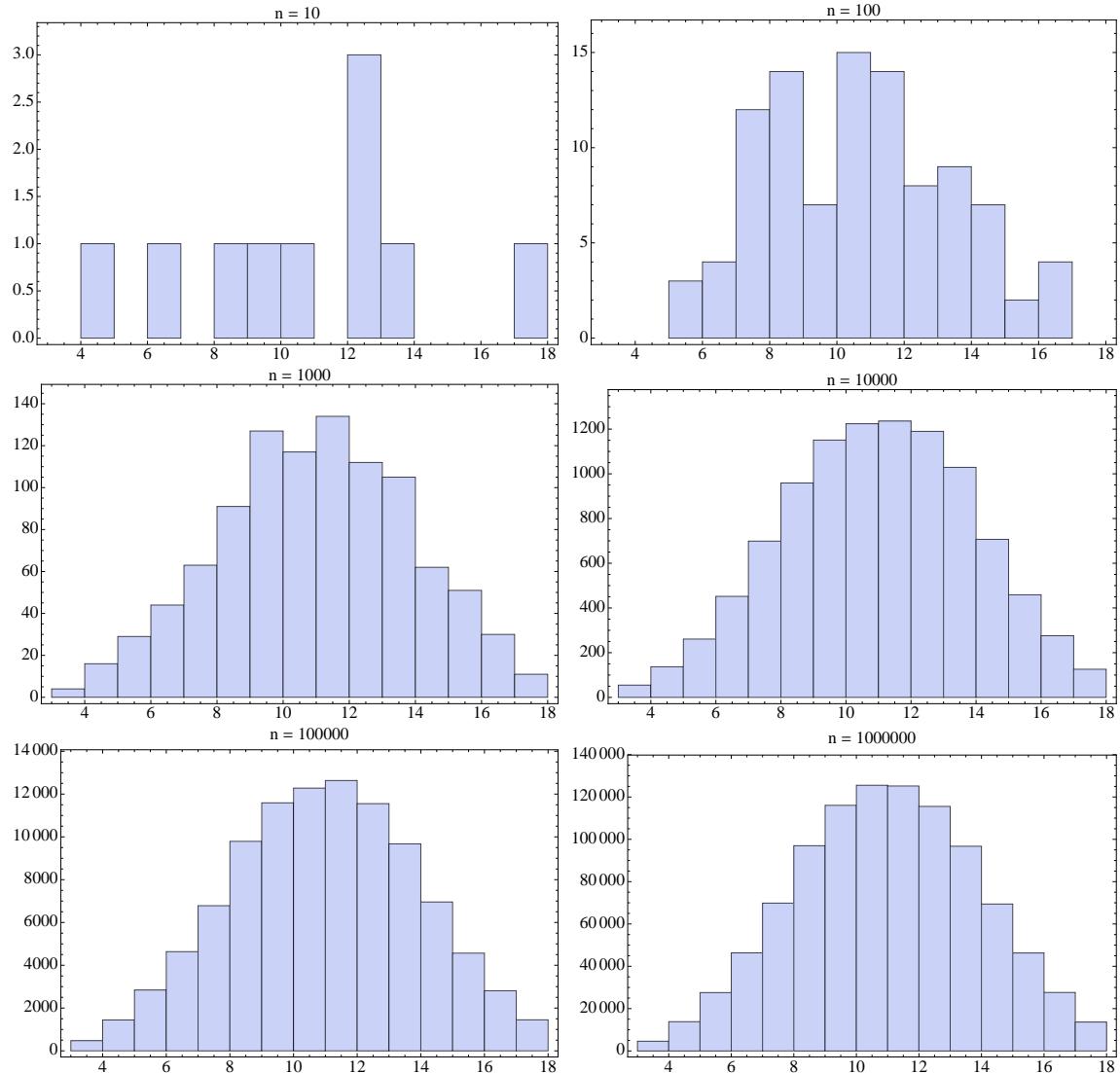


Figure 1.1: Histograms of the sum of three faces in the dice problem of de Méré. As the number of throws  $n$  increases, the histogram look more and more regular.

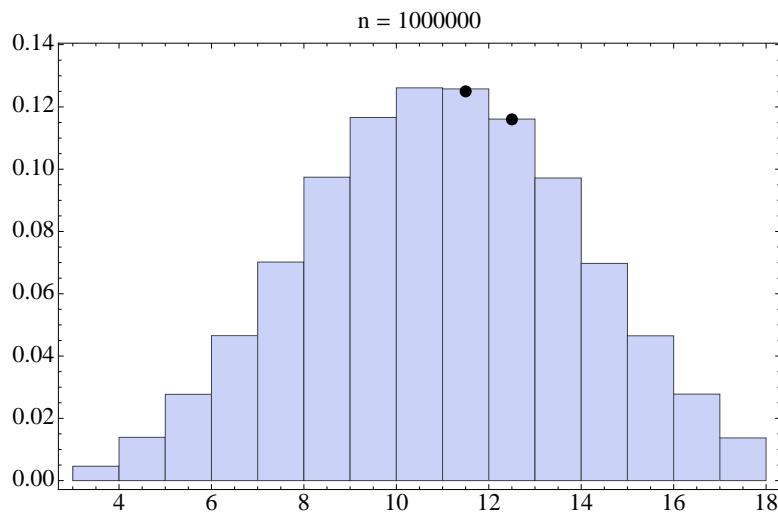


Figure 1.2: Normalized histogram, showing the comparison between the relative frequency values with a large number of throws (here  $n = 10^6$ ) and the expected probability values (indicated by the black dots) in the de Méré dice problem (see text for details). Note that the observed frequency are nearly equal to the theoretical values: this is an illustration of the *law of large numbers*.

The introductory example illustrates the mathematical definition of probability

$$P(A) = \frac{N(A)}{N} \quad (1.1)$$

where  $N$  is the total number of outcomes and  $N(A)$  is the number of *equiprobable* outcomes leading to the occurrence of event A. Note that this definition is based on the intuitive grasp of equiprobable outcomes, and that it *is not* the frequentist definition of probability (the  $N$ 's are not the results of actual experiments). Equiprobability implies a sort of internal symmetry in the problem, and its physical basis in the dice throw problem rests in the chaotic behavior of the physical system of three dice, and on the idea that the system dynamics does not single out any particular configuration.

The *relative frequency*

$$\frac{n(A)}{n} \quad (1.2)$$

where  $n(A)$  is the total number of experiments in which A occurs, out of a total of  $n$ , is an *estimate* of probability  $P(A)$ . It will be shown later that this estimate gets better and better as  $n$  gets larger and larger, and eventually

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n} \quad (1.3)$$

(*Law of large numbers*).

In the *frequentist view*, a similar formula also gives a practical definition of probability

$$P(A) = \frac{n(A)}{n} \quad (1.4)$$

where now  $n$  is the number of actual experiments, and  $n(A)$  is the number of *observed outcomes of A*. The frequentist definition of probability relies on the same law of large numbers. There is a conceptual difference, however: the frequentist formula – although formally the same – is an estimate of the probability of an *assumed probabilistic model* of a physical process. While the mathematical formula is just a statement that is valid in a perfectly determined framework, the frequentist *estimate* of the probability value *assumes* a model which does not necessarily apply to the actual data.

In the *Bayesian view*, probability assignments are *subjective*, i.e., they are determined on the basis of previous information and/or on the basis of a theoretical or heuristic model of the physical world.

The frequentist view is based on the belief that the underlying conceptual model is *true*, while in the Bayesian view there is no such thing as a *true, absolute* model.

Both views concentrate on the physical world and their difference becomes relevant only in statistical inference. In this first part we consider probabilistic models only, and we are not concerned with these subtleties – although we shall be, in the second part of the course.

## 1.2 Rudiments of combinatorics

The assumption of equiprobability of different configurations reduces the calculation of probability to simple counting, and we have seen this concept at work in the de Méré dice throw problem. “Simple” actually is not the right word: although the concept of counting is easy to grasp, in practice it can be an extremely complex task, and combinatorics can lead to very difficult problems, often unsolvable.

Here is a list of simple enumeration problems

**ordered pairs:** the total number of ordered pairs  $(n_1, n_2)$ , where  $n_1 = 1, \dots, n$ ,  $n_2 = 1, \dots, m$ , is  $n \cdot m$ ;

**permutations:** consider a set of  $n$  objects  $a_1, \dots, a_n$ , then there are

$$n! = n(n-1)\dots 2 \cdot 1$$

ways of arranging them;

**pairs taken from a set without replacement:** the total number of (unordered) pairs taken from a set of  $n$  objects is  $n(n-1)/2$ ; this is often useful because it corresponds, e.g., to the number of connections in an undirected graph;

**sampling with replacement:** take a set of  $n$  distinguishable objects, and extract  $r$  times from this set, replacing each time the extracted object back in the set: there are  $n^r$  extraction sequences.

As an example of this kind of sampling, consider the Hawaiian alphabet which has only 12 letters, the vowels  $a, e, i, o, u$  and the consonants  $h, k, l, m, n, p, w$ . How many 4-letter words can be constructed with the Hawaiian alphabet? And how many with the 26-letter English alphabet? How many words can you construct with at most 5 letters with the Hawaiian alphabet? Clearly the answer to the first question is  $12^4 = 20736$ , the answer to the second question is  $26^4 = 456976$ , finally the answer to the third question is

$$\sum_{k=1}^{k=5} 12^k = 271320$$

**sampling without replacement:** as before, take a set of  $n$  distinguishable objects, and extract  $r$  times from this set, however this time do not replace the extracted object back in the set: then there are

$$n(n-1)(n-2)\dots(n-r+1) = \frac{n!}{(n-r)!}$$

ways of doing this;

**combinations:** a subpopulation of  $r$  distinguishable objects can be extracted from a set of  $n$  elements in  $n!/(n-r)!$  ways (there are  $n!$  ways to arrange the  $n$  distinguishable objects, and when we take the first  $r$  objects, the ordering of the remaining  $n-r$  certainly does not matter), however there are  $r!$  orderings for each subpopulation, and therefore, if we neglect ordering inside the subpopulation, there are

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

ways of selecting a subpopulation;

**occupancy problem 1:** consider the case in which  $r$  distinguishable objects must be placed in  $n$  boxes ( $n \geq r$ ), with the condition that there cannot be more than one object in each box; there are  $n$  choices for the first object,  $n-1$  for the second, etc.; eventually there are

$$n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}$$

possible fillings;

**occupancy problem 2:** now assume that multiple fillings of the  $n$  boxes are allowed: in this case with  $r$  objects there are  $n^r$  possible choices, but the order in each box does not matter, therefore, if there are  $n_i$  objects in the  $i$ -th box, such that

$$n_1 + n_2 + \dots + n_n = r$$

we must divide by the total number of redundant arrangements, i.e.,  $n_i!$ , so that the number of possible choices is

$$\frac{n^r}{n_1! \dots n_n!} = \frac{n^r}{\prod_{k=1}^{k=n} n_k!}$$

Here is an interesting, related example problem: a subway train with  $n$  cars is boarded by  $r$  passengers completely at random ( $r \leq n$ ): what is the probability that all passengers end up in an empty car? Clearly there are  $n^r$  possible passenger choices, while there are  $n(n-$

$1) \dots (n - r + 1)$  ways of choosing an empty car, therefore the answer to the problem is

$$P = \frac{n(n-1)\dots(n-r+1)}{n^r}$$

Note that here the equiprobability of the elementary events (the single choices) is an *a priori* hypothesis, spelled out by the qualification that passenger choice is *random*.

**Fermi-Dirac statistics:** this is also an occupancy problem, where  $r$  indistinguishable objects occupy  $n$  boxes ( $n \geq r$ ), with the condition that there can be at most one object in any box. Clearly, there are  $\binom{n}{r}$  ways of selecting the occupied boxes. As an example, notice that misprints follow the Fermi-Dirac statistics!

**Bose-Einstein statistics:** alternatively, a box can be occupied by several objects and there is no constraint between  $n$  and  $r$ . In this case we represent a given configuration in this form  $|000||0|00|||0|$ , where the 0's represent the objects and the bars the box boundaries (external boundaries are fixed, in this case there are 7 boxes and 7 objects). Then we see that we can exchange a total of  $r + n - 1$  objects in  $(r+n-1)!$  ways, and take into account overcounting by factors  $r!$  and  $(n-1)!$ , so that the total number of arrangements is

$$\frac{(r+n-1)!}{r!(n-1)!} = \binom{r+n-1}{r} = \binom{r+n-1}{n-1}$$

### 1.3 Stirling's formula

Factorials are ubiquitous in enumeration problems, and are difficult to evaluate exactly for large arguments. The Stirling formula is an extremely useful approximation of the factorial function. It can be easily justified with a heuristic argument, and proved with more sophisticated mathematical methods. Here we expose the simple argument first and then we move on to a real proof.

#### 1.3.1 Simple, heuristic argument

Take the logarithm of the factorial function

$$\begin{aligned} \ln N! &= \ln \prod_{n=1}^{n=N} n = \sum_{n=1}^{n=N} \ln n \approx \int_1^N \ln x \, dx = (x \ln x - x)|_1^N \\ &= N \ln N - N + 1 \end{aligned} \tag{1.5}$$

where the critical step is the approximation of the sum with an integral. Exponentiating again we can write this formula as

$$N! \approx N^N e^{-(N-1)} \quad (1.6)$$

Table 1 compares the exact value of the factorial with the approximate formula up to  $n = 10$ :

$n$	$\ln n!$	$n \ln n - n + 1$
1	0	0
2	0.693147	0.386294
3	1.79176	1.29584
4	3.17805	2.54518
5	4.78749	4.04719
6	6.57925	5.75056
7	8.52516	7.62137
8	10.6046	9.63553
9	12.8018	11.775
10	15.1044	14.0259

Table 1.1: Comparison between exact values of  $\ln n!$  and the first order Stirling's approximation.

This simple approximation actually underestimates the factorial function (see figure 1.3).

Now note that the area of the  $k$ -th left-out triangle is roughly equal to

$$\frac{1}{2} [\ln(k+1) - \ln k] = \frac{1}{2} \ln \frac{k+1}{k} \quad (1.7)$$

therefore the total correction term is

$$\sum_{n=1}^{N-1} \frac{1}{2} \ln \frac{n+1}{n} = \frac{1}{2} \sum_{n=1}^{N-1} [\ln(n+1) - \ln n] = \frac{1}{2} \ln N \quad (1.8)$$

Thus we find

$$\ln N! \approx N \ln N - N + 1 + \frac{1}{2} \ln N \quad (1.9)$$

which corresponds to the exponentiated form

$$N! \approx N^N \sqrt{N} e^{-(N-1)} \quad (1.10)$$

We can compare the new approximation with the exact values just as before, and we find the results shown in table 2: the new values are closer to the exact ones now, but the approximation is still incomplete.

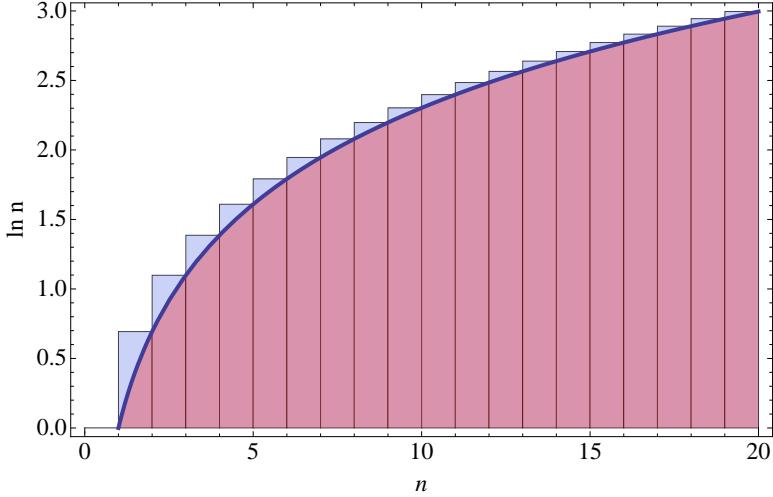


Figure 1.3: Plot of  $\ln n$  vs.  $n$  (bars) and of  $\ln x$  vs.  $x$  (solid line). The integral of  $\ln x$  (red area) underestimates the sum (the nearly triangular areas above the curve are left out).

### 1.3.2 A modern proof

This proof is an adaptation of that due to R. Michel [24]: it is given here skipping many details that are important to show the convergence of the integrals that appear in the proof. First we note that the factorial function can be expressed as a standard integral:

$$n! = \Gamma(n+1) = \int_0^\infty x^n e^{-x} dx \quad (1.11)$$

then, if we make the substitution

$$x = y\sqrt{n} + n \quad (1.12)$$

(so that  $dx = dy\sqrt{n}$  and  $y = (x-n)/\sqrt{n}$ ) we find the modified representation

$$n! = \int_{-\sqrt{n}}^\infty (y\sqrt{n} + n)^n e^{-y\sqrt{n}-n} \sqrt{n} dy = n^n \sqrt{n} e^{-n} \int_{-\sqrt{n}}^\infty \left(1 + \frac{y}{\sqrt{n}}\right)^n e^{-y\sqrt{n}} dy \quad (1.13)$$

The importance of this representation is that it factorizes the  $n$ -dependence, and isolates all constants in the evaluation of the integral: in particular, notice that as  $n \rightarrow \infty$

$$\ln \left(1 + \frac{y}{\sqrt{n}}\right)^n = n \ln \left(1 + \frac{y}{\sqrt{n}}\right) \approx n \left(\frac{y}{\sqrt{n}} - \frac{y^2}{2n}\right) = y\sqrt{n} - \frac{y^2}{2} \quad (1.14)$$

$n$	$\ln n!$	$n \ln n - n + 1 + \frac{1}{2} \ln n$
1	0	0
2	0.693147	0.732868
3	1.79176	1.84514
4	3.17805	3.23832
5	4.78749	4.85191
6	6.57925	6.64644
7	8.52516	8.59433
8	10.6046	10.6753
9	12.8018	12.8736
10	15.1044	15.1771

Table 1.2: Comparison between exact values of  $\ln n!$  and the second order Stirling's approximation: a factor corresponding to a log value  $\approx 0.07$  is still missing.

Therefore

$$\left(1 + \frac{y}{\sqrt{n}}\right)^n e^{-y\sqrt{n}} \approx e^{y\sqrt{n}} e^{-y^2/2} e^{-y\sqrt{n}} = e^{-y^2/2} \quad (1.15)$$

and

$$n! \approx n^n \sqrt{n} e^{-n} \int_{-\infty}^{\infty} e^{-y^2/2} dy = \sqrt{2\pi n} n^n e^{-n} \quad (1.16)$$

The formula

$$n! \approx \sqrt{2\pi n} n^n e^{-n} \quad (1.17)$$

is the complete form of Stirling's approximation. Table 3 shows the comparison of numerical values up to  $n = 100$ .

## 1.4 What's the use of all this?

A clear view of basic probability theory is instrumental in developing more advanced ideas both in theoretical and experimental physics. The Stirling's approximation is an important theoretical tool, which plays an essential role in the evaluation of factorials and binomial coefficients.

$n$	$\ln n!$	$\ln(\sqrt{2\pi n} n^n e^{-n})$
1	0.	-0.0810615
2	0.693147	0.651806
3	1.79176	1.76408
4	3.17805	3.15726
5	4.78749	4.77085
6	6.57925	6.56538
7	8.52516	8.51326
8	10.6046	10.5942
9	12.8018	12.7926
10	15.1044	15.0961
20	42.3356	42.3315
30	74.6582	74.6555
40	110.321	110.319
50	148.478	148.476
60	188.628	188.627
70	230.439	230.438
80	273.673	273.672
90	318.153	318.152
100	363.739	363.739

Table 1.3: Comparison between exact values of  $\ln n!$  and the complete expression of Stirling's approximation.

## Appendix: short history of probability theory

Although basic concepts on chance and randomness date back to ancient times, the very first mathematical developments started only in Renaissance, with the work of Girolamo Cardano and Galileo Galilei. However, their work on probability had little subsequent impact and most accounts of the history of probability start with the exchange of letters among the Chevalier de Mér  , Pascal and Fermat, as in this very short history of probability theory that is excerpted from T. Apostol, [3]. Interested readers can find much more information in the readily available books on the early history of probability “Games, Gods and Gambling: A History of Probability and Statistical Ideas” by F. N. David [7] and “Probability and Statistics: The Science of Uncertainty” by J. Tabak [32].

“A gamblers dispute in 1654 led to the creation of a mathematical theory of probability by two famous French mathematicians, Blaise Pascal and Pierre de Fermat. Antoine Gombaud, Chevalier de M      , a French nobleman with an interest in gaming and gambling questions, called Pascal’s attention to an apparent contradiction concerning a popular dice game. The game consisted in throwing a pair of dice 24 times; the problem was to decide whether or not to bet even money on the occurrence of at least one “double six” during the 24 throws. A seemingly well-established gambling rule led de M   to believe that betting on a double six in 24 throws would be profitable, but his own calculations indicated just the opposite.

This problem and others posed by de M   led to an exchange of letters between Pascal and Fermat in which the fundamental principles of probability theory were formulated for the first time. Although a few special problems on games of chance had been solved by some Italian mathematicians in the 15th and 16th centuries, no general theory was developed before this famous correspondence. The Dutch scientist Christian Huygens, a teacher of Leibniz, learned of this correspondence and shortly thereafter (in 1657) published the first book on probability, entitled *De Ratiociniis in Ludo Aleae*, it was a treatise on problems associated with gambling. Because of the inherent appeal of games of chance, probability theory soon became popular, and the subject developed rapidly during the 18th century. The major contributors during this period were Jakob Bernoulli (1654-1705) and Abraham de Moivre (1667-1754). In 1812 Pierre de Laplace (1749-1827) introduced a host of new ideas and mathematical techniques in his book, *Th  orie Analytique des Probabilit  s*. Before Laplace, probability theory was solely concerned with developing a mathematical analysis of games of chance. Laplace applied probabilistic ideas to many scientific and practical problems. The theory of errors, actuarial mathematics, and statistical mechanics are examples of some of the important applications of probability theory developed in the 19th century. Like so many other branches of

mathematics, the development of probability theory has been stimulated by the variety of its applications. Conversely, each advance in the theory has enlarged the scope of its influence. Mathematical statistics is one important branch of applied probability; other applications occur in such widely different fields as genetics, psychology, economics, and engineering. Many workers have contributed to the theory since Laplace's time; among the most important are Chebyshev, Markov, von Mises, and Kolmogorov.

One of the difficulties in developing a mathematical theory of probability has been to arrive at a definition of probability that is precise enough for use in mathematics, yet comprehensive enough to be applicable to a wide range of phenomena. The search for a widely acceptable definition took nearly three centuries and was marked by much controversy. The matter was finally resolved in the 20th century by treating probability theory on an axiomatic basis. In 1933 a monograph by a Russian mathematician, A. Kolmogorov, outlined an axiomatic approach that forms the basis for the modern theory. (Kolmogorov's monograph is available in English translation as *Foundations of Probability Theory*, Chelsea, New York, 1950.) Since then the ideas have been refined somewhat and probability theory is now part of a more general discipline known as measure theory."

## Chapter 2

# The algebra of probability

### 2.1 Elementary events. The sample space

We have already met the definition of probability

$$P(A) = \frac{N(A)}{N} \quad (2.1)$$

where  $N$  is the total number of outcomes and  $N(A)$  is the number of *equiprobable* outcomes leading to the occurrence of event A. Now we represent the individual outcomes as set elements, and we also assume that there can be an infinite number of them, possibly an uncountable infinity. If we assume that there is an uncountable infinity of elementary outcomes and that they are all equiprobable, then we can represent composite outcomes as subsets of  $\mathbb{R}^n$  and replace equation (2.1) with the ratio of set measures

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} \quad (2.2)$$

where  $A$  represents the set of outcomes that lead to the occurrence of event A,  $\Omega$  represents the whole sample space, the totality of possible outcomes – it is also called *the universe* – and  $\mu$  is the set measure. The usual representation utilizes subsets of  $\mathbb{R}^2$  because they are easy to draw. The standard terminology is as follows (and the corresponding  $\mathbb{R}^2$  representations are shown in figure 2.1):

**certain or sure event:** an event  $A$  that always occurs so that  $A = \Omega$ ;

**impossible event:** an event  $A$  that coincides with the empty set  $A = \emptyset$ ;

**identical or equivalent events:** events  $A_1$  and  $A_2$  are identical whenever the occurrence of one always implies the occurrence of the other,  $A_1 = A_2$ , the event subsets coincide;

**mutually exclusive or incompatible events:** the occurrence of one precludes the occurrence of the other, the event subsets are disjoint;

**complementary event:** this event  $\bar{A}$  occurs whenever  $A$  does not, the event subset is complementary with respect to  $A$ , and using the standard set notation this writes  $\bar{A} = \Omega - A$ ;

**union of events:** the union of two events  $A_1$  and  $A_2$  is the set of elementary outcomes such that either  $A_1$  or  $A_2$  occur, the corresponding set is  $A_1 \cup A_2$ ;

**intersection of events:** the intersection of two events  $A_1$  and  $A_2$  is the set of elementary outcomes such that both  $A_1$  and  $A_2$  occur, the corresponding set is  $A_1 \cap A_2$ . Note that the intersection is often written as a product:  $A_1 A_2 = A_1 \cap A_2$ .

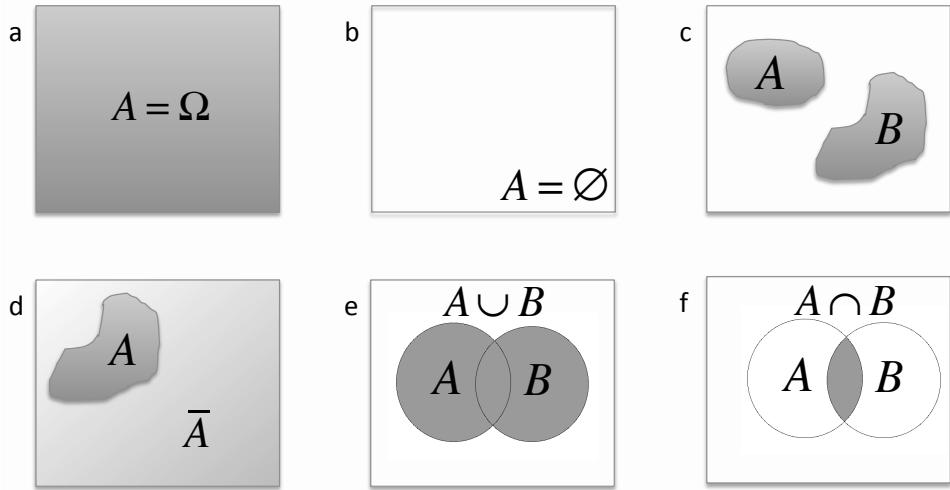


Figure 2.1: A selection of event representations in  $\mathbb{R}^2$ . a) the event  $A$  is certain, it coincides with the whole sample space  $\Omega$ ; b) an empty event has zero measure and vanishing probability (equivalently a non-empty zero-measure event also has vanishing probability); c) mutually exclusive events are represented by disjoint sets; d) the complementary event fills all the sample space outside  $A$ ; e) the union of two events; f) the intersection of events.

### 2.1.1 The Addition Law for Probabilities

If two events  $A_1$  and  $A_2$  are mutually exclusive, then the measure of the set corresponding to the union  $A_1 \cup A_2$  is simply the sum of the individual

measures, i.e.,

$$\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2) \quad (2.3)$$

therefore

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) \quad (2.4)$$

This is easily generalized to a set  $\{A_k\}_{k=1,n}$  of mutually exclusive events, i.e.,

$$P\left(\bigcup_{k=1}^{k=n} A_k\right) = \sum_{k=1}^{k=n} P(A_k) \quad (2.5)$$

(addition law for probabilities).

Note that the addition law for mutually exclusive events, can easily be extended to infinite sets of mutually exclusive events:

$$P\left(\bigcup_{k=1}^{k=\infty} A_k\right) = \sum_{k=1}^{k=\infty} P(A_k) \quad (2.6)$$

A few simple theorems:

1. probabilities lie in the  $[0, 1]$  interval:  $0 \leq N(A) \leq N \Rightarrow 0 \leq P(A) \leq 1$ ;
2.  $A_1 = (A_1 - A_2) \cup (A_1 \cap A_2) \Rightarrow$   
 $P(A_1) = P(A_1 - A_2) + P(A_1 \cap A_2) \Rightarrow$   
 $P(A_1 - A_2) = P(A_1) - P(A_1 \cap A_2)$   
 (remark:  $(A_1 - A_2)$  and  $(A_1 \cap A_2)$  are mutually exclusive);
3. exchanging indexes in the previous result we find:  $P(A_2 - A_1) = P(A_2) - P(A_1 \cap A_2)$ ;
4. addition law in the case of two non-mutually exclusive events:  

$$\begin{aligned} A_1 \cup A_2 &= (A_1 - A_2) \cup (A_2 - A_1) \cup (A_1 \cap A_2) \Rightarrow \\ P(A_1 \cup A_2) &= P(A_1 - A_2) + P(A_2 - A_1) + P(A_1 \cap A_2) \\ &= P(A_1) - P(A_1 \cap A_2) + P(A_2) - P(A_1 \cap A_2) + P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1 A_2) \end{aligned}$$

Exercise: give a graphical representation of these theorems.

Now apply the last theorem to three generic events  $A_1$ ,  $A_2$ , and  $A_3$  (which could be non-mutually exclusive), to find the addition law for three

non-mutually exclusive events:

$$\begin{aligned}
P(A_1 \cup A_2 \cup A_3) &= P((A_1 \cup A_2) \cup A_3) \\
&= P(A_1 \cup A_2) + P(A_3) - P((A_1 \cup A_2)A_3) \\
&= P(A_1) + P(A_2) - P(A_1A_2) + P(A_3) - P(A_1A_3 \cup A_2A_3) \\
&= P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) \\
&\quad + P(A_1A_2A_3) \\
&= \sum_{k=1,3} P(A_k) - \sum_{\substack{k=1,3 \\ l=k+1,3}} P(A_kA_l) + \sum_{\substack{k=1,3 \\ l=k+1,3 \\ n=l+1,3}} P(A_kA_lA_n) \\
&= \sum_{k=1,3} P(A_k) - \sum_{k < l} P(A_kA_l) + \sum_{k < l < n} P(A_kA_lA_n)
\end{aligned}$$

where the notation  $\sum_{k < l}$  is a shorthand for  $\sum_{k=1,3; l=k+1,3}$ , etc.

The last formula is an attempt to generalize the pattern in the addition formulas, and now we surmise that the pattern continues, so that

$$P\left(\bigcup_{n=1,N} A_n\right) = \sum_{n=1,N} (-1)^{n+1} P_n \quad (2.7)$$

where

$$P_n = \sum_{i_1 < i_2 < \dots < i_n} P(A_{i_1}A_{i_2}\dots A_{i_n}) \quad (2.8)$$

i.e.,

$$P\left(\bigcup_{n=1,N} A_n\right) = \sum_{n=1,N} (-1)^{n+1} \sum_{i_1 < i_2 < \dots < i_n} P(A_{i_1}A_{i_2}\dots A_{i_n}) \quad (2.9)$$

Indeed if we assume that the theorem holds for  $N$  events  $\{A_n\}_{n=1,N}$ ,

then if we add one more event  $A_{N+1}$  we find

$$P\left(\bigcup_{n=1,N+1} A_n\right) = P\left(\bigcup_{n=1,N} A_n \cup A_{N+1}\right) \quad (2.10a)$$

$$= P\left(\bigcup_{n=1,N} A_n\right) + P(A_{N+1}) - P\left(\bigcup_{n=1,N} A_n A_{N+1}\right) \quad (2.10b)$$

$$= \sum_{n=1,N} (-1)^{n+1} \sum_{i_1 < i_2 < \dots < i_n} P(A_{i_1} A_{i_2} \dots A_{i_n}) + P(A_{N+1}) \quad (2.10c)$$

$$- \sum_{n=1,N} (-1)^{n+1} \sum_{i_1 < i_2 < \dots < i_n} P(A_{i_1} A_{i_2} \dots A_{i_n} A_{N+1}) \quad (2.10d)$$

$$= \sum_{n=1,N+1} (-1)^{n+1} \sum_{i_1 < i_2 < \dots < i_n} P(A_{i_1} A_{i_2} \dots A_{i_n}) \quad (2.10e)$$

and thus the theorem is proved by induction.

*(addition law for probabilities of non-mutually exclusive events)*

Exercise: look for alternative proofs of this theorem in the literature (hint: search for *exclusion-inclusion principle*).

This kind of reasoning can be applied in several examples, like the “raincoat problem” [27]: given  $n$  students who wear nearly identical raincoats, and leave them on the same rack, what is the probability that at least one raincoat ends up with the original owner?

The problem can be restated as follows: what is the probability that in a given permutation of the set  $\{1, 2, \dots, n\}$  at least one element remains fixed? We denote with  $A_k$  the event that corresponds to keeping the  $k$ -th value fixed, therefore the event that at least one value remains fixed is just

$$A = \bigcup_{k=1}^{k=n} A_k \quad (2.11)$$

The events are non-mutually exclusive, therefore the more complex addition law for probabilities applies:

$$P(A) = P\left(\bigcup_{k=1}^{k=n} A_k\right) = \sum_{k=1,n} (-1)^{k+1} \sum_{i_1 < i_2 < \dots < i_k} P(A_{i_1} A_{i_2} \dots A_{i_k}) \quad (2.12)$$

Moreover  $P(A_{i_1} A_{i_2} \dots A_{i_k})$  is just the probability of fixing  $k$  given positions while leaving the others free, so that

$$P(A_{i_1} A_{i_2} \dots A_{i_k}) = \frac{(n-k)!}{n!} \quad (2.13)$$

(there are  $n!$  permutations, and  $(n-m)!$  permutations that keep  $m$  positions fixed) and there are

$$C_m^n = \binom{n}{k} \quad (2.14)$$

ways of choosing the  $k$  fixed sites, therefore

$$\sum_{i_1 \neq i_2 \neq \dots \neq i_k} P(A_{i_1} A_{i_2} \dots A_{i_k}) = \binom{n}{k} \frac{(n-k)!}{n!} = \frac{1}{k!} \quad (2.15)$$

and

$$\begin{aligned} P(A) &= \sum_{k=1,n} (-1)^{k+1} \frac{1}{k!} = 1 - \left( 1 + \sum_{k=1,n} (-1)^k \frac{1}{k!} \right) = 1 - \sum_{k=0,n} \frac{(-1)^k}{k!} \\ &\approx 1 - e^{-1} \approx 0.6321 \end{aligned} \quad (2.16)$$

## 2.2 Conditional probability

Consider two events, A and B, let  $N(A)$  and  $N(B)$  be the corresponding numbers of elementary (equiprobable) events, a  $N$  be the total number of events (or, equivalently, let  $\mu(A)$  and  $\mu(B)$  be the measures of the events A and B). Then, as before,

$$P(A) = \frac{N(A)}{N} \quad (2.17a)$$

$$P(B) = \frac{N(B)}{N} \quad (2.17b)$$

or

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} \quad (2.18a)$$

$$P(B) = \frac{\mu(B)}{\mu(\Omega)} \quad (2.18b)$$

in the case of uncountable elementary events (in the following paragraphs we settle on the simpler  $N$  notation). We can also select only those elementary events that correspond to the occurrence of both A and B:

$$P(AB) = \frac{N(AB)}{N} \quad (2.19)$$

Now note that  $AB$  is a subset of  $B$ , and that we can limit the sample space to those events that correspond to the occurrence of  $B$ , so that we can define a *conditional probability* of the occurrence of  $A$  in this reduced sample space:

$$P(A|B) = \frac{N(AB)}{N(B)} = \frac{N(AB)}{N} \frac{N}{N(B)} = \frac{P(AB)}{P(B)} \quad (2.20)$$

Clearly we can reverse the roles of  $A$  and  $B$  and we find

$$P(B|A) = \frac{N(AB)}{N(A)} = \frac{N(AB)}{N} \frac{N}{N(A)} = \frac{P(AB)}{P(A)} \quad (2.21)$$

and we can condense both formulas as follows

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \quad (2.22)$$

It is easy to prove a few elementary theorems

1.  $0 \leq N(AB) \leq N(B) \Rightarrow 0 \leq P(A|B) \leq 1$ ;
2. if  $A$  and  $B$  are mutually exclusive, then  $N(AB) = 0$  and  $P(A|B) = P(B|A) = 0$ ;
3. if  $B$  implies  $A$  (i.e.,  $A$  always accompanies the occurrence of  $B$ , although the converse is not necessarily true) so that  $B \subset A$ , then  $N(AB) = N(B)$  and  $P(A|B) = 1$ ;
4. now let  $\{A_k\}_{k=1,n}$  be a set of mutually exclusive events, and let  $A$  be the union of all the  $A_k$ 's. The events  $\{A_kB\}_{k=1,n}$  are incompatible as well, and we can apply the addition law for probabilities of mutually exclusive events

$$\begin{aligned} P(AB) &= P(A|B)P(B) = P\left(\bigcup_{k=1}^{k=n} A_k B\right) = \sum_{k=1}^{k=n} P(A_k B) \\ &= \sum_{k=1}^{k=n} P(A_k|B)P(B) \end{aligned} \quad (2.23)$$

and then

$$P(A|B) = \sum_{k=1}^{k=n} P(A_k|B) \quad (2.24)$$

(*addition law for conditional probabilities*).

### 2.2.1 Statistical independence

Events are statistically independent when they do not influence one another's occurrence: if this is true, then the probability of A occurring when B occurs must be equal to the probability of A occurring at all, i.e.

$$P(A|B) = P(A) \quad (2.25)$$

Then

$$P(AB) = P(A|B)P(B) = P(A)P(B) \quad (2.26)$$

Similarly the events of a set  $\{A_k\}_{k=1,n}$  are mutually independent if all the probabilities of joint events factorize.

Figure 2.2 illustrates how statistical independence can be visualized in a sample space with a simple metric.

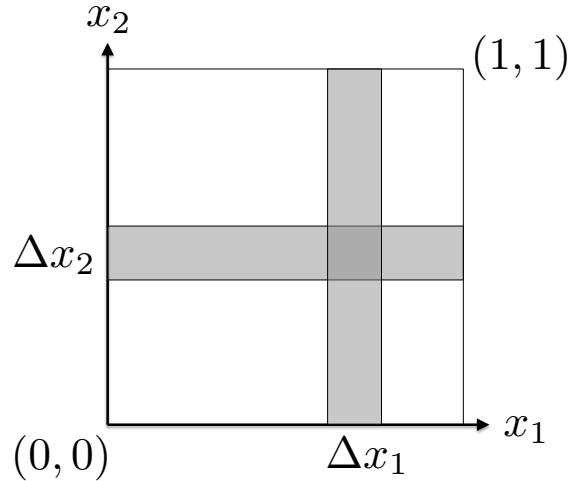


Figure 2.2: The square region represents a sample space. Unlike the previous examples, here we assume that probability corresponds to the simple measure which is induced by the usual metric properties of sets in the Euclidean plane, i.e., the measure of a rectangle with sides parallel to the axes is given by  $\Delta x_1 \cdot \Delta x_2$ , where  $\Delta x_1$  and  $\Delta x_2$  are the lengths of the horizontal and vertical sides and  $x_1, x_2$  are real variables that act as “labels” that identify the elementary events in this sample space. This means that the measures of the two shaded events (sets) are  $\Delta x_1$  and  $\Delta x_2$ , and obviously, their intersection has measure  $\Delta x_1 \cdot \Delta x_2$ . Thus these events are independent.

### 2.2.2 Bayes' theorem

From the formula for conditional probabilities we see that

$$P(A|B)P(B) = P(B|A)P(A) \Rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2.27)$$

(*Bayes' theorem*). Now take mutually exclusive events  $\{A_k\}_{k=1,n}$ , such that their union corresponds to the whole sample space  $\Omega$ , and apply the theorem to one of these events

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B)} \quad (2.28)$$

However we can also write

$$P(B) = P(B \cap \Omega) = P\left(B \cap \bigcup_n A_n\right) = P\left(\bigcup_n A_n B\right) \quad (2.29)$$

Since the events  $A_n B$  are mutually exclusive, we can apply the law for the addition of probabilities of incompatible events and find

$$P(B) = \sum_n P(A_n B) = \sum_n P(B|A_n)P(A_n) \quad (2.30)$$

therefore we find the following modified form for Bayes' theorem

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_n P(B|A_n)P(A_n)} = \frac{P(B|A_k)}{\sum_n P(B|A_n)P(A_n)} P(A_k) \quad (2.31)$$

The whole of Bayesian inference theory is based on this apparently simple theorem.

### 2.2.3 Gambler's ruin

Conditional probabilities are commonplace in the theory of random processes. One important random process is the random walk, of which the gambler's ruin is a relevant subtopic. In this case, at each step  $n$ , the gambler has the conditional probability  $p(x, n) = P(+|x, n)$  or  $q(x, n) = P(-|x, n)$  of winning or losing 1 unit of his/her capital  $x$ . The gambler plays until he either loses all capital ( $x = 0$ ) or wins and reaches the final amount  $x = m$ . Let  $P(x, n)$  be the probability of winning the game with capital  $x$  at step  $n$ , then we can write

$$P(x, n+1) = p(x-1, n)P(x-1, n) + q(x+1, n)P(x+1, n) \quad (2.32)$$

Notice that at each step the normalization condition is  $p(x, n) + q(x, n) = 1$ . Now we assume that conditional probabilities do not depend either on capital

or on step, i.e.,  $p = p(x, n)$  and  $q = q(x, n)$ , and since the ruin probability does not depend on step, but only on capital, we can write  $P(x, n) = P(x)$  and

$$P(x) = pP(x - 1) + qP(x + 1) \quad (2.33)$$

If we take the “fair” game such that  $p = q = 1/2$ , we find

$$P(x) = \frac{1}{2}P(x + 1) + \frac{1}{2}P(x - 1) \quad (2.34)$$

This is complemented by the boundary conditions  $P(0) = 0$ ,  $P(m) = 1$  (also notice that  $P$  is a sort of “discrete” harmonic function, because the function is equal to the average value of neighboring positions). When we try out a linear solution

$$P(x) = C_1 + C_2x \quad (2.35)$$

we find that it satisfies the difference equation, and moreover

$$P(0) = C_1 = 0 \quad (2.36)$$

$$P(m) = C_1 + C_2m = 1 \quad (2.37)$$

so that  $C_1 = 0$ ,  $C_2 = 1/m$ ,  $P(x) = x/m$ . Figure 2.3 shows some realizations of the process in individual games.

This problem can be made more general, taking asymmetric conditional probabilities for winning and losing. In this case the exact solution is considerably more complex (see, e.g., reference [15]), however we can still find an approximate solution using analytic methods. Indeed, the difference equation above is a discretized form of a diffusion equation in a stationary system. And if we assume that each win or each loss corresponds to a capital  $\Delta x$  and we assume that  $x \gg \Delta x$ , then the equation for equal winning and losing probabilities becomes

$$P(x) = \frac{1}{2}P(x + \Delta x) + \frac{1}{2}P(x - \Delta x) \approx P(x) + \frac{1}{2} \frac{\partial^2 P}{\partial x^2} \Delta x^2 \quad (2.38)$$

or also

$$\frac{\partial^2 P}{\partial x^2} = 0 \quad (2.39)$$

In this case the diffusion equation is also the same as the Laplace equation for (1D) electrostatic systems, and the solution of the Laplace problem (a potential) is also the solution of the diffusion problem (a probability density). This can be generalized, and it means that the solution of a deterministic problem (electrostatic potential) can be used to solve a problem of probability theory (for a stationary random process) (see, ref. [19]).

The 1D solution of the stationary diffusion equation is quite simple: integrating twice we find the linear solution

$$P(x) = C_1 + C_2x \quad (2.40)$$

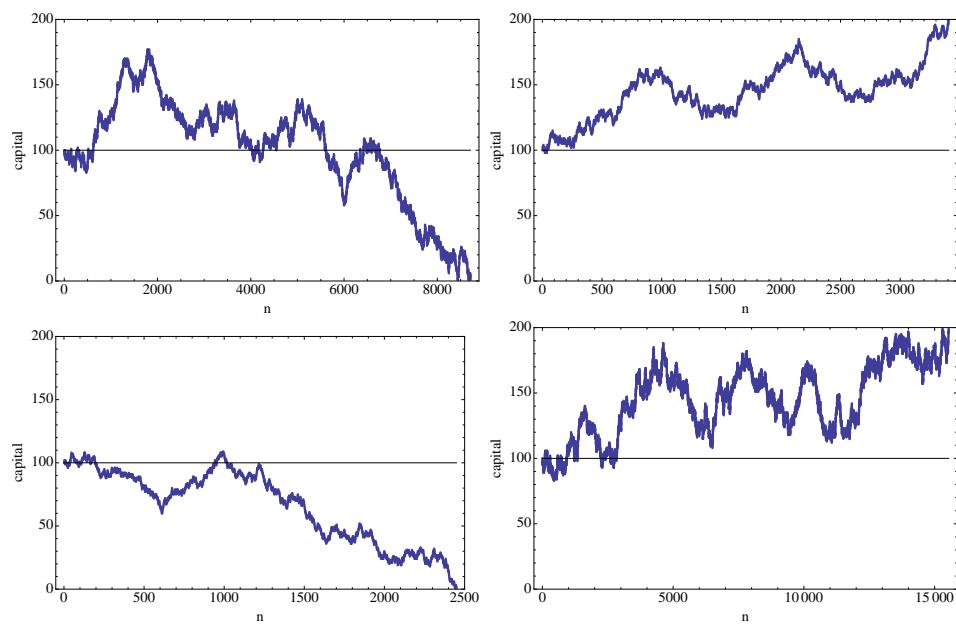


Figure 2.3: Plot of gambler's capital vs.  $n$  in four individual games. The gambler starts with an initial capital of 100 units, and the plot extends up to the gambler's ruin (if the capital vanishes) or success (if the capital reaches 200 units). This process is essentially a *random walk with absorbing barriers*.

just as before. This sort of continuous approximation can be used to solve the case with  $p \neq q$ , which is far more difficult in the discrete case. Indeed,

$$P(x) = pP(x-\Delta x) + qP(x+\Delta x) \approx P(x) - (p-q)\frac{\partial P}{\partial x}\Delta x + \frac{1}{2}\frac{\partial^2 P}{\partial x^2}\Delta x^2 \quad (2.41)$$

and therefore

$$-\frac{(p-q)}{\Delta x}\frac{\partial P}{\partial x} + \frac{1}{2}\frac{\partial^2 P}{\partial x^2} = 0 \quad (2.42)$$

We let  $\Phi = \frac{\partial P}{\partial x}$  and  $\alpha = -(p-q)/\Delta x$ , and the equation becomes

$$\alpha\Phi + \frac{1}{2}\frac{\partial\Phi}{\partial x} = 0 \quad (2.43)$$

which is easy to integrate

$$\frac{\partial P}{\partial x} = Ae^{-2\alpha x} \quad (2.44)$$

We have to perform one final integration of the equation

$$\frac{\partial P}{\partial x} = Ae^{-2\alpha x} \quad (2.45)$$

which yields

$$P(x) = -\frac{1}{2\alpha}Ae^{-2\alpha x} + B \quad (2.46)$$

Using the boundary conditions  $P(0) = 0$  and  $P(m) = 1$  we fix  $A$  and  $B$ :  $B = A/2\alpha$  and  $A = 2\alpha/(1 - e^{-2\alpha m})$ , so that

$$P(x) = \frac{1 - e^{-2\alpha x}}{1 - e^{-2\alpha m}} \quad (2.47)$$

Figure 2.4 shows the resulting  $P(x)$  for different values of the unbalance parameter  $-\alpha = (p-q)/\Delta x$ .

#### 2.2.4 Bertrand's paradox

When we consider a discrete set of elementary events there can be no ambiguity in the definition of *elementary event*. However when we consider events that belong to a set in  $\mathbb{R}^n$ , it is sometimes unclear how to choose the elementary events. This is the origin of some of the so called *probability paradoxes*: here we discuss the famous Bertrand's paradox, which was first introduced by Joseph Bertrand in 1889, in his work *Calcul des probabilités*.

The problem goes as follows: consider an equilateral triangle inscribed inside a circle, and suppose that a chord is chosen at random. What is the probability that the chord is longer than a side of the triangle?

A possible solution goes as follows (see figure 2.5): we take two random points on the circle (figure 2.5, left panel), then we rotate the circle so

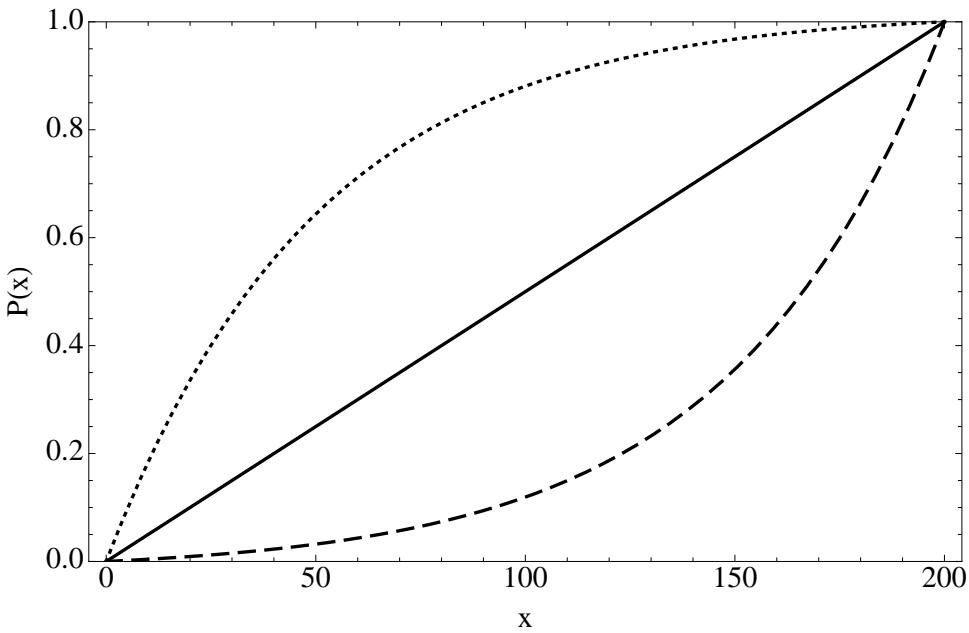


Figure 2.4: Probability of winning the game  $P(x)$  vs. the capital  $x$ , in a game with  $m = 200$ , for different values of the unbalance parameter  $\alpha = -(p - q)/\Delta x$  (solid line,  $\alpha = 0$ ; dashed line,  $\alpha = -0.01$ ; dotted line,  $\alpha = 0.01$ ). These curves show that even a small unbalance can lead to large asymmetries in the probability of winning or losing.

that one of the two points coincides with one of the vertices of the inscribed triangle (figure 2.5, center panel). Thus we see that drawing a random chord is equivalent to taking the first point that defines the chord as one vertex of the triangle while the other is taken “at random” on the circle. Here “at random” means that it is uniformly distributed on the circumference (figure 2.5, right panel). Then we see at once that only those chords that cross the opposite side of the triangle are actually longer than each side. Since the subtended arc is  $1/3$  of the circumference, the probability of drawing a random chord that is longer than one side of the triangle is  $1/3$ .

However, this is not the only way of choosing a random chord. Figure 2.6 shows a different method, where we take first a random radius, and next we choose a random point on this random radius. Then, we take the chord through this point and perpendicular to the radius. When we rotate the triangle so that the radius is perpendicular to one of the sides, we see that half of the points give chords longer than one side of the triangle, therefore the probability is  $1/2$ .

This is surprising: two different, perfectly acceptable “solutions” give different answers. Even more surprising is the fact that there is at least

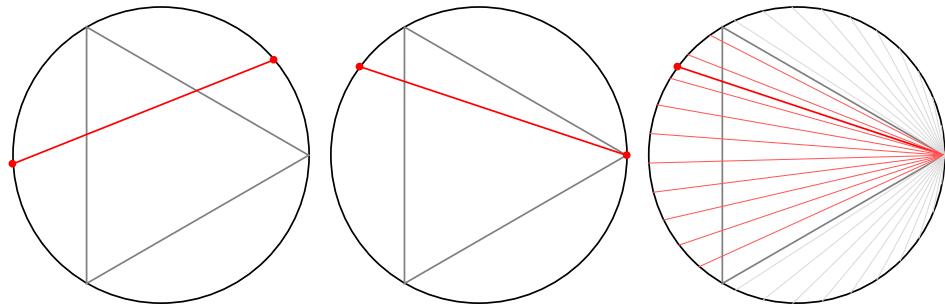


Figure 2.5: Illustration of the first “solution” of Bertrand’s problem (see the main text for the detailed explanation).

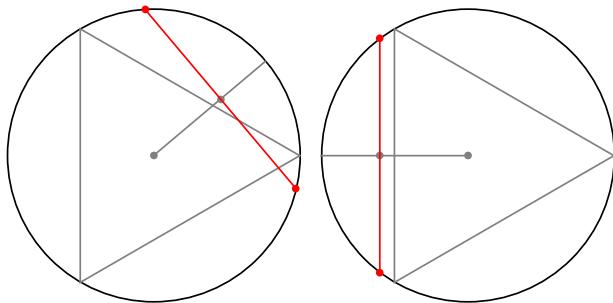


Figure 2.6: Illustration of the second “solution” of Bertrand’s problem (see the main text for the detailed explanation).

another solution, which is again quite acceptable. Consider figure 2.7: we see at once that taking chord midpoints located inside the circle inscribed in the triangle, we obtain chords that are longer than one side of the triangle. Since the ratio of the areas of the two circles is  $1/4$ , we find that now the probability of drawing a long chord is just  $1/4$ .

So, how can we reconcile these different solutions? The answer is just that we are taking different elementary events. We can get further insight into the problem by plotting the random midpoints and the random chords in each case, as shown in figures 2.8-2.10. The figures – especially those that display the distribution of midpoints – show that the distribution of the elementary events is different in each case.

Given these different sample spaces, how should we choose one of them? The problem might seem irrelevant to physics, however it is not, because in many probability models we are faced with the same issue: how should we choose the elementary events? A possible, and quite interesting solution was given by Edwin Jaynes who discussed Bertrand’s paradox in 1973, in the wider context of invariant distribution laws [21]. He noted that the generic

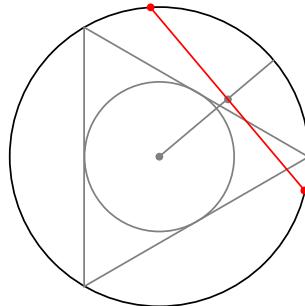


Figure 2.7: Illustration of the third “solution” of Bertrand’s problem (see the main text for the detailed explanation).

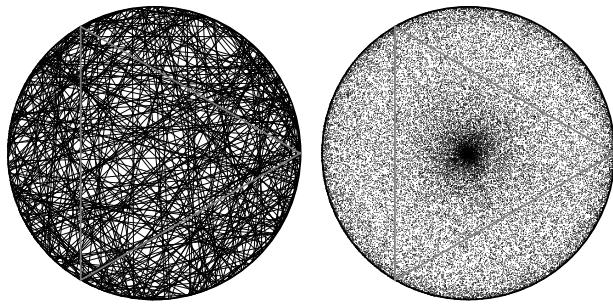


Figure 2.8: Distribution of chords (left panel) and of midpoints (right panel) in the first solution of Bertrand’s paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

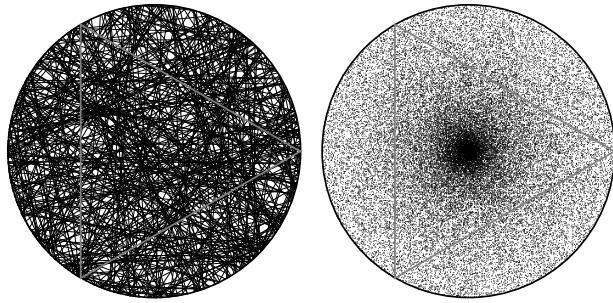


Figure 2.9: Distribution of chords (left panel) and of midpoints (right panel) in the second solution of Bertrand’s paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints).

formulation of the problem implies both rotational and translational invariance, and that these invariances are satisfied only by the second solution.

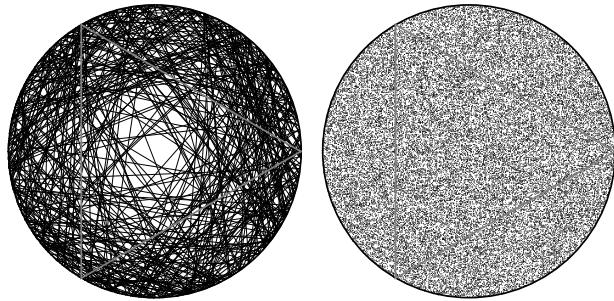


Figure 2.10: Distribution of chords (left panel) and of midpoints (right panel) in the third solution of Bertrand's paradox (the left panel shows 400 chords, the right panel shows 100000 midpoints). Notice that while the distribution of midpoints is uniform, the distribution of the resulting chords is distinctly non-uniform.

We shall discuss Jaynes' solution further on in these notes.

### 2.3 The Borel-Cantelli lemmas

In this section we prove two lemmas which are important in the analysis of rare, exceptional events. The proof of the lemmas proceeds by steps, and we prove first some auxiliary theorems.

- **Theorem 2.3:** Consider a chain (an “increasing sequence”) of events  $A_1 \subset A_2 \subset A_3 \subset \dots$ , and define the new sequence of events

$$\begin{aligned}
 B_1 &= A_1 \\
 B_2 &= A_2 - B_1 \\
 B_3 &= A_3 - (B_1 \cup B_2) \\
 &\dots \\
 B_n &= A_n - \bigcup_{k=1}^{n-1} B_k \\
 &\dots
 \end{aligned}$$

so that each event  $B_k$  is the set of elementary events that are in  $A_k$  but are not contained in anyone of the events  $A_1, \dots, A_{k-1}$ . This also means that all  $B$ 's are mutually exclusive. Moreover we see at once

that

$$\begin{aligned}
 A_1 &= B_1 \\
 A_2 &= B_1 \cup B_2 \\
 A_3 &= B_1 \cup B_2 \cup B_3 \\
 &\dots \\
 A_n &= \bigcup_{k=1}^{k=n} B_k \\
 &\dots
 \end{aligned}$$

therefore

$$\bigcup_{k=1}^{k=n} A_k = \bigcup_{k=1}^{k=n} B_k = A_n \quad (2.48)$$

so that the  $B$ 's are mutually exclusive and their union is the same as the union of the  $A$ 's. This means that

$$P\left(\bigcup_{k=1}^{k=\infty} A_k\right) = P\left(\bigcup_{k=1}^{k=\infty} B_k\right) = \sum_{k=1}^{k=\infty} P(B_k) = \lim_{n \rightarrow \infty} P(A_n) \quad (2.49)$$

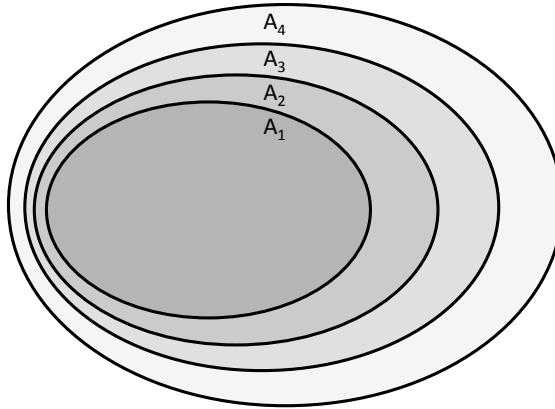


Figure 2.11: Increasing sequence of events as in Theorem 2.3. The  $B_k$  events correspond to the shaded bands.

- **Theorem 2.3':** This theorem is complementary to the previous one, it deals with a “decreasing sequence” of events  $A_1 \supset A_2 \supset A_3 \supset \dots$ . In this case, note that  $\bar{A}_1 \subset \bar{A}_2 \subset \bar{A}_3 \subset \dots$ , so that you can apply the previous theorem and find

$$P\left(\bigcup_{k=1}^{k=\infty} \bar{A}_k\right) = \lim_{n \rightarrow \infty} P(\bar{A}_n) \quad (2.50)$$

Since, for any event  $A$ ,  $P(A) = 1 - P(\bar{A})$ , and

$$\overline{\left( \bigcup_{k=1}^{k=\infty} \bar{A}_k \right)} = \bigcap_{k=1}^{k=\infty} A_k \quad (2.51)$$

we find

$$1 - P\left(\bigcap_{k=1}^{k=\infty} A_k\right) = 1 - \lim_{n \rightarrow \infty} P(A_n) \quad (2.52)$$

i.e.,

$$P\left(\bigcap_{k=1}^{k=\infty} A_k\right) = \lim_{n \rightarrow \infty} P(A_n) \quad (2.53)$$

- **Theorem 2.4:** Now let's go back to the proof of theorem 2.3, and note that  $B_n \subseteq A_n$ , so that  $P(B_n) \leq P(A_n)$ , and

$$P\left(\bigcup_k A_k\right) = \sum_k P(A_k) \leq \sum_k P(A_k) \quad (2.54)$$

### 2.3.1 The first Borel-Cantelli lemma

We consider an infinite set of events  $\{C_n\}_{n=1,\infty}$  (which in general are not mutually exclusive), and we assume that the corresponding probabilities  $p_n = P(C_n)$  are such that

$$\sum_{k=1}^{k=\infty} p_k < \infty \quad (2.55)$$

i.e., the series converges, so that the convergence condition

$$\lim_{n \rightarrow \infty} \sum_{k=n}^{k=\infty} p_k = 0 \quad (2.56)$$

must hold, then the probability of any infinite subset of the events  $C_n$  vanishes. This means that any set of elementary events that belong to an infinite number of  $C_k$ 's has vanishing probability.

Indeed if we define

$$A_n = \bigcup_{k=n}^{k=\infty} C_k \quad (2.57)$$

(i.e., an elementary event in  $A_n$  belongs to at least one of the  $C_k$ 's ( $k \geq n$ )) we obtain a decreasing sequence of events  $A_1 \supset A_2 \supset \dots$ . Moreover, the event

$$A = \bigcap_{k=1}^{k=\infty} A_k \quad (2.58)$$

contains all those events that belong to an infinite number of  $C_k$ 's. From theorem 2.3' and from the convergence condition we find

$$P(A) = P\left(\bigcap_{k=1}^{k=\infty} A_k\right) = \lim_{n \rightarrow \infty} P(A_n) \leq \lim_{n \rightarrow \infty} \sum_{k=n}^{k=\infty} P(C_k) = 0 \quad (2.59)$$

i.e., the probability of occurrence of an infinite number of different  $C_k$ 's vanishes. In other words – with probability 1 – only a finite number of different  $C_k$ 's occurs.

### 2.3.2 The second Borel-Cantelli lemma

Just as in the proof of the first Borel-Cantelli lemma, we consider an infinite set of events  $\{C_n\}_{n=1,\infty}$ , however now we assume that the corresponding probabilities  $p_n = P(C_n)$  are such that

$$\sum_{n=1}^{n=\infty} p_n = \infty \quad (2.60)$$

(the series diverges), and moreover that the  $C_k$ 's are mutually independent. As before, we define a set of auxiliary events

$$A_n = \bigcup_{k=n}^{k=\infty} C_k \quad (2.61)$$

(i.e., the elementary events in  $A_n$  belong to at least one of the  $C_k$ 's ( $k \geq n$ )), and in this way we obtain once again a decreasing sequence of events  $A_1 \supset A_2 \supset \dots$ . Moreover, the elementary events in

$$A = \bigcap_{k=1}^{k=\infty} A_k \quad (2.62)$$

belong to an infinite number of  $C_k$ 's. Taking complements we find:

- the event  $\bar{A}_n$ :

$$\bar{A}_n = \bigcap_{k=n}^{k=\infty} \bar{C}_k \quad (2.63)$$

which is the set of elementary events that do not belong to any of the  $C_k$ 's ( $k \geq n$ );

- the event  $\bar{A}$ , which is the set of elementary events that belong at most to a finite number of  $C_k$ 's:

$$\bar{A} = \bigcup_{k=1}^{k=\infty} \bar{A}_k \quad (2.64)$$

Since the  $C_k$ 's are independent, so are the  $\bar{C}_k$ 's, therefore we find<sup>1</sup>

$$P(\bar{A}_n) = P\left(\bigcap_{k=n}^{k=\infty} \bar{C}_k\right) = \prod_{k=n}^{k=\infty} (1 - p_k) \leq \prod_{k=n}^{k=\infty} \exp(-p_k) = \exp\left(-\sum_{k=n}^{k=\infty} p_k\right) \quad (2.65)$$

Since  $\sum_{k=n}^{k=\infty} p_n = \infty$  (the series diverges), this means

$$0 \leq P(\bar{A}_n) \leq \exp\left(-\sum_{k=n}^{k=\infty} p_k\right) = 0 \quad (2.66)$$

i.e.,  $P(\bar{A}_n) = 0$ , therefore

$$0 \leq P(\bar{A}) \leq \sum_{k=1}^{k=\infty} P(\bar{A}_k) = 0 \quad (2.67)$$

so that  $P(\bar{A}) = 0$  as well, and

$$P(A) = 1 - P(\bar{A}) = 1 \quad (2.68)$$

i.e., an infinite number of  $C_k$ 's occurs with probability 1.

(In this proof we skip some of the technicalities about the convergence of products and sums – see [27] for further details).

## 2.4 What's the use of all this?

Addition formulas for probabilities occur quite often, and are especially useful when dealing with Bayesian statistics. The Borel-Cantelli lemmas can be used to prove the *strong law of large numbers*, they are used in the theory of Markov chains, and their usefulness extends to several topics in theoretical physics, especially those related to Markov chains, like path integrals, dynamical systems, random walks, etc. The Bayes' theorem is a powerful inferential tool, and Bayesian inference is a large, developing field, but unfortunately we do not have the time to explore it in this course. However it is important to note that

- if there are competing, incompatible hypotheses  $\{H_k\}_{k=1,n}$ , and we collect a dataset D, then we can write

$$P(H_k|D) = \frac{P(D|H_k)}{\sum_n P(D|H_n)P(H_n)} P(H_k) \quad (2.69)$$

This formula is interpreted as follows:

---

<sup>1</sup>Here we use the inequality  $e^{-x} \geq 1 - x$ . It is easy to prove that the inequality holds for all  $x$ : indeed  $y = 1 - x$  is the equation of the tangent to  $e^{-x}$  at  $x = 0$ , and  $\frac{d^2 e^{-x}}{dx^2} = e^{-x} > 0$  everywhere, therefore the curvature is upward and the inequality always holds – equality at  $x = 0$ .

- $P(H_k)$  is the *a priori* probability of hypothesis  $H_k$  (i.e., before data collection);
  - $P(H_k|D)$  is the *a posteriori* probability of hypothesis  $H_k$  (i.e., after we have acquired additional information with the collection of data D);
  - $\frac{P(D|H_k)}{\sum_n P(D|H_n)P(H_n)}$  is the so-called *Bayes factor* which depends both on data and on our previous belief, and modifies the *a priori* probability to yield an improved estimate of the probability that hypothesis  $H_k$  is true.
- the same procedure can be used to *learn* from past examples; thus Bayes' theorem is the basis on which we build the theory of statistical learning.



# Chapter 3

## Random variables

### 3.1 Discrete and Continuous Random Variables. Distribution Functions.

This chapter establishes the connection of the generic view of probabilities with actual sample spaces that are number sets, and that can be both discrete or continuous. Moreover elementary events in these sample spaces can have nonuniform distributions and this is expressed by the important concept of probability distributions and probability densities.

This is illustrated by a simple example with 2-dimensional sets of elementary events shown in figure 3.1. In this example, elementary events either are uniformly distributed along both coordinate axes or are non-uniformly distributed according to a simple transformation. This corresponds to different densities of elementary events. The concept of equiprobable elementary events that are associated to the points of a metric space is very general, and it leads to the concept of probability density.

Now we generalize these ideas, starting with one-dimensional sample spaces:

- a)  $\omega$  denotes an elementary event in the sample space  $\Omega$ ;
- b) we associate the elementary event in the abstract sample space to a point on the real line, and we call this association a *random variable* (or *variate*). This is actually a labeling operation: we associate a given label to a random event, and this in turn allows us to define that label as a random label. Usually these labels are real numbers and carry with them all the (metric) properties of the set of random numbers. Thus we view a random variable as a numerical function  $\xi : \Omega \rightarrow \mathbb{R}$ , so that  $x = \xi(\omega)$ . In practice, in this way set a distinction between numerical values  $x$  and the random variable  $\xi$ , e.g., writing  $x' \leq \xi \leq x''$  we mean that the random variable  $\xi$  takes (random) numerical values

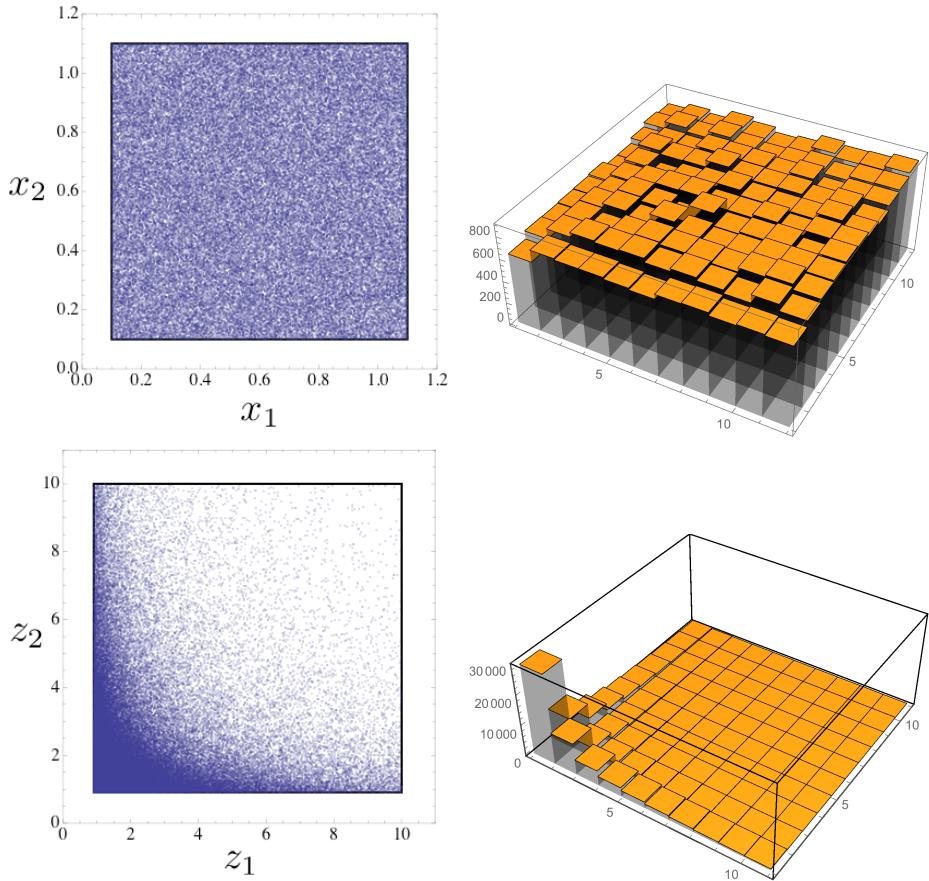


Figure 3.1: 100000 elementary events in the unit square defined by  $x_1 \in (0.1, 1.1)$ ,  $x_2 \in (0.1, 1.1)$ . Upper panel: uniform distribution of the elementary events: when we divide the square in smaller squares and count the number of elementary events in each of them, we find that this number is very nearly the same. Lower panel: here the  $z_i = 1/x_i$ , and the same elementary events, when so labeled, are no longer uniformly distributed over the new, transformed unit square. A different labeling changes the distribution of the elementary events in the (labeled) sample space.

in the interval  $(x', x'')$ ; we shall drop this distinction and simply write  $x$  instead of  $\xi$  as soon as there is no danger of confusing numerical values with the random variables themselves;

- c) let  $P(x' \leq \xi \leq x'')$  be the probability of finding the value  $x = \xi(\omega)$  in the interval  $(x', x'')$ , then  $P$  is called the *probability distribution* of the random variable  $\xi$ ;
- d) the random variable  $\xi$  is *discrete* if it assumes a finite number or a countably infinite number of different values;
- e) in the discrete case,  $P(x' \leq \xi \leq x'') = \sum_{x=x'}^{x''} P(\xi = x)$
- f) the random variable  $\xi$  is *continuous* if it is a continuous function of  $\omega$ ;
- g) in the continuous case one defines a function  $p_\xi(x)$ , which is called the *probability density function* (pdf) of the random variable  $\xi$ , such that  $P(x' \leq \xi \leq x'') = \int_{x'}^{x''} p_\xi(x) dx$ , i.e.,  $p_\xi(x) = \frac{dP}{dx}$ ;
- h) in the continuous case  $P(\xi = x) = 0$ ;  $P(x \leq \xi \leq x + dx) \approx p_\xi(x)dx$ ;
- i) the function  $\Phi(x) = P(\xi \leq x)$  is called the *distribution function*;
- j) in the continuous case,  $\Phi(x') = \int_{-\infty}^{x'} p_\xi(x) dx$ ;
- k) there is a simple relationship between probability distribution and distribution function  $P(x' \leq \xi \leq x'') = \Phi(x'') - \Phi(x')$ ;

These considerations can be extended to two (or more) random variables (and therefore to 2-dimensional – or higher-dimensional – sample spaces):

- l) if there are two random variables  $\xi_1$  and  $\xi_2$  we define a *random vector*  $(\xi_1, \xi_2)$  and a *joint probability distribution*;
- m) if the random variables are discrete, the joint probability distribution is simply  $p_{\xi_1, \xi_2}(x_1, x_2) = P(\xi_1 = x_1, \xi_2 = x_2)$ ;
- n) in the discrete case, the probability that the random point  $(\xi_1, \xi_2)$  belongs to a given set  $B$  is thus  $P((\xi_1, \xi_2) \in B) = \sum_{(x_1, x_2) \in B} p_{\xi_1, \xi_2}(x_1, x_2)$ ;
- o) if the random variables are continuous, we define a *joint probability density*  $p_{\xi_1, \xi_2}$ , as in the scalar case, so that  $P((\xi_1, \xi_2) \in B) = \int \int_B p_{\xi_1, \xi_2}(x_1, x_2) dx_1 dx_2$ ;
- p) given a family of random variables  $\xi_1, \xi_2, \dots, \xi_n$ , such that the events  $\{x'_k \leq \xi_k \leq x''_k\}$  are all independent for arbitrary  $x'_k, x''_k$ , then the random variables are said to be statistically independent; this is generalized to infinite sequences of random variables as well;

- q)* the joint probability distribution of independent variables factorizes, and the individual factors depend on a single random variable;
- r)* in the continuous case, the joint probability density of two independent random variables factorizes as follows:  $p_{\xi_1, \xi_2}(x_1, x_2) = p_{\xi_1}(x_1)p_{\xi_2}(x_2)$ .

### 3.1.1 The uniform distribution

If the weight  $\xi(\omega)$  (the value of the corresponding random variable) attributed to events  $\omega$  is constant, then we say that they are *uniformly distributed*, i.e., they are still equiprobable. Then, the events defined by “the numerical value of the continuous random variable belongs to interval  $(x', x'')$ , have probability proportional to the interval length. Thus

$$P(x' \leq \xi \leq x'') = k(x'' - x') \quad (3.1)$$

and

$$p_\xi(x) = k \quad (3.2)$$

where  $k$  is a constant. Assume that  $\xi$  is constrained to belong to interval  $(a, b)$ , then normalization implies that

$$\int_a^b p_\xi(x) dx = k(b - a) = 1 \quad (3.3)$$

i.e.,

$$k = \frac{1}{b - a} \quad (3.4)$$

and finally the pdf is

$$p_\xi(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in (a, b) \\ 0 & \text{if } x \notin (a, b) \end{cases} \quad (3.5)$$

### 3.1.2 Example: segment length

In this example two points (positions  $x_1$  and  $x_2$ ) are thrown at random on an interval  $[0, L]$ , and we want to determine the probability that the segment length  $|x_2 - x_1| \geq \ell$ .

We assume that throws are independent, then the joint probability density is just the product of the individual probability densities – which are both uniform – i.e.,

$$p_{\xi_1, \xi_2}(x_1, x_2) = \frac{1}{L^2} \quad (3.6)$$

For a given  $\ell$ , the condition  $|x_2 - x_1| \geq \ell$ . corresponds to the shaded regions in figure 3.2. It is easy to see that these two triangles have total area equal to  $(L - \ell)^2$ , and therefore the probability that the length of the

interval determined by the two points is greater or equal to  $l$  is  $(L-l)^2/L^2$ . Equivalently, the probability that interval length does not exceed  $l$  is just

$$1 - \frac{(L-l)^2}{L^2} = \frac{2Ll - l^2}{L^2} \quad (3.7)$$

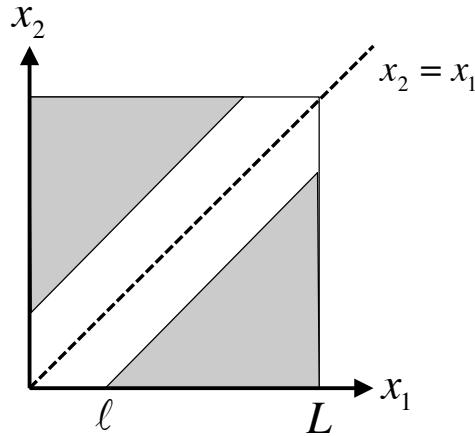


Figure 3.2: Segment length problem: each point in the square represents an acceptable pair of points  $(x_1, x_2)$ , however only the pairs in the shaded regions satisfy the condition  $|x_2 - x_1| \geq \ell$ . Since points are uniformly distributed, and the total area of the shaded regions is  $(L-\ell)^2$ , the probability of segment length  $|x_2 - x_1| \geq \ell$  is  $(L-\ell)^2/L^2$ .

### 3.1.3 Example: Buffon's needle

In 1770 Buffon posed the following problem: find the probability that a needle of length  $l$  dropped on a surface marked by parallel lines with spacing  $\ell$  intersects one of the lines. We can solve this problem as follows: there are two important random variables (see figure 3.3):

- $\xi_1$ , the orientation angle of the needle with respect to the orientation of the lines in the ruling ( $0 \leq \xi_1 \leq \pi$ );
- $\xi_2$ , the distance of the pivoting extremity of the needle from the line immediately below ( $0 \leq \xi_2 \leq \ell$ );

Both variables are uniformly distributed and are independent, therefore the overall distribution is uniform and the normalization constant is  $1/\ell\pi$ .

The length of the needle, projected in the direction perpendicular to the ruled pattern, is

$$l_p = l \sin(\xi_1) \quad (3.8)$$

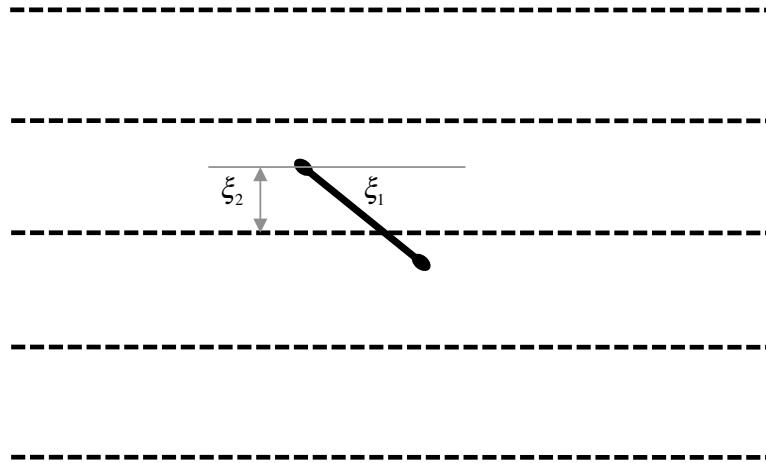


Figure 3.3: Illustration of Buffon’s needle problem: the problem is solved introducing the uniform variates  $\xi_1$ , the orientation angle of the needle with respect to the orientation of the lines in the ruling ( $0 \leq \xi_1 \leq \pi$ ), and  $\xi_2$ , the distance of the pivoting extremity of the needle from the line immediately below ( $0 \leq \xi_2 \leq \ell$ ).

Taking this orientation, the needle intersects the line below it if

$$\xi_2 \leq l_p = l \sin(\xi_1) \quad (3.9)$$

thus the probability of intersection is the area of this arc of sine curve times the normalization constant:

$$P_{\text{intersection}} = 2l/\ell\pi \quad (3.10)$$

Exercise: perform the experiment and use the result to obtain an “experimental” value of  $\pi$ .

Exercise: discuss the solution of Buffon’s needle problem in the two cases:  $l < \ell$  and  $l > \ell$ .

## 3.2 Mathematical Expectation

A probability distribution (or a pdf) places different weights on different regions of sample space: the *mathematical expectation* (or simply *expectation*, *mean value*, *average (value)*) is an estimate of the position of the centroid of the probability distribution (or pdf).

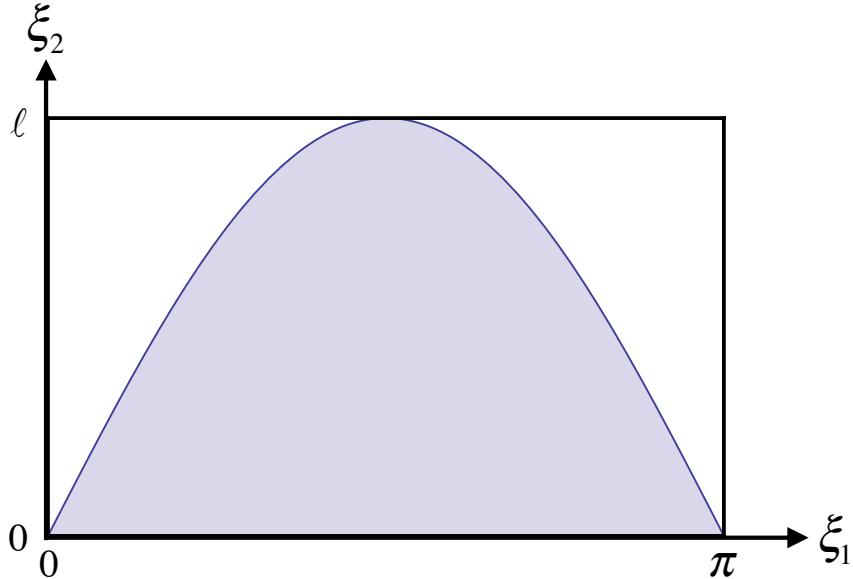


Figure 3.4: The rectangle shows the range of the variables  $\xi_1$  and  $\xi_2$  in Buffon's needle problem with  $l = \ell$ , while the shaded region corresponds to the intersection region. The probability of intersection is just the ratio of the areas of the two regions,  $2l/\ell\pi$ .

The mathematical expectation is indicated here with the usual symbol **E**. This symbol is associated with the purely mathematical operation

$$\mathbf{E}\xi = \sum_{-\infty}^{+\infty} xP_\xi(x) \quad (3.11)$$

for discrete variates, or

$$\mathbf{E}\xi = \int_{-\infty}^{+\infty} xp_\xi(x)dx \quad (3.12)$$

for continuous variates, and is different (though it is related to) the sample mean in descriptive statistics. The notation **E** $\xi$  is seldom used in the physics literature, as  $\langle x \rangle$  is used in its place; however the same notation  $\langle x \rangle$  is often used for the sample mean as well, and this can be confusing, therefore here we stick to **E** $\xi$ .

The extension to the case of two (or more) random variables is rather straightforward:

$$\mathbf{E}\varphi(\xi_1, \xi_2) = \sum_{x_1, x_2} \varphi(x_1, x_2)P_{\xi_1, \xi_2}(x_1, x_2) \quad (3.13)$$

List of properties of the mathematical expectation

- a)** let  $\eta = \varphi(\xi)$  be a new random variable, where  $\varphi(x)$  is some function of  $x$ , then

$$\mathbf{E}\eta = \sum_y y P_\eta(y) = \sum_y y \sum_{x:\varphi(x)=y} P_\xi(x) = \sum_x \varphi(x) P_\xi(x)$$

- b)**  $\mathbf{E}1 = \sum_x P_\xi(x) = 1$

- c)** if  $c$  is an arbitrary constant,  $\mathbf{E}(c\xi) = \sum_x cx P_\xi(x) = c\mathbf{E}\xi$

- d)**  $|\mathbf{E}\xi| = |\sum_x x P_\xi(x)| \leq \sum_x |x| P_\xi(x) = \mathbf{E}|\xi|$

- e)**  $\mathbf{E}(\xi_1 + \xi_2) = \sum_{x_1, x_2} (x_1 + x_2) P_{\xi_1, \xi_2}(x_1, x_2) = \mathbf{E}\xi_1 + \mathbf{E}\xi_2$

- f)** if  $\xi \geq 0$ , then  $\mathbf{E}\xi = \sum_x x P_\xi(x) \geq 0$

- g)** if  $\xi_1 \geq \xi_2$ , then  $\xi_1 - \xi_2 \geq 0$ , so that  $\mathbf{E}(\xi_1 - \xi_2) = \mathbf{E}\xi_1 - \mathbf{E}\xi_2 \geq 0$ , and finally  $\mathbf{E}\xi_1 \geq \mathbf{E}\xi_2$

- h)** if  $\xi_1$  and  $\xi_2$  are independent random variables, then

$$\mathbf{E}(\xi_1 \xi_2) = \sum_{x_1, x_2} x_1 x_2 P_{\xi_1}(x_1) P_{\xi_2}(x_2) = \sum_{x_1} x_1 P_{\xi_1}(x_1) \sum_{x_2} x_2 P_{\xi_2}(x_2) = \mathbf{E}\xi_1 \mathbf{E}\xi_2$$

- i)** if  $\xi_1$  and  $\xi_2$  are independent random variables, then

$$\begin{aligned} \mathbf{E}(\varphi_1(\xi_1)\varphi_2(\xi_2)) &= \sum_{x_1, x_2} \varphi_1(x_1)\varphi_2(x_2) P_{\xi_1}(x_1) P_{\xi_2}(x_2) \\ &= \sum_{x_1} \varphi_1(x_1) P_{\xi_1}(x_1) \sum_{x_2} \varphi_2(x_2) P_{\xi_2}(x_2) \\ &= \mathbf{E}\varphi_1(\xi_1) \mathbf{E}\varphi_2(\xi_2) \end{aligned}$$

The previous properties hold for discrete as well as for continuous distributions: see Rozanov's book [27] for further details on convergence properties of the sums involved in the proofs in the continuous case.

Example 1 (uniform distribution on  $(a, b)$ ):

$$\mathbf{E}\xi = \int_a^b x \frac{1}{b-a} dx = \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{b+a}{2}$$

### 3.3 Variance, standard deviation and covariance

The mean value alone is not enough to characterize a probability distribution, we need an estimator of the spread of values about the centroid of the distribution as well. Here we summarize the important steps that lead to the construction of the estimator and to an important related inequality

- a)** definition of *mean square value*:  $\mathbf{E}\xi^2 = \sum_x x^2 P_\xi(x)$ ;
- b)** definition of *variance*:  $\mathbf{D}\xi = \mathbf{E}(\xi - \mathbf{E}\xi)^2$ ;
- c)** in the physics literature the variance is usually denoted with the symbol  $\text{Var}$  or  $\sigma^2$ ;
- d)**  $\mathbf{D}\xi = \mathbf{E}(\xi - \mathbf{E}\xi)^2 = \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2$
- e)** the previous result implies:  $\mathbf{E}\xi^2 = \mathbf{D}\xi + (\mathbf{E}\xi)^2$
- f)**  $\sigma = \sqrt{\mathbf{D}\xi}$  is the *standard deviation* of the random variable  $\xi$ ;
- g)** if  $c$  is an arbitrary constant,  $\mathbf{D}(c\xi) = \mathbf{E}(c^2\xi^2) - (\mathbf{E}(c\xi))^2 = c^2\mathbf{D}\xi$
- h)** the *covariance* of two random variables  $\xi_1$  and  $\xi_2$  is defined by

$$\text{cov}(\xi_1, \xi_2) = \mathbf{E}[(\xi_1 - \mathbf{E}\xi_1)(\xi_2 - \mathbf{E}\xi_2)]$$

- i)**  $\text{cov}(\xi_1, \xi_2) = \mathbf{E}(\xi_1\xi_2 - \xi_2\mathbf{E}\xi_1 - \xi_1\mathbf{E}\xi_2 + \mathbf{E}\xi_1\mathbf{E}\xi_2) = \mathbf{E}(\xi_1\xi_2) - \mathbf{E}\xi_1\mathbf{E}\xi_2$
- j)** the previous result implies:  $\mathbf{E}(\xi_1\xi_2) = \text{cov}(\xi_1, \xi_2) + \mathbf{E}\xi_1\mathbf{E}\xi_2$
- k)** if  $\xi_1$  and  $\xi_2$  are independent random variables the covariance vanishes

$$\text{cov}(\xi_1, \xi_2) = \mathbf{E}[(\xi_1 - \mathbf{E}\xi_1)(\xi_2 - \mathbf{E}\xi_2)] = \mathbf{E}[(\xi_1 - \mathbf{E}\xi_1) \cdot \mathbf{E}(\xi_2 - \mathbf{E}\xi_2)] = 0$$

- l)** if  $\xi_1$  and  $\xi_2$  are independent random variables, then

$$\begin{aligned} \mathbf{D}(\xi_1 + \xi_2) &= \mathbf{E}[(\xi_1 + \xi_2) - \mathbf{E}(\xi_1 + \xi_2)]^2 \\ &= \mathbf{E}[(\xi_1 - \mathbf{E}(\xi_1)) + (\xi_2 - \mathbf{E}(\xi_2))]^2 \\ &= \mathbf{E}[(\xi_1 - \mathbf{E}(\xi_1))]^2 + \mathbf{E}[(\xi_2 - \mathbf{E}(\xi_2))]^2 - 2\mathbf{E}(\xi_1 - \mathbf{E}(\xi_1))(\xi_2 - \mathbf{E}(\xi_2)) \\ &= \mathbf{D}(\xi_1) + \mathbf{D}(\xi_2) - 2\text{cov}(\xi_1, \xi_2) \\ &= \mathbf{D}(\xi_1) + \mathbf{D}(\xi_2) \end{aligned}$$

- m)** if  $c$  is a constant and  $\xi_1 = c\xi_2$ , then

$$\text{cov}(\xi_1, \xi_2) = c\mathbf{E}(\xi_2 - \mathbf{E}\xi_2)^2 = c\mathbf{D}\xi_2$$

### 3.3.1 Higher moments

Mean value and variance are useful indicators of position and spread of the values of a random variable, however they are still not sufficient to fully characterise the distribution of values of the random variable. A complete description is provided by the full list of moments

$$m_n = \mathbf{E}(\xi^n)$$

for  $n = 1, \dots, \infty$ .  $\mathbf{E}(\xi^n)$  is called the *n-th moment about the origin*<sup>1</sup>. The moments can be used to construct the moment generating function, that we shall encounter again in a later chapter. We anticipate here that it is defined as follows:

$$M_\xi(t) = \mathbf{E}(e^{\xi t}) = \sum_x e^{xt} P_\xi(x) \quad (3.14)$$

$$= \mathbf{E} \left[ \sum_{n=0}^{\infty} \frac{(\xi t)^n}{n!} \right] = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbf{E}(\xi^n) \quad (3.15)$$

The moment generating function has a simple but important property: for i.i.d. random variables  $\{\xi_k\}_{k=1,n}$  and their sum  $\xi = \sum_k \xi_k$ , the following relation holds

$$M_\xi(t) = \mathbf{E}(e^{\sum_k \xi_k t}) = \prod_k \mathbf{E}(e^{\xi_k t}) = [M_{\xi_k}(t)]^n \quad (3.16)$$

(the last  $M_{\xi_k}(t)$  denotes the generic characteristic function of any one of the i.i.d. random variables  $\xi_k$ ).

---

<sup>1</sup>The related expectation values  $\mu_n = \mathbf{E}[(\xi - \mu)^n]$  are called *moments about the mean*.

### 3.4 Chebyshev's inequality, the laws of large numbers and other important inequalities.

#### 3.4.1 Chebyshev's inequality and the weak law of large numbers

Take a random variable  $\xi$  that has average  $\mu = \mathbf{E}\xi$  and variance  $\sigma^2 = \mathbf{E}(\xi - \mu)^2$ , then

$$\sigma^2 = \mathbf{E}(\xi - \mu)^2 = \sum_x (x - \mu)^2 P_\xi(x) \quad (3.17a)$$

$$= \sum_{|x-\mu|>k\sigma} (x - \mu)^2 P_\xi(x) + \sum_{|x-\mu|\leq k\sigma} (x - \mu)^2 P_\xi(x) \quad (3.17b)$$

$$\geq \sum_{|x-\mu|>k\sigma} (x - \mu)^2 P_\xi(x) \quad (3.17c)$$

$$\geq \sum_{|x-\mu|>k\sigma} k^2 \sigma^2 P_\xi(x) \quad (3.17d)$$

$$= k^2 \sigma^2 P(|\xi - \mu| > k\sigma) \quad (3.17e)$$

This means that

$$P(|\xi - \mu| > k\sigma) \leq \frac{1}{k^2} \quad (3.18)$$

i.e., the probability that  $x$  differs from the mean by more than  $k$  standard deviations is less than  $1/k^2$ ; this is Chebyshev's inequality. It is also common to write  $\epsilon = k\sigma$ , so that

$$P(|\xi - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \quad (3.19)$$

Now let  $\eta = \frac{1}{n} \sum_{k=1,n} \xi_k$ , where the  $\xi$ 's are independent, identically distributed (i.i.d.) random variables, each with mean  $\mu$  and variance  $\sigma^2$ . From the theorems of the previous sections, we find that  $\eta$  has the same mean  $\mu$  and variance  $\sigma^2/n$ . Therefore, using Chebyshev's inequality we find

$$P(|\eta - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \quad (3.20)$$

Thus, if we take  $n$  large enough, we can make the probability of finding a deviation larger than  $\epsilon$  as small as desired (*weak law of large numbers*).

Notice that an alternative form for the inequality is

$$P(|\eta - \mu| > \epsilon) = 1 - P(|\eta - \mu| \leq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \Rightarrow P(|\eta - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2} \quad (3.21)$$

In practice, we can use this inequality in cases where we know essentially nothing apart from estimates of mean and variance. Consider the following problem: we want to determine the minimum number of measurements such that the probability of a deviation of  $\eta$  from the mean  $\mu$ , greater than  $\mu/100$  is less than 1% (i.e., “the probability that the *relative error*  $|\eta - \mu|/\mu$  is greater than 1%” is less than 1%). Then

$$P(|\eta - \mu| > \mu/100) \leq \frac{10000\sigma^2}{n\mu^2} \leq 0.01 \quad (3.22)$$

and therefore

$$n \geq \frac{10^6\sigma^2}{\mu^2} \quad (3.23)$$

Now consider events A that occur with probability  $p = P(A)$  (as usual, let also  $q = 1 - p = P(\bar{A})$ ), and define a random variable  $\xi$  that equals 1 if A occurs and 0 if it does not occur (i.e., a Bernoulli random variable which takes the values 0 and 1, with probability  $q$  and  $p$ , respectively, and such that  $p + q = 1$ ). Clearly the expectation values are:

$$\mathbf{E}\xi = p \quad (3.24a)$$

$$\mathbf{E}(\xi^2) = p \quad (3.24b)$$

$$\mathbf{D}\xi = p - p^2 = p(1 - p) = pq \quad (3.24c)$$

Next, consider  $n$  repetitions of an experiment where A can occur, and let  $n(A)$  be the random variable that corresponds to the number of trials in which A occurs. Then the relative frequency of occurrence of A in these experiments is

$$\frac{n(A)}{n} = \frac{\xi_1 + \dots + \xi_n}{n} \quad (3.25)$$

where  $\xi_k$  is the random variable defined above, corresponding to the occurrence of A in the  $k$ -th trial. The mean value of the relative frequency is  $\mu = p = P(A)$ , and its variance is  $pq/n$ . Now we can use Chebishev's inequality and write

$$P\left(\left|\frac{n(A)}{n} - P(A)\right| > \epsilon\right) \leq \frac{pq}{n\epsilon^2} \quad (3.26)$$

i.e., the probability of large deviations of the relative frequency from  $P(A)$  can be made arbitrarily small, taking a large number of repetitions  $n$ .

### 3.4.2 Other inequalities in probability theory

Chebyshev's inequality is just one of many inequalities in probability theory. Here we add a few more.

### Markov's inequality

$$\mathbf{E}(|\xi|) = \sum_x |x| P_\xi(x) = \sum_{|x|>a} |x| P_\xi(x) + \sum_{|x|\leq a} |x| P_\xi(x) \quad (3.27a)$$

$$\geq \sum_{|x|>a} |x| P_\xi(x) \geq a \sum_{|x|>a} P_\xi(x) = a P(|\xi| > a) \quad (3.27b)$$

therefore

$$P(|\xi| > a) \leq \frac{\mathbf{E}(|\xi|)}{a} \quad (3.27c)$$

### Chebyshev's inequality as a special case of Markov's inequality

Let  $\eta = |\xi - \mathbf{E}(\xi)|^2$  then, from Markov's inequality

$$P(|\eta| > a^2) \leq \frac{\mathbf{E}(|\xi - \mathbf{E}(\xi)|^2)}{a^2} = \frac{\text{var } \xi}{a^2} \quad (3.28)$$

Since the square root is a monotonic function, the last inequality can be written as

$$P(|\xi - \mathbf{E}(\xi)| > a) \leq \frac{\text{var } \xi}{a^2} \quad (3.29)$$

which is Chebyshev's inequality again.

### Generalized Chebyshev's inequality

We can easily extend these results to higher powers  $N$  of the random variable:

$$\begin{aligned} \mathbf{E}(|\xi - \mu|^N) &= \sum_x |x - \mu|^N P_\xi(x) \\ &= \sum_{|x-\mu|>a} |x - \mu|^N P_\xi(x) + \sum_{|x-\mu|\leq a} |x - \mu|^N P_\xi(x) \\ &\geq \sum_{|x-\mu|>a} |x - \mu|^N P_\xi(x) \\ &\geq \sum_{|x-\mu|>a} a^N P_\xi(x) \\ &= a^N P(|\xi - \mu| > a) \end{aligned}$$

Therefore

$$P(|\xi - \mu| > a) \leq \frac{\mathbf{E}(|\xi - \mu|^N)}{a^N} \quad (3.30)$$

Notice, however, that this kind of inequalities is useful only as long as the expectation value  $\mathbf{E}(|\xi - \mu|^N)$  exists.

### 3.4.3 The strong law of large numbers

The weak law of large numbers

$$P \left( \left| \frac{1}{n} \sum_{k=1}^n \xi_k - \mu \right| > \epsilon \right) = P \left( \left| \frac{S_n}{n} - \mu \right| > \epsilon \right) \leq \frac{\sigma^2}{n\epsilon^2} \quad (3.31)$$

is weak in the sense that it states that the *sample average*

$$\frac{S_n}{n} = \frac{1}{n} \sum_{k=1}^n \xi_k \quad (3.32)$$

mostly lies in a small range around the mathematical expectation, but does not rule out fluctuations and temporary departures from this average behavior, i.e., the probability of large fluctuations becomes small as  $n$  increases, but although their frequency decreases there can still be – in principle – an infinite number of them (see figure ). Here we prove now a stronger statement, which is appropriately called *the strong law of large numbers*, which assures that – under rather mild conditions – the frequency of fluctuations of the sample mean above any given threshold drops effectively to zero.

To prove the strong law of large numbers, it suffices to show that large fluctuations have vanishing probability, or, in other words, that they cannot occur infinitely often. Indeed we already have a powerful tool to accomplish this task, the first Borel-Cantelli lemma .

First, we simplify slightly the notation, introducing the auxiliary random variables  $\xi'_k = \xi_k - \mu$ , so that now the corresponding sample average  $S'_n/n$  fluctuates about 0, and Chebyshev's inequality reads

$$P \left( \left| \frac{S'_n}{n} \right| > \epsilon \right) = P \left( |S'_n| > n\epsilon \right) \leq \frac{\sigma^2}{n\epsilon^2} \quad (3.33)$$

Now we consider the sample space  $\Omega$  of all possible sequences of values  $\omega = \{x'_1, x'_2, \dots\}$  of the random variables  $\xi'_1, \xi'_2, \dots$ , and the events

$$A_n = \{\omega : |S'_n| > n\epsilon\} \quad (3.34)$$

If we were able to show that

$$\sum_n P(A_n) < \infty \quad (3.35)$$

then we could use the first Borel-Cantelli lemma to show that only finitely many of the events  $A_n$  can occur, and thus we would prove that

$$P(|S'_n| > n\epsilon \text{ i.o.}) = 0 \quad (3.36)$$

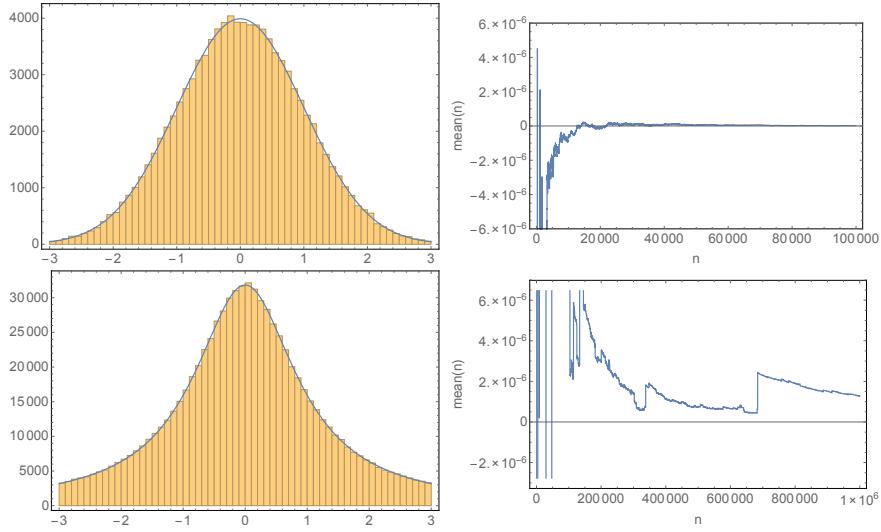


Figure 3.5: Upper panels: 100000 random numbers with Gaussian distribution  $N(0, 1)$ . The upper left panel is the histogram that shows the empirical distribution of the random numbers  $\{x_n\}_{n=1,100000}$ , the solid curve is the theoretical Gaussian pdf. The upper right panel shows the running mean  $(1/n) \sum_{k=1}^n x_k$  vs.  $n$ . Lower panels: 1000000 random numbers with Cauchy distribution  $p(x) = 1/\pi(1 + x^2)$ . The lower left panel is the histogram that shows the empirical distribution of the random numbers  $\{x_n\}_{n=1,1000000}$ , the solid curve is the theoretical Cauchy pdf. The lower right panel shows the running mean  $(1/n) \sum_{k=1}^n x_k$  vs.  $n$ . The Gaussian distributions satisfies the assumption that lead to both the weak law of large numbers and to the strong law of large numbers, and the running mean converges to the mean of the underlying Gaussian. The Cauchy distribution does not satisfy the assumptions, and no matter how many random numbers are drawn, the mean does not converge, and large fluctuations are always possible.

(here “i.o.” stands for “infinitely often”), which is a way to state the strong law of large numbers.

Unfortunately we cannot use the bound derived from Chebyshev’s inequality, because  $P(A_n) = P(|S'_n| > n\epsilon) < \sigma^2/n\epsilon^2$ , and the sum  $\sum_n 1/n$  diverges. There are several ways to proceed, and they all involve additional assumptions to bypass the limitation of Chebyshev’s inequality: here we make the straightforward assumption that the random variables are i.i.d., that  $\sigma^2 = \mathbf{E}\xi'^2 < \infty$ , and finally that  $\mathbf{E}\xi'^4 < \infty$ . The last assumption is slightly too restrictive – the strong law can be proved under milder conditions – but it streamlines the proof considerably.

Since the mathematical expectation  $\mathbf{E}\xi'^4$  exists, we find that the general-

ized Chebyshev's inequality holds

$$P(|S'_n| > n\epsilon) < \frac{\mathbf{E}[(S'_n)^4]}{n^4\epsilon^4} \quad (3.37)$$

Now we must evaluate the expectation value

$$\mathbf{E}[(S'_n)^4] = \mathbf{E}\left[\left(\sum_{k=1}^n \xi'_k\right)^4\right] = \mathbf{E}\left(\sum_{i,j,k,l=1}^n \xi'_i \xi'_j \xi'_k \xi'_l\right) = \sum_{i,j,k,l=1}^n \mathbf{E}(\xi'_i \xi'_j \xi'_k \xi'_l) \quad (3.38)$$

recalling that the random variables are i.i.d. . The indexes in the individual expectations  $\mathbf{E}(\xi'_i \xi'_j \xi'_k \xi'_l)$  can be

**all different:** then  $\mathbf{E}(\xi'_i \xi'_j \xi'_k \xi'_l) = [\mathbf{E}(\xi'_i)]^4 = 0$  (the  $\xi'$  have 0 mean);

**two indexes equal, other indexes different:** then  $\mathbf{E}(\xi'_i \xi'_j \xi'_k \xi'_l) = 0$  just as in the previous case;

**three indexes equal, remaining index different:** again  $\mathbf{E}(\xi'_i \xi'_j \xi'_k \xi'_l) = 0$ ;

**two pairs of equal indexes:** in this case  $\mathbf{E}(\xi'_i \xi'_j \xi'_k \xi'_l) = \mathbf{E}(\xi'^2_i)\mathbf{E}(\xi'^2_j) = \sigma^4$ ; there are  $\binom{4}{2} = 6$  ways of fixing the indexes, and there are  $n$  choices for the first index and  $n - 1$  for the second, therefore there are  $6n(n - 1)$  such (ordered) pairs;

**all indexes are equal:**  $\mathbf{E}(\xi'_i \xi'_j \xi'_k \xi'_l) = \mathbf{E}\xi'^4$ ; clearly there are  $n$  such terms.

Finally we can evaluate the expectation value:

$$\mathbf{E}(S'^4_n) = 6n(n - 1)\sigma^4 + n\mathbf{E}\xi'^4 \quad (3.39)$$

so that for  $n$  large enough we find

$$\mathbf{E}(S'^4_n) < Cn^2 \quad (3.40)$$

and therefore

$$P(|S'_n| > n\epsilon) < \frac{\mathbf{E}(S'^4_n)}{n^4\epsilon^4} < \frac{Cn^2}{n^4\epsilon^4} = \frac{C}{n^2\epsilon^4} \quad (3.41)$$

Then, for  $n$  large enough,

$$\sum_n P(A_n) < \sum_n \frac{C}{n^2\epsilon^4} < \infty \quad (3.42)$$

thus we provide the missing element of the proof, because the series converges, and we find that the strong law of large numbers

$$P(|S'_n| > n\epsilon \text{ i.o.}) = 0 \quad (3.43)$$

really holds.

### 3.4.4 Chernoff bound

The generalized Chebyshev's inequalities of the previous sections put power-law bounds on the tails of the probability distributions, however in some cases it is possible to set much tighter, exponential bounds: one such bound is the Chernoff bound, of which we give standard proofs here.

We assume that  $\xi$  is non-negative, then recalling Markov's inequality, and the moment generating function  $M_\xi(t) = \sum_n \mathbf{E}(\xi^n)t^n/n!$ , we write

$$P(\xi \geq a) = P(e^{t\xi} \geq e^{ta}) \leq \frac{\mathbf{E}(e^{t\xi})}{e^{ta}} = e^{-ta}M_\xi(t) \quad (3.44)$$

for  $t > 0$ , and

$$P(\xi \leq a) = P(e^{t\xi} \geq e^{ta}) \leq \frac{\mathbf{E}(e^{t\xi})}{e^{ta}} = e^{-ta}M_\xi(t) \quad (3.45)$$

for  $t < 0$ ; the pair of inequalities (3.44) and (3.45) is our first version of the Chernoff bounds.

Likewise, when we take i.i.d. non-negative random variables  $\{\xi_k\}_{k=1,n}$  and their sum  $\xi = \sum_k \xi_k$ , we find

$$P(\xi \geq a) \leq e^{-ta}[M_{\xi_k}(t)]^n \quad (3.46)$$

for  $t > 0$ , and

$$P(\xi \leq a) \leq e^{-ta}[M_{\xi_k}(t)]^n \quad (3.47)$$

for  $t < 0$ .

In a very typical application, we take Bernoulli random variables: the mean value of a Bernoulli variate is simply  $p$ , and its moment generating function is

$$M_1(t) = qe^{0 \cdot t} + pe^{1 \cdot t} = 1 - p + pe^t \quad (3.48)$$

A sum  $S_n$  of  $n$  i.i.d. Bernoulli variates has the mean value  $np$  and the moment generating function

$$M(t) = [M_1(t)]^n = (1 - p + pe^t)^n \leq e^{np(e^t - 1)} \quad (3.49)$$

Then

$$P[S_n \geq (1 + \epsilon)np] \leq e^{-t(1+\epsilon)np}[M_{\xi_k}(t)]^n \leq e^{-t(1+\epsilon)np}e^{np(e^t - 1)} \quad (3.50)$$

Now, to get rid of the exponential in the exponent, we take the specific value  $t = \ln(1 + \epsilon)$ , and then

$$P[S_n \geq (1 + \epsilon)np] \leq (1 + \epsilon)^{-(1+\epsilon)np}e^{np\epsilon} \quad (3.51)$$

For small deviations ( $0 < \epsilon < 1$ ) the following inequality holds

$$\frac{e^\epsilon}{(1+\epsilon)^{1+\epsilon}} < e^{-\epsilon^2/3} \quad (3.52)$$

and finally we find

$$P[S_n \geq (1+\epsilon)np] < e^{-npe^2/3} \quad (3.53)$$

We can apply this result to a problem of network traffic. Consider  $n$  packets of information transmitted over a network with  $m$  routers: the mean number of packets through each router is  $n/m$ . If one of the routers receives less than this average number, then at least one of the other routers is above average, and it may even be that *all* of the traffic goes through this single router, and possibly disrupt the network. However this is very unlikely, and we can use the results above to set bounds. We define a set of Bernoulli variates  $\xi_k$  which take the value 0 if the  $k$ -th message does not transit through a given router, and the value 1 if it does. Then, using the Chernoff bound for Bernoulli variates we write

$$\begin{aligned} P^{(+)}(\epsilon) &= P[\text{more than } (1+\epsilon)n/m \text{ packets through one router}] = \\ &P[S_n \geq (1+\epsilon)n/m] < e^{-n\epsilon^2/3m} \end{aligned} \quad (3.54)$$

and likewise

$$\begin{aligned} P^{(-)}(\epsilon) &= P[\text{less than } (1-\epsilon)n/m \text{ packets through one router}] = \\ &P[S_n \leq (1-\epsilon)n/m] < e^{-n\epsilon^2/3m} \end{aligned} \quad (3.55)$$

Then the probability of a fluctuation in the total network traffic is bounded by the inequality

$$\begin{aligned} P[\text{more than } (1+\epsilon)n/m \text{ packets through some router}] &= \\ 1 - [1 - P^{(+)}(\epsilon)]^m &\leq mP^{(+)}(\epsilon) < me^{-n\epsilon^2/3m} \end{aligned} \quad (3.56)$$

Next we prove a version of the Chernoff bound for a sum of variates that are not necessarily non-negative. We shift the mean values so that  $S'_n$  is a zero-mean random variable with variance  $\sigma^2$ , and such that it is the a sum of  $n$  i.i.d. zero-mean random variables  $\xi'_k$ , with the condition that  $|\xi'_k| \leq 1$ , and with variance  $\sigma^2/n$ :

$$S'_n = \sum_{k=1}^n \xi'_k \quad (3.57)$$

Now we prove an easy inequality which holds for all the  $\xi'_i$ 's: we take  $0 \leq t \leq 1$ , and we consider the following expectation value:

$$\mathbf{E}(e^{t\xi'_i}) = \mathbf{E} \left( 1 + t\xi'_i + \sum_{k=2,\infty} \frac{1}{k!} (t\xi'_i)^k \right) \quad (3.58a)$$

$$\leq \mathbf{E} \left( 1 + t\xi'_i + t^2 \xi'^2_i \sum_{k=2,\infty} \frac{1}{k!} \right) \quad (3.58b)$$

$$\leq \mathbf{E} (1 + t\xi'_i + t^2 \xi'^2_i) = 1 + t^2 \text{var } \xi'_i \leq e^{t^2 \sigma^2 / n} \quad (3.58c)$$

Now we go back to the random variable  $S'_n$ , and we note that the probability of a (positive) deviation larger than  $\lambda\sigma$  from the (zero) mean value is

$$P(S'_n \geq \lambda\sigma) = P(tS'_n \geq t\lambda\sigma) = P(e^{tS'_n} \geq e^{t\lambda\sigma}) \quad (3.59)$$

(the equalities hold because of the monotonicity of the transformations). From Markov's inequality, and taking positive deviations from the mean, we find

$$P(S'_n \geq \lambda\sigma) = P(e^{tS'_n} \geq e^{t\lambda\sigma}) \leq \frac{\mathbf{E}(e^{tS'_n})}{e^{t\lambda\sigma}} \quad (3.60)$$

Moreover, from the inequality that we proved above, we obtain

$$\mathbf{E}(e^{tS'_n}) = \mathbf{E}(e^{\sum_i t\xi'_i}) = \mathbf{E} \left( \prod_i e^{t\xi'_i} \right) = \prod_i \mathbf{E}(e^{t\xi'_i}) \leq e^{t^2 \sigma^2} \quad (3.61)$$

therefore

$$P(S'_n \geq \lambda\sigma) \leq \frac{e^{t^2 \sigma^2}}{e^{t\lambda\sigma}} = e^{t^2 \sigma^2 - t\lambda\sigma} \quad (3.62)$$

The last inequality must hold for all  $t \in (0, 1)$ , and in particular it must hold for the minimum of  $t^2 \sigma^2 - t\lambda\sigma$ , which is also the strongest condition for the validity of the inequality, i.e. for  $t_m = \lambda/2\sigma$  and

$$t_m^2 \sigma^2 - t_m \lambda\sigma = -\lambda^2/4 \quad (3.63)$$

so that

$$P(S'_n \geq \lambda\sigma) \leq e^{-\lambda^2/4} \quad (3.64)$$

The argument is symmetric for negative deviations  $\xi < 0$  and yields

$$P(S'_n \leq -\lambda\sigma) \leq e^{-\lambda^2/4} \quad (3.65)$$

and finally we find

$$P(|S'_n| \geq \lambda\sigma) = P(S'_n \leq -\lambda\sigma) + P(S'_n \geq \lambda\sigma) \leq 2e^{-\lambda^2/4} \quad (3.66)$$

which is the Chernoff bound. The Chernoff bound is much stronger than Chebyshev's inequality, however it requires the random variable  $S'_n$  to be the sum of bounded i.i.d. random variables. This bound has applications in hypothesis testing and in quantum mechanics. Notice that, as in Chebyshev's inequality, we can set  $\lambda\sigma = \epsilon$ , so that an alternative form of the bound is

$$P(|S'_n| \geq \epsilon) \leq 2e^{-\epsilon^2/4\sigma^2} \quad (3.67)$$

(this is a variant of the Chernoff bound that is called *Chernoff-Hoeffding bound*).

Finally, if the individual variates have nonzero mean  $\mu$ , and if we let  $\epsilon = k\sigma$ , then the probability of large fluctuations of the sum, and therefore of the average, writes (using the same symbols as in the proof of the strong law of large numbers)

$$P(|S_n - n\mu| \geq k\sigma) = P\left(\left|\frac{S_n}{n} - \mu\right| \geq \frac{k\sigma}{n}\right) \leq 2e^{-k^2/4n^2} \quad (3.68)$$

This estimate of the probability of large fluctuations from the mean falls off very quickly as  $n$  grows, and it is much stronger than that provided by Chebyshev's inequalities. This bonus has been obtained with the important additional assumptions that the individual variates are independent and bounded.

### 3.5 What's the use of all this?

The concepts introduced in this chapter of the book are the backbone of probability models, of descriptive statistics, and also of the basic of statistical inference. While the expectation and variance described here are strictly mathematical concepts, they straightforwardly extend to the related statistical estimators. Further on in the course we shall consider other related statistical estimators, and we shall introduce a particularly important transformation of a set of correlated random variables called “Principal Component Analysis” (PCA).

Buffon's needle can be generalized in several ways, and is a practical model for the filtering of granular materials (see, e.g., [8]).

Finally probabilistic inequalities and the laws of large numbers are important in several ways, both theoretically and practically, since they help in experimental design and in the analysis of traffic networks.

## Chapter 4

# Probability distributions

Here I introduce several important probability distribution and ways to handle them. This is important basic knowledge, since probability distributions are some of the building blocks of statistical modeling.

### 4.1 The uniform distribution

The uniform distribution is totally uninformative, all values of the random variable are equiprobable. Here we consider a continuous scalar random variable uniformly distributed over an interval  $(a, b)$

$$p_\xi(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & x \notin (a, b) \end{cases} \quad (4.1)$$

- average:

$$\mathbf{E}\xi = \int_a^b x p_\xi(x) dx = \frac{1}{2} \frac{b^2 - a^2}{b - a} = \frac{a + b}{2} \quad (4.2)$$

- mean square value:

$$\mathbf{E}(\xi)^2 = \int_a^b x^2 p_\xi(x) dx = \frac{1}{3} \frac{b^3 - a^3}{b - a} = \frac{a^2 + ab + b^2}{3} \quad (4.3)$$

- variance:

$$\begin{aligned} \mathbf{D}\xi &= \mathbf{E}(\xi)^2 - (\mathbf{E}\xi)^2 = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{a^2 - 2ab + b^2}{12} \\ &= \frac{(a - b)^2}{12} \end{aligned} \quad (4.4)$$

- standard deviation:

$$\sigma = \sqrt{\mathbf{D}\xi} = \frac{a - b}{\sqrt{12}} \quad (4.5)$$

- moment generating function:

$$M(t) = \int_a^b \frac{1}{b-a} e^{tx} dx = \frac{1}{t} \cdot \frac{e^{tb} - e^{ta}}{b-a} \quad (4.6)$$

## 4.2 The Bernoulli distribution

The Bernoulli distribution is a discrete probability distribution where the random variable  $\xi$  takes only two different values (usually 1 and 0) with probability, respectively,  $p$  and  $q = 1 - p$ . Then:

- average:

$$\mathbf{E}\xi = 1 \cdot p + 0 \cdot q = p \quad (4.7)$$

- mean square value:

$$\mathbf{E}\xi^2 = 1^2 \cdot p + 0^2 \cdot q = p \quad (4.8)$$

- variance:

$$\mathbf{D}\xi = \mathbf{E}(\xi)^2 - (\mathbf{E}\xi)^2 = p - p^2 = pq \quad (4.9)$$

- standard deviation:

$$\sigma = \sqrt{\mathbf{D}\xi} = \sqrt{pq} \quad (4.10)$$

- moment generating function:

$$M(t) = 1 - p + pe^t \quad (4.11)$$

The Bernoulli distribution is obeyed by *indicator random variables*.

## 4.3 The binomial distribution

In the binomial model, we consider the occurrence of events A in a series of  $n$  trials. The probability of obtaining A in each trial is  $p$ , the probability of obtaining  $\bar{A}$  is  $q = 1 - p$ . Consider the case of  $k$  occurrences in a series of  $n$  trials: each sequence of trials has probability  $p^k q^{n-k}$ , and there are  $\binom{n}{k}$  such sequences, therefore the probability of  $k$  occurrences of A in  $n$  trials is

$$P_\xi(k) = P(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (4.12)$$

The distribution (4.12) is the *binomial distribution*.

- normalization:

$$\sum_{k=0}^{k=n} P_\xi(k) = \sum_{k=0}^{k=n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1 \quad (4.13)$$

- average:

$$\mathbf{E}\xi = \sum_{k=0}^{k=n} k P_\xi(k) = \sum_{k=1}^{k=n} \binom{n}{k} kp^k(1-p)^{n-k} \quad (4.14a)$$

$$= np \sum_{k=1}^{k=n} \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \quad (4.14b)$$

$$= np \sum_{k=0}^{k=n-1} \binom{n-1}{k} p^k(1-p)^{n-k-1} \quad (4.14c)$$

$$= np [p + (1-p)]^{n-1} = np \quad (4.14d)$$

- variance

$$\mathbf{E}\xi^2 = \sum_{k=0}^{k=n} k^2 P_\xi(k) = \sum_k [k(k-1) + k] P_\xi(k) \quad (4.15a)$$

$$= n(n-1) \sum_{k=2}^{k=n} \binom{n-2}{k-2} p^k(1-p)^{n-k} + np \quad (4.15b)$$

$$= n(n-1)p^2 + np \quad (4.15c)$$

Thus

$$\mathbf{D}\xi = \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2 = n(n-1)p^2 + np - n^2p^2 = np - np^2 = npq \quad (4.16)$$

- moment generating function:

$$M(t) = (1 - p + pe^t)^n \quad (4.17)$$

Figure 4.1 shows a binomial distribution with  $p = 0.5$  and  $n = 20$ : the distribution is symmetrical and has a single peak at  $np = 10$ . Figure 4.2 shows two more cases, still with  $n = 20$ : first with  $p = 0.2$ , and then with  $p = 0.01$ .

Notice that we can also write

$$\xi = \sum_{k=1}^n \xi_k \quad (4.18)$$

where the  $\xi_k$ 's are i.i.d. indicator random variables. Therefore we find immediately

$$\mathbf{E}\xi = \sum_{k=1}^n \mathbf{E}\xi_k = np \quad (4.19)$$

and, using independence,

$$\mathbf{D}\xi = \sum_{k=1}^n \mathbf{D}\xi_k = npq \quad (4.20)$$

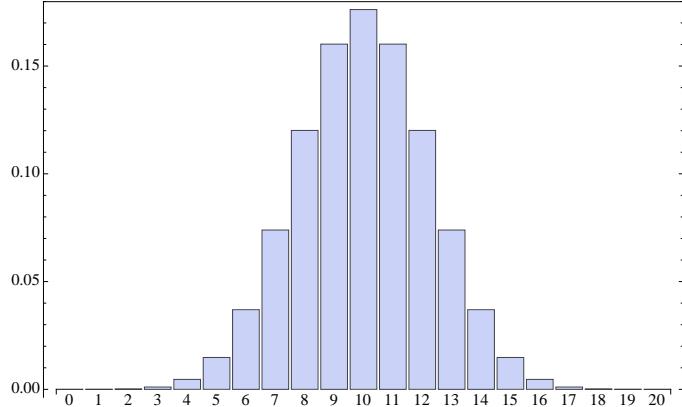


Figure 4.1: Binomial distribution with  $p = 0.5$  and  $n = 20$ : the histogram shows  $P_\xi(k)$  vs.  $k$ . With  $p = 0.5$  the binomial distribution is symmetrical.

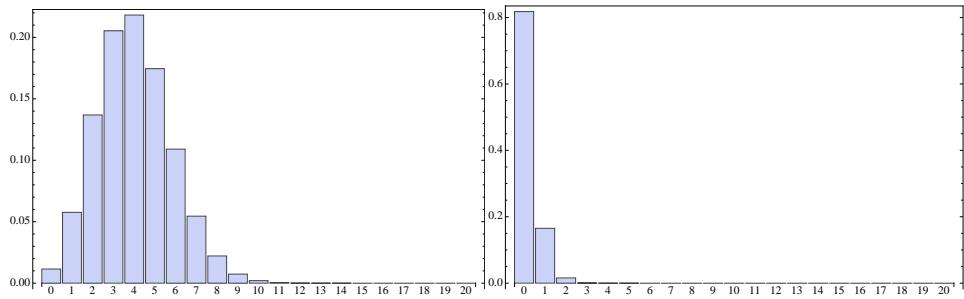


Figure 4.2: Binomial distribution with  $n = 20$  and  $p = 0.2$  (left panel) and  $p = 0.01$  (right panel). Notice that in these cases the distribution is not symmetrical and in the second case it is strongly asymmetrical.

#### 4.4 The multinomial distribution

The binomial distribution can be easily generalized to  $k > 2$  possible outcomes:

$$P_{\xi_1, \xi_2, \dots, \xi_k}(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \quad (4.21)$$

where  $\sum_i n_i = n$  and  $\sum_i p_i = 1$ .

Figure 4.3 shows a multinomial distribution with  $k = 3$  and  $p_1 = p_2 = p_3 = 1/3$ .

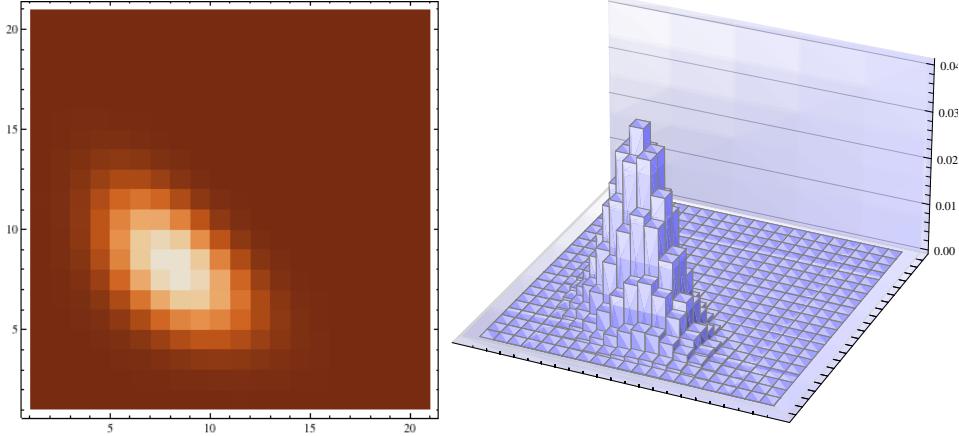


Figure 4.3: Multinomial distribution with  $n = 20$ ,  $k = 3$  and  $p_1 = p_2 = p_3 = 1/3$ , plotted as a function of the independent values  $n_1$  and  $n_2$ . Density plot (left panel) and lego plot (right panel). As an exercise, explain why in this symmetrical case the distribution is not centered in the  $n_1, n_2$  domain, and consider ways to represent multinomial distributions with  $k > 3$ .

- normalization:

$$\begin{aligned} \sum_{n_1+n_2+\dots+n_k=n} P_{\xi_1, \xi_2, \dots, \xi_k}(n_1, n_2, \dots, n_k) &= \sum_{n_1+n_2+\dots+n_k=n} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \\ &= (p_1 + p_2 + \dots + p_k)^n = 1 \end{aligned} \quad (4.22)$$

- averages:

$$\mathbf{E}\xi_i = \sum_{n_1+n_2+\dots+n_k=n} n_i P_{\xi_1, \xi_2, \dots, \xi_k}(n_1, n_2, \dots, n_k) \quad (4.23a)$$

$$= p_i \frac{\partial}{\partial x} \sum_{n_1+n_2+\dots+n_k=n} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots x^{n_i} \dots p_k^{n_k} \Big|_{x=p_i} \quad (4.23b)$$

$$= p_i \frac{\partial}{\partial x} (p_1 + p_2 + \dots + x + \dots + p_k)^n \Big|_{x=p_i} \quad (4.23c)$$

$$= np_i (p_1 + p_2 + \dots + p_k)^{n-1} = np_i \quad (4.23d)$$

- mean square value:

$$\mathbf{E}\xi_i^2 = \sum_{n_1+n_2+\dots+n_k=n} n_i^2 P_{\xi_1,\xi_2,\dots,\xi_k}(n_1, n_2, \dots, n_k) \quad (4.24a)$$

$$= \sum_{n_1+n_2+\dots+n_k=n} [n_i(n_i - 1) + n_i] P_{\xi_1,\xi_2,\dots,\xi_k}(n_1, n_2, \dots, n_k) \quad (4.24b)$$

$$= p_i^2 \frac{\partial^2}{\partial x^2} \sum_{n_1+n_2+\dots+n_k=n} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots x^{n_i} \dots p_k^{n_k} \Big|_{x=p_i} + np_i \quad (4.24c)$$

$$= p_i^2 \frac{\partial^2}{\partial x^2} (p_1 + p_2 + \dots + x + \dots + p_k)^n \Big|_{x=p_i} + np_i \quad (4.24d)$$

$$= n(n-1)p_i^2 (p_1 + p_2 + \dots + p_k)^{n-2} + np_i = n(n-1)p_i^2 + np_i \quad (4.24e)$$

- variance:

$$\sigma_i^2 = \mathbf{E}\xi_i^2 - (\mathbf{E}\xi_i)^2 = n(n-1)p_i^2 + np_i - n^2p_i^2 = np_i(1-p_i) \quad (4.25)$$

- mean cross-product ( $i \neq j$ ):

$$\mathbf{E}(\xi_i\x_j) = \sum_{n_1+n_2+\dots+n_k=n} n_i n_j P_{\xi_1,\xi_2,\dots,\xi_k}(n_1, n_2, \dots, n_k) \quad (4.26a)$$

$$= p_i p_j \frac{\partial^2}{\partial x \partial y} \sum_{n_1+n_2+\dots+n_k=n} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots x^{n_i} \dots y^{n_j} \dots p_k^{n_k} \Big|_{x=p_i, y=p_j} \quad (4.26b)$$

$$= p_i p_j \frac{\partial^2}{\partial x \partial y} (p_1 + p_2 + \dots + x + \dots + y + \dots + p_k)^n \Big|_{x=p_i, y=p_j} \quad (4.26c)$$

$$= n(n-1)p_i p_j \quad (4.26d)$$

- covariance ( $i \neq j$ ):

$$\text{cov}(\xi_i, \xi_j) = \mathbf{E}(\xi_i\x_j) - \mathbf{E}\xi_i \mathbf{E}\xi_j = n(n-1)p_i p_j - n^2 p_i p_j = -np_i p_j \quad (4.27)$$

- moment generating function:

$$M(t_1, \dots, t_n) = \sum_{n_1+n_2+\dots+n_k=n} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} e^{n_1 t_1} p_2^{n_2} e^{n_2 t_2} \dots p_k^{n_k} e^{n_k t_k} \\ = (p_1 e^{t_1} + p_2 e^{t_2} + \dots + p_k e^{t_k})^n \quad (4.28)$$

## 4.5 The Poisson distribution

Now take a binomial distribution with  $n \gg 1$  and  $p \ll 1$ , and with the condition that  $a = np$  is not negligibly small. We start from the recursive formula for binomial probability

$$\frac{P_\xi(k)}{P_\xi(k-1)} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{(k-1)!(n-k+1)!}{n!} \frac{1}{p^{k-1}(1-p)^{n-k+1}} \quad (4.29a)$$

$$= \frac{n-k+1}{k} \frac{p}{1-p} \quad (4.29b)$$

We can approximate this as follows for  $k \ll n$

$$P_\xi(k) \approx \frac{np}{k} P_\xi(k-1) = \frac{a}{k} P_\xi(k-1) \quad (4.30)$$

and since

$$P_\xi(0) = (1-p)^n = \left(1 - \frac{a}{n}\right)^n \approx e^{-a} \quad (4.31)$$

we find

$$\begin{aligned} P_\xi(0) &= e^{-a} \\ P_\xi(1) &= ae^{-a} \\ P_\xi(2) &= \frac{a^2}{2} e^{-a} \\ P_\xi(3) &= \frac{a^3}{2 \cdot 3} e^{-a} \\ &\dots \\ P_\xi(k) &= \frac{a^k}{k!} e^{-a} \\ &\dots \end{aligned}$$

The probability distribution  $P_\xi(k) = \frac{a^k}{k!} e^{-a}$  is the *Poisson distribution*. Figure 4.4 shows the shape of the Poisson distribution for different  $a$  values.

In the foregoing text we have obtained the Poisson distribution from a limiting behavior of the binomial model, however another important probability model leads to the same distribution. Consider a random process in time (events occur at random times), and let  $\lambda$  be the event rate, so that the average number of events in a time interval  $\Delta t$  is  $\lambda\Delta t$ . When  $\Delta t$  is very small, so that  $\lambda\Delta t < 1$  this is also the probability that an event occurs in the time interval  $\Delta t$  (exercise: explain why this is so). Assume also that the events are independent, so that the occurrence times are uncorrelated. The

probability  $P_\xi(k, t)$  of the occurrence of  $k > 0$  events in a time interval  $(0, t)$  depends on  $t$ , and this time-dependence must be such that

$$P_\xi(k, t + dt) = P_\xi(k, t)(1 - \lambda dt) + P_\xi(k - 1, t)\lambda dt \quad (4.33)$$

while for  $k = 0$

$$P_\xi(0, t + dt) = P_\xi(0, t)(1 - \lambda dt) \quad (4.34)$$

where we assume that for short time intervals the probability of occurrence of an event is equal to the average number of events in that interval. Thus we obtain the following differential equation for  $k = 0$

$$\frac{dP_\xi(0, t)}{dt} = -\lambda P_\xi(0, t) \quad (4.35)$$

and for  $k > 0$

$$\frac{dP_\xi(k, t)}{dt} = -\lambda P_\xi(k, t) + \lambda P_\xi(k - 1, t) \quad (4.36)$$

where the initial conditions are  $P_\xi(0, 0) = 1$  and  $P_\xi(k, 0) = 0$ . The solution of the first equation is trivial

$$P_\xi(0, t) = e^{-\lambda t} \quad (4.37)$$

The other equations are solved by rewriting them as follows

$$\frac{dP_\xi(k, t)}{dt} + \lambda P_\xi(k, t) = e^{-\lambda t} \frac{d}{dt} (e^{\lambda t} P_\xi(k, t)) = \lambda P_\xi(k - 1, t) \quad (4.38)$$

therefore

$$e^{-\lambda t} \frac{d}{dt} (e^{\lambda t} P_\xi(1, t)) = \lambda e^{-\lambda t} \quad (4.39)$$

which has the solution

$$P_\xi(1, t) = \lambda t e^{-\lambda t} \quad (4.40)$$

Similarly

$$e^{-\lambda t} \frac{d}{dt} (e^{\lambda t} P_\xi(2, t)) = \lambda^2 t e^{-\lambda t} \quad (4.41)$$

which has the solution

$$P_\xi(2, t) = \frac{\lambda^2 t^2}{2} e^{-\lambda t} \quad (4.42)$$

We can surmise that in general

$$P_\xi(k, t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (4.43)$$

and indeed this is easily proved by substitution in the generic differential equation (do it as an exercise!), and this way the general expression is proved by induction.

Exercise: find a way to relate the two probability models that lead to the Poisson distribution.

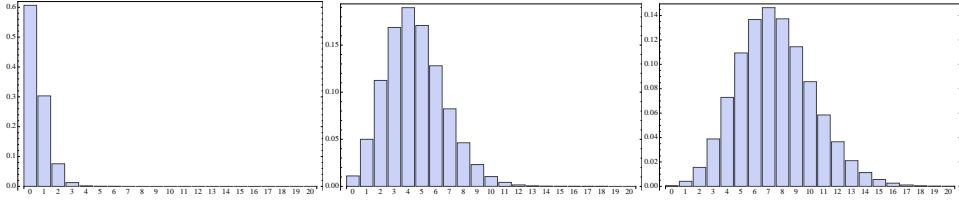


Figure 4.4: Poisson distribution with  $a = 0.5$  (left panel),  $a = 4.5$  (center panel), and  $a = 7.5$  (right panel). The distribution with low average rate is strongly asymmetrical, and evolves toward a more symmetrical shape as  $a$  grows.

- normalization:

$$\sum_{k=0}^{k=\infty} P_\xi(k) = \sum_{k=0}^{k=\infty} \frac{a^k}{k!} e^{-a} = 1 \quad (4.44)$$

- average:

$$\mathbf{E}\xi = \sum_{k=0}^{k=\infty} k P_\xi(k) = \sum_{k=0}^{k=\infty} k \frac{a^k}{k!} e^{-a} \quad (4.45a)$$

$$= ae^{-a} \sum_{k=1}^{k=\infty} \frac{a^{k-1}}{(k-1)!} = ae^{-a} \sum_{k=0}^{k=\infty} \frac{a^k}{(k)!} = a \quad (4.45b)$$

- mean square value:

$$\mathbf{E}(\xi)^2 = \sum_{k=0}^{k=\infty} k^2 P_\xi(k) = \sum_{k=0}^{k=\infty} k^2 \frac{a^k}{k!} e^{-a} = e^{-a} \sum_{k=0}^{k=\infty} [k(k-1) + k] \frac{a^k}{k!} \quad (4.46a)$$

$$= a^2 e^{-a} \sum_{k=2}^{k=\infty} \frac{a^{k-2}}{(k-2)!} + a = a^2 + a \quad (4.46b)$$

- variance:

$$\mathbf{D}\xi = \mathbf{E}(\xi)^2 - (\mathbf{E}\xi)^2 = a^2 + a - a^2 = a \quad (4.47)$$

i.e., the variance of the Poisson distribution is equal to its mean value.

- moment generating function:

$$M(t) = \sum_n \frac{e^{tn} a^n}{n!} e^{-a} = e^{a(e^{-1})} \quad (4.48)$$

### 4.5.1 The lottery ticket problem

Consider a lottery that has  $M$  winning tickets out of  $N$ , so that the probability that any given ticket is winning is  $p = M/N$  (this probability is usually very small!). Draws can be considered independent if we assume that the number of draws is much less than the number of tickets, and the probability model is Poissonian. Then, if one buys  $n$  tickets, the average number of wins is  $nM/N$ , and the probability of no win is  $P(0) = e^{-nM/N}$ , so that the probability of having at least one winning ticket is  $1 - P(0) = 1 - e^{-nM/N}$ . Take, e.g.,  $N = 10^7$ , and  $M = 100$ , so that  $M/N = 10^{-5}$ . Then if you want a 10% chance to win, you must have  $1 - P(0) = 0.1$ , i.e.,  $e^{-nM/N} = 0.9 \Rightarrow n = -(N/M) \ln 0.9 \approx 10^4$ .

### 4.5.2 Cell plating statistics

What is the probability that one culture well contains at least one cell if you prepare the mixture of cells and culture medium putting  $n$  cells, and use it to prepare  $N$  culture wells (see figure 4.5)? Obviously the average number of cells per well is just  $n/N$ , and therefore the probability of finding none is  $e^{-n/N}$ , so that the probability of at least one cell in a well is  $1 - e^{-n/N}$ .



Figure 4.5: Biological culture wells.

## 4.6 The exponential distribution

Now we consider the second probability model that leads to the Poisson distribution, and we concentrate on the waiting time between two successive events. We have found the following equation for the probability of no event in a given time interval

$$P_\xi(0, t + dt) = P_\xi(0, t)(1 - \lambda dt) \quad (4.49)$$

Similarly the probability of no event in  $(0, t)$  and of 1 event in  $(t, t + dt)$  is

$$P(t \leq \text{waiting time} \leq t + dt) = dP(t) = P_\xi(0, t)\lambda dt = \lambda e^{-\lambda t} dt \quad (4.50)$$

therefore the pdf is

$$p(t) = \lambda e^{-\lambda t} = \frac{1}{\tau} e^{-t/\tau} \quad (4.51)$$

where  $\tau = 1/\lambda$ .

- normalization:

$$\int_0^\infty p(t)dt = \int_0^\infty \frac{1}{\tau} e^{-t/\tau} dt = 1 \quad (4.52)$$

- average:

$$\mathbf{E}t = \int_0^\infty tp(t)dt = \int_0^\infty \frac{t}{\tau} e^{-t/\tau} dt = \tau \quad (4.53)$$

- mean square value:

$$\mathbf{E}(t^2) = \int_0^\infty t^2 p(t)dt = \int_0^\infty \frac{t^2}{\tau} e^{-t/\tau} dt = 2\tau^2 \quad (4.54)$$

- variance:

$$\mathbf{Dt} = \mathbf{E}(t^2) - (\mathbf{Et})^2 = 2\tau^2 - \tau^2 = \tau^2 \quad (4.55)$$

- moment generating function:

$$M(s) = \int_0^\infty e^{st} \frac{1}{\tau} e^{-t/\tau} dt = \frac{1}{st - 1} \quad (4.56)$$

#### 4.6.1 Detector dead time

Consider pulse counting detectors: there are two main types of devices, those that stop functioning for a time  $\tau_D$  after a successful detection, and then resume normal operation (*nonparalyzable detectors*), and those that can resume functioning only if there are no pulses for at least a time  $\tau_D$  (i.e., they keep being paralyzed by pulses, *paralyzable detectors*). Thus, the *dead time*  $\tau_D$  is the time required to process a single event; see figure 4.6 for an illustration of paralyzable and nonparalyzable detectors.

Now we let  $n$  be the mean rate of events, we assume that the event flow is a Poisson process, and finally we let  $m$  be the detection rate in a nonparalyzable detector: then  $n\tau_D$  is the mean number of events in a dead time, and the mean rate of lost events is  $mn\tau_D$ , i.e.,

$$\text{lost event rate} = n - m = mn\tau_D \quad (4.57)$$

This shows that we can use the measured rate  $m$  to find the true rate  $n$ :

$$n = \frac{m}{1 - m\tau_D} \quad (4.58)$$

Notice that we can also find the measured rate as a function of the true rate:

$$m = \frac{n}{1 + n\tau_D} \quad (4.59)$$

and from this formula we see that the measured rate reaches the saturation value  $m_0 = 1/\tau_D$  if the pulse rate  $n \rightarrow \infty$ .

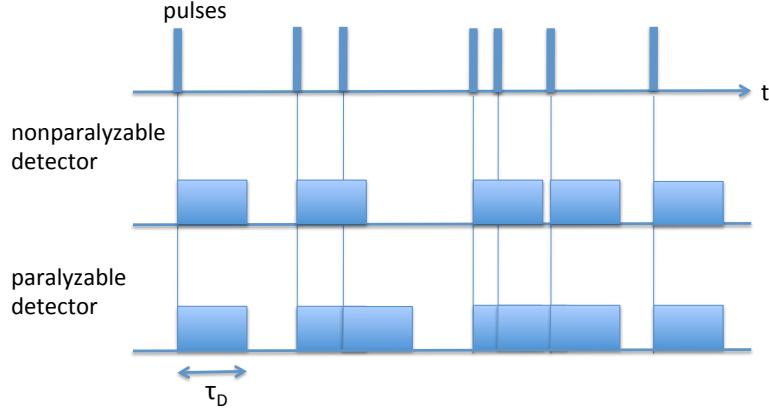


Figure 4.6: Schematic action of nonparalyzable and paralyzable detectors. It takes a time  $\tau_D$  to process one pulse in both detectors. However non-paralyzable detectors are insensitive to pulses occurring during a dead time, while paralyzable detectors require a time  $\tau_D$  to regain responsiveness, even after pulses occurring during a dead time. In this example, there are 7 pulses, the nonparalyzable detector detects 5 pulses, while the paralyzable detector detects only 4 pulses.

A paralyzable detector cannot resume normal operation, unless a time interval longer than  $\tau_D$  occurs, and this happens with probability

$$\int_{\tau_D}^{\infty} ne^{-nt} dt = e^{-n\tau_D} \quad (4.60)$$

This means that the measured rate is

$$m = ne^{-n\tau_D} \quad (4.61)$$

In this case  $m$  vanishes for a very high pulse rate, and the measured rate reaches the peak value

$$m = \frac{e^{-1}}{\tau_D} \quad (4.62)$$

for  $n = 1/\tau_D$ .

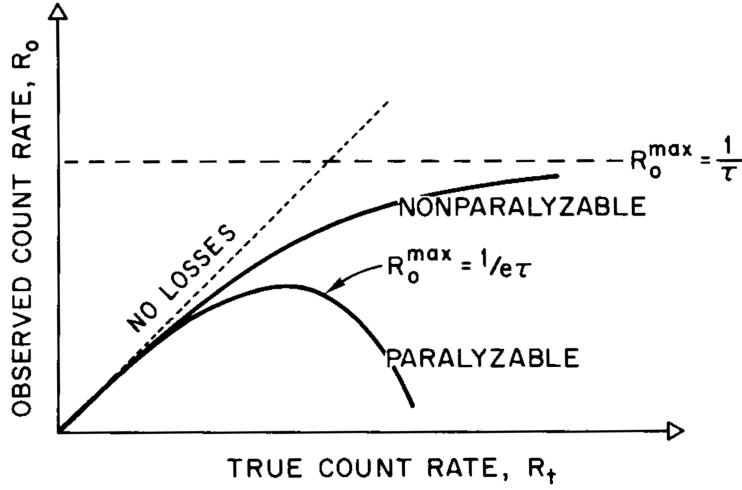


Figure 4.7: Plot of measured rate vs. true rate for paralyzable and nonparalyzable detectors.

## 4.7 The De Moivre-Laplace theorem. The normal distribution

Now notice that binomial coefficients can be approximated as follows for large  $n$ , using the Stirling approximation:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{\sqrt{2\pi n}}{\sqrt{2\pi k}} \frac{n^n e^{-n}}{\sqrt{2\pi(n-k)}(n-k)^{n-k} e^{-n+k}} \quad (4.63a)$$

$$= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \frac{n^n}{k^k (n-k)^{n-k}} \quad (4.63b)$$

therefore the binomial probability is approximately equal to

$$P(k) = \binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} \quad (4.64)$$

Now define a new random variable,

$$x = \frac{k - \mathbf{E}k}{\sqrt{\mathbf{D}k}} = \frac{k - np}{\sqrt{npq}} \quad (4.65)$$

so that

$$k = np + x\sqrt{npq} \quad (4.66a)$$

$$n - k = nq - x\sqrt{npq} \quad (4.66b)$$

The smallest change of  $k$  is  $\Delta k = 1$ , therefore the smallest change of  $x$  is  $\Delta x = 1/\sqrt{npq}$ .

Then, as  $n \rightarrow \infty$  and  $x$  is held fixed

$$\sqrt{\frac{n}{k(n-k)}} = \sqrt{\frac{n}{n^2pq + o(n^2)}} \rightarrow \frac{1}{\sqrt{npq}} = \Delta x \quad (4.67)$$

(the “small  $o$ ” notation means that  $n^2$  is greater than the neglected terms for  $n \gg 1$ ). Similarly

$$\ln \left( \frac{np}{k} \right)^k = - (np + x\sqrt{npq}) \ln \left( \frac{np + x\sqrt{npq}}{np} \right) \quad (4.68a)$$

$$= - (np + x\sqrt{npq}) \ln \left( 1 + x\sqrt{\frac{q}{np}} \right) \quad (4.68b)$$

$$\approx - (np + x\sqrt{npq}) \left( x\sqrt{\frac{q}{np}} - \frac{1}{2}x^2 \frac{q}{np} \right) \quad (4.68c)$$

$$\approx -x\sqrt{npq} - \frac{q}{2}x^2 \quad (4.68d)$$

and

$$\ln \left( \frac{nq}{n-k} \right)^{n-k} = - (nq - x\sqrt{npq}) \ln \left( \frac{nq - x\sqrt{npq}}{nq} \right) \quad (4.69a)$$

$$= - (nq - x\sqrt{npq}) \ln \left( 1 - x\sqrt{\frac{p}{nq}} \right) \quad (4.69b)$$

$$\approx - (nq - x\sqrt{npq}) \left( -x\sqrt{\frac{p}{nq}} - \frac{1}{2}x^2 \frac{p}{nq} \right) \quad (4.69c)$$

$$\approx x\sqrt{npq} - \frac{p}{2}x^2 \quad (4.69d)$$

so that Then, as  $n \rightarrow \infty$  and  $x$  is held fixed

$$\ln \left[ \left( \frac{np}{k} \right)^k \left( \frac{nq}{n-k} \right)^{n-k} \right] \approx -\frac{x^2}{2} \quad (4.70)$$

and Then, as  $n \rightarrow \infty$  and  $x$  is held fixed

$$\left( \frac{np}{k} \right)^k \left( \frac{nq}{n-k} \right)^{n-k} \approx e^{-x^2/2} \quad (4.71)$$

Finally, as  $n \rightarrow \infty$ , the probability distribution is Then, as  $n \rightarrow \infty$  and  $x$  is held fixed

$$P_\xi(x \leq \xi < x + \Delta x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Delta x \quad (4.72)$$

and the corresponding pdf is Then, as  $n \rightarrow \infty$  and  $x$  is held fixed

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (4.73)$$

This is the *De Moivre-Laplace theorem*, and the pdf is called *normal* or *Gaussian* probability density. (see [27] for a more careful treatment of the limiting approximations). The more general Gaussian distribution with mean  $\mu$  and variance  $\sigma^2$ . Then, as  $n \rightarrow \infty$  and  $x$  is held fixed

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad (4.74)$$

is often denoted with  $N(\mu, \sigma)$ .

The normalization, mathematical expectation and variance are easily found<sup>1</sup>:

- normalization:

$$\int_{-\infty}^{+\infty} p(x) dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx = 1 \quad (4.75)$$

- average:

$$\mathbf{E}\xi = \int_{-\infty}^{+\infty} xp(x) dx = \mu \quad (4.76)$$

(from an obvious symmetry argument).

- variance:

$$\mathbf{D}\xi = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (x - \mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx \quad (4.77a)$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y^2 e^{-y^2/2} dy \quad (4.77b)$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y d(-e^{-y^2/2}) \quad (4.77c)$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-y^2/2} dy = \sigma^2 \quad (4.77d)$$

<sup>1</sup>Here we use the well known integral

$$I = \int_{-\infty}^{+\infty} e^{-x^2/2} dx = \sqrt{2\pi}$$

This result can easily be found as follows: first we square the original integral

$$I^2 = \int_{-\infty}^{+\infty} e^{-x^2/2} dx \int_{-\infty}^{+\infty} e^{-y^2/2} dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)/2} dx dy$$

then we transform the double integral to use polar coordinate, so that  $x^2 + y^2 = r^2$ ,  $dx dy = rdr d\theta$

$$I^2 = \int_0^{2\pi} d\theta \int_0^{\infty} r e^{-r^2/2} dr = 2\pi \int_0^{\infty} d(-e^{-r^2/2}) = 2\pi$$

and finally  $I = \sqrt{2\pi}$ .

(this is obtained by setting  $y = (x - \mu)/\sigma$  and then integrating by parts).

- moment generating function:

$$M(t) = e^{\mu^2/2\sigma^2} \exp \left[ -\frac{(t + \mu/\sigma^2)^2}{2/\sigma^2} \right] \quad (4.78)$$

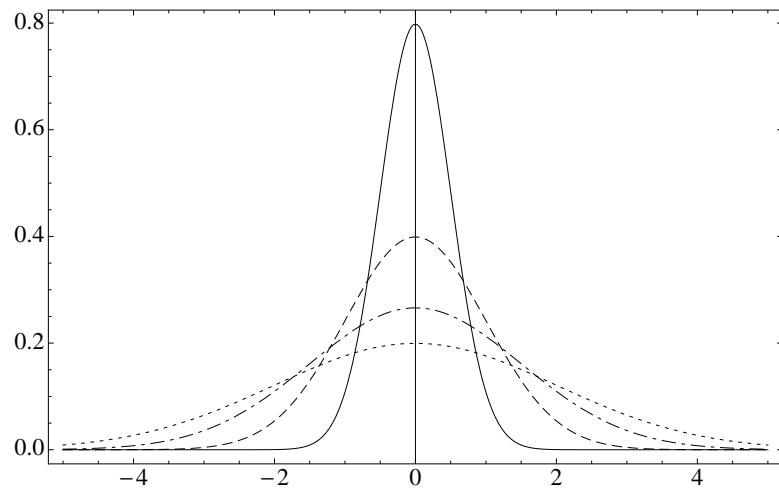


Figure 4.8: Plot of normal distributions with zero mean and standard deviation ranging from 0.5 to 2 (solid line:  $\sigma = 0.5$ ; dashed line:  $\sigma = 1$ ; dashed-dotted line:  $\sigma = 1.5$ ; dotted line:  $\sigma = 2$ )

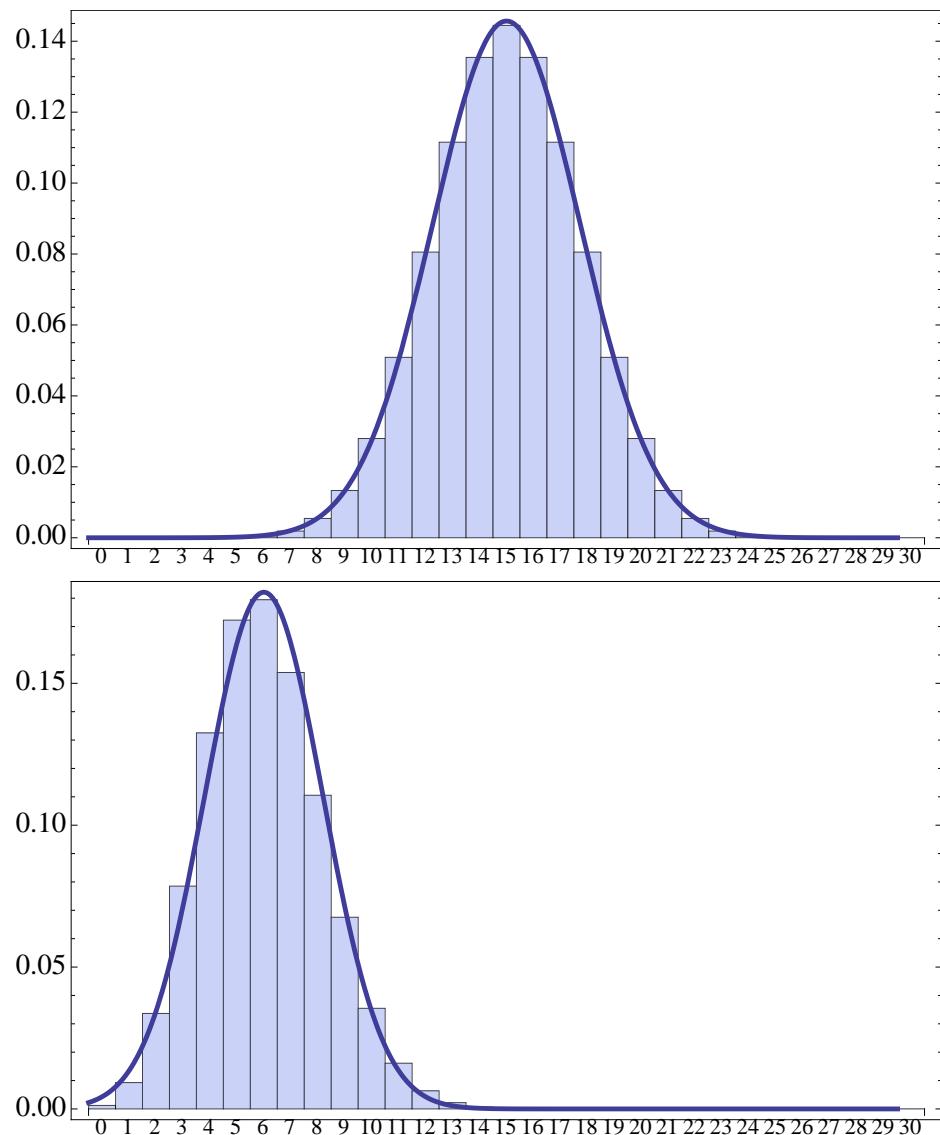


Figure 4.9: Comparison between binomial distributions with  $n = 30$  and  $p = 0.5$  (upper panel), and  $n = 30$  and  $p = 0.2$  (lower panel), and the corresponding normal pdf's with the same mean and standard deviation.

### 4.7.1 The error function

The Gaussian distribution function is

$$\Phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-t^2/2\sigma^2} dt \quad (4.79)$$

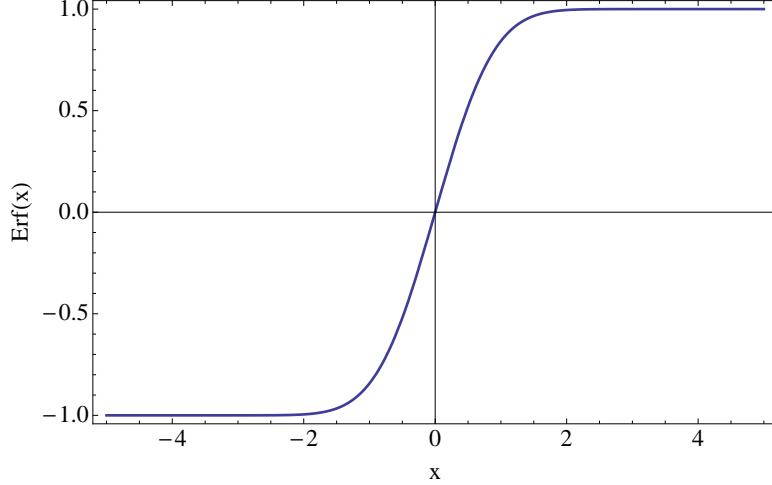


Figure 4.10: Plot of the error function  $\text{Erf}(x)$  vs.  $x$ .

Given the importance of the Gaussian distribution and of the Gaussian distribution function, and since the integral cannot be expressed as an elementary function, a new special function is introduced, the so called *error function*

$$\text{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (4.80)$$

(see figure 4.10).

It is easy to see that the relationship between the Gaussian distribution function and the error function is

$$\begin{aligned} \Phi(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-t^2/2\sigma^2} dt = \frac{1}{2} \left( 1 + \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2/2\sigma^2} \frac{dt}{\sqrt{2\sigma^2}} \right) \\ &= \frac{1}{2} \left[ 1 + \text{Erf} \left( \frac{x}{\sqrt{2\sigma^2}} \right) \right] \end{aligned} \quad (4.81)$$

### 4.7.2 High-rate limit of the Poisson distribution

Now we consider a Poisson distribution with large rate  $np$ , and recall that

$$P_\xi(k) = \frac{(np)^k}{k!} e^{-np} \quad (4.82)$$

We use the lowest order Stirling's approximation and we find

$$\ln P_\xi(k) = k \ln(np) - k \ln k + k - np \quad (4.83)$$

Next, we treat  $k$  as a continuous variable, and find the position of the maximum (modal value) of the distribution:

$$\frac{d \ln P_\xi}{dk} = \ln(np) - \ln k = 0 \Rightarrow k = np \quad (4.84)$$

The second derivative at the maximum is

$$\left. \frac{d^2 \ln P_\xi}{dk^2} \right|_{k=np} = -\frac{1}{k} \Big|_{k=np} = -\frac{1}{np} \quad (4.85)$$

therefore

$$\ln P_\xi(k) \approx \ln P_\xi(k_{max}) - \frac{(k - np)^2}{2np} \Rightarrow P_\xi(k) \approx P_\xi(k_{max}) e^{-(k-np)^2/2np} \quad (4.86)$$

Since

$$P_\xi(k_{max}) \approx \frac{(np)^{np}}{\sqrt{2\pi np} (np)^{np} e^{-np}} = \frac{1}{\sqrt{2\pi np}} \quad (4.87)$$

eventually we find

$$P_\xi(k) \approx \frac{1}{\sqrt{2\pi np}} e^{-(k-np)^2/2np} \quad (4.88)$$

i.e., a high-rate Poisson distribution is close to a Gaussian with average and variance both equal to  $np$ .

## 4.8 Transformations of random variables

### 4.8.1 Sum of two random variables

This is a simple but important example. The random variable  $\eta = \xi_1 + \xi_2$  is the sum of two independently distributed random variables with pdf's  $p_{\xi_1}(x)$ ,  $p_{\xi_2}(x)$ . The joint probability density of the two variables is the product  $p_{\xi_1}(x_1)p_{\xi_2}(x_2)$ , therefore the probability distribution of  $\eta$  is

$$P(y' \leq \eta \leq y'') = \int \int_{y' \leq x_1 + x_2 \leq y''} p_{\xi_1}(x_1)p_{\xi_2}(x_2) dx_1 dx_2 \quad (4.89)$$

The double integral can be rearranged taking the new variables  $u = (x_1 + x_2)$ , and  $v = x_1$ , so that the Jacobian determinant of the transformation equals 1, and the transformed integral becomes

$$P(y' \leq \eta \leq y'') = \int_{y' \leq x_1 + x_2 \leq y''} du \int_{-\infty}^{+\infty} p_{\xi_1}(v)p_{\xi_2}(u-v) dv \quad (4.90)$$

This means that the pdf of  $\eta$  is the convolution of the original pdf's

$$p_\eta(u) = \int_{-\infty}^{+\infty} p_{\xi_1}(v)p_{\xi_2}(u-v)dv \quad (4.91)$$

In this context it is common to consider the sum of two uniform variates with the same pdf

$$p(x) = \begin{cases} 1 & \text{if } x \in (0, 1) \\ 0 & \text{if } x \notin (0, 1) \end{cases} \quad (4.92)$$

Then the pdf of the sum is nonzero in the  $(0, 2)$  interval, and for a given value  $u$ , the product  $p(v)p(u-v)$  in the integral is either equal to 1 (when  $0 \leq u \leq 1$  and  $v \leq u$  or  $1 \leq u \leq 2$  and  $v \geq u-1$ ) or 0 (elsewhere), thus

$$p_\eta(u) = \begin{cases} \int_0^u dv = u & \text{if } 0 \leq u \leq 1 \\ \int_{u-1}^1 dv = 2-u & \text{if } 1 \leq u \leq 2 \\ 0 & \text{if } u \notin (0, 2) \end{cases} \quad (4.93)$$

which is a triangular pdf.

#### 4.8.2 General considerations on continuous random variable transformations

Now consider a transformation of random variables  $\xi_1, \dots, \xi_n$  into a new set  $\eta_1, \dots, \eta_n$ , then the probability integral transforms as follows

$$P(\xi \in A) = \int_A p_\xi(x_1, \dots, x_n) dx_1 \dots dx_n \quad (4.94a)$$

$$= \int_B p_\xi[x_1(u_1, \dots, u_n), \dots, x_n(u_1, \dots, u_n)] \left| \det \frac{\partial x_i}{\partial u_j} \right| du_1 \dots du_n \quad (4.94b)$$

$$= \int_B p_\eta(u_1, \dots, u_n) du_1 \dots du_n \quad (4.94c)$$

where B is the image of the A under the coordinate transformation. Thus

$$p_\eta(u_1, \dots, u_n) = \sum p_\xi[x_1(u_1, \dots, u_n), \dots, x_n(u_1, \dots, u_n)] \left| \det \frac{\partial x_i}{\partial u_j} \right| \quad (4.95)$$

where the last sum is over all the disconnected volumes in  $\xi$ -space that are mapped into the same volume in  $\eta$ -space by the coordinate transformation.

#### A 1D example with disconnected regions

Consider a normal variate  $x$  with zero mean and variance  $\sigma^2$ , and the variable transformation  $y = x^2$ , then

$$\left| \frac{dx}{dy} \right| = \frac{1}{2\sqrt{y}} \quad (4.96)$$

and therefore

$$\begin{aligned} p_y(y) &= \sum_{y|y=x^2} p_x(x(y)) \left| \frac{dx}{dy} \right| = 2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y}{2\sigma^2}\right) \frac{1}{2\sqrt{y}} \\ &= \frac{y^{-1/2}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y}{2\sigma^2}\right) \end{aligned} \quad (4.97)$$

Here the sum is carried out over the two regions, for positive and negative  $x$  that map into the same range of  $y$  (see figure 4.11).

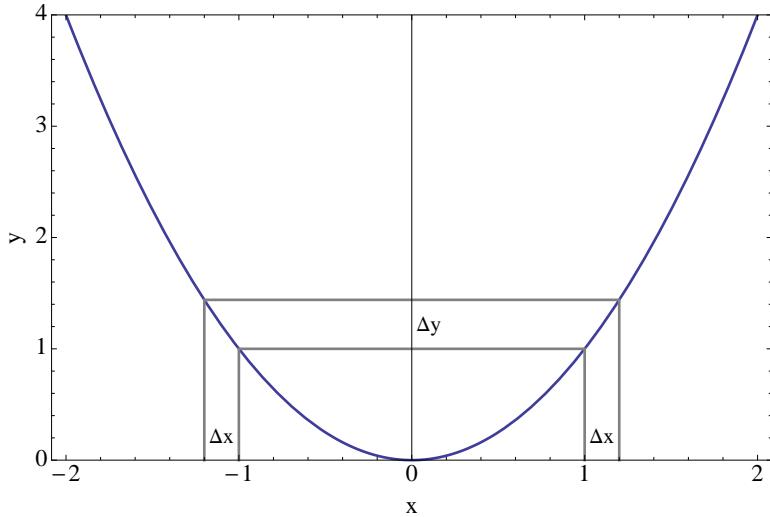


Figure 4.11: Random variable transformation with two disconnected regions: both the negative and the positive ranges of  $x$  map into the same (positive) range of  $y$ .

#### 4.8.3 Probability density function of a generic function of several random variables

Consider a function of random variables  $y = y_1 = f(x_1, \dots, x_n)$ , and define the auxiliary random variables  $y_k = x_k$ ,  $k = 2, \dots, n$ . Then the first column of the Jacobian matrix has the elements

$$\frac{\partial x_k}{\partial y_1} = \left( \frac{\partial y_1}{\partial x_k} \right)^{-1} = \left( \frac{\partial f}{\partial x_k} \right)^{-1} \quad (4.98)$$

while all the remaining matrix elements are zero, except the diagonal elements that are equal to 1. Therefore the Jacobian determinant is

$$|J| = \left| \frac{\partial f}{\partial x_1} \right|^{-1} \quad (4.99)$$

and the pdf of  $y$  is

$$p_y(y) = \int p_{y_1, \dots, y_n}(y, y_2, \dots, y_n) dy_2 \dots dy_n \quad (4.100a)$$

$$= \int p_{x_1, \dots, x_n} [x_1(y, y_2, \dots, y_n), y_2, \dots, y_n] \left| \frac{\partial f}{\partial x_1} \right|^{-1} dy_2 \dots dy_n \quad (4.100b)$$

The dummy variates  $y_1, \dots, y_n$  are usually called *nuisance variables*, and we get rid of them by integrating them out.

### Product of two independent random variables

As an example, consider the independent, non-negative random variables  $x_1$  and  $x_2$  with pdf  $p_1(x_1)$  and  $p_2(x_2)$ , and determine the pdf of their product  $x_1 x_2$ .

Take the new random variables  $y = y_1 = x_1 x_2$ ,  $y_2 = x_2$ , so that  $x_2 = y_2$ ,  $x_1 = y_1/y_2$  and the Jacobian determinant is  $J = 1/|y_2|$ . Then, using equation (4.100b), we find

$$p_y(y) = \int p_1 \left( \frac{y_1}{y_2} \right) p_2(y_2) \frac{1}{|y_2|} dy_2 = \quad (4.101)$$

Note that if we let  $u = \ln y_1 = \ln y$ ,  $v = \ln y_2$ , then

$$p_1 \left( \frac{y_1}{y_2} \right) p_2(y_2) = p_1[\exp(u - v)] p_2[\exp(v)] = f_1(u - v) f_2(v)$$

and therefore

$$p_y(y) = \int f_1(u - v) f_2(v) dv \quad (4.102)$$

which has the structure of a convolution. Thus, the original equation (4.101) is a sort of convolution between logarithmic variables, and this special convolution is called the *Mellin convolution* of  $p_1$  and  $p_2$ .

#### 4.8.4 Error propagation

Consider the case where we have  $n$  random variables  $x_1, \dots, x_n$ , which have some joint pdf  $p(\mathbf{x})$ , and assume that the pdf is not completely known and that we know only the averages  $\mu_i$  and the covariance matrix  $V_{ij} = \text{cov}(x_i, x_j)$ . Then if we take a function  $y(\mathbf{x})$  we cannot use the transformation formulas described in the previous sections, however we can still determine the average and the variance of  $y$ . Indeed we can expand  $y$  about the average of the random variables  $x_i$ :

$$y(\mathbf{x}) \approx y(\boldsymbol{\mu}) + \sum_{i=1}^{i=n} \frac{\partial y}{\partial x_i} \Big|_{\mathbf{x}=\boldsymbol{\mu}} (x_i - \mu_i) \quad (4.103)$$

Then:

- average:

$$\mathbf{E}[y(\mathbf{x})] \approx y(\boldsymbol{\mu}) \quad (4.104)$$

- variance:

$$\begin{aligned} \mathbf{D}[y(\mathbf{x})] &= \mathbf{E}\{[y(\mathbf{x}) - y(\boldsymbol{\mu})]^2\} = \mathbf{E} \sum_{i,j=1}^n \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} (x_i - \mu_i)(x_j - \mu_j) \\ &\quad (4.105a) \end{aligned}$$

$$= \sum_{i,j=1}^n V_{i,j} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} \quad (4.105b)$$

- case of uncorrelated  $x_i$ 's ( $V_{i,j} = 0$  if  $i \neq j$ ):

$$\mathbf{D}[y(\mathbf{x})] = \sum_{i=1}^n V_{i,i} \left[ \frac{\partial y}{\partial x_i} \right]_{\mathbf{x}=\boldsymbol{\mu}}^2 \quad (4.106)$$

- covariance: similarly, if there are more than one dependent variables:

$$U_{l,m} = \text{cov}(y_l, y_m) = \sum_{i,j=1}^n V_{i,j} \left[ \frac{\partial y_l}{\partial x_i} \frac{\partial y_m}{\partial x_j} \right]_{\mathbf{x}=\boldsymbol{\mu}} \quad (4.107)$$

- matrix form: the previous formula can be recast in matrix form if we let

$$A_{i,j} = \left. \frac{\partial y_i}{\partial x_j} \right|_{\mathbf{x}=\boldsymbol{\mu}} \quad (4.108)$$

so that

$$U_{l,m} = \sum_{i,j=1}^n A_{l,i} V_{i,j} A_{m,j} \quad (4.109)$$

and

$$U = AVA^T \quad (4.110)$$

- important notice: the approximation breaks down for highly nonlinear  $y$ 's !

#### 4.8.5 Orthogonal transformations of random variables

##### General linear transformations

The procedure of the previous section is approximate and we use it whenever it is difficult to find the actual probability distribution of the transformed random variables  $y$ 's. However there is an important special case, the linear

transformation of random variables, where we easily obtain exact results. In this case

$$y_i = \sum_j A_{i,j} x_j \quad (4.111)$$

or also, in matrix form,

$$\mathbf{y} = A\mathbf{x} \quad (4.112)$$

Clearly, in this case

$$\frac{\partial y_i}{\partial x_j} = A_{i,j} \quad (4.113)$$

and the formulas of the previous section are exact, rather than approximate. If  $V$  is the covariance matrix of the  $x$ 's, as above, the covariance matrix of the new set of random variables  $y$  is

$$U = AVA^T \quad (4.114)$$

### Orthogonal transformations

When we consider orthogonal transformations we can obtain a further simplification, i.e., we can transform our original variates into a set of independent variates. Statistical independence leads to considerable simplifications, and it is a highly desirable property, and here we discuss how to achieve it with an orthogonal linear transformation. Notice first that the covariance matrix

$$V_{i,j} = \text{cov}(x_i, x_j) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} (x_i - \mu_i)(x_j - \mu_j) p(\mathbf{x}) dx_1 \dots dx_n \quad (4.115)$$

is symmetric and non-singular, and therefore it can be diagonalized as follows

- first solve the eigenvalue problem

$$V\mathbf{e}^{(i)} = \lambda^{(i)}\mathbf{e}^{(i)} \quad (4.116)$$

- note that if the eigenvalues are not degenerate, the eigenvectors  $\mathbf{e}^{(i)}$  are mutually orthogonal; start from the following equation

$$(V\mathbf{e}^{(i)})^T (V\mathbf{e}^{(j)}) = (\mathbf{e}^{(i)})^T V^T V \mathbf{e}^{(j)} = (\mathbf{e}^{(i)})^T V^2 \mathbf{e}^{(j)} \quad (4.117a)$$

$$= (\lambda^{(j)})^2 \mathbf{e}^{(i)} \cdot \mathbf{e}^{(j)} \quad (4.117b)$$

$$= \lambda^{(i)} \lambda^{(j)} \mathbf{e}^{(i)} \cdot \mathbf{e}^{(j)} \quad (4.117c)$$

then

$$\lambda^{(j)} (\lambda^{(i)} - \lambda^{(j)}) \mathbf{e}^{(i)} \cdot \mathbf{e}^{(j)} = 0 \quad (4.118)$$

therefore  $\mathbf{e}^{(i)} \cdot \mathbf{e}^{(j)} = 0$  if  $i \neq j$ ;

- if some eigenvalues are degenerate, then the previous derivation does not hold, however in the subspace corresponding to the degenerate eigenvectors we can still choose orthogonal eigenvectors; thus at the end of the procedure we have a set of orthogonal directions, and we can still normalize the set of eigenvectors;
- the matrix  $A_{i,j} = e_j^{(i)}$  (the matrix whose rows are the eigenvectors of  $V$ ) diagonalizes  $V$ ; indeed

$$U_{i,j} = (AVA^T)_{i,j} = \sum_{k,l} A_{i,k} V_{k,l} A_{j,l} = \sum_{k,l} e_k^{(i)} V_{k,l} e_l^{(j)} \quad (4.119a)$$

$$= \sum_k \lambda^{(j)} e_k^{(i)} e_k^{(j)} = \lambda^{(j)} \delta_{i,j} \quad (4.119b)$$

- the matrix  $A$  defined above is orthogonal

$$(AA^T)_{i,j} = \sum_k A_{i,k} A_{j,k} = \sum_k e_k^{(i)} e_k^{(j)} = \mathbf{e}^{(i)} \cdot \mathbf{e}^{(j)} = \delta_{i,j} \quad (4.120)$$

We conclude that using the matrix  $A$  found in this procedure we can define a new set of statistically independent random variables, which are generally easier to deal with than the original variates.

## 4.9 Multivariate Gaussian distribution

The multivariate Gaussian distribution nicely illustrates the power of the concepts introduced in the previous section. A multivariate Gaussian distribution of a set of zero-mean random variables is defined by the probability density

$$p(x_1, \dots, x_n) = N \exp\left(-\frac{1}{2} \mathbf{x}^T V^{-1} \mathbf{x}\right) \quad (4.121)$$

where  $N$  is the normalization constant, and  $\mathbf{x}^T V^{-1} \mathbf{x}$  is a symmetric quadratic form which generalizes the usual  $x^2/2\sigma^2$  exponent found in the 1D Gaussian distribution (and therefore the matrix  $V$  is akin to variance).

- normalization: we introduce first an orthogonal linear transformation of the random variables  $\mathbf{x}$ :

$$\mathbf{y} = A\mathbf{x} \quad (4.122)$$

that diagonalizes the  $V^{-1}$  matrix:

$$AV^{-1}A^T = \text{diag}(\lambda^{(1)}, \dots, \lambda^{(n)}) \quad (4.123)$$

so that

$$\mathbf{x}^T V^{-1} \mathbf{x} = \mathbf{x}^T A^T A V^{-1} A^T A \mathbf{x} = \mathbf{y}^T A V^{-1} A^T \mathbf{y} = \sum_{k=1}^n \lambda^{(k)} y_k^2 \quad (4.124)$$

The Jacobian determinant is

$$|J| = \left| \det \left( \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right| = \left| \det (A^{-1}) \right| = 1 \quad (4.125)$$

(the matrix  $A$  is orthogonal) therefore

$$1 = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p(x_1, \dots, x_n) dx_1 \dots dx_n \quad (4.126a)$$

$$= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p(y_1, \dots, y_n) |J| dy_1 \dots dy_n \quad (4.126b)$$

$$= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} N \prod_{k=1}^n e^{-\lambda^{(k)} y_k^2 / 2} dy_1 \dots dy_n \quad (4.126c)$$

$$= N(2\pi)^{n/2} \prod_{k=1}^n \left| \frac{1}{\lambda^{(k)}} \right|^{1/2} \quad (4.126d)$$

$$= N(2\pi)^{n/2} |\det V|^{1/2} \quad (4.126e)$$

and finally

$$N = \frac{1}{(2\pi)^{n/2} |\det V|^{1/2}} \quad (4.127)$$

and

$$p(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} |\det V|^{1/2}} \exp \left( -\frac{1}{2} \mathbf{x}^T V^{-1} \mathbf{x} \right) \quad (4.128)$$

- covariance matrix:

$$\text{cov}(x_i, x_j) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} x_i x_j N \exp \left( -\frac{1}{2} \mathbf{x}^T V^{-1} \mathbf{x} \right) dx_1 \dots dx_n \quad (4.129a)$$

$$= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \sum_{k,l} A_{i,k}^{-1} A_{j,l}^{-1} y_k y_l N \prod_n e^{-\lambda^{(n)} y_n^2 / 2} dy_1 \dots dy_n \quad (4.129b)$$

$$= \sum_{k,l} A_{i,k}^{-1} A_{j,l}^{-1} (\lambda^{(k)})^{-1} \delta_{k,l} \quad (4.129c)$$

$$= \sum_k A_{i,k}^{-1} (\lambda^{(k)})^{-1} A_{k,j} = V_{i,j} \quad (4.129d)$$

Thus the  $V$  matrix is just the symmetric covariance matrix, which is often written in the form

$$V = \begin{pmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,n}\sigma_1\sigma_n \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,n}\sigma_1\sigma_n & 0 & \cdots & \sigma_n^2 \end{pmatrix} \quad (4.130)$$

where  $\rho_{i,j}$  is the *correlation coefficient* between the  $i$ -th and the  $j$ -th variate. Moreover, when we remove the zero mean assumption, the more general form of the distribution is

$$p(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} |\det V|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T V^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \quad (4.131)$$

## 4.10 The log-normal distribution

This is the probability distribution of a variable whose logarithm is normally distributed, i.e.,

$$p_{\ln \xi}(\ln x) = \frac{1}{\sqrt{2\pi} \sigma^2} e^{-(\ln x - \mu)^2 / 2\sigma^2} \quad (4.132)$$

Since

$$p_{\ln \xi}(\ln x) d \ln x = p_{\ln \xi}(\ln x) \frac{dx}{x} = p_\xi(x) dx \quad (4.133)$$

we find

$$p_\xi(x) = \frac{p_{\ln \xi}(\ln x)}{x} = \frac{1}{x \sqrt{2\pi} \sigma^2} e^{-(\ln x - \mu)^2 / 2\sigma^2} \quad (4.134)$$

The log-normal distribution is characterized by an asymmetric shape and by a long tail (see figure 4.12). This distribution has important applications, that we shall discuss further on during this course.

- average:

$$\mathbf{E}\xi = \int_0^{+\infty} x p_\xi(x) dx = \int_{-\infty}^{+\infty} e^{\ln x} p_{\ln \xi}(\ln x) d \ln x \quad (4.135a)$$

$$= \int_{-\infty}^{+\infty} e^y \frac{1}{\sqrt{2\pi} \sigma^2} e^{-(y - \mu)^2 / 2\sigma^2} dy \quad (4.135b)$$

Now notice that

$$y - \frac{(y - \mu)^2}{2\sigma^2} = -\frac{y^2 - 2(\mu + \sigma^2)y + \mu^2}{2\sigma^2} = -\frac{[y - (\mu + \sigma^2)]^2 - (2\mu\sigma^2 + \sigma^4)}{2\sigma^2} \quad (4.136)$$

then:

$$\mathbf{E}\xi = e^{\mu + \sigma^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi} \sigma^2} e^{-[y - (\mu + \sigma^2)]^2 / 2\sigma^2} dy = e^{\mu + \sigma^2/2} \quad (4.137)$$

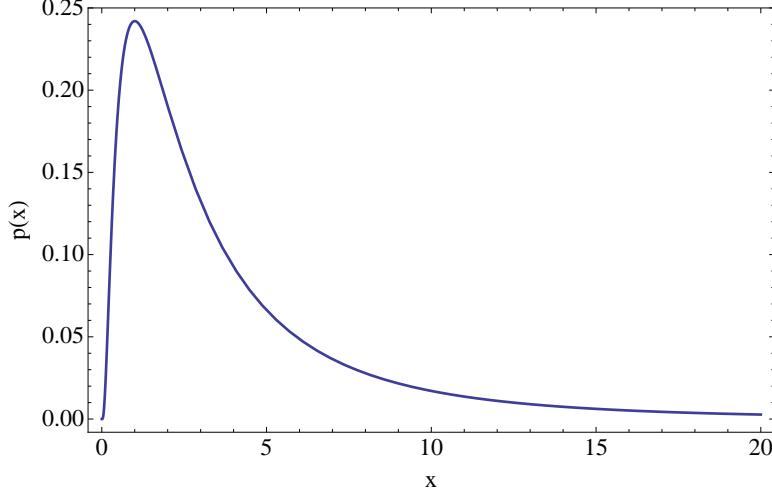


Figure 4.12: Plot of a lognormal distribution for  $\sigma = \mu = 1$ .

- mean square value:

$$\mathbf{E}(\xi^2) = \int_0^{+\infty} x^2 p_\xi(x) dx = \int_{-\infty}^{+\infty} e^{2 \ln x} p_{\ln \xi}(\ln x) d \ln x \quad (4.138a)$$

$$= \int_{-\infty}^{+\infty} e^{2y} \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(y-\mu)^2/2\sigma^2} dy \quad (4.138b)$$

Just as before, notice that

$$2y - \frac{(y-\mu)^2}{2\sigma^2} = -\frac{y^2 - 2(\mu + 2\sigma^2)y + \mu^2}{2\sigma^2} = -\frac{[y - (\mu + 2\sigma^2)]^2 - (4\mu\sigma^2 + 4\sigma^4)}{2\sigma^2} \quad (4.139)$$

then:

$$\mathbf{E}(\xi)^2 = e^{2\mu+2\sigma^2} \quad (4.140)$$

- variance:

$$\mathbf{D}\xi = \mathbf{E}(\xi)^2 - (\mathbf{E}\xi)^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \quad (4.141)$$

## 4.11 The gamma distribution

Now we consider a case where the probability density is proportional to  $x^{k-1}e^{-x/\vartheta}$ , for  $\theta, k > 0$  – i.e.,  $p_\xi(x) = Cx^{k-1}e^{-x/\vartheta}$  – in the range  $x \in (0, \infty)$ :

- normalization:

$$\int_0^\infty p_\xi(x) dx = \int_0^\infty Cx^{k-1}e^{-x/\vartheta} dx = C\theta^k \Gamma(k) = 1 \quad (4.142)$$

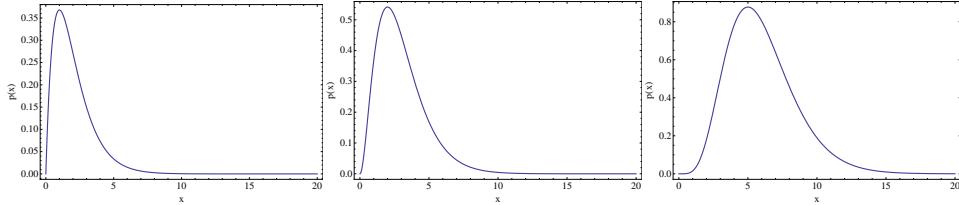


Figure 4.13: Gamma distribution for  $\theta = 1$  and  $k = 1$  (left panel),  $k = 2$  (center panel), and  $k = 5$  (right panel).

i.e.,  $C = \theta^{-k}/\Gamma(k)$ , and

$$p_\xi(x) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)} \quad (4.143)$$

- average:

$$\mathbf{E}\xi = \int_0^\infty x p_\xi(x) dx = \int_0^\infty \frac{x^k e^{-x/\theta}}{\theta^k \Gamma(k)} dx = \frac{\theta^{k+1} \Gamma(k+1)}{\theta^k \Gamma(k)} = \theta k \quad (4.144)$$

where we used the identity  $\Gamma(k+1) = k\Gamma(k)$ .

- mean square value:

$$\begin{aligned} \mathbf{E}(\xi^2) &= \int_0^\infty x^2 p_\xi(x) dx = \int_0^\infty \frac{x^{k+1} e^{-x/\theta}}{\theta^k \Gamma(k)} dx = \frac{\theta^{k+2} \Gamma(k+2)}{\theta^k \Gamma(k)} \\ &= \theta^2 k(k+1) \end{aligned} \quad (4.145)$$

- variance:

$$\mathbf{D}\xi = \mathbf{E}(\xi^2) - (\mathbf{E}\xi)^2 = \theta^2 k(k+1) - \theta^2 k^2 = \theta^2 k \quad (4.146)$$

## 4.12 The Cauchy-Lorentz (Breit-Wigner) distribution

The Cauchy-Lorentz distribution appears in dual space (frequency space) whenever there is some resonance phenomenon. In this case the probability density is proportional to  $1/(\Gamma^2/4 + (x - x_0)^2)$ , where  $\Gamma > 0$  (i.e.,  $p_\xi(x) = C/(\Gamma^2/4 + (x - x_0)^2)$ )

- normalization<sup>2</sup>:

$$\int_{-\infty}^{+\infty} p_\xi(x) dx = C \int_{-\infty}^{+\infty} \frac{1}{\Gamma^2/4 + (x - x_0)^2} dx = \frac{\pi}{\Gamma/2} C \quad (4.147)$$

---

<sup>2</sup>Here we use the indefinite integral

$$\int \frac{1}{a^2 + b^2 x^2} dx = \frac{1}{ab} \arctan \frac{bx}{a}$$

and therefore

$$p_\xi(x) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2} \quad (4.148)$$

- average:

$$\mathbf{E}\xi = \int_{-\infty}^{+\infty} xp_\xi(x)dx = \frac{1}{\pi} \int_{-\infty}^{+\infty} [(x - x_0) + x_0] \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2} dx \quad (4.149a)$$

$$= \frac{\Gamma/2}{\pi} \int_{-\infty}^{+\infty} \frac{1}{2} d \ln(\Gamma^2/4 + (x - x_0)^2) + x_0 \quad (4.149b)$$

$$= \frac{\Gamma/2}{2\pi} \ln(\Gamma^2/4 + (x - x_0)^2) \Big|_{-\infty}^{+\infty} + x_0 \quad (4.149c)$$

$$= \text{undefined} \quad (4.149d)$$

- variance:

$$\mathbf{D}\xi = \int_{-\infty}^{+\infty} (x - x_0)^2 p_\xi(x)dx = \text{undefined} \quad (4.150)$$

Both average and variance are undefined.

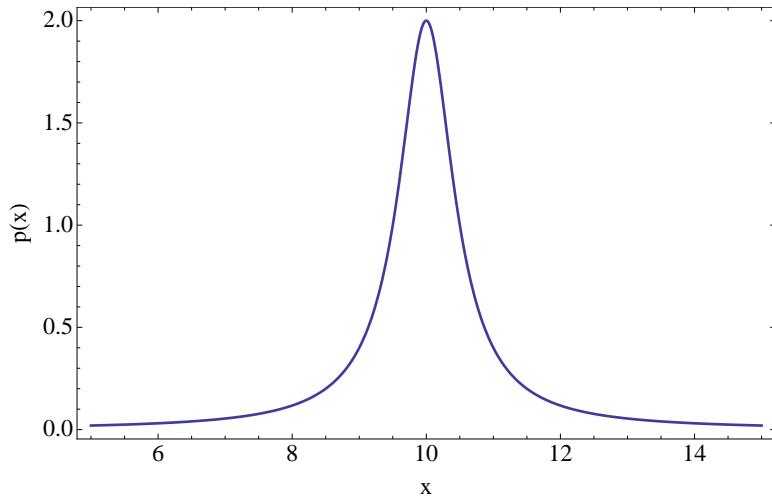


Figure 4.14: Plot of a Cauchy-Lorentz distribution for  $\Gamma = 1$ ,  $x_0 = 10$ .

---

so that

$$\int_{-\infty}^{+\infty} \frac{1}{a^2 + b^2 x^2} dx = \frac{\pi}{ab}$$

### 4.13 The Landau distribution

In 1944 Landau found the pdf of the energy loss distribution for a charged particle traversing a layer of matter, in the form of a Laplace Transform

$$p(\lambda) = \int_{c-i\infty}^{c+i\infty} e^{s \ln s + \lambda s} ds \quad (4.151)$$

where  $c$  is any real positive number,  $\lambda = R(E - E_p)$ ,  $E_p$  is the most probable energy loss, and  $R$  is a constant dependent on the absorbing material (this result was generalized by Vavilov in 1956). The Landau pdf can be approximated by the following expression, found by Moyal in 1955:

$$p(\lambda) \approx \frac{1}{\sqrt{2\pi}} e^{-(\lambda + e^{-\lambda})/2} \quad (4.152)$$

The Landau distribution is important in the analysis of detector response: here we do not discuss it further, and we refer to other, specific texts and courses (e.g., F. Sauli, *Principles of operation of multiwire proportional and drift chambers* [29]).

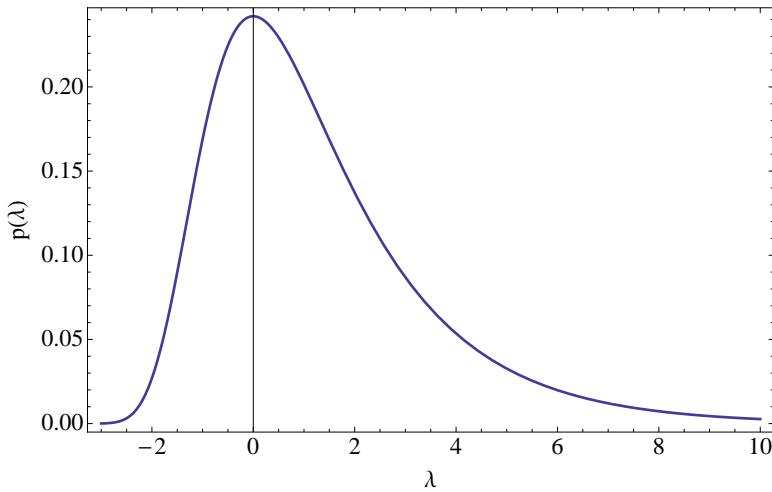


Figure 4.15: Plot of a Landau distribution: this distribution is characterized by an asymmetrical shape and by a long tail.

### 4.14 The Rayleigh distribution

When we consider the bivariate Gaussian distribution

$$p_{\xi_1, \xi_2}(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x_1^2 + x_2^2}{2\sigma^2}\right) \quad (4.153)$$

the phase angle of the vector  $(x_1, x_2)$  is often unimportant and we are only interested in the distribution of the vector length  $\xi = \sqrt{\xi_1^2 + \xi_2^2}$ ; this is usually the case in the analysis of power spectra. The pdf is easy to find

$$p_\xi(r) = 2\pi r \frac{1}{2\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) \quad (4.154)$$

and this defines the Rayleigh distribution.

- normalization:

$$\int_0^\infty \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr = \int_0^\infty x \exp(-x^2/2) dx = - \int_0^\infty d \exp(-x^2/2) = 1 \quad (4.155)$$

- average:

$$\mathbf{E}\xi = \int_0^\infty \frac{r^2}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr = \sigma \int_0^\infty x^2 \exp(-x^2/2) dx \quad (4.156a)$$

$$= \sigma\sqrt{2} \int_0^\infty y^{3/2-1} \exp(-y) dy = \sigma\sqrt{2} \Gamma(3/2) = (\sigma\sqrt{2}) \frac{\sqrt{\pi}}{2} = \frac{1}{2}\sqrt{2\pi\sigma^2} \quad (4.156b)$$

- mean square value:

$$\mathbf{E}\xi^2 = \int_0^\infty \frac{r^3}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr = \sigma^2 2 \int_0^\infty \frac{x^2}{2} \exp\left(-\frac{x^2}{2}\right) d\left(\frac{x^2}{2}\right) \quad (4.157a)$$

$$= 2\sigma^2 \int_0^\infty y \exp(-y) dy = 2\sigma^2 \quad (4.157b)$$

- variance:

$$\mathbf{D}\xi = \mathbf{E}(\xi^2) - (\mathbf{E}\xi)^2 = \frac{4 - \pi}{2}\sigma^2 \quad (4.158)$$

## 4.15 A short selection of other important probability distributions

The distributions in the previous sections are the most common, but they are not the only important ones. Here is a handful of other relevant probability distributions:

**Beta distribution:** this distribution is important in the Bayesian analysis of data that follow a binomial distribution:

$$p_\xi(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (4.159)$$

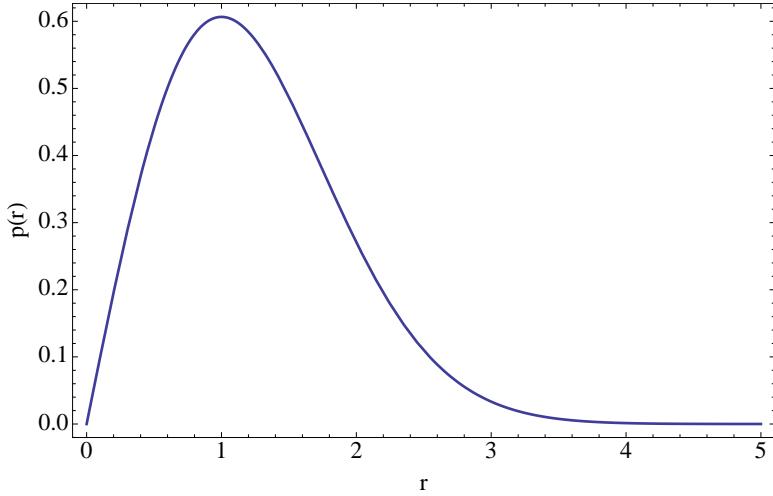


Figure 4.16: Plot of a Rayleigh distribution with  $\sigma = 1$ .

where the beta function is defined by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx \quad (4.160)$$

**Laplace distribution:** consider two i.i.d. exponential variates  $\xi_1, \xi_2$  with decay parameter  $\lambda$ , and their difference  $\xi = \xi_1 - \xi_2$ , so that  $\xi_2 = \xi_1 - \xi$ . Then, defining the new pair of variates  $(\xi_1, \xi)$ , we find that the Jacobian determinant is 1 (see also figure 4.17 for the layout of the integration region) and

$$p_\xi(z) = \lambda^2 \int_0^\infty \exp(-\lambda x) \exp[-\lambda(x + |z|)] dx = \frac{\lambda}{2} \exp(-\lambda|z|) \quad (4.161)$$

This is often written in the slightly different form

$$p_\xi(z) = \frac{1}{2a} \exp\left(-\frac{|z - \mu|}{a}\right) \quad (4.162)$$

which also includes the location parameter  $\mu$ .

**Logistic distribution:** the sigmoid function

$$P(x) = \frac{1}{1 + \exp(-x/s)} \quad (4.163)$$

which is often used in the theory of neural networks, is very similar to a distribution function. If we do assume that it is indeed a distribution function, then the corresponding pdf is

$$p(x) = \frac{\exp(-x/s)}{s[1 + \exp(-x/s)]^2} \quad (4.164)$$

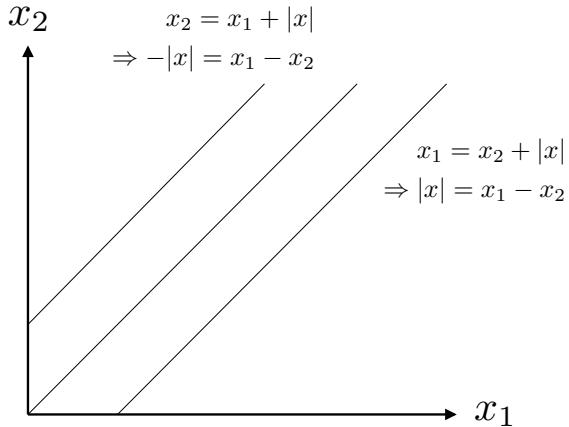


Figure 4.17: Layout of the integration regions to obtain the distribution of the difference of two exponential variates.

and if we also include a location parameter  $\mu$ , we obtain the full expression

$$p(x) = \frac{\exp[-(x - \mu)/s]}{s\{1 + \exp[-(x - \mu)/s]\}^2} \quad (4.165)$$

**Exercise:** Plot the pdf's of the beta, Laplace, and logistic distributions for different parameter values.

## 4.16 Bertrand's paradox, again

In this section we go back to Bertrand's paradox, in particular we consider the anticipated Jaynes' solution of the paradox [21], analyzing it with the tools developed in this chapter.

In his 1973 paper, Jaynes noted that there is one implicit assumption in the statement of Bertrand's paradox, namely that the distribution of chords is invariant with respect to rotations, translations and scale transformations. Figure 4.18 shows two examples with translated and scaled circles.

Since a chord is defined by the position of its center, we follow Jaynes and introduce the distribution  $f(r, \theta)$  of the chord centers inside the circle.

### 4.16.1 Rotational invariance

In a reference frame which is at an angle  $\alpha$  with respect to the original frame, i.e., the new angle  $\theta' = \theta - \alpha$ , the distribution of centers is given by a different distribution function  $g(r, \theta') = g(r, \theta - \alpha)$ . Since we require

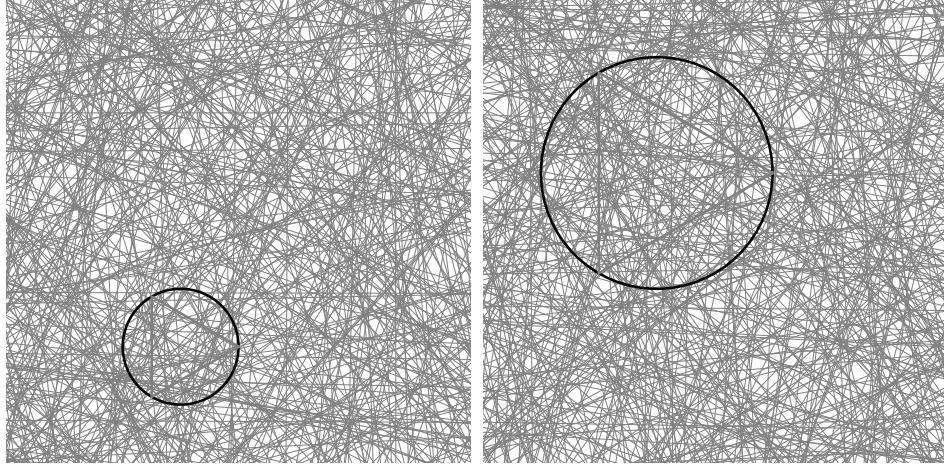


Figure 4.18: Bertrand's circle against a background with a random distribution of straight lines. According to Jaynes the solution of the paradox must be invariant with respect to rotations, translations and scale transformations. These example figures show translations and scale transformations.

rotational invariance

$$f(r, \theta) = g(r, \theta - \alpha) \quad (4.166)$$

with the condition  $g(r, \theta)|_{\alpha=0} = f(r, \theta)$ , and this must hold for every angle  $\alpha$ , so the only possibility is that there is no dependence on  $\theta$ , and  $f(r, \theta) = g(r, \theta) = f(r)$ .

#### 4.16.2 Scale invariance

When we consider a circle with radius  $R$ , the normalization of the distribution  $f(r)$  is given by the integral

$$\int_0^{2\pi} \int_0^R f(r) r dr d\theta = 2\pi \int_0^R f(r) r dr = 1 \quad (4.167)$$

The same distribution induces a similar distribution  $h(r)$  on a smaller concentric circle with radius  $aR$  ( $0 < a < 1$ ), such that  $h(r)$  is proportional to  $f(r)$ , i.e.,  $h(r) = Kf(r)$ , and

$$1 = 2\pi \int_0^{aR} h(u) u du = 2\pi \int_0^{aR} Kf(u) u du = 2\pi K \int_0^{aR} f(u) u du \quad (4.168)$$

i.e.,

$$K^{-1} = 2\pi \int_0^{aR} f(u) u du \quad (4.169)$$

and

$$f(r) = 2\pi h(r) \int_0^{aR} f(u) u du \quad (4.170)$$

inside the smaller circle. Now we invoke the assumed scale invariance: the probability of finding a center in an annulus with radii  $r$  and  $r + dr$  in the original circle, must be equal to the probability of finding a center in the scaled down annulus,

$$h(ar)(ar)d(ar) = f(r)rdr \quad (4.171)$$

and therefore

$$a^2 h(ar) = f(r) \quad (4.172)$$

Equation (4.172) can also be rewritten in the form

$$h(r) = \frac{1}{a^2} f\left(\frac{r}{a}\right) \quad (4.173)$$

and inserting this into equation (4.170) we find

$$a^2 f(ar) = 2\pi f(r) \int_0^{aR} f(u) u du \quad (4.174)$$

We solve equation (4.174) taking first its derivative with respect to  $a$ : the relation that we find must hold for all  $a$ 's, and therefore also for  $a = 1$  (no scaling), and we find the differential equation

$$rf'(r) = (2\pi R^2 f(R) - 2) f(r) \quad (4.175)$$

i.e.,

$$rf'(r) = (q - 2) f(r) \quad (4.176)$$

where the constant  $q = 2\pi R^2 f(R)$  is unknown. However, we can still solve the equation and find

$$f(r) = Ar^{q-2} \quad (4.177)$$

The constant  $A$  is easy to find from the normalization condition:  $A = q/2\pi R^q$ , and therefore

$$f(r) = \frac{qr^{q-2}}{2\pi R^q} \quad (4.178)$$

### 4.16.3 Translational invariance

Now we investigate the consequences of translational invariance. Figure 4.19 illustrates the situation: the original circle is crossed by a straight line which defines the chord. This circle is displaced by the amount  $b$ , and the new radius and angle that define the midpoint of the chord are

$$r' = |r - b \cos \theta| \quad (4.179)$$

$$\theta' = \theta \text{ (if } r \geq b \cos \theta \text{)} \text{ or } \theta' = \theta + \pi \text{ (if } r < b \cos \theta \text{)} \quad (4.180)$$

Now consider a region  $\Gamma$  surrounding the midpoint in the original circle, which is transformed into a region  $\Gamma'$  by the translation. The probability of finding a chord with the midpoint in the region  $\Gamma$  is

$$\int_{\Gamma} f(r) r dr d\theta = \int_{\Gamma} \frac{qr^{q-1}}{2\pi R^q} dr d\theta = \frac{q}{2\pi R^q} \int_{\Gamma} r^{q-1} dr d\theta \quad (4.181)$$

Likewise, the same probability for the translated circle is

$$\frac{q}{2\pi R^q} \int_{\Gamma'} (r')^{q-1} dr' d\theta' = \frac{q}{2\pi R^q} \int_{\Gamma} |r - b \cos \theta|^{q-1} dr d\theta \quad (4.182)$$

where the Jacobian of the transformation is 1. Equating expression (4.181) and (4.182), and given the arbitrariness of the region  $\Gamma$ , we see that the integrand must be a constant, and therefore  $q = 1$ , and

$$f(r, \theta) = \frac{1}{2\pi Rr} \quad (r \leq R; \ 0 \leq \theta < 2\pi) \quad (4.183)$$

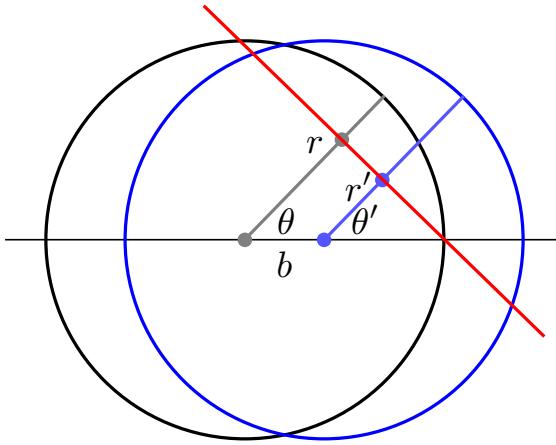


Figure 4.19: Geometrical construction for the discussion of translational invariance. The original circle (black) is crossed by a straight line (red) which defines a chord. The translated circle is shown in blue.

Using this distribution, we find that the probability of finding a midpoint inside the circle with radius  $R/2$  – i.e., the probability of finding a chord longer than the side of the triangle in Bertrand's paradox – is

$$\int_0^{2\pi} d\theta \int_0^{R/2} f(r, \theta) r dr = 2\pi \int_0^{R/2} \frac{1}{2\pi Rr} r dr = \frac{1}{2} \quad (4.184)$$

which corresponds to the second alternative in the discussion of Bertrand's paradox in section 2.2.4.

Obviously this is not a proof that the second solution correspond to this probability density, only that they yield the same probability of finding a chord as required by Bertrand's conditions. However we see at once that in the second solution  $2\pi r dr f(r) = \text{constant} \times dr$  so that  $f(r) \propto 1/r$ . Finally the normalization condition leads to  $f(r) = 1/(2\pi Rr)$ , and the second solution does indeed correspond to the invariant pdf (4.183).

Bertrand's paradox is a highly idealized problem, which nevertheless leads to a very interesting exploration of the principle of indifference which is at the very basis of the idea of equally probable elementary events. This principle of indifference is far from abstract because it is continuously applied in physics: for instance we assign equal probabilities to microstates in statistical ensembles, and we do this on the basis of implicit, unstated, symmetry principles. For a more detailed discussion of these topics, see Jayne's paper [21].

## 4.17 Probability in Quantum Mechanics

Although the role of probability in all modern physical theories is very deep and wide, in Quantum Mechanics (QM) it assumes a puzzling and fundamental character. Indeed, QM is all about complex probability amplitudes, and probability only appears as the squared modulus of these complex numbers. Here I illustrate very briefly how this leads to a strange and important result that tells us something of the nature of reality – although the message sent by Nature is unclear and still hotly debated.

Consider two *identical* objects, i.e., *objects that share the same general properties*, and assume that

1. the properties are multivalued, and that their values are predetermined and are not influenced by measurement (an example of such a property could be the *color* of a set of “identical” balls: the color can take different values such as *red* or *blue*);
2. *locality* holds, i.e., when the objects are spatially separated, the determination of the value of the property of one object does not influence the other one.

We assume that these two objects share 3 different binary properties, A, B, and C, which can take the values 0 and 1, and we experiment with them. We enclose each of them in an opaque box, so that we cannot observe it directly, and because of locality, we know that a measurement on object 1 cannot influence a measurement on object 2. We can also consider probabilities for each particular property value or combination of property

values, such as  $P(A_1 = B_2)$ , which is that the probability that the value of property A for object 1 is the same as the value of property B for object 2, i.e., that they are both 0 or both 1. For example, *if we assume* that the two objects share exactly the same property values for each property, then

$$P(A_1 = A_2) = P(B_1 = B_2) = P(C_1 = C_2) = 1 \quad (4.185)$$

From this assumption we deduce that

$$P(A_1 = B_2) + P(A_1 = C_2) + P(B_1 = C_2) \geq 1 \quad (4.186)$$

and this is demonstrated graphically in figure 4.20. The inequality (4.186) is known as *Bell's inequality*, and it is violated by quantum systems: this means that one or more of the basic assumptions is also violated.

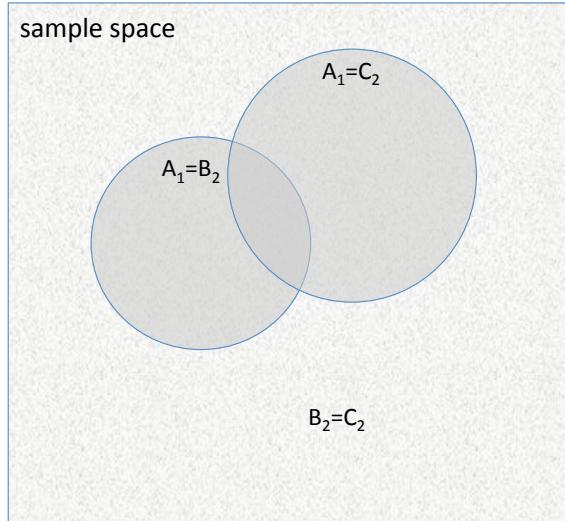


Figure 4.20: Sample space representation of the events that contribute to Bell's inequality (4.186). The overlap region shared by the two regions marked  $A_1 = B_2$  and  $A_1 = C_2$  clearly contains events such that  $B_2 = C_2$ . The same happens in the external region where  $A_1 \neq B_2$  and  $A_1 \neq C_2$ , because of the binary nature of the properties A, B, and C. However  $B_1 = B_2$  by assumption, therefore  $B_2 = C_2$  implies  $B_1 = C_2$ , and finally we see that the total probability of the events  $A_1 = B_2$ ,  $A_1 = C_2$ , and  $B_1 = C_2$ , exceeds 1.

In order to prove the violation, we construct a quantum mechanical counterexample: we consider two two-level quantum systems in the entangled state

$$|\phi^+\rangle = \frac{|00\rangle + |11\rangle}{\sqrt{2}}$$

and the two-valued properties A, B, and C obtained by projecting the entangled state on the rotated orthogonal bases

$$A : \begin{cases} |a_0\rangle = |0\rangle \\ |a_1\rangle = |1\rangle \end{cases}$$

$$B : \begin{cases} |b_0\rangle = \frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}|1\rangle \\ |b_1\rangle = \frac{\sqrt{3}}{2}|0\rangle - \frac{1}{2}|1\rangle \end{cases}$$

$$C : \begin{cases} |c_0\rangle = \frac{1}{2}|0\rangle - \frac{\sqrt{3}}{2}|1\rangle \\ |c_1\rangle = \frac{\sqrt{3}}{2}|0\rangle + \frac{1}{2}|1\rangle \end{cases}$$

It is easy to check that these are orthonormal bases, and that

$$|\phi^+\rangle = \frac{|a_0a_0\rangle + |a_1a_1\rangle}{\sqrt{2}} = \frac{|b_0b_0\rangle + |b_1b_1\rangle}{\sqrt{2}} = \frac{|c_0c_0\rangle + |c_1c_1\rangle}{\sqrt{2}}$$

so that

$$\langle a_0a_0|\phi^+\rangle = \langle a_1a_1|\phi^+\rangle = \frac{1}{\sqrt{2}}$$

therefore

$$P(A_1 = 0 \text{ and } A_2 = 0) = P(A_1 = 1 \text{ and } A_2 = 1) = 1/2$$

and finally

$$P(A_1 = A_2) = 1$$

A similar calculation with properties B and C yields the same result, and therefore this quantum system satisfies the condition (4.185).

Now notice that we can rewrite the vectors A as linear combinations of the vectors B:

$$|a_0\rangle = \frac{1}{2}|b_0\rangle + \frac{\sqrt{3}}{2}|b_1\rangle$$

$$|a_1\rangle = \frac{\sqrt{3}}{2}|b_0\rangle - \frac{1}{2}|b_1\rangle$$

and therefore

$$|\phi^+\rangle = \frac{|a_0\rangle(|b_0\rangle + \sqrt{3}|b_1\rangle) + |a_1\rangle(\sqrt{3}|b_0\rangle - |b_1\rangle)}{2\sqrt{2}}$$

This means that

$$\langle a_0b_0|\phi^+\rangle = \langle a_1b_1|\phi^+\rangle = \frac{1}{2\sqrt{2}}$$

and therefore

$$P(A_1 = 0 \text{ and } B_1 = 0) = P(A_1 = 1 \text{ and } B_2 = 1) = 1/8$$

i.e.

$$P(A_1 = B_2) = 1/4$$

This can be replicated to obtain

$$P(A_1 = C_2) = P(B_1 = C_2) = 1/4$$

so that finally we find that in this two-level quantum system

$$P(A_1 = B_2) + P(A_1 = C_2) + P(B_1 = C_2) = 3/4 < 1$$

and the inequality (4.186) is violated.

The inequality presented in this section is just one of many versions that have been produced since Bell's discovery in 1964. The violation is surprising, it is confirmed by experiments, and it indicates that there is something amiss in our understanding of the physical world. It is still unclear how all this comes about, what is the origin of the violation. It is common to indicate locality as the culprit, however this was disfavoured by John Bell, the originator of this line of thinking.



Figure 4.21: John Stewart Bell (Belfast, June 28th 1928 – Belfast, October 10th 1990), theoretical physicist at CERN. Bell gave important contributions to the physics of particle accelerators, to quantum field theory, and to the foundations of quantum mechanics.

## 4.18 What's the use of all this?

Probability distributions are one of the fundamental tools of both probability theory and statistics, they are really at the heart of both disciplines.

Given their importance in computer simulations of all kinds, excellent lists of distributions with a lot of documentation and methods to generate pseudorandom numbers that follows those distributions can be found in the documentation of the main numerical libraries, like Boost

[http://www.boost.org/doc/libs/1\\_54\\_0/libs/math/doc/html/dist.html](http://www.boost.org/doc/libs/1_54_0/libs/math/doc/html/dist.html)

or the GNU Scientific Library

[http://www.gnu.org/software/gsl/manual/html\\_node/](http://www.gnu.org/software/gsl/manual/html_node/)

[Random-Number-Distributions.html#Random-Number-Distributions](http://www.gnu.org/software/gsl/manual/html_node/Random-Number-Distributions.html#Random-Number-Distributions)

## Appendix: volume element in $n$ -dimensional space

A volume element in  $n$ -dimensional space is defined by the directions of the coordinate axes, and can be computed with a simple formula. Now take a set of vectors  $\{\mathbf{v}_i\}$ : it is well-known and easy to prove that in 2-dimensional space the volume (area) defined by two such vectors (see figure 4.22) is

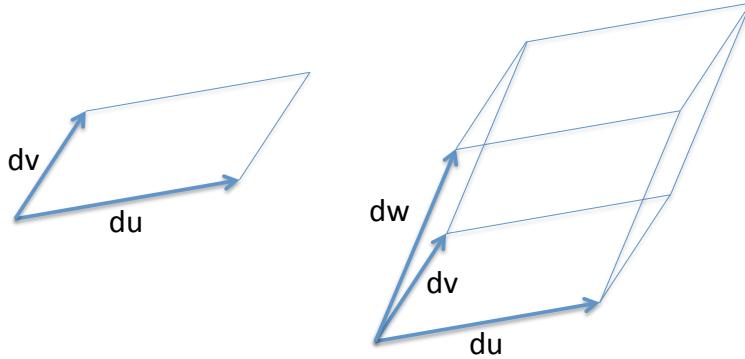


Figure 4.22: The figure shows how a pair of vectors defines an area in 2-space and a triple defines a volume in 3-space.

$$V_2 = |\mathbf{v}_1 \times \mathbf{v}_2| = |\epsilon_{jk} v_{1,j} v_{2,k}| \quad (4.187)$$

and similarly in 3-space, the volume is

$$V_3 = |\epsilon_{jkl} v_{1,j} v_{2,k} v_{3,l}| \quad (4.188)$$

Now we prove by induction that in  $N$ -space

$$V_N = |\epsilon_{jkl\dots mn} v_{1,j} v_{2,k} v_{3,l} \dots v_{N,m}| \quad (4.189)$$

Indeed the vector with components

$$x_n = \epsilon_{jkl\dots mn} v_{1,j} v_{2,k} v_{3,l} \dots v_{N-1,m} \quad (4.190)$$

– where the sums run over all the indexes of  $\epsilon_{jkl\dots mn}$  except  $n$  – has magnitude  $V_{N-1}$  (= volume of  $N-1$  dimensional region, by the induction hypothesis), and is orthogonal to all the vectors  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{N-1}\}$ , since

$$\mathbf{x} \cdot \mathbf{v}_i = \epsilon_{jkl\dots mn} v_{1,j} v_{2,k} v_{3,l} \dots, v_{i,l}, v_{i,s}, \dots v_{N-1,m} = 0 \quad (4.191)$$

( $v_{i,l}, v_{i,s}$  is symmetrical with respect to the exchange of indices  $l$  and  $s$ , and because of the antisymmetrical tensor  $\epsilon$  the sum vanishes), therefore if we let  $\mathbf{v}_N = \mathbf{v}_\perp + \mathbf{v}_\parallel$ ,  $\mathbf{x} \cdot \mathbf{v}_N = \mathbf{x} \cdot \mathbf{v}_\perp = V_N$ .

Now consider coordinates  $x_1, x_2, \dots, x_n$  and the transformed coordinates

$$u_i = u_i(x_1, x_2, \dots, x_n) \quad (4.192)$$

then the elementary volume element in  $x$ -space is defined by vectors with  $u$ -coordinates

$$d\mathbf{x}_i = \left( \frac{\partial x_i}{\partial u_1} du_1, \dots, \frac{\partial x_i}{\partial u_n} du_n \right) \quad (4.193)$$

and the elementary volume is

$$dx_1 \dots dx_n = \left| \epsilon_{ijk\dots} \frac{\partial x_1}{\partial u_i} du_i \frac{\partial x_2}{\partial u_j} du_j \frac{\partial x_k}{\partial u_k} du_k \dots \right| = \left| \det \frac{\partial x_i}{\partial u_j} \right| du_1 \dots du_n \quad (4.194)$$

(the contraction with the  $\epsilon$  tensor returns the oriented (signed) volume, and we take the absolute value to drop the orientation sign).

# Chapter 5

## The Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) is another important law of large numbers, since it specifies the behavior of sums of random variables under rather general conditions. Here I give a standard proof that utilizes the important formalism of generating functions. Therefore this chapter starts with a short introduction to generating functions and some of their applications to physics.

### 5.1 Generating Functions

#### 5.1.1 Definition and simple examples

Now we consider a (infinite) sequence  $a_0, a_1, \dots, a_k, \dots$ : this sequence can be used to define a *generating function*

$$A(z) = \sum_k a_k z^k \quad (5.1)$$

where  $z$  is a complex variable. If the series  $A(z)$  can be (at least formally) summed and given in a compact closed form, then the generating function can have a great value, because it can be used to simplify complex calculations. In this section we take a quick look at some of the important features of generating functions: for much more in-depth information have a look at H. S. Wilf's book *generatingfunctionology* [34].

#### Examples

- geometric sequence: In this case  $a_k = \alpha^k$  ( $k = 0, \dots, \infty$ ), then

$$A(z) = \sum_k \alpha^k z^k = \frac{1}{1 - \alpha z} \quad (5.2)$$

- a two term recurrence: generating functions can be used to find closed formulas for sequences that are defined only by recurrence formulas, as in this example

$$a_{n+1} = 2a_n + 1 \quad (n \geq 0; a_0 = 0) \quad (5.3)$$

Then, multiplying by  $z^n$  and summing, we find

$$\sum_{n=0,\infty} a_{n+1} z^n = \sum_{n=0,\infty} (2a_n + 1) z^n = 2A(z) + \frac{1}{1-z} \quad (5.4)$$

Since

$$\begin{aligned} \sum_{n=0,\infty} a_{n+1} z^n &= \frac{1}{z} \sum_{n=0,\infty} a_{n+1} z^{n+1} = \frac{1}{z} \sum_{n=1,\infty} a_n z^n \\ &= \frac{A(z) - a_0}{z} = \frac{A(z)}{z} \end{aligned} \quad (5.5)$$

we find the equation

$$\frac{A(z)}{z} = 2A(z) + \frac{1}{1-z} \quad (5.6)$$

and

$$\begin{aligned} A(z) &= \frac{z}{(1-2z)(1-z)} = z \left( \frac{2}{1-2z} - \frac{1}{1-z} \right) \\ &= z \sum_{n=0,\infty} (2(2z)^n - z^n) = z \sum_{n=0,\infty} (2^{n+1} - 1) z^n \end{aligned} \quad (5.7)$$

then  $a_n = 2^n - 1$ , which is a closed formula for  $a_n$ .

- a harder two-term recurrence problem: the previous problem could easily be solved by induction, while the present one is not so easy to attack with elementary methods

$$a_{n+1} = 2a_n + n \quad (n \geq 0; a_0 = 1) \quad (5.8)$$

As before

$$\sum_{n=0,\infty} a_{n+1} z^n = \sum_{n=0,\infty} (2a_n + n) z^n = 2A(z) + \sum_{n=0,\infty} nz^n \quad (5.9)$$

and we see that we have to evaluate a new term

$$\begin{aligned} \sum_{n=0,\infty} nz^n &= \sum_{n=1,\infty} nz^n = z \frac{d}{dz} \sum_{n=0,\infty} z^n = z \frac{d}{dz} \frac{1}{1-z} = \frac{z}{(1-z)^2} \end{aligned} \quad (5.10)$$

Therefore

$$\frac{A(z) - 1}{z} = 2A(z) + \frac{z}{(1-z)^2} \quad (5.11)$$

and

$$A(z) = \frac{1-2z+2z^2}{(1-2z)(1-z)^2} = \frac{2}{1-2z} - \frac{1}{(1-z)^2} \quad (5.12)$$

Since

$$\frac{2}{1-2z} = 2 \sum_{n=0,\infty} (2z)^n = \sum_{n=0,\infty} 2^{n+1} z^n \quad (5.13)$$

and

$$\frac{1}{(1-z)^2} = \frac{1}{z} \sum_{n=1,\infty} n z^n = \sum_{n=0,\infty} (n+1) z^n \quad (5.14)$$

then

$$A(z) = \sum_{n=0,\infty} 2^{n+1} z^n - \sum_{n=0,\infty} (n+1) z^n \quad (5.15)$$

and

$$a_n = 2^{n+1} - (n+1) \quad (5.16)$$

- Fibonacci recurrence:

$$F_{n+1} = F_n + F_{n-1} \quad (n \geq 0; F_0 = 0; F_1 = 1) \quad (5.17)$$

In this case we define the generating function

$$F(z) = \sum_{n=0,\infty} F_n z^n \quad (5.18)$$

and we find

$$\sum_{n=1,\infty} F_{n+1} z^{n-1} = \sum_{n=1,\infty} F_n z^{n-1} + \sum_{n=1,\infty} F_{n-1} z^{n-1} \quad (5.19)$$

i.e.,

$$\sum_{n=0,\infty} F_{n+2} z^n = \sum_{n=0,\infty} F_{n+1} z^n + \sum_{n=0,\infty} F_n z^n \quad (5.20)$$

and also

$$\frac{F(z) - F_1 z - F_0}{z^2} = \frac{F(z) - F_0}{z} + F(z) \quad (5.21)$$

Substituting the first two values of the recurrence formula we find

$$\frac{F(z) - z}{z^2} = \frac{F(z)}{z} + F(z) \quad (5.22)$$

and

$$F(z) = \frac{z}{1-z-z^2} = \frac{z}{(1-r_+ z)(1-r_- z)} \quad (5.23)$$

where

$$r_{\pm} = \frac{1 \pm \sqrt{5}}{2} \quad (5.24)$$

(prove this as an exercise ...). The generating function can be manipulated further, so that

$$F(z) = \frac{1}{\sqrt{5}} \left( \frac{1}{1 - r_+ z} - \frac{1}{1 - r_- z} \right) = \frac{1}{\sqrt{5}} \left( \sum_{n=0,\infty} (r_+ z)^n - \sum_{n=0,\infty} (r_- z)^n \right) \quad (5.25)$$

and finally

$$F_n = \frac{r_+^n - r_-^n}{\sqrt{5}} \quad (5.26)$$

(Binet's formula)

### 5.1.2 Probability generating functions

Now consider a discrete random variable  $\xi$  with probabilities  $P_{\xi}(k) = P(\xi = k)$ : we define the corresponding probability generating function (pgf):

$$F_{\xi}(z) = \sum_k z^k P_{\xi}(k) = \mathbf{E}(z^{\xi}) \quad (5.27)$$

*a)* average and variance:

$$F'_{\xi}(1) = \sum_k k P_{\xi}(k) = \mathbf{E}\xi \quad (5.28)$$

$$F''_{\xi}(1) = \sum_k k(k-1) P_{\xi}(k) = \mathbf{E}(\xi^2) - \mathbf{E}\xi \quad (5.29)$$

therefore

$$\mathbf{E}\xi = F'_{\xi}(1); \quad \mathbf{D}\xi = F''_{\xi}(1) + F'_{\xi}(1) - (F'_{\xi}(1))^2 \quad (5.30)$$

*b)* higher order derivatives of the pgf:

$$F'''_{\xi}(1) = \sum_k k(k-1)(k-2) P_{\xi}(k) = \mathbf{E}(\xi^3) - 3\mathbf{E}(\xi^2) + 2\mathbf{E}(\xi) \quad (5.31)$$

therefore

$$\mathbf{E}(\xi^3) = F'''_{\xi}(1) + 3F''_{\xi}(1) + F'_{\xi}(1) \quad (5.32)$$

*c)* example, the Poisson distribution: the Poisson distribution has probabilities

$$P_{\xi}(k) = \frac{a^k}{k!} e^{-a} \quad (5.33)$$

therefore

$$F_\xi(z) = \sum_{k=0}^{\infty} z^k \frac{a^k}{k!} e^{-a} = e^{a(z-1)} \quad (5.34)$$

From this pgf we find

$$F'_\xi(1) = a; \quad F''_\xi(1) = a^2; \quad \dots \quad F_\xi^{(n)}(1) = a^n \quad (5.35)$$

so that

$$\mathbf{E}\xi = a; \quad \mathbf{D}\xi = a; \quad \mathbf{E}(\xi^3) = a^3 + 3a^2 + a \quad (5.36)$$

- d)** sum of independent random variables: let  $\xi = \sum_k \xi_k$ , where the  $\xi_k$  are independent random variables, then

$$F_\xi(z) = \mathbf{E}(z^\xi) = \mathbf{E}(z^{\sum_k \xi_k}) = \prod_k \mathbf{E}(z^{\xi_k}) = \prod_k F_{\xi_k}(z) \quad (5.37)$$

In the case of  $n$  i.i.d. variates we find

$$F_\xi(z) = [F(z)^n] \quad (5.38)$$

- e)** example, the binomial distribution: the sum  $\xi = \sum_k \xi_k$  where  $\xi_k$  is either 0 or 1 with probability, respectively  $p$  and  $q = 1 - p$ , has a binomial distribution. The individual variates  $\xi_k$  have generating function  $F(z) = q + pz$ , therefore the sum of  $n$  such random variables has generating function  $F(z) = (q + pz)^n$ .

From this pgf we find

$$F'_\xi(1) = np; \quad F''_\xi(1) = n(n-1)p^2; \quad F'''_\xi(1) = n(n-1)(n-2)p^3 \quad (5.39)$$

therefore

$$\mathbf{E}\xi = np; \quad \mathbf{D}\xi = npq; \quad \mathbf{E}(\xi^3) = npq \quad (5.40)$$

and

$$\mathbf{E}(\xi^3) = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np \quad (5.41)$$

- f)** by manipulating pgf's we can solve fairly difficult problems: consider the probability generating function of a sum  $\xi = \sum_k \xi_k$  of  $N$  i.i.d. variates  $\xi_k$ , where  $N$  itself corresponds to a random variable  $\nu$  (independent from the other variates), then, using the basic definitions, we can write

$$P_\xi(k) = \sum_N P(k, N) = \sum_N P(k|N)P_\nu(N) \quad (5.42)$$

therefore the generating function is

$$F_\xi(z) = \sum_k P_\xi(k) z^k = \sum_k \sum_N P(k|N) P_\nu(N) z^k \quad (5.43)$$

$$= \sum_N P_\nu(N) \sum_k P(k|N) z^k \quad (5.44)$$

$$= \sum_N P_\nu(N) [F_{\xi_1}(z)]^N \quad (5.45)$$

$$= F_\nu[F_{\xi_1}(z)] \quad (5.46)$$

where  $F_{\xi_1}(z)$  is the pgf of each  $\xi_i$ , and from this pgf we can compute all the moments of the distribution of  $\xi$ .

- g)** weak convergence of probability distributions: in the previous example we have defined a sequence of probability distributions (the probabilities  $\{P_n(k)\}$  of the occurrence of  $k$  in  $n$  trials). If there exists a limiting probability distribution  $P(k)$  such that

$$\lim_{n \rightarrow \infty} P_n(k) = P(k) \quad (5.47)$$

for all  $k$ , then we say that the sequence  $P_n(k)$  converges weakly to the limiting distribution  $P(k)$ .

- h)** theorem on weak convergence: a sequence  $P_n(k)$  converges weakly to a limiting distribution  $P(k)$  if and only if the corresponding sequence of generating functions  $F_n(z)$  converges to the generating function  $F(z)$  of the sequence  $P(k)$ . This is theorem 6.2 in [27]. The proof is straightforward and we skip it here.

- i)** weak convergence of the binomial distribution to the Poisson distribution: suppose that for a binomial distribution the following condition holds:

$$\lim_{n \rightarrow \infty} np = a \quad (5.48)$$

then

$$\lim_{n \rightarrow \infty} (q+pz)^n = \lim_{n \rightarrow \infty} (1-p(1-z))^n = \lim_{n \rightarrow \infty} \left(1 + \frac{a(z-1)}{n}\right)^n = e^{a(z-1)} \quad (5.49)$$

thus the generating function of the binomial distribution converges to the generating function of the Poisson distribution, and therefore the binomial distribution converges weakly to the Poisson distribution.

- j)** the Poisson distribution, again: consider the occurrence of random, uncorrelated events in a time interval  $\Delta t$ , and let  $\xi = \xi(\Delta t)$  be the random variable that represents the number of events that occur in

this time interval. Let  $\lambda$  be the rate of occurrence, and now subdivide finely the time interval, splitting it into  $n$  smaller time intervals  $\Delta t/n$ . If you make these intervals small enough, the average number of events that occur in a subinterval is quite small, and the probability that more than one event actually occurs is negligible, i.e., the occurrence of an event is modeled by the binomial probabilities

$$p = \frac{\lambda\Delta t}{n} + o\left(\frac{\Delta t}{n}\right); \quad q = 1 - \frac{\lambda\Delta t}{n} + o\left(\frac{\Delta t}{n}\right) \quad (5.50)$$

therefore the generating function for the probabilities of any one of the indicator random variables that corresponds to each subinterval of duration  $\Delta t/n$  is

$$\left(1 - \frac{\lambda\Delta t}{n}\right) + \frac{\lambda\Delta t}{n}z + o\left(\frac{\Delta t}{n}\right) \quad (5.51)$$

Since

$$\xi(\Delta t) = \sum_{i=1,n} \xi_i \left(\frac{\Delta t}{n}\right) \quad (5.52)$$

(the number of events in the long interval is the sum of the number of events in the subintervals) the generating function of the probability distribution of  $\xi(\Delta t)$  is

$$F(z) = \lim_{n \rightarrow \infty} F_n(z) = \lim_{n \rightarrow \infty} \left(1 + \frac{\lambda\Delta t}{n}(z - 1)\right)^n = \exp(\lambda\Delta t(z - 1)) \quad (5.53)$$

which is the generating function of a Poisson distribution with average  $\lambda\Delta t$ .

### 5.1.3 Photomultiplier noise statistics

As an example of the usefulness of generating functions in physics, we study the photomultiplier noise statistics (more information on these processes can be found in [17], and in [10]).

Figure 5.1 shows the working principle of photomultipliers, where an initial photon extracts an electron from a photocathode and the electron is accelerated toward a chain of dynodes. Electrons impinging on dynodes extract additional electrons and eventually the collecting anode receives a large charge (usually the amplification factor is in the range  $10^5$ - $10^8$  electrons for each initial photoelectron).

The development of the electron cascade is an example of a *branching process*, fluctuations at each stage are amplified by all successive stages, and it is important to find the size of the fluctuations at the collecting anode.

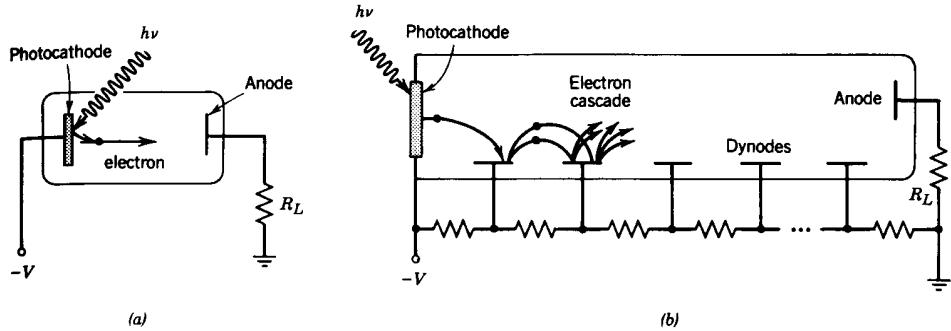


Figure 5.1: (left panel) A photocell has a photocathode where impinging light can extract electrons by photoelectric effect. The extracted electrons then drift to a collecting anode. The photocurrent is proportional to light intensity. (right panel) Photomultipliers usually have a transparent photocathode (metal oxides with low extraction work are coated on a glass substrate); the emitted photoelectrons accelerate toward a nearby electrode (a dynode) which has higher electric potential. Each colliding electron extracts more than one electron, and the total number of electrons grows. This process is repeated at each successive dynode, until the electron cascade reaches the collecting anode. The amplification factor is often of the order of  $10^6$ .

This kind of process was first studied – although in a different context – by Bienaymé, and, independently by Galton and Watson in 1874, and is usually called the *Galton-Watson process*. Later on it was found that it can be applied to different problems (nuclear chain reactions, electromagnetic showers, etc.). The mathematics of the Galton-Walton process was rediscovered during World War II, and for some time it was classified information [18].

Now let  $p_k$  be the probability that a single electron extracts  $k$  electrons when it hits a dynode then the probability generating function is

$$F(z) = \sum_{k=0}^{\infty} p_k z^k \quad (5.54)$$

Clearly this is also the probability generating function for the number of electrons extracted from the first dynode.

At the next step – i.e., at the second dynode – the probability of extracting  $n$  electrons is

$$p_n^{(2)} = \sum_{m=0}^{\infty} P(n|m)P(m) = \sum_{m=0}^{\infty} P(n|m)p_m \quad (5.55)$$

where  $P(n|m)$  is the probability of extracting  $n$  electrons with  $m$  primary electrons, and  $P(m) = p_m$ , so that the PGF is

$$F_2(z) = \sum_n p_n^{(2)} z^n = \sum_{n=0}^{\infty} z^n \sum_{m=0}^{\infty} P(n|m)p_m = \sum_{m=0}^{\infty} p_m \sum_{n=0}^{\infty} P(n|m)z^n \quad (5.56)$$

We start with the evaluation of  $P(n|2)$ : the probability that the first electron extracts  $k$  electrons, while the second electron extracts  $n - k$  electrons is  $p_k p_{n-k}$ , because the two extraction processes are independent. Thus the probability that – in all – these two electrons extract  $n$  electrons is

$$P(n|2) = \sum_{k=0}^n p_k p_{n-k} \quad (5.57)$$

Then the corresponding contribution to the generating function is<sup>1</sup>

$$\begin{aligned} \sum_{n=0}^{\infty} P(n|2)z^n &= \sum_{n=0}^{\infty} \sum_{k=0}^n p_k p_{n-k} z^n = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} p_k p_n z^{k+n} \\ &= \sum_{k=0}^{\infty} p_k z^k \sum_{n=0}^{\infty} p_n z^n = \left( \sum_{k=0}^{\infty} p_k z^k \right)^2 \end{aligned} \quad (5.58)$$

The contributions from higher values of  $m$  are evaluated in much the same way, so that the PGF at the second dynode is

$$F_2(z) = \sum_{m=0}^{\infty} p_m \left( \sum_{k=0}^{\infty} p_k z^k \right)^m = \sum_{m=0}^{\infty} p_m (F(z))^m = F[F(z)] \quad (5.59)$$

All this can be iterated for the other dynodes: we note that at the  $j$ -th step

$$p_n^{(j)} = \sum_{m=0}^{\infty} P(n|m)p_m^{(j-1)} \quad (5.60)$$

---

<sup>1</sup>This can be pictured in the following way: the array of terms in the inner sum is

$$\begin{array}{cccc} p_0 p_0 z^0 & & & \\ p_0 p_1 z^1 & p_1 p_0 z^1 & & \\ p_0 p_2 z^2 & p_1 p_1 z^2 & p_2 p_0 z^2 & \\ p_0 p_3 z^3 & p_1 p_2 z^3 & p_2 p_1 z^3 & p_3 p_0 z^3 \\ \dots & & & \end{array}$$

which can be rearranged as follows

$$\begin{array}{ccccc} p_0 p_0 z^0 & p_1 p_0 z^1 & p_2 p_0 z^2 & p_3 p_0 z^3 & \dots \\ p_0 p_1 z^1 & p_1 p_1 z^2 & p_2 p_1 z^3 & p_3 p_1 z^4 & \dots \\ p_0 p_2 z^2 & p_1 p_2 z^3 & p_2 p_2 z^4 & p_3 p_2 z^5 & \dots \\ p_0 p_3 z^3 & p_1 p_3 z^4 & p_2 p_3 z^5 & p_3 p_3 z^6 & \dots \\ \dots & & & & \end{array}$$

and this clearly correspond to the intermediate sum.

and therefore

$$\begin{aligned} F_j(z) &= \sum_n p_n^{(j)} z^n = \sum_{n=0}^{\infty} z^n \sum_{m=0}^{\infty} P(n|m) p_m^{(j-1)} = \sum_{m=0}^{\infty} p_m^{(j-1)} \sum_{n=0}^{\infty} P(n|m) z^n \\ &= \sum_{m=0}^{\infty} p_m^{(j-1)} \left( \sum_{k=0}^{\infty} p_k z^k \right)^m = \sum_{m=0}^{\infty} p_m^{(j-1)} (F(z))^m = F_{j-1}[F(z)] \end{aligned} \quad (5.61)$$

Obviously,  $F_{j-1}[F(z)] = F[F_{j-1}(z)]$ , and we define the sequence of generating functions

$$F_0(z) = z \quad (5.62a)$$

$$F_1(z) = F(z) = \sum_{k=0}^{\infty} p_k z^k \quad (5.62b)$$

$$F_2(z) = F[F(z)] \quad (5.62c)$$

...

$$F_{n+1}(z) = F[F_n(z)] \quad (5.62d)$$

...

The average number of electrons extracted at the first dynode is

$$\mu = \mathbf{E}(k) = F'(1) = \mu_1 \quad (5.63)$$

and the variance is

$$\sigma^2 = \mathbf{D}(k) = F''(1) + F'(1) - [F'(1)]^2 = \sigma_1^2 \quad (5.64)$$

Likewise, the mean at the  $n$ -th dynode is

$$\begin{aligned} \mu_n &= F'_n(1) = \frac{d}{dz} F[F_{n-1}(z)] \Big|_{z=1} \\ &= F'[F_{n-1}(z)] \frac{d}{dz} F_{n-1}(z) \Big|_{z=1} = F'(1) F'_{n-1}(1) = \mu \cdot \mu_{n-1} \end{aligned} \quad (5.65)$$

so that

$$\mu_n = \mu^n \quad (5.66)$$

Similarly, the variance is

$$\sigma_n^2 = F''_n(1) + F'_n(1) - [F'_n(1)]^2 \quad (5.67)$$

i.e.,

$$F''_n(1) = \sigma_n^2 - \mu_n + \mu_n^2 = \sigma_n^2 - \mu^n + \mu^{2n} \quad (5.68)$$

and since

$$\begin{aligned} F_n''(x) &= \frac{d}{dz} \{ F'[F_{n-1}(z)]F'_{n-1}(z) \} \\ &= F''[F_{n-1}(z)] [F'_{n-1}(z)]^2 + F'[F_{n-1}(z)]F''_{n-1}(z) \end{aligned} \quad (5.69)$$

we find

$$F_n''(1) = F''(1) [F'_{n-1}(1)]^2 + F'(1)F''_{n-1}(1) \quad (5.70a)$$

$$= (\sigma^2 - \mu + \mu^2)\mu_{n-1}^2 + \mu(\sigma_{n-1}^2 - \mu_{n-1} + \mu_{n-1}^2) \quad (5.70b)$$

$$= (\sigma^2 - \mu + \mu^2)\mu^{2n-2} + \mu(\sigma_{n-1}^2 - \mu^{n-1} + \mu^{2n-2}) \quad (5.70c)$$

$$= \sigma^2\mu^{2n-2} - \mu^{2n-1} + \mu^{2n} + \sigma_{n-1}^2\mu - \mu^n + \mu^{2n-1} \quad (5.70d)$$

$$= \sigma^2\mu^{2n-2} + \mu^{2n} + \sigma_{n-1}^2\mu - \mu^n \quad (5.70e)$$

Equating the two expressions for  $F_n''(1)$  we find

$$\sigma_n^2 - \mu^n + \mu^{2n} = \sigma^2\mu^{2n-2} + \mu^{2n} + \sigma_{n-1}^2\mu - \mu^n \quad (5.71)$$

i.e.,

$$\sigma_n^2 = \mu\sigma_{n-1}^2 + \sigma^2\mu^{2n-2} \quad (5.72)$$

or also

$$\sigma_{n+1}^2 = \mu\sigma_n^2 + \sigma^2\mu^{2n} \quad (5.73)$$

This is a new recurrence formula, which we can solve with the same generating function formalism that we have introduced in the first section. Let  $G(z)$  be the generating function of the sequence  $\sigma_n^2$

$$G(z) = \sum_{n=1}^{\infty} \sigma_n^2 z^n \quad (5.74)$$

with the initial condition  $\sigma_1^2 = \sigma^2$ , then we find

$$\frac{G(z)}{z} - \sigma^2 = \mu G(z) + \frac{\sigma^2\mu^2 z}{1 - \mu^2 z} \quad (5.75)$$

and

$$G(z) = \frac{\sigma^2}{\mu - \mu^2} \left( \frac{1}{1 - \mu z} - \frac{1}{1 - \mu^2 z} \right) = \frac{\sigma^2}{\mu - \mu^2} \left( \sum_{n=0}^{\infty} (\mu^n - \mu^{2n}) z^n \right) \quad (5.76)$$

and finally

$$\sigma_n^2 = \frac{\mu^{2n} - \mu^n}{\mu^2 - \mu} \sigma^2 \quad (5.77)$$

(exercise: justify all passages of the derivation of this formula).

It is interesting to note that since  $\sigma_n^2 \neq \mu_n = \mu^n$  the fluctuations in the cascade are not Poissonian.

## 5.2 Characteristic functions

The generating function works only for discrete random variables; the characteristic function is a sort of generalization that works with continuous random variables as well.

- a) definition: the characteristic function of a probability distribution is defined by the following formula

$$f_\xi(t) = \mathbf{E}(e^{i\xi t}) = \sum_x e^{ixt} P_\xi(x) \quad (5.78)$$

where  $t$  is a real variable.

- b) connection with the generating function formalism: consider a discrete random variable  $\xi$  and let  $z = e^{it}$ , then

$$f_\xi(t) = \mathbf{E}(e^{i\xi t}) = \mathbf{E}(z^\xi) = F_\xi(z) = F_\xi(e^{it}) \quad (5.79)$$

- c) connection with the Fourier transform formalism: if  $p_\xi(x)$  is the pdf of the continuous random variable  $\xi$ , then

$$f_\xi(t) = \int_{-\infty}^{+\infty} p_\xi(x) e^{ixt} dx \quad (5.80)$$

and conversely

$$p_\xi(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f_\xi(t) e^{-ixt} dt \quad (5.81)$$

- d) characteristic function of a sum of  $N$  i.i.d. random variables: let  $f_n(t)$  be the characteristic function of the  $n$ -th variate, then the characteristic function of the sum is

$$f_{\text{sum}}(t; N) = \mathbf{E}\left(e^{i\sum_n \xi_n t}\right) = \mathbf{E}\left(\prod_n e^{i\xi_n t}\right) = \prod_n \mathbf{E}\left(e^{i\xi_n t}\right) = \prod_n f_n(t) \quad (5.82)$$

In the case of  $N$  i.i.d. random variables, with characteristic function  $f(t)$ , we find  $f_{\text{sum}}(t; N) = [f(t)]^N$ .

- e) moments about the origin: the  $n$ -th moment about the origin is defined by the formula

$$m_n = \mathbf{E}(\xi^n) = \sum_x x^n P_\xi(x) \quad (5.83)$$

Notice that  $\mu = m_1$  is the mean value.

- f)** moments about the mean:  $n$ -th moment about the mean is defined by the formula

$$\mu_n = \mathbf{E} [(\xi - \mu)^n] = \sum_x (x - \mu)^n P_\xi(x) \quad (5.84)$$

Notice that the second moment about the mean is the variance.

- g)** series expansion of the characteristic function: take the definition of the characteristic function and expand the exponential:

$$\begin{aligned} f_\xi(t) &= \sum_x e^{ixt} P_\xi(x) = \sum_x \left( \sum_{n=0,\infty} \frac{(ixt)^n}{n!} \right) P_\xi(x) \\ &= \sum_{n=0,\infty} \frac{(it)^n}{n!} \sum_x x^n P_\xi(x) = \sum_{n=0,\infty} \frac{(it)^n}{n!} m_n \end{aligned} \quad (5.85)$$

therefore the  $n$ -th coefficient of the series expansion about the origin is  $i^n \mu_n$ . The same result can also be obtained by direct differentiation:

$$\frac{d^n f_\xi}{dt^n} \Big|_{t=0} = \frac{d^n}{dt^n} \sum_x e^{ixt} P_\xi(x) \Big|_{t=0} = \sum_x (ix)^n P_\xi(x) = i^n m_n \quad (5.86)$$

**Knowledge of all the moments about the origin unambiguously determines the characteristic function, i.e., the Fourier transform of the probability distribution, therefore the same knowledge unambiguously determines the probability distribution.**

- h)** we already met the moments in an earlier chapter, and there we defined the *moment generating function*. Note that

$$f_\xi(t) = M_\xi(it) \quad (5.87)$$

- i)** notice that in particular

$$f_\xi(0) = 1 \quad (5.88)$$

because it coincides with the normalization sum,

$$f'_\xi(0) = \frac{df_\xi}{dt} \Big|_{t=0} = im_1 = i\mu \quad (5.89)$$

and

$$f''_\xi(0) = \frac{d^2 f_\xi}{dt^2} \Big|_{t=0} = -m_2 = -\mathbf{E}(\xi^2) = -\sigma^2 - \mu^2 \quad (5.90)$$

i.e.,

$$\mu = -if'_\xi(0); \quad \sigma^2 = -f''_\xi(0) + [f'_\xi(0)]^2 \quad (5.91)$$

j) moments of the distribution: the first few moments about the mean have special names and meanings:

- (a) mean (average, mean value, expectation value):  $\mathbf{E}(\xi) = \mu_1$ ;
- (b) variance:  $\mathbf{E}[(\xi - \mu)^2] = \sigma^2$  (standard deviation:  $\sigma$ );
- (c) skewness:  $\frac{\mathbf{E}[(\xi - \mu)^3]}{\sigma^3}$ ;
- (d) kurtosis:  $\frac{\mathbf{E}[(\xi - \mu)^4]}{\sigma^4}$ .

As an instance, consider the skewness of the Poisson distribution with mean  $a$  (and thus with variance  $\sigma^2 = a$ ). The third moment about the mean is

$$\mu_3 = \sum_{n=0}^{\infty} (n-a)^3 \frac{a^n}{n!} e^{-a} = a \quad (5.92)$$

(check this as an exercise), therefore

$$\text{skewness} = \frac{\mu_3}{\sigma^3} = \frac{a}{a^{3/2}} = \frac{1}{\sqrt{a}} \quad (5.93)$$

i.e., the skewness of the Poisson distribution decreases as the mean value increases.

Next consider the kurtosis of the normal distribution  $N(0, \sigma)$ : in this case we have to evaluate the integral

$$\begin{aligned} \mu_4 &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} t^4 e^{-t^2/2\sigma^2} dt = \frac{2\sigma^4}{\sqrt{2\pi}} \int_0^{+\infty} x^3 e^{-x^2/2} d\left(\frac{x^2}{2}\right) \\ &= \frac{2\sigma^4}{\sqrt{2\pi}} \int_0^{+\infty} (2y)^{3/2} e^{-y} dy = \frac{4\sigma^4}{\sqrt{\pi}} \Gamma\left(\frac{5}{2}\right) = \frac{4\sigma^4}{\sqrt{\pi}} \cdot \frac{3\sqrt{\pi}}{4} = 3\sigma^4 \end{aligned} \quad (5.94)$$

therefore kurtosis =  $\mu_4/\sigma^4 = 3$ .

Figure 5.2 shows how shape is connected to kurtosis.

Similarly a uniform distribution on the interval  $(-a, a)$  (the interval is chosen so that it has zero mean, to simplify calculations) has

$$\mu_4 = \int_{-a}^a x^4 \frac{dx}{2a} = \frac{1}{5} a^4 \quad (5.95)$$

Since

$$\sigma^2 = \int_{-a}^a x^2 \frac{dx}{2a} = \frac{1}{3} a^2 \quad (5.96)$$

kurtosis =  $\mu_4/\sigma^4 = 9/5 = 1.8$ .

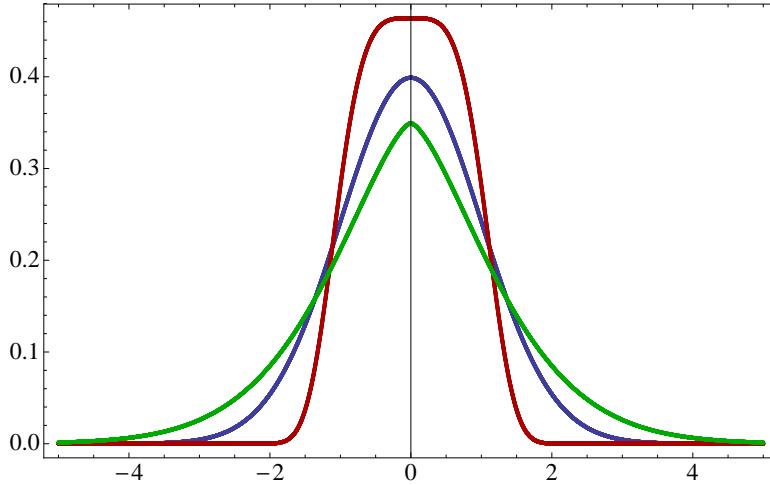


Figure 5.2: The curves represent normalized pdf's. The blue curve is a Gaussian pdf (kurtosis = 3); the other two pdf's have kurtosis = 1.05 (red curve) and = 6.36 (green curve).

- k)** mode, median: the moments are often complemented by other estimators of the shape of probability distributions. Two common estimators are
  - (a) mode or modal value: this is the value of the random variable that maximizes the probability distribution (discrete variates) or the probability density (continuous variates). Notice that pdf's can have multiple peaks, and in such cases the distribution is *multimodal*.
  - (b) median: this is the value  $x_m$  of the random variable such that

$$\int_{-\infty}^{x_m} p(x)dx = \int_{x_m}^{+\infty} p(x)dx = \frac{1}{2} \quad (5.97)$$

- l)** characteristic functions of a few selected distributions:

- (a) uniform distribution: here we consider a uniform distribution on the interval  $(a, b)$ , then

$$\begin{aligned}
 f_\xi(t) &= \int_a^b \frac{1}{b-a} e^{ixt} dx = \frac{1}{it} \frac{e^{ibt} - e^{iat}}{b-a} \\
 &= \frac{e^{i(b+a)t/2}}{it} \frac{e^{i(b-a)t/2} - e^{-i(b-a)t/2}}{b-a} \\
 &= e^{i(b+a)t/2} \operatorname{sinc} \left[ \frac{(b-a)t}{2} \right] \quad (5.98)
 \end{aligned}$$

- (b) binomial distribution: this is a discrete distribution, and using the connection with the generating function we find

$$f_\xi(t) = F_\xi(e^{it}) = (q + pe^{it})^n = [1 + p(e^{it} - 1)]^n \quad (5.99)$$

- (c) Poisson distribution: once again this is a discrete distribution, and using the connection with the generating function, we find (taking the average =  $a$ )

$$f_\xi(t) = F_\xi(e^{it}) = e^{a(e^{it}-1)} \quad (5.100)$$

- (d) exponential distribution:

$$f_\xi(t) = \int_0^\infty \frac{1}{\tau} e^{-x/\tau} e^{ixt} dx = \frac{1}{1-it\tau} \quad (5.101)$$

- (e) Gaussian distribution:

$$f_\xi(t) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} e^{ixt} dx \quad (5.102)$$

Since

$$-\frac{(x-\mu)^2}{2\sigma^2} + ixt = -\frac{[(x-it\sigma^2) - \mu]^2}{2\sigma^2} + \left(it\mu - \frac{t^2\sigma^2}{2}\right) \quad (5.103)$$

then

$$\begin{aligned} f_\xi(t) &= \exp\left(it\mu - \frac{t^2\sigma^2}{2}\right) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-it\sigma^2)-\mu]^2/2\sigma^2} dx \\ &= e^{it\mu - t^2\sigma^2/2} \end{aligned} \quad (5.104)$$

(the integral is just the normalization integral, and therefore it equals 1 – check this as an exercise). This characteristic function becomes quite simple for a  $N(0, 1)$  distribution:  $f_\xi(t) = e^{-t^2/2}$ .

- m)** sum of normal variates: consider a sum of independent normal variates  $\xi_k$  with mean  $\mu_k$  and variance  $\sigma_k^2$ , so that their characteristic function is  $e^{it\mu_k - t^2\sigma_k^2/2}$ , then the characteristic function of the sum is

$$\prod_k e^{it\mu_k - t^2\sigma_k^2/2} = e^{it(\sum_k \mu_k) - t^2(\sum_k \sigma_k^2)/2} \quad (5.105)$$

and therefore the sum is normally distributed with mean  $\sum_k \mu_k$  and variance  $\sum_k \sigma_k^2$ .

- n)** the previous result on the sum of normal variates can be trivially extended to a linear combination of normal variates: if

$$\xi = \sum_k a_k \xi_k$$

then

$$\begin{aligned} f_\xi(t) &= \mathbf{E} e^{it \sum_k a_k \xi_k} = \prod_k \mathbf{E} e^{ita_k \xi_k} = \prod_k f_k(a_k t) \\ &= \prod_k e^{ita_k \mu_k - t^2(a_k \sigma_k)^2/2} = e^{it(\sum_k a_k \mu_k) - t^2(\sum_k a_k^2 \sigma_k^2)/2} \end{aligned} \quad (5.106)$$

Therefore the linear combination is again a normal variate, with mean value  $\sum_k a_k \mu_k$  and variance  $\sum_k a_k^2 \sigma_k^2$ .

A variate is said to be stable (or to have a *stable distribution*) if it has the property that a linear combination of independent copies of the variate has the same distribution, up to location (mean) and scale parameters (variances). We have just shown that normal variates are stable.

- o)** generating function of more than one variate: there is a straightforward extension to  $k$  variates

$$f_\xi(\mathbf{t}) = \mathbf{E}(e^{i\xi \cdot \mathbf{t}}) = \sum_x e^{i\mathbf{x} \cdot \mathbf{t}} P_\xi(\mathbf{x}) \quad (5.107)$$

so that now

$$\left. \frac{\partial^n f_\xi}{\partial t_1^{n_1} \partial t_2^{n_2} \dots \partial t_k^{n_k}} \right|_{\mathbf{t}=0} = \mathbf{E}(x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}) \quad (5.108)$$

and the function generates the full array of moments and correlation functions.

- p)** theorem on weak convergence: this is the analog of the corresponding theorem for generating functions. A sequence  $P_n(k)$  converges weakly to a limiting distribution  $P(k)$  if and only if the corresponding sequence of characteristic functions  $f_n(t)$  converges to the generating function  $f(t)$  of the sequence  $P(k)$ . This is theorem 6.2' in [27]. Proof and details are straightforward and we skip them here.

- q)** series expansion:

$$\begin{aligned} f_\xi(t) &= \mathbf{E}(e^{i\xi t}) = \mathbf{E} \left( 1 + i\xi t - \frac{\xi^2 t^2}{2} + \sum_{n=3}^{\infty} \frac{(i\xi t)^n}{n!} \right) \\ &= 1 + i\mu t - \frac{\mathbf{E}(\xi^2)t^2}{2} + R(t) \end{aligned} \quad (5.109)$$

where

$$R(t) = \mathbf{E} \sum_{n=3}^{\infty} \frac{(i\xi t)^n}{n!} = \mathbf{E} \left( -i\xi^3 t^3 \sum_{n=0}^{\infty} \frac{(i\xi t)^n}{(n+3)!} \right) \quad (5.110)$$

then

$$|R(t)| \leq \mathbf{E} \left| -i\xi^3 t^3 \sum_{n=0}^{\infty} \frac{(i\xi t)^n}{(n+3)!} \right| \leq \mathbf{E} \left( |\xi|^3 |t|^3 \left| \sum_{n=0}^{\infty} \frac{(i\xi t)^n}{n!} \right| \right) = \mathbf{E} |\xi|^3 \cdot |t|^3 \quad (5.111)$$

i.e.,

$$|R(t)| \leq \mathbf{E} |\xi|^3 \cdot |t|^3 \quad (5.112)$$

### 5.3 The Central Limit Theorem (CLT)

*Theorem (CLT): consider a sequence of independent variates  $\xi_n$  with means  $\mu_n$  and variances  $\sigma_n^2$  that satisfies the Lyapunov condition*

$$\lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n \mathbf{E} |\xi_k - \mu_k|^3}{B_n^3} = 0$$

where

$$B_n^2 = \mathbf{D}S_n = \sum_{k=1}^n \sigma_k^2$$

and

$$S_n = \sum_{k=1}^n \xi_k,$$

then the sequence of the distributions of the normalized sums

$$S_n^* = \frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}} = \frac{S_n - \mathbf{E}S_n}{B_n}$$

converges weakly to the normal distribution  $N(0, 1)$ .

Note first that

$$S_n^* = \frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}} = \frac{S_n - \mathbf{E}S_n}{B_n} = \sum_{k=1}^n \frac{(\xi_k - \mu_k)}{B_n} \quad (5.113)$$

therefore the normalized sums are actually sums of zero-mean independent random variables  $\eta_{k,n} = (\xi_k - \mu_k)/B_n$  with variance  $\sigma_k^2/B_n^2$ . The characteristic functions of these variates are

$$f_{k,n}(t) = 1 - \frac{t^2 \sigma_k^2}{2B_n^2} + R_{k,n}(t) \quad (5.114)$$

where

$$|R_{k,n}(t)| \leq \mathbf{E} \left| \frac{\xi_k - \mu_k}{B_n} \right|^3 |t|^3 = \frac{\mathbf{E} |\xi_k - \mu_k|^3}{B_n^3} |t|^3 \quad (5.115)$$

and the characteristic function of the normalized sum is

$$f_n(t) = \prod_{k=1}^n \left( 1 - \frac{t^2 \sigma_k^2}{2B_n^2} + R_{k,n}(t) \right) \quad (5.116)$$

Taking logarithms, we write

$$\begin{aligned} \ln f_n(t) &= \sum_{k=1}^n \ln \left( 1 - \frac{t^2 \sigma_k^2}{2B_n^2} + R_{k,n}(t) \right) \\ &= \sum_{k=1}^n \left( -\frac{t^2 \sigma_k^2}{2B_n^2} + R_{k,n}(t) + O\left(\frac{t^4}{B_n^4}\right) \right) \end{aligned} \quad (5.117)$$

Moreover

$$0 \leq \sum_{k=1}^n |R_{k,n}(t)| \leq \frac{\sum_{k=1,n} \mathbf{E} |\xi_k - \mu_k|^3}{B_n^3} |t|^3 \xrightarrow{n \rightarrow \infty} 0 \quad (5.118)$$

therefore, because of the Lyapunov condition,

$$\ln f_n(t) \xrightarrow{n \rightarrow \infty} -\frac{t^2}{2} \sum_{k=1}^n \frac{\sigma_k^2}{B_n^2} = -\frac{t^2}{2} \quad (5.119)$$

and

$$f_n(t) \xrightarrow{n \rightarrow \infty} e^{-t^2/2} \quad (5.120)$$

i.e., the characteristic function converges weakly to the characteristic function of a  $N(0, 1)$  pdf.

### 5.3.1 Convergence rate

Notice that if the  $\xi$ 's are i.i.d. and  $\alpha = \mathbf{E} |\xi_k - \mu_k|^3$  exists, then

$$B_n^2 = \sum_{k=1,n} \sigma^2 = n\sigma^2 \quad (5.121)$$

and therefore

$$\frac{\sum_{k=1,n} \mathbf{E} |\xi_k - \mu_k|^3}{B_n^3} = \frac{n\alpha}{n^{3/2}\sigma^3} = \frac{\alpha}{\sqrt{n}\sigma^3} \xrightarrow{n \rightarrow \infty} 0 \quad (5.122)$$

and the Lyapunov condition is automatically satisfied. This is interesting on two accounts: first because it shows that the theorem works in an important

special case, second, because it gives an estimate of the speed of convergence (the deviation from Gaussianity should be roughly proportional to  $1/\sqrt{n}$ ).

**Exercise:** show that in the case of the binomial distribution,  $\alpha = p - p^3$ .

Here we state without proof the Berry-Esseen theorem, that sets a bound on the speed of convergence in the case of i.i.d. random variables:

*Theorem (Berry-Esseen): in the case of i.i.d. variates*

$$\sup_x \left| P(S_n^* < x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy \right| \leq \frac{C\alpha}{\sqrt{n}\sigma^3} \quad (5.123)$$

where  $C$  is a constant  $C < 0.7915$ .

### 5.3.2 Additive and multiplicative processes

The Gaussian distribution arises in the context of additive processes. If a process is multiplicative, i.e.,

$$S_n = \prod_{k=1}^n \xi_k \quad (5.124)$$

we should consider

$$\ln S_n = \sum_{k=1}^n \ln \xi_k \quad (5.125)$$

and thus it is the logarithm of the sum that is normally distributed: this leads naturally to the lognormal distribution.

Multiplicative processes arise in many interesting situations, consider, e.g., the following problem proposed by Kolmogorov: rocks in ore extraction are processed and crushed by heavy machinery; let  $x_0$  be the initial rock size, then after every crush only a fraction of rock is left (we follow just one fragment, all fragments have a similar fate), i.e., if  $x_n$  is the rock size at the  $n$ -th step

$$x_{n+1} = r_n \cdot x_n = \left( \prod_{k=0}^n r_k \right) x_0 \quad (5.126)$$

where  $r_n$  is a random variable  $r_n \in (0, 1)$ , and

$$\ln x_{n+1} = \sum_{k=0}^n \ln r_k + \ln x_0 \quad (5.127)$$

Clearly the random variable  $\prod_{k=0}^{\infty} r_k$  has a lognormal distribution

$$p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\ln x - \mu)^2/2\sigma^2} \quad (5.128)$$

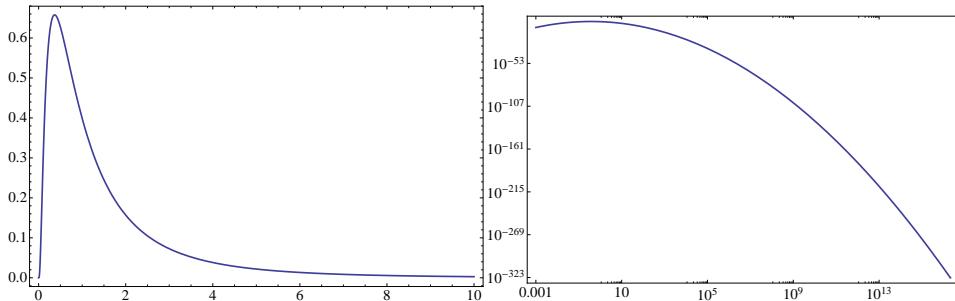


Figure 5.3: (left panel) plot of a lognormal distribution ( $p(x)$  vs.  $x$ ) with  $\mu = 10^{-5}$  and  $\sigma = 1$ ; (right panel) same distribution on a log-log scale.

Part of the interest in this kind of distribution stems also from the following consideration: if we plot a lognormal pdf on a log-log scale we find something like the curve in the right panel of figure 5.3, which is nearly straight for large values of  $x$ . This is much like the behavior of *power laws*, which appear ubiquitously in many parts of physics: indeed if

$$p(x) = \alpha x^\beta \quad (5.129)$$

( $\alpha$  and  $\beta$  are fixed parameters,  $\beta$  is a negative exponent), then

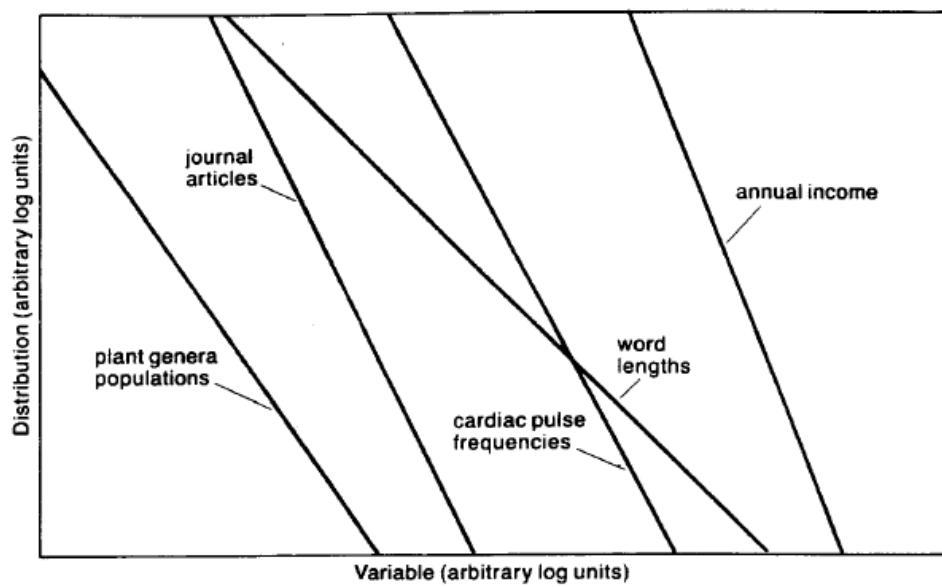
$$\ln p(x) = \ln \alpha + \beta \ln x \quad (5.130)$$

and the power law is a straight line in a log-log plot.

An example of a strange and puzzling power law is the *Zipf's law*, which states that the frequency of words is inversely proportional to their length (see, e.g., figure 5.5).

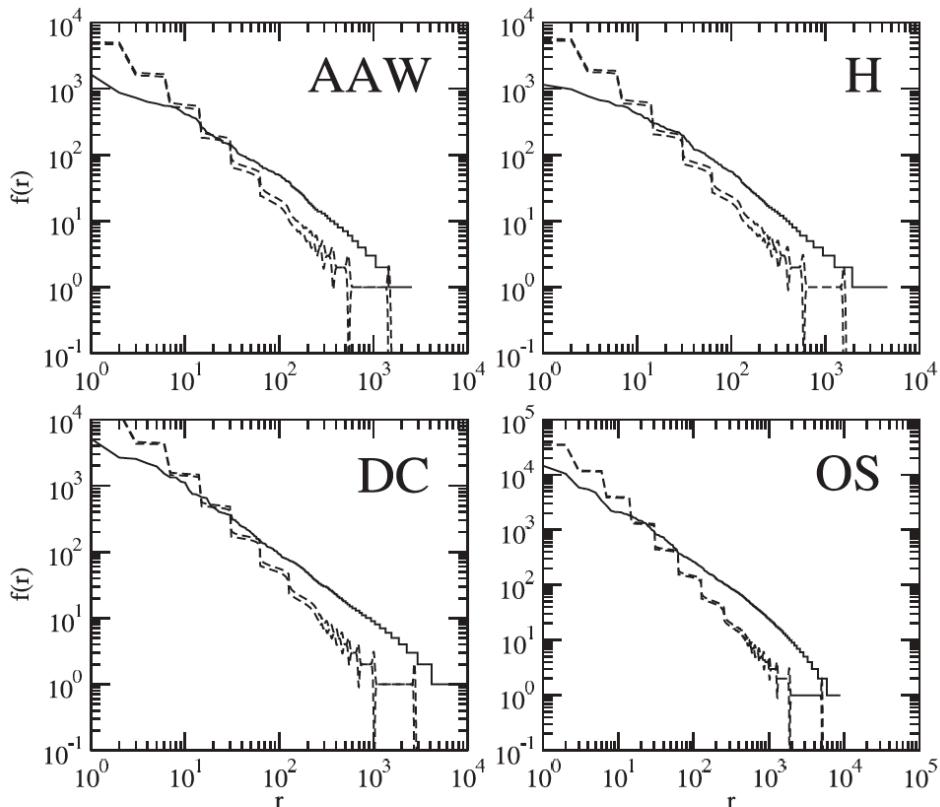
## 5.4 What's the use of all this?

Characteristic functions and generating functions have a wide applicability in many sectors of statistics and probability. The CLT is a basic theorem that is applicable to many situations that actually occur in physics.



**Figure 3.** Inverse power laws govern, at least in part, a surprising number of sociobiological phenomena. Five disparate phenomena are plotted above on logarithmic axes with arbitrary units so that an inverse power law results in a straight line. From the left: the populations within plant genera plotted against the number of genera with such populations (16); the number of journal articles published plotted against the number of scientists publishing that many (17); the frequencies contained in a cardiac pulse plotted against the occurrence of those frequencies (18); word length in the English language plotted against the usage of each word (19); income plotted against the number of people with such income (20). As opposed to Shockley, Lotka (17) observed an inverse power law behavior in scientific publishing because he used a larger sampling.

Figure 5.4: A short list of power laws (taken from [33]).



**Figure 1. The rank histograms of English texts versus that of random texts ( $RT_1$ ).** A comparison of the real rank histogram (thin black line) and two control curves with the  $3\sigma$  upper and lower bounds of the expected histogram of a random text of the same length in words (dashed lines) involving four English texts.  $f(r)$  is the frequency of the word of rank  $r$ . For the random text we use the model  $RT_1$  with alphabet size  $N=2$ . The expected histogram of the random text is estimated averaging over the rank histograms of  $10^4$  random texts. For ease of presentation, the expected histogram is cut off at expected frequencies below 0.1. AAW: Alice's adventures in wonderland. H: Hamlet. DC: David Crockett. OS: The origin of species.  
doi:10.1371/journal.pone.0009411.g001

Figure 5.5: Comparison of word distributions in several texts (taken from [13]).



# Chapter 6

## Markov processes

In this chapter I give a short introduction to both discrete and continuous time Markov processes. We start with the discrete time Markov processes, which are better known as *Markov chains*<sup>1</sup>.

### 6.1 Transition Probabilities

Consider a system such that

- the system can occupy a finite or countably infinite set of states  $S_n$ ;
- the system changes state randomly at discrete times  $t = 1, 2, \dots$ ;
- if the system is in state  $S_i$ , then the probability that the system goes into state  $S_j$  is

$$p_{ij} = P[S(n+1) = S_j | S(n) = S_i] \quad i, j = 1, 2, \dots$$

i.e., this probability depends only on the previous state, and is independent of all previous states (this is the *Markov property*);

- the *transition probabilities*  $p_{ij}$  do not depend on time  $n$ .

Such a system is a special type of discrete time stochastic process, which is called *Markov chain*.

The assumption that the transition probabilities do not depend on time is not strictly necessary, however many interesting processes fall in this category of *homogeneous Markov chains*.

Consider the following example, taken from Kemeny, Snell, and Thompson *Introduction to Finite Mathematics*, 3rd ed. (Prentice-Hall, 1974): in

---

<sup>1</sup>These notes are partly based on ch. 6, ref. [27]; additional material is taken from [5, 6, 16]. Please note that the notation used in these notes differs slightly from the notation in [27].

in the Land of Oz they never have two nice days in a row, rather, after a sunny day it either rains or snows. If they have a nice day, they are just as likely to have snow as rain the next day. If they have snow or rain, they have an even chance of having the same the next day. If there is change from snow or rain, only half of the time is this a change to a nice day. This means – if we denote the three states with the symbols N (Nice), R (Rain), or S (Snow) – that the transition probabilities are:

$$\begin{aligned} p_{NN} &= 0; & p_{NR} &= 1/2; & p_{NS} &= 1/2 \\ p_{RN} &= 1/4; & p_{RR} &= 1/2; & p_{RS} &= 1/4 \\ p_{SN} &= 1/4; & p_{SR} &= 1/4; & p_{SS} &= 1/2 \end{aligned}$$

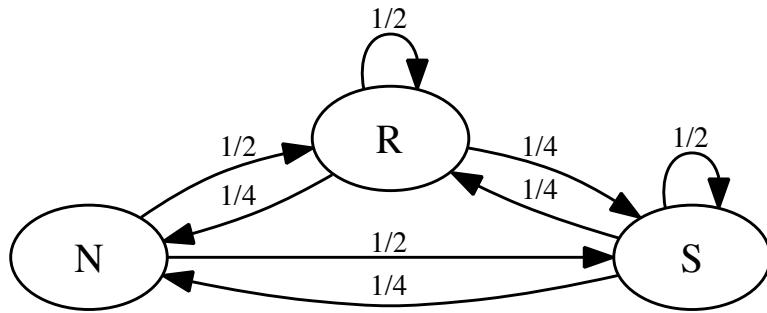


Figure 6.1: Directed graph representation of the states and transition probabilities of the three-state Markov chain that describes weather in the Land of Oz.

The whole system can be represented as a directed graph (see figure 6.1), however there is a more useful representation: we can set up a matrix of transition probabilities (also called *transition kernel*):

$$\mathbf{P} = \begin{pmatrix} p_{NN} & p_{NR} & p_{NS} \\ p_{RN} & p_{RR} & p_{RS} \\ p_{SN} & p_{SR} & p_{SS} \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$$

This is a *row stochastic matrix*, i.e., all rows are such that  $\sum_j p_{ij} = 1$ . There are also *column stochastic matrices* (all columns are such that  $\sum_i p_{ij} = 1$ ), and *doubly stochastic matrices* (both row and column stochastic). Notice that doubly stochastic matrices are necessarily square: indeed if such a matrix has  $n$  rows and  $m$  columns, then

$$\sum_{i=1}^n \sum_{j=1}^m p_{ij} = \sum_{i=1}^n 1 = n \tag{6.1}$$

and

$$\sum_{j=1}^m \sum_{i=1}^n p_{ij} = \sum_{j=1}^m 1 = m \quad (6.2)$$

however the two double sums are equal and therefore  $m = n$ .

Now let  $\pi_i^{(n)} = P[S(n) = S_i]$  be the probability that at time  $t = n$  the system is in state  $S_i$ , and since the probabilities of being in any given state are mutually exclusive, we can write

$$\begin{aligned} \pi_j^{(n+1)} &= \sum_i P[S(n+1) = S_j | S(n) = S_i] P[S(n) = S_i] \\ &= \sum_i p_{ij} \pi_i^{(n)} \end{aligned} \quad (6.3)$$

Equation (6.3) can be rewritten in matrix form, defining the *probability vector*<sup>2</sup>  $\boldsymbol{\pi}^{(n)} = \{\pi_i^{(n)}\}$ , and using the *transition probability matrix* defined above  $\mathbf{P} = \{p_{ij}\}$ :

$$\boldsymbol{\pi}^{(n+1)} = \boldsymbol{\pi}^{(n)} \mathbf{P} \quad (6.4)$$

Thus we find a recursive formula for the probability distribution (the probability vector) at the  $n$ -th step, and to solve it we must define an initial probability vector  $\boldsymbol{\pi}(0)$ . Clearly the solution involves the powers of the transition probability matrix:

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} \mathbf{P}^n \quad (6.5)$$

The matrix  $\mathbf{P}^n$  is also called the *n-step transition kernel*, and its entries are the probabilities  $p_{ij}^{(n)}$  of a transition from state  $i$  to state  $j$  in  $n$  steps.

---

<sup>2</sup>Obviously, the probability vector obeys the normalization condition  $\sum_i \pi_i^{(n)} = 1$ .

For example, if we take the transition probability matrix for the weather in the Land of Oz, we find:

$$\begin{aligned}\mathbf{P}^2 &= \begin{pmatrix} 0.25 & 0.375 & 0.375 \\ 0.1875 & 0.4375 & 0.375 \\ 0.1875 & 0.375 & 0.4375 \end{pmatrix} \\ \mathbf{P}^5 &= \begin{pmatrix} 0.199219 & 0.400391 & 0.400391 \\ 0.200195 & 0.400391 & 0.399414 \\ 0.200195 & 0.399414 & 0.400391 \end{pmatrix} \\ \mathbf{P}^{10} &= \begin{pmatrix} 0.200001 & 0.4 & 0.4 \\ 0.2 & 0.400001 & 0.4 \\ 0.2 & 0.4 & 0.400001 \end{pmatrix} \\ \mathbf{P}^{20} &= \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix} \\ \mathbf{P}^{100} &= \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}\end{aligned}$$

The matrix power converges to a fixed matrix, and the rows are all equal. This means that whatever the initial probability vector, the distribution converges to that specified by any one row: indeed, if  $\mathbf{P}^n \xrightarrow[n \rightarrow \infty]{} \mathbf{P}_\infty$  such that  $(\mathbf{P}_\infty)_{i,j} = f_j$ , then

$$\sum_i \pi_i^{(0)} (\mathbf{P}_\infty)_{i,j} = \sum_i \pi_i^{(0)} f_j = f_j \quad (6.6)$$

### 6.1.1 The discrete-time, discrete-space random walk

An important example of a Markov chain is the random walk. Consider a 1D string of sites where a particle can jump, moving either left or right at each time step. The state of the system is given by the position of the particle, and the probabilities are

$$p_{i,i+1} = p; \quad p_{i,i-1} = q$$

and 0 otherwise. This process is regulated by a binomial distribution, and if after a total of  $n$  steps,  $k$  is the number of steps to the right, then  $n - k$  is the number of steps to the left, and a random walker that starts at the origin reaches position  $x = k - (n - k) = 2k - n$ . Then, it is easy to see that the mean position of a particle that starts at the origin, after  $n$  steps is  $n(p - q) = n(2p - 1)$ , while the variance of the position is  $4npq$ . Figure 6.2 shows a pair of realizations of the random walk. Figure 6.3 shows the graph representation of the states and of the transition probabilities.

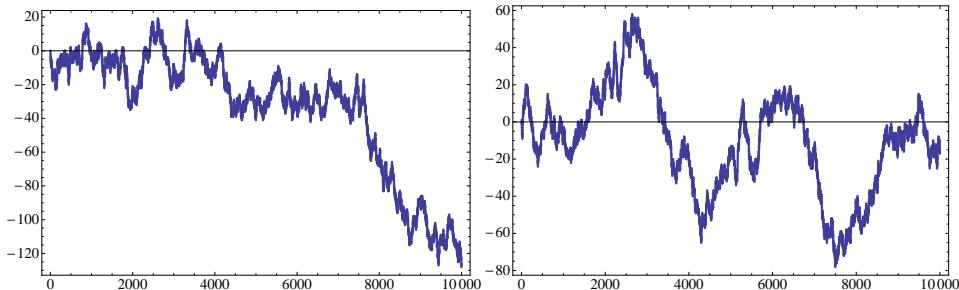


Figure 6.2: Plots of two random walks with  $p = q = 1/2$  (position vs. timestep), for 10000 time steps. Generic representations such as this are called *sample paths*. The Mathematica code used to generate these sample paths is listed in Appendix A.

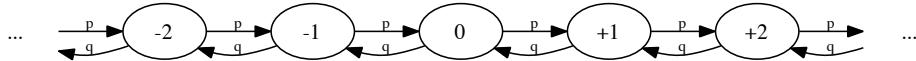


Figure 6.3: Graph representation of the states and transition probabilities of the 1D random walk.

## 6.2 The Ehrenfest urn model

In the Ehrenfest urn model there are two urns and  $n$  balls. The balls are numbered, and at each step one ball is chosen at random, and is placed in the other urn. The state of the system is the number of balls in one – fixed – urn. Figure 6.4 shows a pair of simulations of the model: the system is started in an unbalanced situation – all balls are either in one or in the other urn – and after a while the balls are evenly distributed in the two boxes, and the system fluctuates about this equilibrium condition.

If the system has  $k$  balls in the first urn (i.e., it is in state  $k$ ), then the probability of extracting one ball from that urn is  $k/n$ , therefore the transition probabilities are

$$p_{k,k-1} = \frac{k}{n} \quad p_{k,k+1} = \frac{n-k}{n} \quad (6.7)$$

and 0 otherwise.

This model was proposed by the Ehrenfest's in an effort to understand irreversibility and approach to equilibrium in statistical mechanics [9].

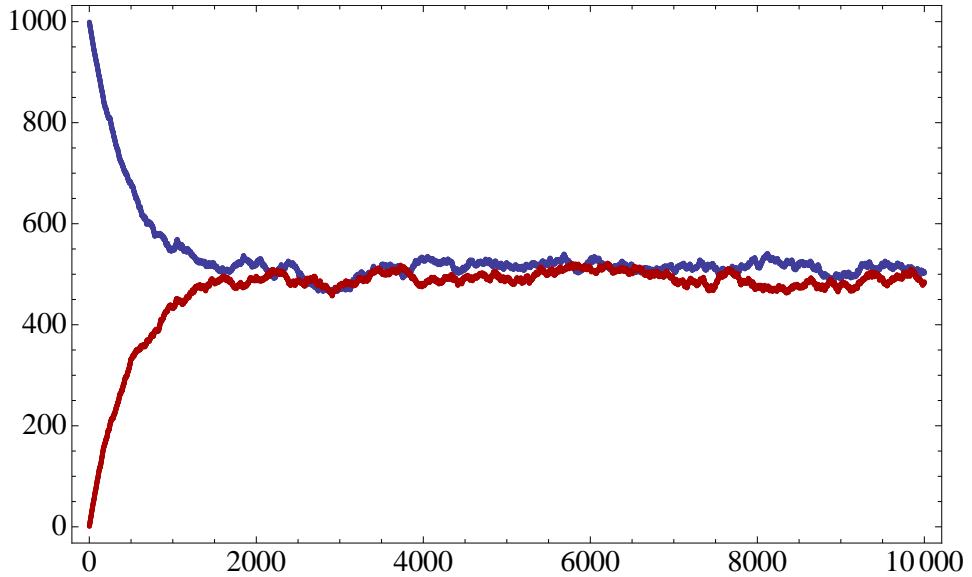


Figure 6.4: Simulations of the Ehrenfest urn model with 1000 balls, 10000 time steps (number of balls in the first urn vs. time). The system is started in two different out-of-equilibrium conditions: all balls in the first urn (blue curve), no balls in the first urn (red curve). In both cases the system reaches equilibrium and fluctuates about it. The Mathematica code used to generate these sample paths is listed in Appendix C.

### 6.3 Classification of States. Persistent and Transient States

The system may or may not return to a given state, and more generally we can classify states as in table 1. Figure 6.5 illustrates the meaning of transient and persistent states in a finite Markov chain. The same figure also shows that state D leads to state E (E is *accessible* from D) but not vice versa, while E leads to F and F leads to E: we say that E and F *communicate*, while D and E do not communicate. On the other hand D leads to A, and A leads to D (after 3 steps), thus A communicates with D. In general we say that two states  $i, j$  communicate if after a finite number of steps  $i$  leads to  $j$  and  $j$  leads to  $i$ . The communication between states establishes equivalence classes: the Markov chain of figure 6.5 has two equivalence classes: (A, B, C, D) and (E, F). A Markov chain with just one class, such that all states communicate, is said to be *irreducible*. A state is periodic if it is persistent, and if the system returns to it at times  $t, 2t, 3t, \dots$ , where  $t > 1$ . Clearly all the periodic states in the same equivalence class have the same period. For the same reason all the periodic states in an irreducible Markov chain

have the same period.

Type of state	Definition of state (assuming, where applicable, that the state is initially occupied)
Periodic	Return to state possible only at times $t, 2t, 3t, \dots$ , where $t > 1$
Aperiodic	Not periodic
Recurrent/Persistent	Eventual return to state certain
Transient	Eventual return to state uncertain
Ephemeral	Is a state $j$ such that $p_{ij} = 0$ for every $i$
Positive-recurrent	Recurrent/persistent, finite mean recurrence time
Null-recurrent	Recurrent, infinite mean recurrence time
Ergodic	Aperiodic, positive-recurrent

Table 6.1: Classification of states in a Markov chain (from Cox and Miller).

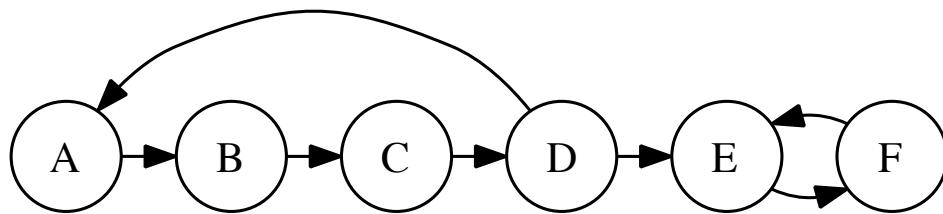


Figure 6.5: Persistent and transient states. This graph represents the states and the transition probabilities of a finite Markov chain with 6 states. The arrows correspond to nonzero transition probabilities. If the chain starts with any one of states A, B, C or D, it can loop around these four states until a transition  $D \rightarrow E$  occurs, then the system is locked in the E-F loop. States A, B, C, and D are transient, while states E and F are persistent (and periodic, with period 2). This Markov chain is not irreducible.

As shown by the example in figure 6.5, a persistent state is characterized by an infinite number of returns. Now start from state  $i$  and let  $u_n = p_{ii}^{(n)}$  be the probability of a return after  $n$  steps. Also, let  $v_n$  be the probability of a return after *precisely*  $n$  steps. Then, we can write

$$\begin{aligned}
u_n &= u_0 v_n + u_1 v_{n-1} + u_2 v_{n-2} + \cdots + u_{n-2} v_2 + u_{n-1} v_1 + u_n v_0 \\
&= \sum_{k=0}^n u_k v_{n-k} \quad (6.8)
\end{aligned}$$

for  $n \geq 1$  (for  $n = 0$  the equation is simply  $u_0 = u_0$ ), where we set  $u_0 = 1$  and  $v_0 = 0$  by definition. Thus we have a convolution between the sequence of  $u$ 's and  $v$ 's, and if we use the corresponding generating functions

$$U(z) = \sum_{k=0}^{\infty} u_k z^k; \quad V(z) = \sum_{k=0}^{\infty} v_k z^k \quad (6.9)$$

we can write

$$U(z) - u_0 = \sum_{n=1}^{\infty} u_n z^n = \sum_{n=0}^{\infty} \left( \sum_{k=0}^n u_k v_{n-k} \right) z^n = U(z)V(z) \quad (6.10)$$

and therefore

$$U(z) = \frac{u_0}{1 - V(z)} = \frac{1}{1 - V(z)} \quad (6.11)$$

The sum

$$v = \sum_{n=0}^{\infty} v_n \quad (6.12)$$

is the probability that the system returns sooner or later to the original state. The state is persistent if  $v = 1$ , and is transient if  $v < 1$ . Notice that if the system returns to the original state, then this applies again and therefore if  $v = 1$  there are infinite returns to the original state. Now notice that

$$v = \sum_{n=0}^{\infty} v_n = \lim_{z \rightarrow 1} V(z) \quad (6.13)$$

therefore

$$\lim_{z \rightarrow 1} U(z) = \lim_{z \rightarrow 1} \frac{1}{1 - V(z)} \quad (6.14)$$

and if the state is persistent, then

$$\lim_{z \rightarrow 1} U(z) = \sum_{n=0}^{\infty} u_n = \infty \quad (6.15)$$

It is also rather obvious that the reverse holds as well. We can see what this means also from a different point of view: if the condition

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \sum_{n=0}^{\infty} u_n < \infty \quad (6.16)$$

holds (the series converges), then

$$\lim_{z \rightarrow 1} U(z) = \sum_{n=0}^{\infty} u_n < \infty \quad (6.17)$$

and

$$v = \lim_{z \rightarrow 1} V(z) \neq 1 \quad (6.18)$$

and since  $v \leq 1$ , it must be  $v < 1$ . Thus

*Theorem: a state is persistent if and only if  $\sum_{n=0}^{\infty} u_n = \infty$ .*

We can easily prove a few more theorems on persistent and transient states. Notice first that if the probability of return is  $p_1 = v$ , because of the time invariance of the description of the chain, the probability of returning twice is just  $p_2 = v^2$  and in general the probability of returning  $k$  times is  $p_k = v^k$ . Therefore

$$\sum_{k=1}^{\infty} p_k = \sum_{k=1}^{\infty} v^k \quad (6.19)$$

and if  $v < 1$

$$\sum_{k=1}^{\infty} p_k = \sum_{k=1}^{\infty} v^k = \frac{1}{1-v} < \infty \quad (6.20)$$

and then – from the first Borel-Cantelli lemma – we know that only a finite number of return events can take place, i.e., if a state is transient there is only a finite number of returns. On the other hand, this does not hold the state is persistent, i.e., if  $v = 1$ . In this case the sum diverges, and since the return events are independent from one another, the second Borel-Cantelli lemma applies and thus a persistent state recurs an infinite number of times.

It is also quite easy to see that if a state  $i$  is accessible from a persistent state  $j$ , then the persistent state is also accessible from  $i$ , and  $i$  is persistent as well. Indeed if  $j$  were not accessible from  $i$ , as soon as the system would be in state  $i$  it could no longer reach  $j$ , which could no longer recur: thus  $j$  would not be persistent.

Moreover since  $i$  is accessible from  $j$  there must be an integer  $M$  such that  $p_{ji}^{(M)} > 0$ , and likewise, since  $j$  is accessible from  $i$  there must be an  $N$  such that  $p_{ij}^{(N)} > 0$ . This means that

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} \geq \sum_{n=M+N}^{\infty} p_{ii}^{(n)} = \sum_{n=0}^{\infty} p_{ii}^{(n+M+N)} \quad (6.21a)$$

$$= \sum_{n=0}^{\infty} [\mathbf{P}^{n+M+N}]_{ii} \geq \sum_{n=0}^{\infty} p_{ij}^{(N)} p_{jj}^{(n)} p_{ji}^{(M)} \quad (6.21b)$$

$$= p_{ij}^{(N)} p_{ji}^{(M)} \sum_{n=0}^{\infty} p_{jj}^{(n)} = \infty \quad (6.21c)$$

and state  $i$  is persistent as well.

Now notice that in a finite Markov chain there must exist persistent states, because the system must return to at least one of them infinitely often. Then if a finite chain is irreducible, so that all states are accessible from any other state, all states are persistent.

### 6.3.1 Example: the simple random walk

A simple 1D random walk has infinite states and is irreducible. Moreover a return can take place only after an even number of time steps, and the probability of taking  $n$  positive and  $n$  negative steps is (for any  $i$ )

$$\begin{aligned} p_{ii}^{(2n)} &= \binom{2n}{n} p^n q^n = \frac{(2n)!}{(n!)^2} p^n q^n \approx \frac{\sqrt{4n\pi}(2n)^{2n} e^{-2n}}{2n\pi n^{2n} e^{-2n}} p^n q^n \\ &= \frac{1}{\sqrt{\pi n}} (4pq)^n \end{aligned} \quad (6.22)$$

for large  $n$ . Thus

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = 1 + \sum_{n=1}^{\infty} p_{ii}^{(2n)} \approx 1 + \sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}} (4pq)^n \quad (6.23)$$

Now notice that

$$4pq = (p+q)^2 - (p-q)^2 = 1 - (2p-1)^2 \quad (6.24)$$

and clearly

$$0 \leq 4pq \leq 1 \quad (6.25)$$

with  $4pq = 1$  if  $p = q = 1/2$ . Then, if  $p = q = 1/2$

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} \approx 1 + \sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}} \geq 1 + \frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{1}{n} = \infty \quad (6.26)$$

and all states are persistent.

However if  $p \neq q$ ,

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} \approx 1 + \sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}} (4pq)^n < \sum_{n=0}^{\infty} (4pq)^n = \frac{1}{1-4pq} < \infty \quad (6.27)$$

and all states are transient.

### 6.3.2 Example: higher dimensional random walks on a lattice

We turn now to higher dimensional random walks.

## 2-dimensional random walks

Again, there can be a return to the origin only when the number of steps is even, but now the random walker can take steps both in the  $x$  and the  $y$  direction, and in each direction the number of positive steps must equal the number of negative steps. With a total of  $2n$  steps, there are  $\binom{2n}{2n_x}$  ways of choosing  $2n_x$  steps along  $x$  and  $2n_y = 2n - 2n_x$  steps along  $y$ , therefore the probability of returning to the site at  $(x, y)$  at step  $2n$  is<sup>3</sup>

$$p^{(2n)}(x, y) = \sum_{\substack{n_x=0,n \\ n_y=n-n_x}} \binom{2n}{2n_x} \binom{2n_x}{n_x} \binom{2n_y}{n_y} \frac{1}{4^{2n}} = \sum_{\substack{n_x=0,n \\ n_y=n-n_x}} \frac{(2n)!}{(n_x!)^2 (n_y!)^2} \frac{1}{4^{2n}} \quad (6.28a)$$

$$= \frac{(2n)!}{(n!)^2} \frac{1}{4^{2n}} \sum_{n_x=0}^n \left[ \frac{n!}{n_x!(n-n_x)!} \right]^2 = \binom{2n}{n} \frac{1}{4^{2n}} \sum_{n_x=0}^n \left[ \binom{n}{n_x} \right]^2 \quad (6.28b)$$

$$= \left[ \binom{2n}{n} \right]^2 \frac{1}{4^{2n}} \quad (6.28c)$$

This is just the square of the probabilities that we computed in the case of 1D random walks, therefore the same expansion with the Stirling approximation applies, and we find

$$\sum_{n=0}^{\infty} p^{(2n)}(x, y) \approx 1 + \sum_{n=1}^{\infty} \frac{1}{\pi n} = \infty \quad (6.29)$$

and once again we conclude that the probability of eventual return is 1 and that there is an infinity of returns.

## 3- and higher dimensional random walks

Similarly, the return probability to any position  $(x, y, z)$  in a 3-dimensional random walk on a cubic lattice is

$$p^{(2n)}(x, y, z) = \sum_{n_x+n_y+n_z=n} \frac{(2n)!}{(n_x!)^2 (n_y!)^2 (n_z!)^2} \frac{1}{6^{2n}} \quad (6.30a)$$

$$= \binom{2n}{n} \sum_{n_x=0}^n \sum_{n_y=0}^{n_x} \left[ \frac{n!}{(n_x!)(n_y!)(n-n_x-n_y)!} \right]^2 \frac{1}{6^{2n}} \quad (6.30b)$$

---

<sup>3</sup>Here we use the binomial identity

$$\sum_{k=0}^n \left[ \binom{n}{k} \right]^2 = \binom{2n}{n}$$

Prove it as an exercise in combinatorics (or look at the proof in Appendix A).

In the higher-dimensional case (dimension  $d$ ) we extend this calculation in a straightforward way:

$$p^{(2n)}(x_1, \dots, x_d) = \sum_{\sum_{k=1,d} n_k = n} \frac{(2n)!}{\prod_{k=1,d} (n_k!)^2} \frac{1}{(2d)^{2n}} \quad (6.31a)$$

$$= \binom{2n}{n} \sum_{\sum_{k=1,d} n_k = n} \left[ \frac{n!}{\prod_{k=1,d} (n_k!)^2} \right]^2 \frac{1}{(2d)^{2n}} \quad (6.31b)$$

Since the logarithm of the multinomial coefficient is

$$s = \ln \left[ \frac{n!}{\prod_{k=1,d} (n_k!)^2} \right] \approx n \ln n - n - \sum_{k=1,d} (n_k \ln n_k - n_k) \quad (6.32a)$$

$$= n \ln n - \sum_{k=1,d} n_k \ln n_k \quad (6.32b)$$

and this can be maximized with one Lagrange multiplier  $\lambda$ , which takes care of the constraint  $\sum_{k=1,d} n_k = n$ :

$$\frac{\partial}{\partial n_i} \left[ s + \lambda \left( \sum_{k=1,d} n_k - n \right) \right] = -\ln n_i - 1 + \lambda = 0 \quad (6.33)$$

we find  $n_i = \exp(\lambda - 1)$ , i.e.,  $n = d \exp(\lambda - 1)$ , and the multinomial coefficient is maximized when all the  $n_i$ 's have the same value  $n_i = n/d$ , so that the maximum is

$$\frac{n!}{[(n/d)!]^d} \approx \frac{\sqrt{2\pi n} n^n e^{-n}}{(\sqrt{2\pi n/d} (n/d)^{(n/d)} e^{-n/d})^d} = \frac{d^{d/2}}{(2\pi n)^{(d-1)/2}} d^n \quad (6.34)$$

Therefore:

$$\begin{aligned} p^{(2n)}(x_1, \dots, x_d) &< \\ &< \binom{2n}{n} \sum_{\sum_{k=1,d} n_k = n} \left[ \frac{d^{d/2}}{(2\pi n)^{(d-1)/2}} d^n \right] \left[ \frac{n!}{\prod_{k=1,d} (n_k!)^2} \right] \frac{1}{(2d)^{2n}} \end{aligned} \quad (6.35)$$

Since

$$\sum_{\sum_{k=1,d} n_k = n} \frac{n!}{\prod_{k=1,d} (n_k!)^2} = d^n \quad (6.36)$$

and

$$\binom{2n}{n} \approx \frac{4^n}{\sqrt{\pi n}} \quad (6.37)$$

we obtain eventually

$$p^{(2n)}(x_1, \dots, x_d) < \frac{4^n}{\sqrt{\pi n}} \left[ \frac{d^{d/2}}{(2\pi n)^{(d-1)/2}} d^n \right] d^n \frac{1}{(2d)^{2n}} = \sqrt{\frac{2}{(2\pi)^d}} \frac{d^{d/2}}{n^{d/2}} \quad (6.38)$$

and the sum

$$\sum_n p^{(2n)}(x_1, \dots, x_d) < \sum_n \sqrt{\frac{2}{(2\pi)^d}} \frac{d^{d/2}}{n^{d/2}} \quad (6.39)$$

converges for all  $d > 2$ . Therefore all states of the  $d$ -dimensional random walks on cubic lattices with  $d > 2$  are transient.

## 6.4 Limiting Probabilities and Stationary Distributions

We have found earlier – in the example of the weather in the Land of Oz – that the  $n$ -step transition probability matrix converges to a matrix with rows that are all equal, and that each one of these rows represents the final probability vector, for any choice of the initial probability vector. Now we prove that these features are very general, subject only to few restrictive conditions.

Consider a finite Markov chain, and assume that for some  $N$

$$\min_{i,j} p_{ij}^{(N)} = \delta > 0 \quad (6.40)$$

so that the chain is also irreducible.

Let

$$r_j^{(n)} = \min_i p_{ij}^{(n)}; \quad R_j^{(n)} = \max_i p_{ij}^{(n)} \quad (6.41)$$

be the minimum and the maximum values of the  $j$ -th column vector in the  $n$ -step transition matrix  $\mathbf{P}^n$ , then

$$r_j^{(n+1)} = \min_i p_{ij}^{(n+1)} = \min_i \mathbf{P}_{ij}^{n+1} = \min_i (\mathbf{P}\mathbf{P}^n)_{ij} = \min_i \sum_k p_{ik} p_{kj}^{(n)} \quad (6.42a)$$

$$\geq \min_i \sum_k p_{ik} r_j^{(n)} = r_j^{(n)} \quad (6.42b)$$

and similarly

$$R_j^{(n+1)} = \max_i p_{ij}^{(n+1)} = \max_i \mathbf{P}_{ij}^{n+1} = \max_i (\mathbf{P}\mathbf{P}^n)_{ij} = \max_i \sum_k p_{ik} p_{kj}^{(n)} \quad (6.43a)$$

$$\leq \max_i \sum_k p_{ik} R_j^{(n)} = R_j^{(n)} \quad (6.43b)$$

This means that as  $n$  grows, the minimum and the maximum values in a column vector get closer and closer (the components of the column vector get closer and closer).

How close do the  $r_j^{(n)}$  and the  $R_j^{(n)}$  get for larger and larger  $n$ ? Do they eventually approach the same limit? This question is answered by the difference

$$R_j^{(n)} - r_j^{(n)} = \max_i p_{ij}^{(n)} - \min_k p_{kj}^{(n)} = \max_{i,k} [p_{ij}^{(n)} - p_{kj}^{(n)}] \quad (6.44)$$

In addition, shifting the index  $n \rightarrow n + N$  the same difference writes

$$\begin{aligned} R_j^{(n+N)} - r_j^{(n+N)} &= \max_{i,k} [p_{ij}^{(n+N)} - p_{kj}^{(n+N)}] \\ &= \max_{i,k} \left\{ \sum_l [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} \right\} \end{aligned} \quad (6.45)$$

Now notice that by splitting the sum over the positive and negative values of the difference in square brackets, and denoting by  $\sum^+$  ( $\sum^-$ ) the sum over positive (negative) values we find

$$\begin{aligned} \sum_l [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} &= \sum_l^+ [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} + \sum_l^- [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} \\ &\quad (6.46a) \end{aligned}$$

$$\begin{aligned} &\leq \sum_l^+ [p_{il}^{(N)} - p_{kl}^{(N)}] R_j^{(n)} + \sum_l^- [p_{il}^{(N)} - p_{kl}^{(N)}] r_j^{(n)} \\ &\quad (6.46b) \end{aligned}$$

Now consider the structure of the positive sum

$$\sum_l^+ [p_{il}^{(N)} - p_{kl}^{(N)}] = \sum_l^+ p_{il}^{(N)} - \sum_l^+ p_{kl}^{(N)} \leq \sum_l^+ p_{il}^{(N)} - \delta = 1 - \delta \quad (6.47)$$

Clearly the same argument applies to the negative sum, and therefore

$$\sum_l [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} < (1 - \delta) [R_j^{(n)} - r_j^{(n)}] \quad (6.48)$$

Therefore if we take strides of  $N$  steps at a time, we find

$$R_j^{(kN)} - r_j^{(kN)} < (1 - \delta)^k [R_j^{(N)} - r_j^{(N)}] \xrightarrow{k \rightarrow \infty} 0 \quad (6.49)$$

because  $0 < (1 - \delta) < 1$ . Then the difference between maximum and minimum vanishes, the column values converge to a single value  $p_j^*$ , and the

matrix structure is the same as in the example (i.e., all rows are equal).

Then we can write

$$p_{ij}^* = \lim_{n \rightarrow \infty} [\mathbf{P}^n]_{ij} = p_j^* \quad (6.50)$$

and therefore, if we start from the initial probability vector  $\pi_j^{(0)}$ , we find the asymptotic probability vector

$$\pi_j^* = \sum_k \pi_k^{(0)} p_{kj}^* = \sum_k \pi_k^{(0)} p_j^* = p_j^* \quad (6.51)$$

i.e., the asymptotic probability vector coincides with the rows of the asymptotic transition probability matrix, and does not depend on the initial probability vector.

This asymptotic distribution is stable, indeed from

$$\pi_j^{(n)} = \sum_k \pi_k^{(n-1)} p_{kj} \quad (6.52)$$

we find

$$[\pi^* \mathbf{P}]_j = \sum_k \pi_k^* p_{kj} = \sum_k p_k^* p_{kj} = \sum_k p_{ik}^* p_{kj} = p_{ij}^* = p_j^* = \pi_j^* \quad (6.53)$$

or, in matrix form

$$\pi^* = \pi^* \mathbf{P} \quad (6.54)$$

i.e., the asymptotic probability vector is the left eigenvector with eigenvalue 1 of the transition probability matrix. The distribution expressed by the probability vector  $\pi^*$  is called *invariant distribution* or *stationary distribution*.

## 6.5 Time reversal. Detailed balance

The chain can be time-reversed, and we find

$$\begin{aligned} P[S(n) = S_i | S(n+1) = S_j] \\ = P[S(n+1) = S_j | S(n) = S_i] \frac{P[S(n) = S_i]}{P[S(n+1) = S_j]} \end{aligned} \quad (6.55)$$

i.e.,

$$P[S(n) = S_i | S(n+1) = S_j] = p_{ij} \frac{\pi_i^{(n)}}{\pi_j^{(n+1)}} \quad (6.56)$$

which shows that the time-reversed process is time-dependent (the reversed chain is inhomogeneous even if the original chain is homogeneous). However if states are distributed according to the invariant distribution, we have

$$P[S(n) = S_i | S(n+1) = S_j] = p_{ij} \frac{\pi_i^*}{\pi_j^*} \quad (6.57)$$

which means that the backward transition probabilities are again time-independent, and in particular they must coincide with the forward transition probabilities, i.e.,

$$p_{ji}\pi_j^* = p_{ij}\pi_i^* \quad (6.58)$$

this condition is called *detailed balance*.

We have found that if the probability distribution of states is stationary, then the detailed balance condition holds. The converse is also true: indeed

$$\pi_j^{(n+1)} = \sum_i \pi_i^{(n)} p_{ij} = \sum_i \pi_j^{(n)} p_{ji} = \pi_j^{(n)} \sum_i p_{ji} = \pi_j^{(n)} \quad (6.59)$$

i.e., if the detailed balance condition holds, then the probability distribution of states is stationary.

## 6.6 Continuous time Markov processes

Just as in the discrete case we introduce the probability of finding the system in state  $S_n$  at time  $t$

$$P(S_n, t) = P[S(t) = S_n] \quad (6.60)$$

where  $t$  is now a continuous time variable, and we note that in general

$$\begin{aligned} P(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \dots; S_{i_0}, t_0) &= \\ &= P(S_{i_k}, t_k | S_{i_{k-1}}, t_{k-1}; \dots; S_{i_0}, t_0) P(S_{i_{k-1}}, t_{k-1}; \dots; S_{i_0}, t_0) \end{aligned} \quad (6.61)$$

where the time ordering is  $t_0 \leq t_1 \leq \dots \leq t_{n-1} \leq t_n$ , i.e., the state at time  $n$  depends on the system behavior at *all* previous times.

The case of a memoryless system is particularly simple, as the probability does not depend on any previous time:

$$P(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \dots; S_{i_0}, t_0) = P(S_{i_k}, t_k) \quad (6.62)$$

The continuous Markov process corresponds to the discrete Markov chain in the sense that it depends only on the latest time in the sequence:

$$P(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \dots; S_{i_0}, t_0) = P(S_{i_k}, t_k | S_{i_{k-1}}, t_{k-1}) P(S_{i_{k-1}}, t_{k-1}) \quad (6.63)$$

Now notice that, given  $t_1 \leq t \leq t_2$ ,

$$\begin{aligned} P(S_{i_2}, t_2; S_{i_1}, t_1) &= P(S_{i_2}, t_2 | S_{i_1}, t_1) P(S_{i_1}, t_1) \\ &= \sum_j P(S_{i_2}, t_2 | S_j, t) P(S_j, t | S_{i_1}, t_1) P(S_{i_1}, t_1) \end{aligned} \quad (6.64)$$

i.e.,

$$P(S_{i_2}, t_2 | S_{i_1}, t_1) = \sum_j P(S_{i_2}, t_2 | S_j, t) P(S_j, t | S_{i_1}, t_1) \quad (6.65)$$

which is the *Chapman-Kolmogorov equation*: the Chapman-Kolmogorov equation is an integral relationship between *transition probabilities*.

Now we note that the following equation also holds

$$\begin{aligned} P(S_n, t + \Delta t) &= P(S_n, t) + \\ &+ \sum_j [P(S_n, t + \Delta t | S_j, t) P(S_j, t) - P(S_j, t + \Delta t | S_n, t) P(S_n, t)] \end{aligned} \quad (6.66)$$

which is called the *master equation*, and if we assume that the transition probabilities are time-invariant and we define the transition rates  $T$

$$P(S_n, t + \Delta t | S_j, t) = T_{n,j} \Delta t \quad (6.67)$$

we find

$$\frac{d}{dt} P(S_n, t) = \sum_j [T_{n,j} P(S_j, t) - T_{j,n} P(S_n, t)] \quad (6.68)$$

which is a differential form of the master equation.

As an example, consider the 1D random walk on a discrete array of sites on the real line, so that  $S_n$  represents the position  $n\Delta x$ , and let  $T_{n,n\pm 1} = (1/2\Delta t)$  while all the other transition probabilities vanish. Then

$$\frac{d}{dt} P(n\Delta x, t) = \frac{1}{2\Delta t} [P((n+1)\Delta x, t) + P((n-1)\Delta x, t) - 2P(n\Delta x, t)] \quad (6.69)$$

and if we let  $x = n\Delta x$  and  $\Delta x \rightarrow 0$ ,  $\Delta t \rightarrow 0$  we find the *diffusion equation*

$$\frac{\partial}{\partial t} P(x, t) = D \frac{\partial^2}{\partial x^2} P(x, t) \quad (6.70)$$

where

$$D = \lim_{\substack{\Delta x \rightarrow 0 \\ \Delta t \rightarrow 0}} \frac{(\Delta x)^2}{\Delta t} \quad (6.71)$$

is the *diffusion coefficient*. The diffusion equation is a special form of *Fokker-Planck equation*. Moreover we notice that a nonzero, finite, diffusion coefficient, requires  $(\Delta x)^2 = O(\Delta t)$ .

### 6.6.1 The H-theorem and approach to equilibrium

Using the master equation we can prove a version of Boltzmann's H-theorem which demonstrates the entropy increase in continuous-time Markov processes. We let  $\pi_i(t) = P(S_i, t)$ , and rewrite the master equation in the more compact form

$$\frac{d\pi_n}{dt} = \sum_j [T_{n,j}\pi_j(t) - T_{j,n}\pi_n(t)] \quad (6.72)$$

We also assume that the physics of microscopic processes is time-reversible, and therefore  $T_{n,j} = T_{j,n}$ : it follows that the master equation takes the simpler form

$$\frac{d\pi_n}{dt} = \sum_j T_{n,j} [\pi_j(t) - \pi_n(t)] \quad (6.73)$$

and that at equilibrium it yields the equality

$$\sum_j T_{n,j} (\pi_j^* - \pi_n^*) = 0 \quad (6.74)$$

Since the equality (6.74) holds for any set of transition rates, it implies  $\pi_j^* = \pi_n^*$  for any pair  $(j, n)$ , i.e., the states  $S$  are all equally probable.

Now we turn our attention to the following sum

$$H = \sum_n \pi_n \ln \pi_n \quad (6.75)$$

which is proportional to Gibbs' entropy, and which we shall meet again in due time. Using the master equation (6.73), we find a differential equation for  $H$ :

$$\begin{aligned} \frac{dH}{dt} &= \sum_n \frac{d}{dt} (\pi_n \ln \pi_n) = \sum_n \frac{d\pi_n}{dt} (\ln \pi_n + 1) \\ &= \sum_{n,j} T_{n,j} (\pi_j - \pi_n) (\ln \pi_n + 1) \end{aligned} \quad (6.76)$$

Obviously the equation holds with exchanged indexes as well, i.e.,

$$\frac{dH}{dt} = \sum_{n,j} T_{n,j} (\pi_n - \pi_j) (\ln \pi_j + 1) \quad (6.77)$$

and therefore, by summing, we obtain

$$\frac{dH}{dt} = \frac{1}{2} \sum_{n,j} T_{n,j} (\pi_n - \pi_j) (\ln \pi_j - \ln \pi_n) \quad (6.78)$$

The logarithm is an monotonic, increasing function of its argument, therefore the products

$$(\pi_n - \pi_j) (\ln \pi_j - \ln \pi_n) \leq 0$$

and finally

$$\frac{dH}{dt} \leq 0 \quad (6.79)$$

since the transition rates are  $T_{n,j} \geq 0$ . The derivative vanishes at equilibrium, and we find that it is a stable point for  $H$ . Since  $H$  is essentially the negative of Gibbs' entropy, the theorem states that the entropy of a Markov chain increases up to a maximum which is reached at equilibrium.

## 6.7 An important application: the Hidden Markov Models (HMM)

Consider the following problem: a physical source emits photons and the emission process is Poissonian in time. This means that it is a memoryless process, where the time intervals between successive emissions follow the exponential distribution, or also, that the number of photons observed in a given time span is a Poisson variate.

For some reason we believe that the photon source switches randomly between different rates, i.e., that there are several states that correspond to the different rates, and that the switching process is described by transitions between the states of a Markov chain.

In general we only observe the photon rates, while we do not know the number of states, nor the transition rates between the states: the Markov chain is *hidden*.

A Hidden Markov Model (HMM)  $\lambda$  is defined by the matrix  $\mathbf{A}$  of transition rates, by the probability model  $\mathbf{B}$  that corresponds to each state (in this case a Poisson distribution, with the corresponding photon emission rate, but it could be another discrete or continuous distribution), and finally by the initial probability distribution  $\boldsymbol{\pi}_0$  assigned to the different states – in symbols  $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}_0)$ . In general we do not know how many hidden states are there, and we must find ways of guessing the number of states  $N$ .

In practice we observe for a total time  $T$ , and obtain sequences of symbols  $\mathbf{O} = O_1, O_2, O_3, \dots, O_T$ . Then, for a given HMM, there are three outstanding problems:

**Problem 1** : Find the probability  $P(\mathbf{O}|\lambda)$  of a given output sequence  $\mathbf{O} = O_1, O_2, O_3, \dots, O_T$ , given a model  $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}_0)$ .

**Problem 2** : Given the observation sequence  $\mathbf{O} = O_1, O_2, O_3, \dots, O_T$  and the model  $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}_0)$ , find the corresponding sequence of hidden states  $\mathbf{I} = I_1, I_2, \dots, I_T$  which is – in some sense – optimal.

**Problem 3 :** Adjust the model parameters of  $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}_0)$  to maximise  $P(\mathbf{O}|\lambda)$ .

Here we consider only the solution to the first problem. If we assume that the system goes through the sequence of states  $\mathbf{I} = I_1, I_2, \dots, I_T$ , then the probability of the given sequence of states is

$$P(I_1, I_2, \dots, I_T | \lambda) = \pi_{I_1} a_{I_1, I_2} a_{I_2, I_3} \dots a_{I_{T-1}, I_T}$$

and the probability of the observed symbols is

$$P(O_1, O_2, \dots, O_T | \mathbf{I}, \lambda) = \prod_{n=1}^T b_{I_n}(O_n)$$

where  $b_{I_n}(O_n)$  is the probability of observing  $O_n$  when the state is  $I_n$ . Therefore

$$\begin{aligned} P(O_1, O_2, \dots, O_T; I_1, I_2, \dots, I_T | \lambda) \\ = P(O_1, O_2, \dots, O_T | \mathbf{I}, \lambda) P(I_1, I_2, \dots, I_T | \lambda) \\ = \pi_{I_1} b_{I_1} a_{I_1, I_2} b_{I_2} a_{I_2, I_3} b_{I_3} \dots a_{I_{T-1}, I_T} b_{I_T} \end{aligned} \quad (6.80)$$

Finally, since the actual sequence of hidden states is unknown, we find

$$\begin{aligned} P(\mathbf{O} | \lambda) &= \sum_{\langle \mathbf{I} \rangle} P(\mathbf{O} | \mathbf{I}, \lambda) P(\mathbf{I} | \lambda) \\ &= \sum_{\langle \mathbf{I} \rangle} \pi_{I_1} b_{I_1} a_{I_1, I_2} b_{I_2} a_{I_2, I_3} b_{I_3} \dots a_{I_{T-1}, I_T} b_{I_T} \end{aligned} \quad (6.81)$$

where  $\langle \mathbf{I} \rangle$  indicates the set of all possible hidden state sequences.

The sum (6.81) is the *likelihood* of the sequence  $\mathbf{O}$ , and it is important when we try to estimate the model parameters. Indeed, in the following chapters we will be mostly concerned with parameter estimates, and likelihoods such as (6.81) will be one outstanding tool.

## 6.8 What's the use of all this?

Markov chains are spread throughout science, and they are used to model countless different processes. They are also a basic ingredient of many important statistical and modeling tools, like the Markov Chain Monte Carlo, which have a widespread use, e.g. as numerical tools in the framework of Bayesian statistics. The Hidden Markov Models are used to model the human voice, DNA basepair sequences, and many other complex phenomena.



Figure 6.6: Images from the life of Andrei Andreevich Markov (June 14, 1856 - July 20, 1922), who first studied Markov chains. Upper left: a young Markov; upper right: his wife, Maria; lower left: wife and son; lower right: a middle-aged Markov (taken from [4]).

## Appendix A: Proof of the Vandermonde convolution

The identity

$$\sum_{k=0}^n \left[ \binom{n}{k} \right]^2 = \binom{2n}{n}$$

used in section 6.3.1 is a special case of the Vandermonde convolution. Several proofs of the Vandermonde convolution exist, and here we settle on a simple algebraic proof. We start from the identity

$$(1+x)^n = (1+x)^{n-p}(1+x)^p$$

Expanding the l.h.s. we find

$$(1+x)^n = \sum_m \binom{n}{m} x^m$$

while the r.h.s. yields

$$\begin{aligned} (1+x)^{n-p}(1+x)^p &= \sum_{k=0}^{n-p} \binom{n-p}{k} x^k \sum_{j=0}^p \binom{p}{j} x^j \\ &= \sum_{j=0}^p \sum_{k=0}^{n-p} \binom{p}{j} \binom{n-p}{k} x^{j+k} \\ &= \sum_{j=0}^p \sum_{m=j}^n \binom{p}{j} \binom{n-p}{m-j} x^m \\ &= \sum_{m=0}^n \left[ \sum_{j=0}^m \binom{p}{j} \binom{n-p}{m-j} \right] x^m \end{aligned}$$

Therefore, equating the coefficients of  $x^m$ , we find the Vandermonde convolution

$$\binom{n}{m} = \sum_{j=0}^m \binom{p}{j} \binom{n-p}{m-j}$$

When we let  $n \rightarrow 2n$ ,  $p \rightarrow n$ ,  $m \rightarrow n$ , and  $j \rightarrow k$ , we obtain the identity used in section 6.3.1

$$\binom{2n}{n} = \sum_{k=0}^n \binom{n}{k} \binom{n}{n-k} = \sum_{k=0}^n \left[ \binom{n}{k} \right]^2$$

## Appendix B: Mathematica code for the simulation of a random walk

This is the small Mathematica program used to simulate a simple 1D random walk. All statements are commented for clarity (comments in Mathematica are represented by the string `\* <comment> \*`, and can be placed anywhere in the program).

```
\* Simulation of a simple random walk *\

\* system initialization *
ntot = 10000; \* this is the total number of time steps \*
x = Table[0, ntot]; \* this initializes the position vector *\

\* main simulation loop *
For[n = 2, n <= ntot,
    x[[n]] = x[[n - 1]] + (2*RandomInteger[] - 1);
    n++];

\* this plots the position vs. timestep \*
ListPlot[x, Frame -> True, PlotRange -> All, ImageSize -> 8*72,
FrameStyle -> 18]
```

## Appendix C: Mathematica code for the simulation of the Ehrenfest urn model

This is the small Mathematica program used to simulate the Ehrenfest urn model.

```
\* Simulation of Ehrenfest urn model *\

\* system initialization *
ntot = 10000; \* this is the total number of time steps \*
nballs = 1000; \* this is the total number of balls \*
where = Table[0, nballs]; \* list of ball positions *\

\* (0 or 1 identifies the urn) *\

\* the next statement initializes the record of ball numbers in
the first urn (0) \*
nrec = Table[0, ntot];

\* main simulation loop *
For[k = 1, k <= ntot,
```

```
    ball = RandomInteger[1, nballs]; /* here we choose the ball number
*\
    where[[ball]] = 1 - where[[ball]]; /* this exchanges ball position
*\
    nrec[[k]] = nballs - Apply[Plus, where]; /* this counts the number
                                              of balls in urn 0 *\'
    k++];

/* this plots the state of the system vs. timestep */
ListPlot[nrec, Frame -> True, PlotRange -> All, ImageSize -> 8*72,
FrameStyle -> 18];
```

# Chapter 7

## Descriptive statistics

### 7.1 Descriptive statistics

#### 7.1.1 Sample mean

We start from  $n$  independent samples  $x_k$  drawn from the same distribution with mean  $\mu$  and variance  $\sigma^2$ , i.e., the  $x$ 's are i.i.d. random variables. We already know that the sample mean

$$\mu_S = \frac{1}{n} \sum_{k=1}^n x_k \quad (7.1)$$

has the expectation value

$$\mathbf{E}(\mu_S) = \mu \quad (7.2)$$

and that the distribution of  $\mu_S$  is increasingly concentrated around  $\mu$  for larger and larger  $n$  (weak law of large numbers), therefore we take  $\mu_S$  is an *estimator* of the mean  $\mu$ . Such an estimate does not depend on any model of the distribution of the samples, and in particular it does not depend on any parameter, therefore this is also a model-free, nonparametric estimate.

#### 7.1.2 Sample variance

Now we guess that the variance can be estimated with an expression like

$$\frac{1}{n} \sum_{k=1}^n (x_k - \mu_S)^2 \quad (7.3)$$

where we use the estimate of the mean instead of the mean. Now notice that

$$\mathbf{E} (x_k - \mu_S)^2 = \mathbf{E} \left( x_k^2 - \frac{2}{n} \sum_{j=1}^n x_k x_j + \frac{1}{n^2} \sum_{i,j=1}^n x_i x_j \right) \quad (7.4a)$$

$$= \mathbf{E}(x_k^2) - \frac{2}{n} \sum_{j=1}^n \mathbf{E}(x_k x_j) + \frac{1}{n^2} \sum_{i,j=1}^n \mathbf{E}(x_i x_j) \quad (7.4b)$$

and

$$\mathbf{E}(x_i x_j) = \sigma^2 \delta_{i,j} + \mu^2 \quad (7.5)$$

because of the statistical independence of different samples. Then

$$\mathbf{E} (x_k - \mu_S)^2 = (\sigma^2 + \mu^2) - \frac{2}{n} \sum_{j=1}^n (\sigma^2 \delta_{k,j} + \mu^2) + \frac{1}{n^2} \sum_{i,j=1}^n (\sigma^2 \delta_{i,j} + \mu^2) \quad (7.6a)$$

$$= (\sigma^2 + \mu^2) - \frac{2}{n} (\sigma^2 + n\mu^2) + \frac{1}{n^2} \left( \sum_{i=1}^n \sigma^2 + n^2 \mu^2 \right) \quad (7.6b)$$

$$= \sigma^2 + \mu^2 - \frac{2}{n} \sigma^2 - 2\mu^2 + \frac{1}{n} \sigma^2 + \mu^2 \quad (7.6c)$$

$$= \sigma^2 - \frac{1}{n} \sigma^2 = \quad (7.6d)$$

$$= \frac{n-1}{n} \sigma^2 \quad (7.6e)$$

$$(7.6f)$$

and finally

$$\mathbf{E} \left[ \frac{1}{n} \sum_{k=1}^n (x_k - \mu_S)^2 \right] = \frac{n-1}{n} \sigma^2 \quad (7.7)$$

or also

$$\sigma^2 = \mathbf{E} \left[ \frac{1}{n-1} \sum_{k=1}^n (x_k - \mu_S)^2 \right] \quad (7.8)$$

This means that the sample variance

$$\sigma_S^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \mu_S)^2 \quad (7.9)$$

is an *unbiased estimator* of the variance (unbiased: mean of estimator equal to the value that we wish to estimate), while the quantity

$$\frac{1}{n} \sum_{k=1}^n (x_k - \mu_S)^2 \quad (7.10)$$

has mean value

$$\frac{n-1}{n}\sigma^2 \quad (7.11)$$

and coincides with the variance only for large  $n$  (this is an example of a *biased estimator*).

### 7.1.3 Variance of sample mean

The variance of the sample mean is

$$\text{var } \mu_S = \frac{1}{n^2} \sum_{k=1}^n \text{var}_S x_k \quad (7.12)$$

and if we use  $\sigma_S^2$  to estimate the individual variances of the i.i.d. variates  $x_k$ , we find the estimate for the variance of the sample mean

$$\text{var}_S \mu_S = \frac{1}{n^2} \sum_{k=1}^n \sigma_S^2 = \frac{\sigma_S^2}{n} = \frac{1}{n(n-1)} \sum_{k=1}^n (x_k - \mu_S)^2 \quad (7.13)$$

(the notation  $\text{var}_S$  indicates the sample variance, i.e., the estimate of the variance obtained from samples).

### 7.1.4 Sample covariance

The estimate of other moments proceeds much in the same way. For example, if the data samples are  $(x, y)$  pairs, we find

$$\mathbf{E} (x_k - \mu_{x,S}) (y_k - \mu_{y,S}) = \mathbf{E} \left( x_k y_k - \frac{1}{n} \sum_{j=1}^n x_k y_j - \frac{1}{n} \sum_{j=1}^n y_k x_j + \frac{1}{n^2} \sum_{i,j=1}^n x_i y_j \right) \quad (7.14a)$$

$$= \mathbf{E}(x_k y_k) - \frac{1}{n} \sum_{j=1}^n \mathbf{E}(x_k y_j) - \frac{1}{n} \sum_{j=1}^n \mathbf{E}(y_k x_j) + \frac{1}{n^2} \sum_{i,j=1}^n \mathbf{E}(x_i y_j) \quad (7.14b)$$

and

$$\mathbf{E}(x_i y_j) = \text{cov}(xy)\delta_{i,j} + \mu_x \mu_y \quad (7.15)$$

because of the statistical independence of different samples. Then

$$\begin{aligned}\mathbf{E} (x_k - \mu_{x,S})(y_k - \mu_{y,S}) &= (\text{cov}(xy) + \mu_x\mu_y) - \frac{2}{n} \sum_{j=1}^n (\text{cov}(xy)\delta_{k,j} + \mu_x\mu_y) \\ &\quad + \frac{1}{n^2} \sum_{i,j=1}^n (\text{cov}(xy)\delta_{i,j} + \mu_x\mu_y) \end{aligned}\tag{7.16a}$$

$$= \frac{n-1}{n} \text{cov}(xy) \tag{7.16b}$$

This means that

$$\text{cov}_S(x, y) = \frac{1}{n-1} \sum_{k=1}^n (x_k - \mu_{x,S})(y_k - \mu_{y,S}) \tag{7.17}$$

is an unbiased estimator of the covariance  $\text{cov}(xy)$ .

### 7.1.5 The correlation coefficient

Here we examine in greater detail the meaning of the sample covariance introduced above, and we add another estimator, the correlation coefficient.

Consider pairs of i.i.d. values  $x_k, y_k$  that are somehow related to each other, and assume the linear relationship

$$y = ax + b \tag{7.18}$$

We introduce the following procedure – that we shall justify in a later chapter – to estimate the values of the parameters  $a$  and  $b$ : we define first a mean square deviation

$$S = \sum_{k=1}^n [y_k - (ax_k + b)]^2 \tag{7.19}$$

and we minimize it with respect to  $a$  and  $b$ . To this end, as usual, we compute the derivatives and set them to zero, and we solve the resulting linear system:

$$\frac{\partial S}{\partial a} = -2 \sum_{k=1}^n x_k [y_k - (ax_k + b)] = 0 \tag{7.20a}$$

$$\frac{\partial S}{\partial b} = -2 \sum_{k=1}^n [y_k - (ax_k + b)] = 0 \tag{7.20b}$$

The linear system can be recast in matrix form

$$\begin{pmatrix} \langle x^2 \rangle & \mu_{x,S} \\ \mu_{x,S} & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \langle xy \rangle \\ \mu_{y,S} \end{pmatrix} \tag{7.21}$$

where  $\mu_{x,S} = \frac{1}{n} \sum_{k=1}^n x_k$  and  $\mu_{y,S} = \frac{1}{n} \sum_{k=1}^n y_k$  are the sample means, and  $\langle x^2 \rangle = \frac{1}{n} \sum_{k=1}^n x_k^2$ , and  $\langle xy \rangle = \frac{1}{n} \sum_{k=1}^n x_k y_k$ . The solution of the linear system is

$$a = \frac{\langle xy \rangle - \mu_{x,S} \mu_{y,S}}{\langle x^2 \rangle - \mu_{x,S}^2} = \frac{\text{cov}_S(x, y)}{\sigma_{x,S}^2} \quad (7.22\text{a})$$

$$b = \frac{\langle x^2 \rangle \mu_{y,S} - \langle xy \rangle \mu_{x,S}}{\langle x^2 \rangle - \mu_{x,S}^2} = \mu_{y,S} - \frac{\text{cov}_S(x, y)}{\sigma_{x,S}^2} \mu_{x,S} \quad (7.22\text{b})$$

These formulas are further parameterized with the introduction of the *correlation coefficient*  $r$  (also called *Pearson's correlation coefficient*), which is abstractly defined as<sup>1</sup>

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

In general  $-1 \leq r \leq 1$  (this is proved in the Appendix to this chapter).

The corresponding sample correlation coefficient is

$$r_S = \frac{\text{cov}_S(x, y)}{\sigma_{x,S} \sigma_{y,S}} \quad (7.23)$$

so that

$$a = \frac{r_S \sigma_{y,S}}{\sigma_{x,S}} \quad (7.24\text{a})$$

$$b = \mu_{y,S} - \frac{r_S \sigma_{y,S}}{\sigma_{x,S}} \mu_{x,S} \quad (7.24\text{b})$$

Notice also that the estimator of the correlation coefficient is

$$r_S = \frac{\text{cov}_S(x, y)}{\sigma_{x,S} \sigma_{y,S}} = \frac{\sum_{k=1}^n (x_k - \mu_{x,S})(y_k - \mu_{y,S})}{\sqrt{\sum_{i=1}^n (x_i - \mu_{x,S})^2 \sum_{j=1}^n (y_j - \mu_{y,S})^2}} \quad (7.25)$$

While  $r$  can indeed be a useful estimator of linear correlation, it can be misleading if used carelessly (see figure 7.1, taken from the Wikipedia article on *Correlation and dependence*).

## 7.2 Probability distributions of estimators in descriptive statistics

Now we explore in detail a specific example where samples are exponentially distributed.

---

<sup>1</sup>the correlation coefficient is usually denoted with the symbol  $r$ , and sometimes with the symbol  $\rho$

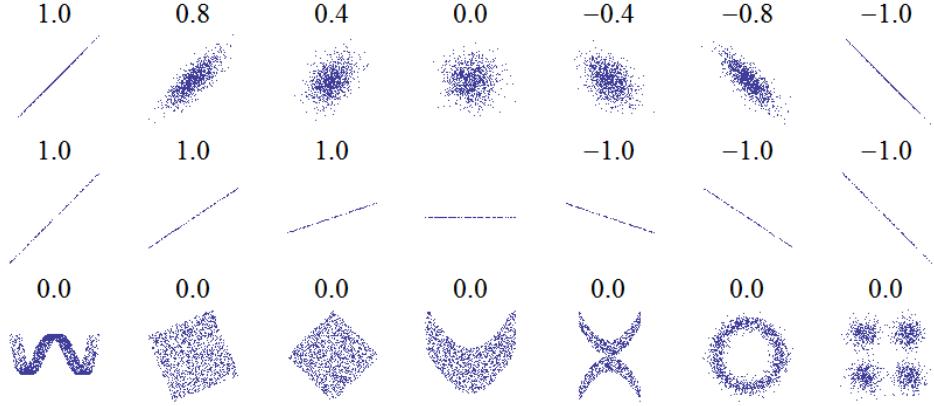


Figure 7.1: Examples of distributions of two random variables, with the corresponding values of the correlation coefficient. While the correlation coefficient can be quite useful in detecting a linear correlation, it can fail badly to detect nonlinear correlations.

### 7.2.1 Distribution of the sample mean of exponentially distributed samples

Estimators like those of the previous sections often have very complex probability distributions. Consider a simple example, where the samples  $t_k$  have an exponential pdf

$$p(t) = \frac{1}{\tau} e^{-t/\tau} \quad (7.26)$$

In this case we can easily find the pdf of the sample mean using the characteristic functions. Recall that the characteristic function of an exponential pdf is

$$f(z) = \int_0^\infty \frac{1}{\tau} e^{-t/\tau} e^{izt} dt = \frac{1}{1 - iz\tau} \quad (7.27)$$

and that the cf of a the sample mean  $\mu_S = (1/n) \sum_k t_k$  of i.i.d. random variables is the product of the individual cf's

$$f(z) = \mathbf{E}(e^{iz\mu_S}) = \mathbf{E}(e^{iz(1/n) \sum_k t_k}) = \prod_{k=1}^n \mathbf{E}(e^{izt_k/n}) = \prod_{k=1}^n f_k\left(\frac{z}{n}\right) \quad (7.28)$$

and therefore in this case the cf of the sample mean is

$$f_{\mu_S}(z) = \prod_{k=1}^n f_k\left(\frac{z}{n}\right) = \frac{1}{(1 - iz\tau/n)^n} \quad (7.29)$$

The pdf of the sample mean is the inverse Fourier transform of this expression

$$p_{\mu_S}(\mu_S) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{(1 - iz\tau/n)^n} e^{-iz\mu_S} dz \quad (7.30)$$

Now we note that the complex function

$$F(z) = \frac{e^{-iz\mu_S}}{(1 - iz\tau/n)^n} = \left(\frac{in}{\tau}\right)^n \frac{e^{-iz\mu_S}}{(z + in/\tau)^n} \quad (7.31)$$

has a pole of order  $n$  in  $z = -in/\tau$  (in the lower half-plane), and the associated residue is

$$\text{Res}(F, -in/\tau) = \frac{1}{(n-1)!} \frac{d^{n-1}}{dz^{n-1}} [(z + in/\tau)^n F(z)]|_{z=-in/\tau} \quad (7.32a)$$

$$= \frac{1}{(n-1)!} \frac{d^{n-1}}{dz^{n-1}} \left[ \left(\frac{in}{\tau}\right)^n e^{-iz\mu_S} \right]|_{z=-in/\tau} \quad (7.32b)$$

$$= \frac{1}{(n-1)!} \left(\frac{in}{\tau}\right)^n (-i\mu_S)^{n-1} e^{-n\mu_S/\tau} \quad (7.32c)$$

Thus closing the integration path in the lower half-plane, we find

$$p_{\mu_S}(\mu_S) = \frac{1}{2\pi} (-2\pi i) \text{Res}(F, -in/\tau) = -i \frac{1}{(n-1)!} \left(\frac{in}{\tau}\right)^n (-i\mu_S)^{n-1} e^{-n\mu_S/\tau} \quad (7.33a)$$

$$= \frac{n^n}{(n-1)!} \left(\frac{\mu_S^{n-1}}{\tau^n}\right) e^{-n\mu_S/\tau} \quad (7.33b)$$

This is a special case of the Gamma distribution

$$p(x) = \frac{x^{n-1} e^{-x/\theta}}{\theta^n \Gamma(n)} \quad (7.34)$$

with  $\theta = \tau/n$ . Figure 7.2 shows the Gamma distributions for different values of  $n$ .

### 7.2.2 Mean of the sample mean distribution

The mean value of the sample mean is

$$\mathbf{E}\mu_S = \int_0^\infty \frac{x^n e^{-x/\theta}}{\theta^n \Gamma(n)} \Big|_{\theta=\tau/n} = \left(\frac{\tau}{n}\right) n = \tau \quad (7.35)$$

i.e., the mean value of the sample mean does not depend on  $n$ . Likewise, the variance of the the sample mean distribution can be found from the formula for the Gamma distribution

$$\mathbf{D}\mu_S = \theta^2 n = \left(\frac{\tau}{n}\right)^2 n = \frac{\tau^2}{n} \quad (7.36)$$

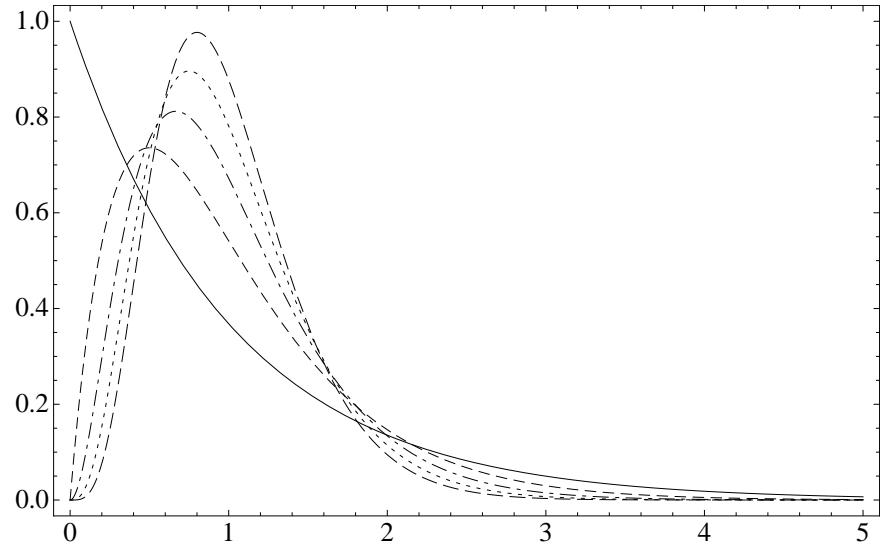


Figure 7.2: Plot of the Gamma probability density function ( $p(x)$  vs.  $x$ ) for  $\tau = 1$  and  $n = 1$  (solid line);  $n = 2$  (dashed line);  $n = 3$  (dashed-dotted line);  $n = 4$  (dotted line);  $n = 5$  (long dashes).

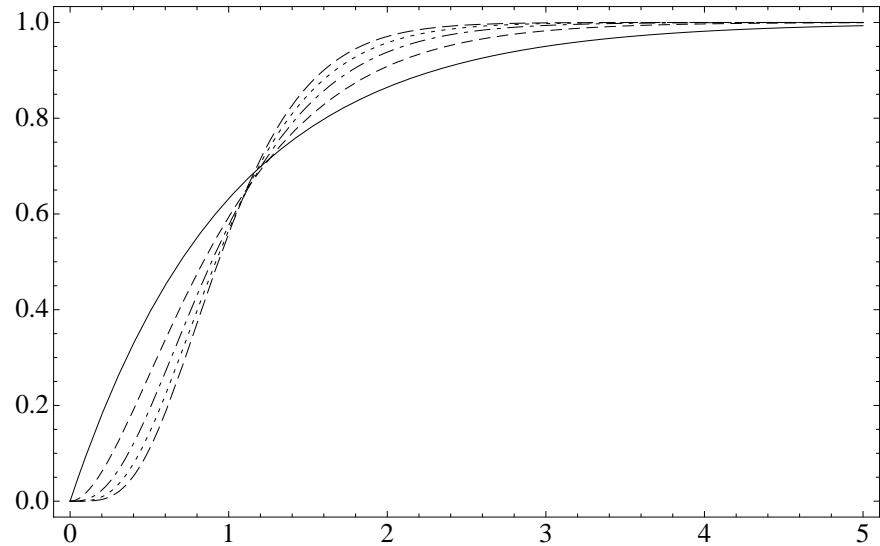


Figure 7.3: Plot of the Gamma distribution ( $F(x)$  vs.  $x$ ) for  $\tau = 1$  and  $n = 1$  (solid line);  $n = 2$  (dashed line);  $n = 3$  (dashed-dotted line);  $n = 4$  (dotted line);  $n = 5$  (long dashes).

The probability distribution function is

$$F(x) = \int_0^x p(x')dx' = \int_0^x \frac{x'^{n-1}e^{-x'/\theta}}{\theta^n \Gamma(n)} dx' \quad (7.37a)$$

$$= \frac{1}{\Gamma(n)} \int_0^{x/\theta} z^{n-1} e^{-z} dz = \frac{\Gamma(n) - \Gamma(n, x/\theta)}{\Gamma(n)} = \frac{\Gamma(n) - \Gamma(n, xn/\tau)}{\Gamma(n)} \quad (7.37b)$$

where

$$\Gamma(n, x) = \int_x^\infty z^{n-1} e^{-z} dz \quad (7.38)$$

is the *incomplete Gamma function*. Figure 7.3 shows the Gamma distribution functions for different values of  $n$ .

### 7.2.3 Example with exponentially distributed samples

Now take a small dataset: from this dataset we determine a sample mean and a sample variance. We know that the sample mean is an unbiased estimator of the mean value of the exponential distribution, however there are fluctuations and we know that this assertion is true only on average. As an example, consider the 20-sample dataset histogrammed in figure 7.4: the histogram is far from regular. How do we infer the parameter  $\tau$  of the underlying distribution?

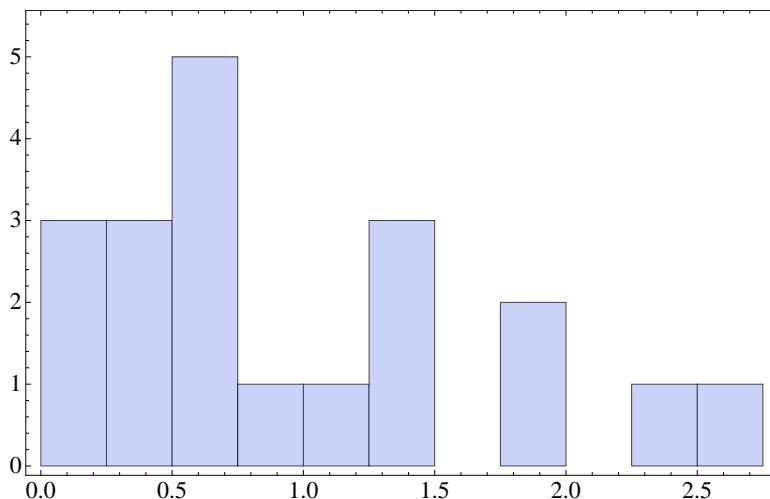


Figure 7.4: Histogram of a small dataset  $\{0.540766, 0.174497, 0.58976, 1.27419, 0.662204, 1.97348, 0.429274, 0.103007, 1.40585, 1.38147, 0.819439, 1.99229, 0.542477, 1.22707, 0.704053, 2.28692, 0.0488034, 2.74663, 0.313817, 0.407119\}$  of 20 exponentially distributed values. The sample mean is 0.98 and the sample standard deviation is 0.17. The corresponding standard deviation of the sample mean is 0.04.

### 7.2.4 Gaussian approximation

Each sample has an exponential pdf, therefore the variance is  $\tau^2$ , and the variance of the sample mean is  $\tau^2/n$ , because the samples are i.i.d. random variables. This – unsurprisingly – coincides with the result of the previous

section. If we use the Central Limit Theorem, we can also assert that in the large  $n$  limit, the distribution of the sample mean is well approximated by a Gaussian distribution  $N(\tau, \tau/\sqrt{n})$ . The convergence of the Gamma pdf discussed above to this Gaussian pdf is shown in figure 7.5.

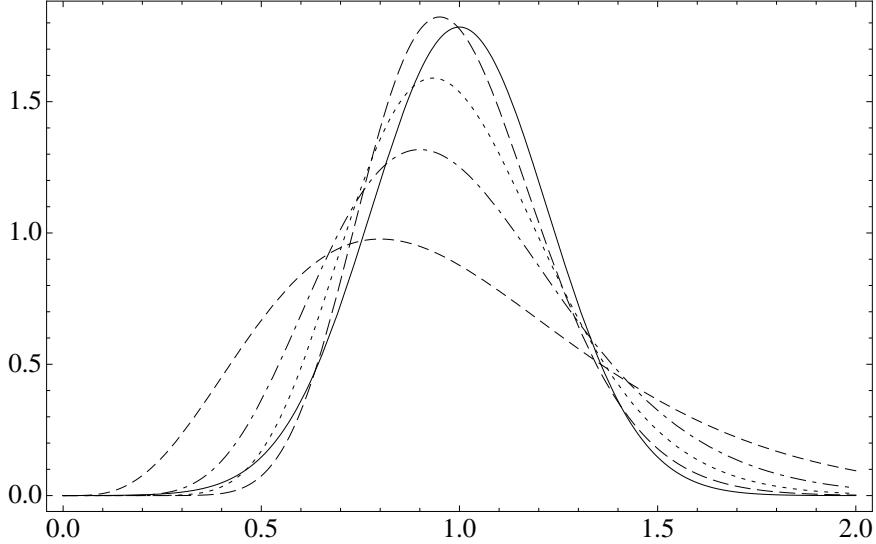


Figure 7.5: Plot of the Gamma probability density function ( $p(x)$ ) vs.  $x$  for  $\tau = 1$  and  $n = 5$  (dashed line);  $n = 10$  (dashed-dotted line);  $n = 15$  (dotted line);  $n = 20$  (long dashes). These pdf's should be compared with the Gaussian approximation (solid line).

### 7.3 The german tank problem

Statistical inference is a powerful tool that is used in many fields besides physics, and even during wars. There are many legendary stories of actual applications during the Second World War, and one of them deals with the estimate of the actual number  $N$  of tanks manufactured by the Germans [28]. The tanks were labeled with serial numbers  $x_k$ , and the Allies collected them. Eventually, a large number  $n$  of serials was known:

$$x_1 < x_2 < x_3 < \dots < x_n$$

how can we use these serials to estimate  $N$ ? Certainly  $N \geq x_n$ , therefore a very simple estimate is  $N_0 = x_n$ . Clearly  $N \geq x_n \geq n$ , and there are  $\binom{x_n - 1}{n - 1}$  ways of selecting a given  $x_n$  and randomly selecting the other serials. Since there are  $\binom{N}{n}$  ways of selecting  $n$  serials, we find that the probability that

$x_n$  is the highest serial found is

$$p(x_n) = \frac{\binom{x_n-1}{n-1}}{\binom{N}{n}} = \frac{n!(N-n)!}{N!} \frac{(x_n-1)!}{(n-1)!(x_n-n)!} \quad (7.39)$$

and thus the expectation value of this estimate is

$$\mathbf{E}(N_0) = \sum_{x_n=n,N} x_n p(x_n) \quad (7.40a)$$

$$= n \frac{(N-n)!}{N!} \sum_{x_n=n,N} x_n \frac{(x_n-1)!}{(x_n-n)!} \quad (7.40b)$$

$$= n \frac{(N-n)!}{N!} \frac{(N+1)!}{(N-n)!(n+1)} \quad (7.40c)$$

$$= \frac{n(N+1)}{(n+1)} \quad (7.40d)$$

$$= N - \frac{N-n}{n+1} \quad (7.40e)$$

and we conclude that this is a biased estimator<sup>2</sup>. Moreover this estimator can be quite off the mark, for instance if  $n = 1$  (just one serial) and  $N \gg 1$ , then  $\mathbf{E}(N_0) \approx N/2$ .

We can use the same methods to compute the higher moments of the estimator  $N_0$ , e.g.,

$$\mathbf{E}(N_0^2) = \sum_{x_n=n,N} x_n^2 p(x_n) \quad (7.41a)$$

$$= n \frac{(N-n)!}{N!} \sum_{x_n=n,N} x_n^2 \frac{(x_n-1)!}{(x_n-n)!} \quad (7.41b)$$

$$= n \frac{(N-n)!}{N!} \frac{(N+1)(Nn+N+n)N!}{(N-n)!(n+1)(n+2)} \quad (7.41c)$$

$$= \frac{n(N+1)(Nn+N+n)}{(n+1)(n+2)} \quad (7.41d)$$

and therefore

$$\begin{aligned} \text{Var}N_0 &= \frac{n(N+1)(Nn+N+n)}{(n+1)(n+2)} - \left[ \frac{n(N+1)}{(n+1)} \right]^2 \\ &= \frac{n(N+1)(N-n)}{(n+1)^2(n+2)} \end{aligned} \quad (7.42)$$

---

<sup>2</sup>Sums like those in this expression can be evaluated with the techniques explained in the book “A=B” by M. Petkovsek, H. S. Wilf, and D. Zeilberger, downloadable at <http://www.math.upenn.edu/~wilf/AeqB.html>. These techniques are implemented in advanced computer algebra programs like *Mathematica*.



Figure 7.6: A destroyed tank rots along the roadside during the Second World War.

Notice that for  $n = 1$  (just one serial) and  $N \gg 1$ , this variance is very close to  $N^2/12$ , which is the value expected for a uniform distribution.

We can easily find other estimators. We notice that if the serial numbers are uniformly distributed between 1 and  $N$ , then the sample average (or the sample median) should lie halfway, there should be  $\mu_S - 1$  values below the average, and just as many above the average, so that  $N = (\mu_S - 1) + 1 + (\mu_S - 1)$ , and therefore we define the estimators

$$N_1 = 2\mu_S - 1 \quad (7.43)$$

and likewise

$$N_2 = 2m_S - 1 \quad (7.44)$$

where  $m_S$  is the sample median.

We can also argue that the number of unobserved serials below  $x_1$ , i.e.,  $x_1 - 1$  is on average the same as the number of unobserved serials above  $x_n$ , i.e.,  $N - x_n$ , so that

$$N - x_n = x_1 - 1 \quad (7.45)$$

and therefore we find the estimator

$$N_3 = x_n + x_1 - 1 \quad (7.46)$$

We can try to make this more precise arguing that the number of unobserved serials above  $x_n$  can be estimated by the average of the unobserved serials between the observed serials:

$$N - x_n = \frac{1}{n} [(x_1 - 1) + (x_2 - x_1 - 1) + (x_3 - x_2 - 1) + \cdots + (x_n - x_{n-1} - 1)] \quad (7.47a)$$

$$= \frac{1}{n} (x_n - n) \quad (7.47b)$$

and therefore we obtain the estimator

$$N_4 = \frac{n+1}{n} x_n - 1 \quad (7.48)$$

We have already found  $p(x_n)$ , and we can readily analyze the properties of  $N_4$ :

$$\mathbf{E}(N_4) = \frac{n+1}{n} \mathbf{E}(N_0) - 1 = \frac{n+1}{n} \frac{n(N+1)}{(n+1)} - 1 = N \quad (7.49)$$

our first result is that  $N_4$  is an unbiased estimator of  $N$ . Similarly,

$$\begin{aligned} \text{Var}N_4 &= \frac{(n+1)^2}{n^2} \text{Var}N_0 = \frac{(n+1)^2}{n^2} \frac{n(N+1)(N-n)}{(n+1)^2(n+2)} \\ &= \frac{n(N+1)(N-n)}{n(n+2)} \end{aligned} \quad (7.50)$$

and therefore  $\text{Var}N_4 > \text{Var}N_0$ .

Exercise: find the average and the variance of the estimators  $N_1$ ,  $N_2$  and  $N_3$ .

### 7.3.1 Order statistics

In the German Tank problem we met a special instance of *order statistics*. In general when we consider a set of  $N$  measurements  $x_1, \dots, x_N$  which are extracted independently and which correspond to variates with the same probability density  $p(x)$  and distribution function  $F(x)$ , we can arrange them in increasing order

$$y_1 < y_2 < \cdots < y_N$$

for instance  $y_1 = \min_i x_i$  and  $y_N = \max_i x_i$ .

Now we set about finding the probability density of  $y_N$ : since  $F(x) = P(x_i < x)$  for all  $i$ 's, then

$$F_{y_N}(x) = \prod_i F_i(x) = [F(x)]^N \quad (7.51)$$

and therefore

$$p_{y_N}(x) = \frac{d}{dx} F_{y_N}(x) = N[F(x)]^{N-1} p(x) \quad (7.52)$$

Similarly, we find the probability density of  $y_1$ : since  $P(x > x_i) = 1 - F(x)$  for all  $i$ 's, then

$$F_{y_1}(x) = 1 - [1 - F(x)]^N \quad (7.53)$$

and therefore

$$p_{y_1}(x) = \frac{d}{dx} F_{y_1}(x) = N[1 - F(x)]^{N-1} p(x) \quad (7.54)$$

Now divide the  $x$ -range into three intervals,  $(-\infty, x)$ ,  $(x, x + dx)$ , and  $(x + dx, +\infty)$ , and consider the probabilities of finding a value of in one of these intervals, which are  $F(x)$ ,  $p(x)dx$  and  $1 - F(x + dx) = 1 - F(x)$ , respectively. Then the probability of finding  $y_k$  in the middle interval  $(x, x + dx)$  is given by the multinomial expression

$$p_{y_k}(x)dx = \frac{N!}{(k-1)!1!(N-k)!}[F(x)]^{k-1}[p(x)dx][1 - F(x)]^{N-k} \quad (7.55)$$

and therefore

$$\begin{aligned} p_{y_k}(x) &= \frac{N!}{(k-1)!(N-k)!}[F(x)]^{k-1}[1 - F(x)]^{N-k}p(x) \\ &= N \binom{N-1}{k-1} [F(x)]^{k-1}[1 - F(x)]^{N-k}p(x) \end{aligned} \quad (7.56)$$

Likewise, when we divide the  $x$ -range into five intervals,

$$\begin{aligned} &(-\infty, x') \\ &(x', x' + dx') \\ &(x' + dx', x'') \\ &(x'', x'' + dx'') \\ &(x'' + dx'', +\infty) \end{aligned}$$

and consider the probabilities of finding a value of in one of these intervals, which are  $F(x')$ ,  $p(x')dx'$ ,  $F(x'') - F(x')$ ,  $p(x'')dx''$ , and  $1 - F(x'' + dx'') = 1 - F(x'')$ , respectively, we find that the joint probability density is given by the multinomial expression

$$\begin{aligned} &p_{y_j, y_k}(x', x'') \\ &= \frac{N!}{(j-1)!(k-j-1)!(N-k)!}[F(x')]^{j-1}[F(x'') - F(x')]^{k-j-1}[1 - F(x'')]^{N-k}p(x')p(x'') \end{aligned} \quad (7.57)$$

As an example of order statistics, consider the uniform distribution on the  $(0, 1)$  interval, then

$$p(x) = 1; \quad F(x) = x$$

and therefore, in this case, we find

$$p_{y_k}(x) = \frac{N!}{(k-1)!(N-k)!} x^{k-1} (1-x)^{N-k} \quad (7.58)$$

which is a beta distribution. In particular, consider the median of the distribution, i.e., the value that corresponds to index  $k = N/2$  for even  $N$  and to  $k = N/2 + 1/2$  for odd  $N$ , then we find:

- Even  $N$ :

$$p_{y_{N/2}}(x) = \frac{N!}{(N/2-1)!(N/2)!} x^{N/2-1} (1-x)^{N/2} \quad (7.59)$$

- Odd  $N$ :

$$p_{y_{(N+1)/2}}(x) = \frac{N!}{[(N-1)/2]![((N-1)/2)]!} x^{(N-1)/2} (1-x)^{(N-1)/2} \quad (7.60)$$

Figure 7.7 shows the PDF's  $p_{y_1}(x)$ ,  $p_{y_n}(x)$ , and  $p_{y_{N/2}}(x)$  for the uniform distribution and  $n = 10$ .

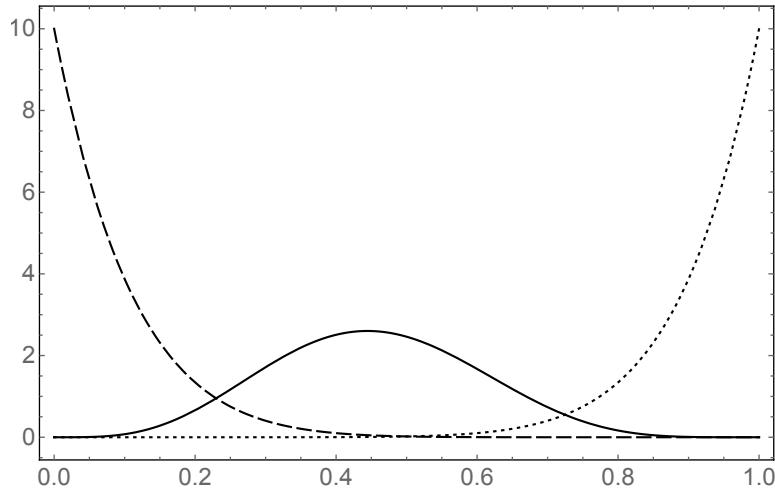


Figure 7.7: PDF's of the order statistics  $p_{y_1}(x)$  (dashed curve),  $p_{y_n}(x)$  (dotted curve), and  $p_{y_{N/2}}(x)$  (solid curve) for the uniform distribution and  $n = 10$ .

This is only a very short introduction to order statistics, which is an important example of *non parametric statistics*.

## 7.4 Inadequacies of the standard deviation

The standard deviation is usually taken as an indicator of the spread of a probability distribution, however it has obvious limitations. These inadequacies are easily dealt with in the usual practice of statistics – for instance, the indefiniteness of the variance of the Cauchy distribution prompts us to use the full-width at half-maximum in its place – however they may become troublesome if standard deviation is used uncritically in related fields (e.g., it may lead to inconsistent results in studies of the uncertainty relations, which are based on mean square deviation). Here is a simple example that utilizes a discrete distribution to illustrate the most common failing: consider a discrete random variable  $\xi$  that takes the values  $0, \dots, N-2$  and  $\alpha(N-1)$  with probability  $p_0, \dots, p_{N-1}$  and let  $p_0 = \dots = p_{N-2}$  and  $p_{N-1} = 1/\alpha$ . Then from the normalization condition

$$\sum_{k=0}^{N-2} p_k = 1 - \frac{1}{\alpha} \quad (7.61)$$

we find

$$p_k = \frac{1}{N-1} \left( 1 - \frac{1}{\alpha} \right) \quad (7.62)$$

Therefore, the mean value is

$$\begin{aligned} \mathbf{E}\xi &= \sum_{k=0}^{N-1} kp_k = \frac{1}{N-1} \left( 1 - \frac{1}{\alpha} \right) \sum_{k=0}^{N-2} k + \alpha(N-1) \frac{1}{\alpha} \\ &= \frac{1}{2}(N-2) \left( 1 - \frac{1}{\alpha} \right) + (N-1) \end{aligned} \quad (7.63)$$

and

$$\lim_{\alpha \rightarrow \infty} \mathbf{E}\xi = \frac{1}{2}(N-2) + (N-1) \quad (7.64)$$

We can set up a similar calculation for the mean square value

$$\mathbf{E}(\xi^2) = \sum_{k=0}^{N-1} k^2 p_k = \frac{1}{N-1} \left( 1 - \frac{1}{\alpha} \right) \sum_{k=0}^{N-2} k^2 + \alpha^2(N-1)^2 \frac{1}{\alpha} \quad (7.65a)$$

$$= \frac{1}{6}(N-2)(2N-3) \left( 1 - \frac{1}{\alpha} \right) + \alpha(N-1)^2 \quad (7.65b)$$

however this shows that now

$$\lim_{\alpha \rightarrow \infty} \mathbf{E}(\xi^2) = \infty \quad (7.66)$$

and therefore variance and standard deviation are undefined, even though the outlier value  $\alpha(N-1)$  has a vanishing probability.

The full-width at half-maximum is a reasonable substitute of standard deviation when the distribution is unimodal, however this is not always the case (as in the example above), and other estimators of spread have been proposed, and in one proposal, Shannon's entropy has been suggested as an alternative estimator. Recall that Shannon's entropy is

$$H = - \sum_k p_k \ln p_k \quad (7.67)$$

(the usual definition utilizes  $\log_2$ , here I use natural logarithms to avoid the introduction of unimportant constants). It is interesting to maximize this entropy using only the normalization constraint  $\sum_k p_k = 1$ : we must introduce one Lagrange multiplier  $\lambda$  and maximize the resulting function

$$H' = H - \lambda \left( \sum_k p_k - 1 \right) = - \sum_k p_k \ln p_k - \lambda \left( \sum_k p_k - 1 \right) \quad (7.68)$$

i.e.,

$$\frac{\partial H'}{\partial p_j} = -\ln p_j - 1 - \lambda = 0 \quad (7.69)$$

Solving these equations we find

$$p_j = \exp(-\lambda - 1) \quad (7.70)$$

and because of the normalization constraint, this means  $p_j = 1/N$ , i.e., Shannon's entropy is maximized by the uniform distribution, and any distribution that is more concentrated has a smaller entropy. The maximal entropy value is

$$H = \sum_k \frac{1}{N} \ln N = \ln N \quad (7.71)$$

The entropy of the distribution in the example discussed above is

$$H = \left\{ - \sum_{k=0}^{N-2} \frac{1}{N-1} \left( 1 - \frac{1}{\alpha} \right) \ln \left[ \frac{1}{N-1} \left( 1 - \frac{1}{\alpha} \right) \right] + \frac{1}{\alpha} \ln \alpha \right\} \xrightarrow{\alpha \rightarrow \infty} \ln(N-1) \quad (7.72)$$

and thus the entropy depends only on the extension of the bulk of the distribution and not on the position of the outlier, moreover it is only marginally smaller than the entropy of the uniform distribution. The conclusion is that Shannon's entropy conveys information on the shape of the distribution that is complementary to the combination of mean and standard deviation.

### 7.4.1 Sample variance of correlated data. The Allan variance.

Consider a zero-mean time series  $\{x_n\}_{n=0,\dots,N-1}$ . The sample variance of this time series is

$$\sigma_S^2 = \frac{1}{N-1} \sum_{n=0}^{N-1} x_n^2 = \frac{N}{N-1} \sum_{k=0}^{N-1} S_k \quad (7.73)$$

where  $S_k$  is the power spectrum of the time series, and the equality represents the discrete version of Parseval's theorem<sup>3</sup>. A problem arises in this definition of the sample variance when the power spectrum  $S_k$  has a divergence (or a near divergence) at low frequency, as it happens in the case of the  $1/f^\alpha$  noise processes (this means that  $S_k \sim k^{-\alpha}$ ). This is especially troublesome when dealing with precise timekeeping, where one would like to use sample variance as an estimator of clock precision.

To be more specific, consider the phase of a clock,  $\Phi(t) = 2\pi\nu_0 t + \varphi(t)$ , where  $\varphi(t)$  is the deviation from the regular, linear phase increase  $2\pi\nu_0 t$ . Then the fractional instantaneous frequency

$$\frac{\nu(t)}{\nu_0} = \frac{1}{2\pi\nu_0} \frac{d\Phi}{dt} = 1 + \frac{1}{2\pi\nu_0} \frac{d\varphi}{dt} \quad (7.74)$$

has a fractional frequency deviation

$$\frac{1}{2\pi\nu_0} \frac{d\varphi}{dt} \quad (7.75)$$

which is approximated by the time series

$$y_n = \frac{\varphi(n\Delta t + \tau) - \varphi(n\Delta t)}{2\pi\nu_0\tau} \quad (7.76)$$

where  $\Delta t$  is the sampling interval,  $\tau$  is the time interval used in the estimate of the derivative (the instantaneous frequency), and  $\nu_0$  is the central frequency of the clock. It is important to define an estimators of the fluctuations of clock frequency, and the (averaged) sample variance is a natural candidate

$$\sigma_y^2(M, \Delta t, \tau) = \left\langle \frac{1}{M-1} \sum_{n=0}^{M-1} \left( y_n - \frac{1}{M} \sum_{m=1}^{M-1} y_m \right)^2 \right\rangle \quad (7.77)$$

---

<sup>3</sup>Here I use the following definition for the discrete Fourier transform

$$F_k = \sum_{n=0}^{N-1} x_n \exp\left(-\frac{2\pi i}{N} k n\right)$$

so that the power spectrum is

$$S_k = \frac{|F_k|^2}{N^2}$$

where the angle brackets  $\langle \cdot \rangle$  denote an average over several ensembles of  $M$  samples. The averaged sample variance  $\sigma_y(M, \Delta t, \tau)$  is also called *M-point variance*, and since both clock phase and frequency suffer from random low-frequency drifts ( $1/f^\alpha$  noises), it behaves badly for large  $M$ : to avoid the problems associated to spectral divergences it is customary to use the 2-point variance with  $\Delta t = \tau$ , – also called *Allan variance* – as an estimator of clock stability

$$\text{AVAR} = \sigma_y^2(2, \tau, \tau) = \left\langle \left[ y_0 - \frac{1}{2}(y_0 + y_1) \right]^2 + \left[ y_1 - \frac{1}{2}(y_0 + y_1) \right]^2 \right\rangle \quad (7.78)$$

or, equivalently, the *Allan deviation*

$$\text{ADEV} = \text{AVAR}^{1/2} = \sigma_y(2, \tau, \tau) \quad (7.79)$$

Figure 7.8 shows the relationship between Allan variance and noise spectral density<sup>4</sup>, and figures 7.9, 7.10 show the spectral density and the Allan deviation for an oven-controlled crystal oscillator. Notice that because of the large low-frequency fluctuations of some power-law noises, the Allan deviation reaches a minimum and then rises again as the observation time increases.

## 7.5 What's the use of all this?

Descriptive, nonparametric statistics is a basic tool which is used to describe datasets in a model-free way. It is universally used in data analysis, both with real and with simulated data.

## Appendix: Cauchy-Schwartz's inequality and the correlation coefficient

Take two random variables  $\xi_1$  and  $\xi_2$  with means  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$ , and the corresponding zero-mean variables  $\eta_i = \xi_i - \mu_i$  and consider the following expectation

$$S(\lambda) = \mathbf{E}(\eta_1 + \lambda\eta_2)^2 = \mathbf{E}(\eta_1^2) + \lambda^2\mathbf{E}(\eta_2^2) + 2\lambda\mathbf{E}(\eta_1\eta_2) \geq 0 \quad (7.80)$$

---

<sup>4</sup>In interpreting the figure recall that the Fourier transform of the derivative of a signal  $f(t)$  is related to the Fourier transform of the signal  $F(\omega)$  by a simple relation:

$$\int_{-\infty}^{+\infty} f'(t)e^{-i\omega t} dt = f(t)e^{-i\omega t} \Big|_{-\infty}^{+\infty} + i\omega \int_{-\infty}^{+\infty} f(t)e^{-i\omega t} dt = i\omega F(\omega)$$

where we assume that the signal  $f(t)$  vanishes for large negative and positive times. Therefore

$$S_{f'(t)}(\omega) = \omega^2 S_{f(t)}(\omega)$$

and this explains why in figure 7.8 we step from a white noise in frequency fluctuations to a flicker noise in phase fluctuations.

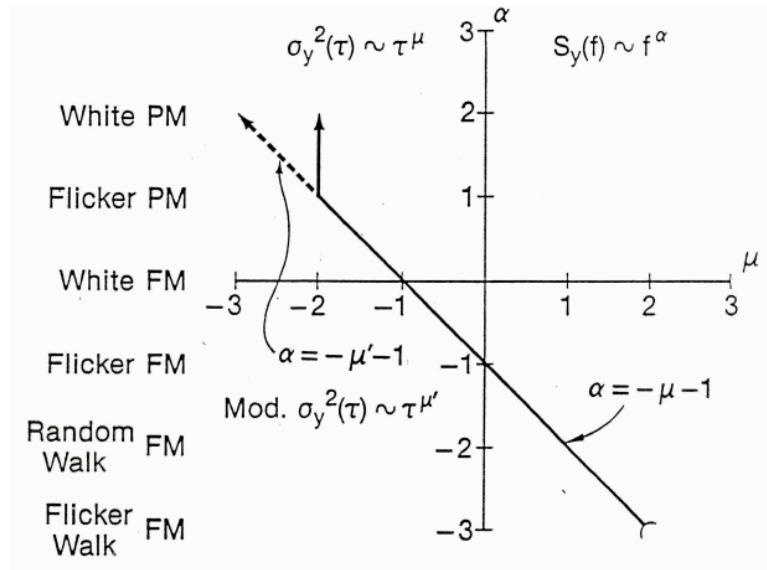


Figure 7.8: Noise spectral density and Allan Variance. Here  $\tau$  is the total observation time.

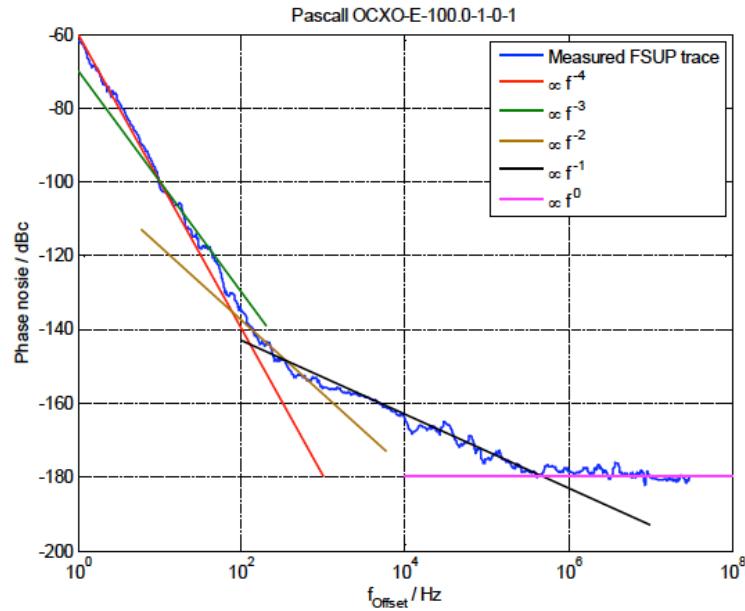


Figure 7.9: Spectral density of phase noise in an oven-controlled crystal oscillator by Rohde & Schwartz. The spectral density displays contributions from different power-law noises.

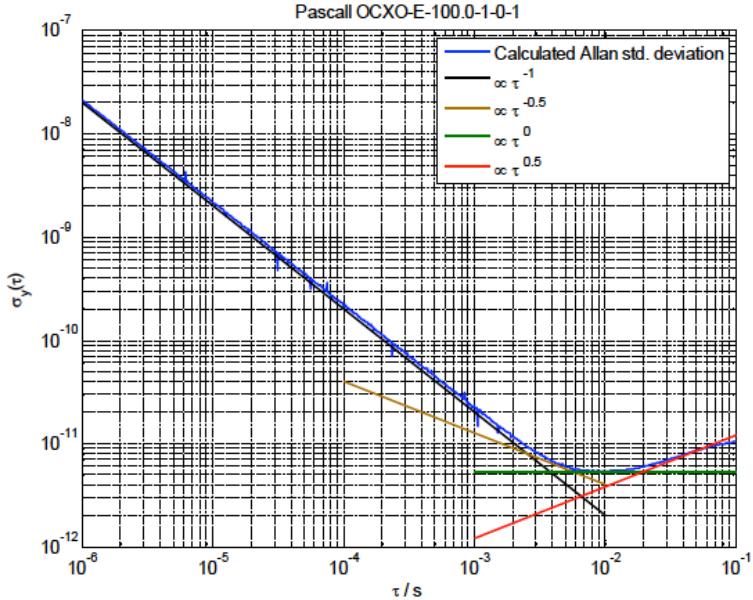


Figure 7.10: Allan Variance corresponding to the spectral density in figure 7.9. Notice that the Allan variance flattens at high frequency and then increases again because of the contributions from power-law noises with large fluctuations at low frequencies: power-law spectra thus defy the rule that "longer observation times mean higher precision".

where  $\lambda$  is a real parameter. The minimum of  $S(\lambda)$  is obtained for

$$\frac{\partial S}{\partial \lambda} = 2\lambda \mathbf{E}(\eta_2^2) + 2\mathbf{E}(\eta_1\eta_2) = 0 \quad (7.81)$$

i.e.

$$\lambda = -\frac{\mathbf{E}(\eta_1\eta_2)}{\mathbf{E}(\eta_2^2)} \quad (7.82)$$

and thus

$$S_{min} = \mathbf{E}(\eta_1^2) - \frac{[\mathbf{E}(\eta_1\eta_2)]^2}{\mathbf{E}(\eta_2^2)} \geq 0 \quad (7.83)$$

Finally we find

$$\mathbf{E}(\eta_1^2)\mathbf{E}(\eta_2^2) \geq [\mathbf{E}(\eta_1\eta_2)]^2 \quad (7.84)$$

i.e.,

$$\sigma_1^2\sigma_2^2 \geq \text{cov}(\xi_1, \xi_2) \quad (7.85)$$

or also

$$\sigma_1^2\sigma_2^2 \geq r^2\sigma_1^2\sigma_2^2 \quad (7.86)$$

which means that the correlation coefficient satisfies the inequality

$$|r| \leq 1 \quad (7.87)$$

# Chapter 8

## Monte Carlo Methods

In this chapter I introduce and discuss the principles of the Monte Carlo method (here also MC for short). For a brief history of the method, see Appendix A of this chapter.

### 8.1 Introduction to Monte Carlo

Consider the following simple example: we want to evaluate the area of Bernoulli's lemniscate, i.e., the curve defined by the following equation in polar coordinates

$$r^2 = a^2 \cos 2\theta \quad (8.1)$$

where  $\cos 2\theta \geq 0$ , and shown in figure 8.1

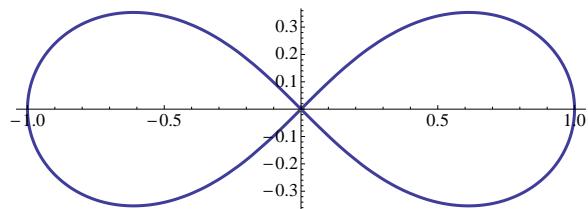


Figure 8.1: Plot of a lemniscate with  $a = 1$ , and with the main axis rotated along the horizontal direction.

The curve is certainly circumscribed by the square with vertices in  $(a, a)$ ,  $(-a, a)$ ,  $(-a, -a)$ , and  $(a, -a)$ , as shown in figure 8.2

If we generate  $n$  pairs of numbers  $(x, y)$ , such that both  $x$  and  $y$  are uniformly distributed on the interval  $(-1, 1)$ , then these pairs are uniformly distributed over the square, and we can estimate the area of the lemniscate relative to that of the square by taking the ratio  $n_L/n$  where  $n_L$  are the pairs contained in the lemniscate<sup>1</sup>. Figure 8.3 illustrates the procedure.

---

<sup>1</sup>the Mathematica program used to generate points and estimate areas is listed in

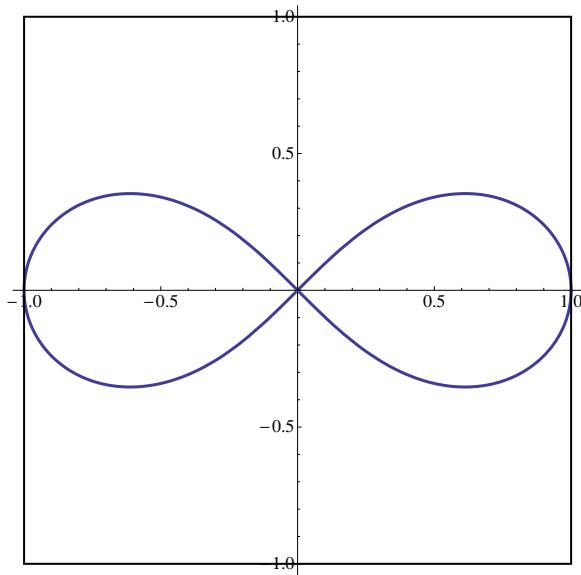


Figure 8.2: Lemniscate with  $a = 1$ , circumscribed by the square with vertices in  $(1, 1)$ ,  $(-1, 1)$ ,  $(-1, -1)$ , and  $(1, -1)$ .

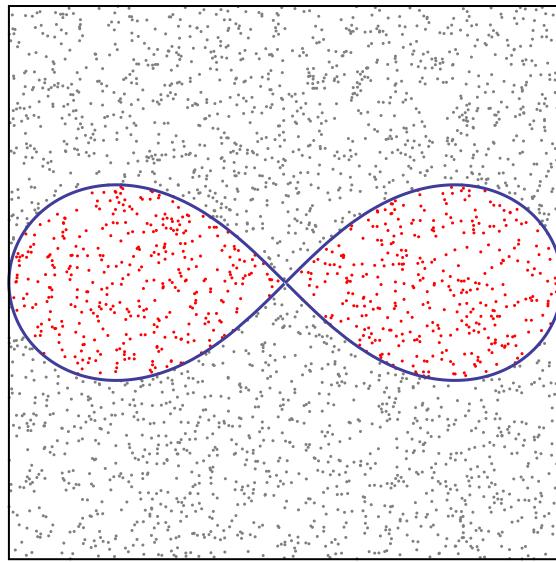


Figure 8.3: 3000 points uniformly distributed over the square. The points that fall inside the lemniscate are identified and counted (and drawn in red in this figure). In this case  $(\text{area of lemniscate}) = (\text{area of square}) \times (n_L/n) = 0.992$ .

The distribution of points over the square shows clumps and random irregularities: it is quite intuitive that to obtain a more precise result one should increase the number of points.

We can estimate the variability of the result repeating many times our numerical experiment. Figure 8.4 shows the histogram of 1000 area estimates, and – using sample mean and sample variance – the overall result of the area estimate is  $0.9992 \pm 0.0011$ . An exact calculation of the area yields  $a^2$ , which means 1 in this case (since  $a = 1$ ).

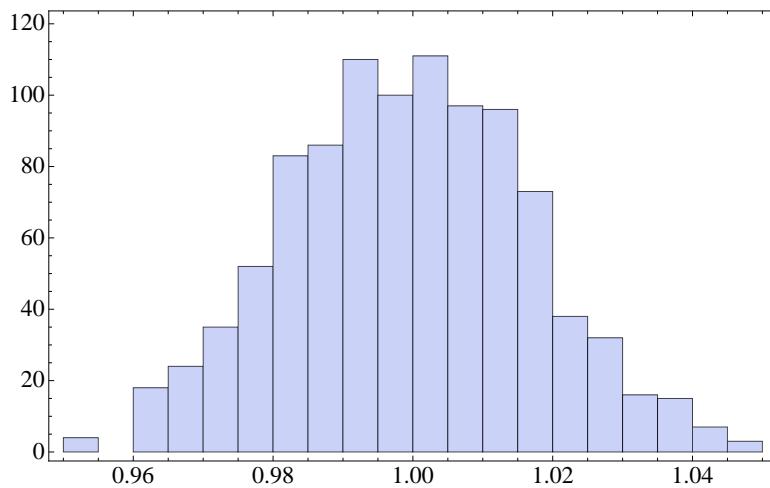


Figure 8.4: Histogram of the area estimates obtained in 1000 repetitions of the numerical experiment of figure 8.3. The estimated mean value is 0.999242, and the estimated standard deviation of the distribution of values is 0.036, therefore the estimated standard deviation of the mean is 0.0011, i.e., the result of the area estimate is  $0.9992 \pm 0.0011$ .

A few additional remarks:

- the area of the enclosing square is too large: we could make the MC estimate more efficient simply taking a rectangle that exactly circumscribes the lemniscate;
- since, in the example, the area of the square is 4, and the (exact) area of the lemniscate is 1 (prove it as an exercise), the event “point falls inside lemniscate” has probability  $p = 1/4$  in this case, and the whole set of  $n = 3000$  generated points in a single MC realization follows a binomial distribution with mean value  $\mu_{n_L} = np = 750$  and standard

deviation  $\sigma_{nL} = \sqrt{np(1-p)} \approx 23.717$ . The area is given by

$$(\text{area of lemniscate}) = (\text{area of square}) \times \left(\frac{n_L}{n}\right)$$

therefore the mean area of lemniscate is 1 and its standard deviation is

$$\sigma_{\text{area}} = (\text{area of square}) \times \frac{\sqrt{np(1-p)}}{n} \approx \frac{4 \times 23.717}{3000} \approx 0.032$$

this is very close to the value actually found in the MC simulation;

## 8.2 Empirical probability distributions

We have seen in chapter 5 that the knowledge of the moments of a distribution is equivalent to the knowledge of the distribution itself, i.e., to the knowledge of the analytical shape of the pdf. When we only have sample points, we can approximate this knowledge by computing the moments of the empirical distribution, i.e., the moments of the histogram obtained from the sample points. Thus a set of sample points carries with it the basic information contained in the pdf, and all we have to do is compute the moments of the distribution, i.e., evaluate integrals. Take, e.g., a discrete random variable  $\xi$  with values  $X_0, X_1, \dots, X_{N-1}$  and the corresponding probabilities  $p_0, p_1, \dots, p_{N-1}$  then the mean is

$$\mu = \mathbf{E}(\xi) = \sum_{n=0}^{N-1} X_n p_n \quad (8.2)$$

If we take  $M$  measurements  $x_k$ , and we detect  $m_n$  times the value  $X_n$ , then we find the empirical frequency  $f_n = m_n/M$ , and the sample mean is

$$\mu_S = \sum_{n=0}^{N-1} f_n X_n = \frac{1}{M} \sum_{n=0}^{N-1} m_n X_n = \frac{1}{M} \sum_{k=1}^M x_k \quad (8.3)$$

Likewise, for a continuous variate

$$\mu = \int_{-\infty}^{+\infty} X p(X) dX \approx \frac{1}{M} \sum_{k=1}^M x_k \quad (8.4)$$

and the sum (the sample mean)

$$\mu_S = \frac{1}{M} \sum_{k=1}^M x_k \quad (8.5)$$

is used to estimate the value of the mean  $\mu$ . Notice also that we know how to estimate the accuracy of the method, because the variance of the sample mean is

$$\text{var}\mu_S = \frac{1}{M(M-1)} \sum_{k=1}^M (x_k - \mu_S)^2 \quad (8.6)$$

Instead of actual measurements we can use Monte Carlo generated data, and we can estimate means and any other moment of the distribution, or the mean of any function of the random variable, or any such integral, by means of these approximated sums.

### 8.3 Uniformly distributed pseudo-random numbers

The MC method needs uniformly distributed numbers, so we have to find a way to produce such a sequence. This is wide and deep topic, and here we can only give a few general indications of how to proceed in practical cases. You can find a lot more information in the following references:

- S. K. Park and K. W. Miller, *Random Number Generators: Good Ones Are Hard To Find*, Communications of the ACM **31** (1988) 1192.  
The suggestions given in the paper are now outdated, but the general discussion is interesting and tackles many basic problems of pseudo-random number generation.
- W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing, Third Edition* (Cambridge Univ. Press, 2007).  
This is a well known introductory book on numerical methods, that includes a long chapter on random numbers. The book does not delve in theory, but offers a broad view of several important algorithms. It is no different in its treatment of random numbers, and one gets a fairly good overview of the field. The book also lists C++ code: unfortunately this code does not respect the usual C++ standards, and it is covered by copyright, therefore when coding one should rather turn to one of the many good GNU scientific libraries.
- D. E. Knuth, *The Art of Computer Programming: Seminumerical Algorithms* (Vol 2, 3rd Ed) (Addison-Wesley, 1997). There is an extensive discussion in this book which is one of the great classics of computer programming and numerical algorithms. Brilliant and engaging, although some of the suggested algorithms for the generation of pseudo-random numbers should be avoided.
- P. L'Ecuyer, *Random Number Generation*, Chapter 4 of the Handbook on Simulation, Jerry Banks Ed. (Wiley, 1998) pp. 93–137. This is a

good review work by one of the contributors of some excellent free software for random number generation.

- <http://www.gnu.org/software/gsl/>: web address of the GNU Scientific Library, which includes random number generation and random number distributions.
- <http://www.netlib.org/>: web address of the Netlib free software repository, which includes several random number libraries, and among them `ranlib`, both in C and Fortran, which is well documented in P. L'Ecuyer and S. Côté, ACM Transactions on Mathematical Software **17** (1991) 98.
- <http://www.boost.org>; web address of the Boost C++ Library, which includes another implementation of random number generation and random number distributions.

The problem with algorithms for random numbers is that they are *deterministic*, i.e., non random, and therefore they can only *mimic* true random numbers. For this reason they are called *pseudo-random numbers*, and the generators are called *pseudo random number generators* (PRNG).

### 8.3.1 Randomness vs. chaotic dynamics

The basic idea behind the generation of pseudo-random numbers is the substitution of randomness with a chaotic dynamics. Consider for instance the so-called *standard map*, where one defines a discrete dynamics

$$p_{n+1} = p_n - \frac{k}{2\pi} \sin 2\pi q_n \quad (8.7a)$$

$$q_{n+1} = p_{n+1} + q_n \quad (8.7b)$$

and then wraps the values on the 2D torus (mathematically this means summing or subtracting an integer that brings the  $p$ 's and  $q$ 's back on the  $(0,1)$  interval). Here  $k$  is a parameter that tunes the behavior of the map. Figures 8.5 and 8.6 show sets of orbits for two different values of the parameter  $k$  in the phase space  $(q, p)$ : a low value of  $k$  produces orderly regions and chaotic regions, and a larger  $k$  leads to an expansion of the chaotic regions. When we take few steps along a single orbit with a starting point in a chaotic region in figure 8.6, we find a 2D distribution that mimics a uniform distribution, see figure 8.7. Clearly distribution is not random, as it bears the mark of the orderly regions, that although reduced, are still present. There is also a short range correlation between points along the same orbit. Chaoticity is the standard substitute for randomness in pseudo-random number generators: clearly we must choose the dynamical process with great care, to prevent the underlying determinism to show up in some unexpected way in our simulations.

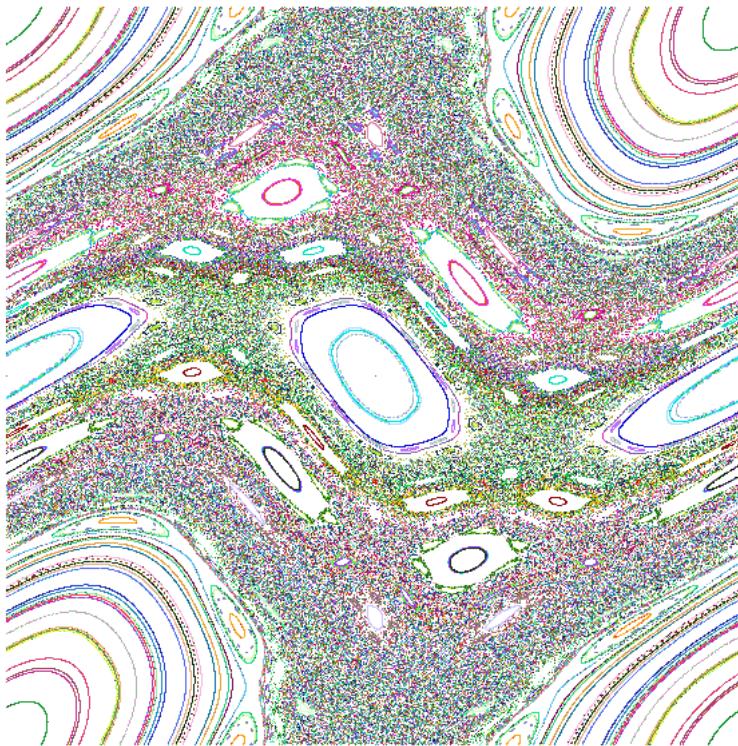


Figure 8.5: Phase space of the standard map for  $k = 1.1$ . Different orbits are labeled with different colors.

### 8.3.2 Standard methods

The dynamical systems used to generate pseudo-random numbers exist in many different flavors, and the most common is still the Linear Congruential Generator (LCG), defined by the recurrence

$$x_n = (ax_{n-1} + c) \mod m \quad (8.8)$$

where  $m > 0$ ,  $a > 0$ , and  $c$  are integers called the *modulus*, the *multiplier*, and the *additive constant*, respectively. When  $c = 0$  the generator is called a Multiplicative Linear Congruential Generator (MLCG).

The LCG has a seemingly chaotic dynamics; since it can generate at most  $m$  different numbers, it is periodic, and the period cannot exceed  $m$ . It is not easy to choose triples  $a$ ,  $m$  and  $c$  that yield the maximal period, and moreover the generator is not always reliable – in the sense that its distribution fails tests of randomness. An infamous case is the RANDU LCG generator, introduced in the 1960's on the IBM mainframe computers, which generates strongly correlated numbers. RANDU is a MLCG defined by  $a = 65539$  and  $m = 2^{31}$ , and it generates triples of points that cluster around a

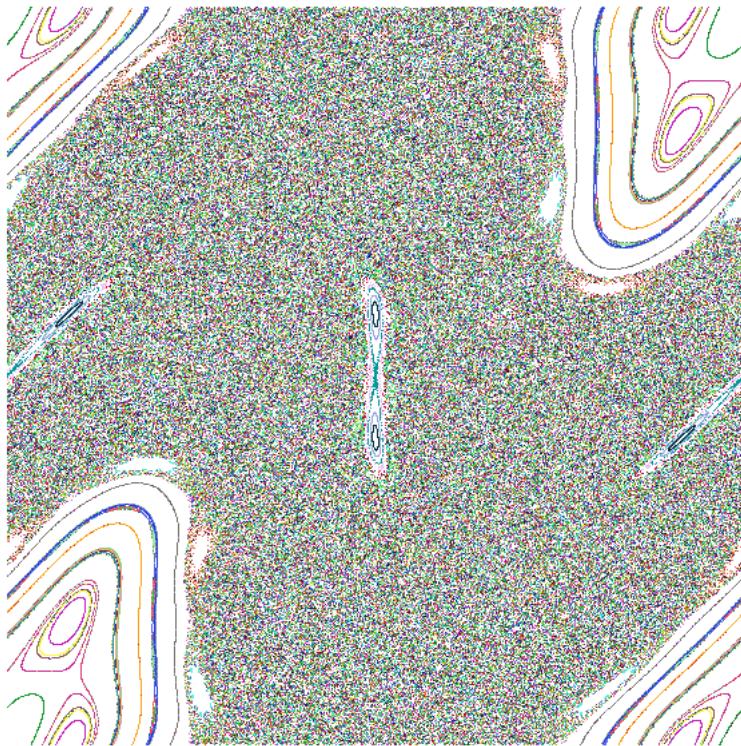


Figure 8.6: Phase space of the standard map for  $k = 2.1$ . Different orbits are labeled with different colors.

few well defined planes. So the lesson is: be very careful when you choose a uniform random number generator!

Indeed it is very difficult to define the criteria for randomness in a pseudo-random sequence, and the opinions of the practitioners and researchers in the field are often very personal. We could say that a sequence is pseudo-random when it *looks* random. This means that numbers must not follow evident patterns, that they must be evenly distributed, that they must not be correlated, etc.

Several tests of randomness have been standardized: here I mention the “Diehard battery of tests of randomness”, developed by George Marsaglia (and freely available on the web: <http://www.stat.fsu.edu/pub/diehard/>). The tests in this suite are<sup>2</sup>:

**Birthday spacings** : Choose random points on a large interval. The spacings between the points should be asymptotically exponentially dis-

---

<sup>2</sup>list adapted from the `tests.txt` file that comes with the CD-rom distributed by Marsaglia

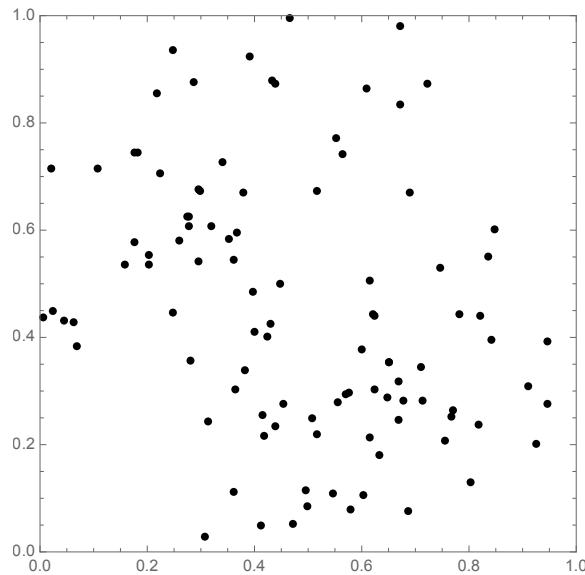


Figure 8.7: The first 100 steps along an orbit in a chaotic region of the phase space of the standard map for  $k = 2.1$ .

tributed. This is clearly related to the memoryless character of true Poisson processes, and the name is based on the “birthday paradox”, whereby the birthdays of several unconnected people seem to cluster.

**Overlapping permutations** : Analyze sequences of five consecutive random numbers. The 120 possible orderings (permutations of the ordered set of the 5 consecutive numbers) should occur with statistically equal probability.

**Ranks of matrices** : Select some number of bits from some number of random numbers to form a  $31 \times 31$  matrix over the field 0,1, then determine the rank of the matrix. That rank can be from 0 to 31, but ranks  $< 28$  are rare. This is repeated for  $32 \times 32$  matrices, and for  $6 \times 8$  matrices.

**Bitstream test** : The sequence of random numbers is treated as a bit stream, and bits are taken 20 at a time, to form 20-bit long words. The test searches for missing words, forms a statistic, and compares it with the expected result.

**Monkey (or Gorilla) tests** : Treat sequences of some number of bits as words. Count the overlapping words in a stream. The number of words that don't appear should follow a known distribution.

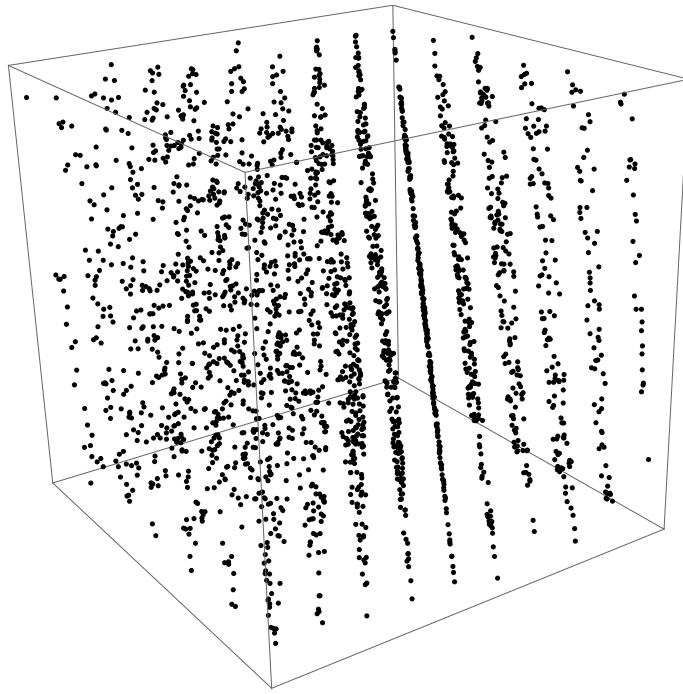


Figure 8.8: 2500 triples of numbers obtained with RANDU: the random vectors are not randomly distributed in space, instead they cluster in a limited number of planes. This is an infamous example of a flawed PRNG.

**Count the 1s :** Count the 1 bits in each of either successive or chosen bytes. Convert the counts to letters, and count the occurrences of five-letter words.

**Parking lot test :** Randomly place unit circles in a  $100 \times 100$  square. If the circle overlaps an existing one, try again. After 12,000 tries, the number of successfully parked circles should follow a certain normal distribution.

**Minimum distance test :** Randomly place 8,000 points in a  $10,000 \times 10,000$  square, then find the minimum distance between the pairs. The square of this distance should be exponentially distributed with a certain mean.

**The 3D spheres test :** Randomly choose 4,000 points in a cube of edge 1,000. Center a sphere on each point, whose radius is the minimum distance to another point. The smallest sphere's volume should be exponentially distributed with a certain mean.

**The squeeze test :** Multiply  $2^{31} = 2147483647$  by random floats on  $[0,1]$

until you reach 1. Repeat this 100,000 times. The number of floats needed to reach 1 should follow a certain distribution.

**Overlapping sums test** : Generate a long sequence of random floats on  $[0, 1)$ . Add sequences of 100 consecutive floats. The sums should be normally distributed with characteristic mean and sigma.

**Runs test** : Generate a long sequence of random floats on  $[0, 1)$ . Count ascending and descending runs. This example shows how runs are counted: .123,.357,.789,.425,.224,.416,.95, contains an up-run of length 3, a down-run of length 2 and an up-run of (at least) 2, depending on the next values. The covariance matrices for the runs-up and runs-down are well known.

**The craps test** : Play 200,000 games of craps, counting the wins and the number of throws per game. Each count should follow a certain distribution.

### 8.3.3 RANLUX

Present users of CERNLIB can pick RANLUX as their pseudo-random number generator: according to the description in the CERNLIB manual, *RANLUX ... is an advanced pseudo-random number generator based on the RCARRY algorithm proposed in 1991 by Marsaglia and Zaman. RCARRY used a subtract-and-borrow algorithm with a period on the order of  $10^{171}$  but still had detectable correlations between numbers. Martin Luescher proposed the RANLUX algorithm in 1993; RANLUX generates pseudo-random numbers using RCARRY but throws away numbers to destroy correlations. Thus RANLUX trades execution speed for quality; by choosing a larger luxury setting one gets better random numbers slower. By the tests available at the time it was proposed, RANLUX at the default luxury setting appears to be a significant advance quality over previous generators.*

The RANLUX (*random numbers with luxury*) PRNG starts from the RCARRY algorithm of Marsaglia and Zaman, which generates a sequence of integers  $x_0, x_1, x_2, \dots$  and a sequence of “carry bits”  $c_0, c_1, c_2, \dots$ . The algorithm requires an arbitrary (large) integer  $b$  which is called *the base*, and two other integers  $r > s > \geq 1$ , *the lags*: for  $n \geq r$  one computes the difference

$$\Delta_n = x_{n-s} - x_{n-r} - c_{n-1}$$

and then determines  $x_n$  and  $c_n$  as follows

$$\begin{aligned} x_n &= \Delta_n & c_n &= 0 & \text{if } \Delta_n \geq 0 \\ x_n &= \Delta_n + b & c_n &= 1 & \text{if } \Delta_n < 0 \end{aligned}$$

This algorithm depends on the initialisation sequence  $x_0, x_1, x_2, \dots, x_r$  and on the carry bit  $c_{r-1}$ , and it outputs  $x_n$  such that  $0 \leq x_n < b$  with an intricate chaotic dynamics. The algorithm has several “certified” good properties (see the original paper by M. Lüscher) but it also has short range correlations that must be destroyed: in RANLUX this is done by reading  $r$  numbers and discarding the next  $p - r$  in the sequence, then reading the next  $r$  and so on. RANLUX allows for different choices of the number  $p$  which define different “luxury levels” (we take the luxury of improving the quality at the expense of efficiency).

In what follows we assume that whatever system you use, some good-quality pseudo-random number generator (PRNG) is available, and we use it to obtain non-uniformly distributed numbers. The next sections deal with some standard methods to achieve this goal.

## 8.4 The transformation method

Taking for granted that we can generate uniform random numbers, we must still generate numbers with other distribution functions. Here we discuss the transformation method that was first proposed by Stanislaw Ulam. Consider two random variables  $x$  and  $y$  with pdf  $p_x(x)$  and  $p_y(y)$ , respectively, related by a transformation  $y = y(x)$ . Then the distribution function of  $y$  is

$$F(y) = \int_{-\infty}^{y(x)} p_y(y') dy' = \int_{-\infty}^x p_x(x') dx' \quad (8.9)$$

If  $x$  is uniformly distributed between 0 and 1, then  $p_x(x) = 1$ , and we get

$$F(y) = \int_{-\infty}^{y(x)} p_y(y') dy' = x \quad (8.10)$$

and finally

$$y = F^{-1}(x) \quad (8.11)$$

Consider, e.g., the exponential distribution

$$p_y(y) = \lambda \exp(-\lambda y) \quad (8.12)$$

then

$$F(y) = \int_0^y \lambda \exp(-\lambda y') dy' = 1 - \exp(-\lambda y) \quad (8.13)$$

therefore we must solve

$$1 - \exp(-\lambda y) = x \quad (8.14)$$

i.e.,

$$y = -\frac{1}{\lambda} \ln(1 - x) \quad (8.15)$$

or also

$$y = -\frac{1}{\lambda} \ln x \quad (8.16)$$

(because both  $x$  and  $1 - x$  are uniformly distributed over  $(0, 1)$ ).

This method can also be used with empirical distribution functions.

#### 8.4.1 Gaussian variates with the transformation method (Box-Müller)

The transformation method works also in the multivariate case. Consider a pair of i.i.d. Gaussian variates  $y_1$  and  $y_2$ , with the target pdf

$$p(y_1, y_2) = \frac{1}{2\pi} \exp\left(-\frac{y_1^2 + y_2^2}{2}\right) \quad (8.17)$$

Now let

$$x_1 = \exp\left(-\frac{y_1^2 + y_2^2}{2}\right) \quad (8.18a)$$

$$x_2 = \frac{1}{2\pi} \arctan \frac{y_2}{y_1} \quad (8.18b)$$

Both numbers are in the  $(0, 1)$  interval,  $x_1$  maps the distance from the origin on the interval  $(0, 1)$ , while  $x_2$  maps the uniform angular distribution over  $(0, 2\pi)$  on the same  $(0, 1)$  interval, and moreover

$$y_1 = \sqrt{-2 \ln x_1} \cos 2\pi x_2 \quad (8.19a)$$

$$y_2 = \sqrt{-2 \ln x_1} \sin 2\pi x_2 \quad (8.19b)$$

It is easy to check that if  $x_1$  and  $x_2$  are uniformly distributed, then the pair  $y_1$  and  $y_2$  each have a  $N(0, 1)$  pdf. Indeed the Jacobian determinant is

$$J = \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} = \begin{vmatrix} -x_1 y_1 & -x_1 y_2 \\ -\frac{1}{2\pi} \frac{y_2/y_1^2}{1+y_2^2/y_1^2} & \frac{1}{2\pi} \frac{1/y_1}{1+y_2^2/y_1^2} \end{vmatrix} = -\frac{x_1}{2\pi} \quad (8.20)$$

therefore

$$|J| = \frac{1}{2\pi} \exp\left(-\frac{y_1^2 + y_2^2}{2}\right) \quad (8.21)$$

and the  $y$  variates are normally distributed.

This method is seldom used now, there are better, more efficient ways to generate Gaussian variates.

## 8.5 The acceptance-rejection method

This method was first proposed by John von Neumann, and it is schematically illustrated in figure 8.9. The method works for multivariate distributions as well, and here is a simple, very general proof, adapted from [14].

The proof consists of three short lemmas and a conclusion:

- Lemma 1: let  $\mathbf{x}$  be a  $p$ -dimensional random variable with pdf  $g(\mathbf{x})$  defined over a subset  $S$  of  $\mathbb{R}^p$  ( $g$  may be unnormalized), let  $c > 0$  be a real constant, and let  $y$  be a real random variable such that its conditional pdf – given  $\mathbf{x}$  – is uniform in  $(0, cg(\mathbf{x}))$ . Then the joint pdf of the  $(p+1)$ -dimensional random variable  $\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix}$  is uniform in the set

$$\mathcal{A} = \left\{ \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} : \mathbf{x} \in S, y \in (0, cg(\mathbf{x})) \right\} \quad (8.22)$$

Indeed, if  $h(\mathbf{x}, y)$  is the joint pdf, then if  $\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} \in \mathcal{A}$

$$h(\mathbf{x}, y) = p(y|\mathbf{x})p_{\mathbf{x}}(\mathbf{x}) \propto \frac{1}{cg(\mathbf{x})}g(\mathbf{x}) = \frac{1}{c} \quad (8.23)$$

and 0 otherwise.

- Lemma 2: let  $V(\cdot)$  represent the  $m$ -dimensional volume of a set. If a random variable  $\mathbf{z}$  has a uniform distribution in  $\mathcal{A} \subset \mathbb{R}^m$ , where  $0 < V(\mathcal{A}) < \infty$ , and if  $\mathcal{B}$  is another set such that  $\mathcal{B} \subset \mathcal{A}$  with  $V(\mathcal{B}) > 0$ , then  $\mathbf{z}$  has a uniform conditional distribution in  $\mathcal{B}$ . Indeed:

$$p(z \in \mathcal{B}|z \in \mathcal{A}) = \frac{p(z \in \mathcal{B})}{p(z \in \mathcal{A})} = \frac{V(\mathcal{B})}{V(\mathcal{A})} \quad (8.24)$$

- Lemma 3: if the  $(p+1)$ -dimensional random variable  $\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix}$  has a uniform distribution in the set

$$\mathcal{B} = \left\{ \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} : \mathbf{x} \in S, y \in (0, f(\mathbf{x})) \right\} \quad (8.25)$$

where  $f(\mathbf{x})$  is a pdf, then the pdf of  $\mathbf{x}$  is  $f(\mathbf{x})$ . Indeed, because of Lemma 2,  $y$  has a conditional uniform distribution for every  $\mathbf{x}$ , i.e.,  $p(y|\mathbf{x}) = 1/f(\mathbf{x})$ , and therefore

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(y|\mathbf{x})} \propto f(\mathbf{x}) \quad (8.26)$$

Putting the lemmas together we find

$$p_{\mathbf{x}}(\mathbf{x}) = f(\mathbf{x}) \quad (8.27)$$

if  $f(\mathbf{x})$  is properly normalized.

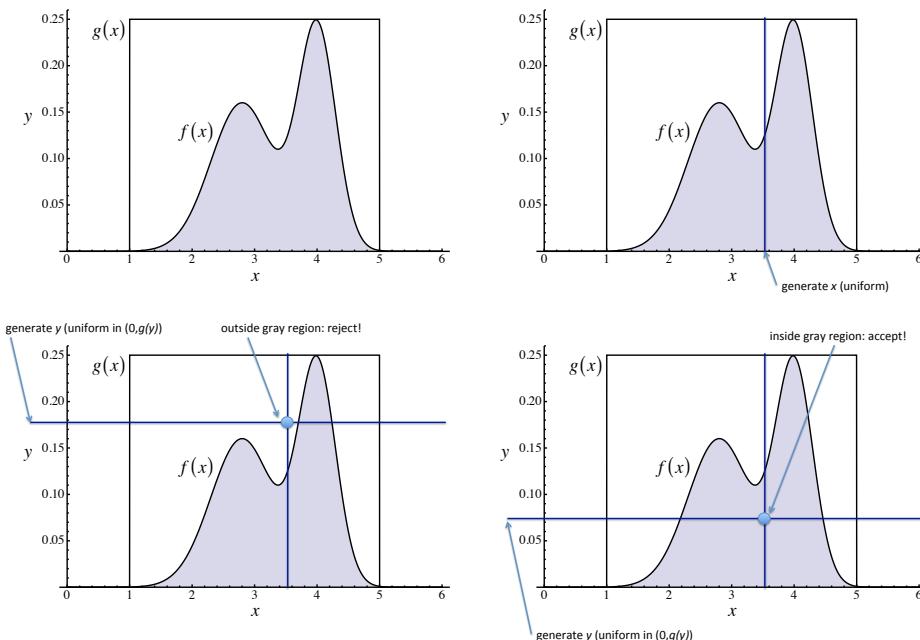


Figure 8.9: Schematic illustration of the acceptance-rejection method. Start from a pdf  $f(x)$  defined in the interval  $(x_{min}, x_{max})$ , and take an enclosing pdf  $g(x)$ : notice that in general these are unnormalized pdf's. Now generate  $x$  according to the pdf  $g(x)$  – in the case shown in the figure this means uniform in  $(x_{min}, x_{max})$  – (upper right panel), and  $y$ , uniform in  $(0, g(x))$ . If  $y > f(x)$ , reject  $x$  (lower left panel), otherwise accept it (lower right panel). The accepted values  $x$  have pdf  $\propto f(x)$ .

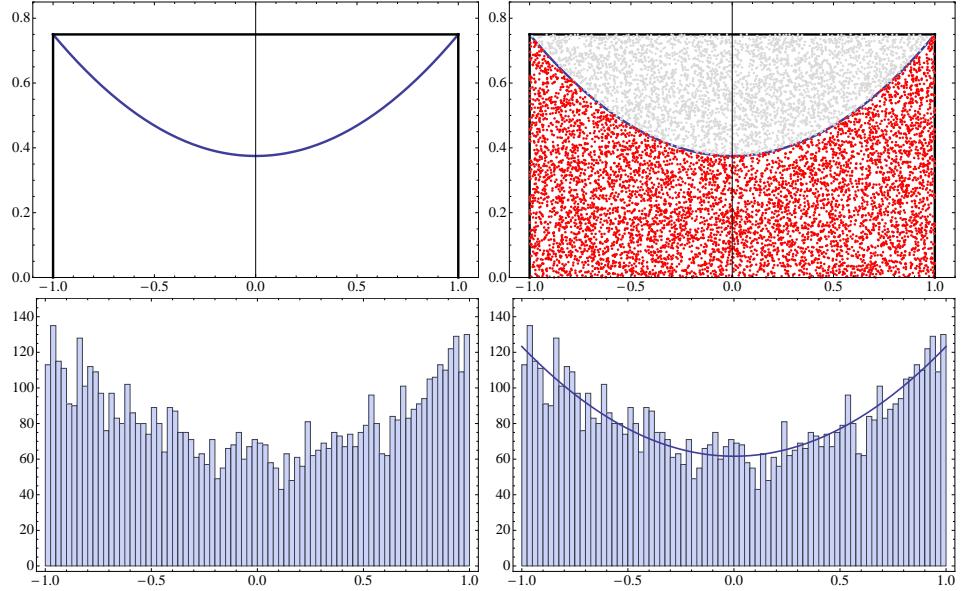


Figure 8.10: Example of generation of angles for the process  $e^+e^- \rightarrow \mu^+\mu^-$ . (Upper left panel) Plot of  $f(x)$  vs.  $x$ , where  $x = \cos \theta$ . The upper line marks the containing, unnormalized uniform distribution  $g(x) = 3/4$ . (Upper right panel) Uniform distribution of random points: only those below the curve  $f(x)$  are accepted (red dots). (Lower left panel) Histogram of the generated  $x$  values. (Lower right panel) Comparison with the distribution  $f(x)$ .

### 8.5.1 Examples: angular distributions in $e^+e^-$ collider physics

Consider first the following example: we want to simulate the angular distribution of muons in the  $e^+e^- \rightarrow \mu^+\mu^-$  process in a collider, and we know that the differential cross section is

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{4s} (1 + \cos^2 \theta) \quad (8.28)$$

(see, e.g., [26]). For a fixed CM energy, the only random variables are the zenithal and azimuthal angles  $\theta$  and  $\phi$ , and since  $d\Omega = d(\cos \theta)d\phi$  this means that the cross section above corresponds to a uniform distribution of  $\phi$  in  $(0, 2\pi)$ , and that the normalized pdf of  $\cos \theta$  is

$$f(x) = \frac{3}{8}(1 + x^2) \quad (8.29)$$

where  $x = \cos \theta$ , and  $-1 \leq x \leq 1$ . Figure 8.10 shows how the angles are drawn from this distribution  $f(x)$ .

Next consider the  $e^+e^- \rightarrow e^+e^-$  process (Bhabha scattering): the tree-level Feynman diagrams are shown in figure 8.11, and the differential cross

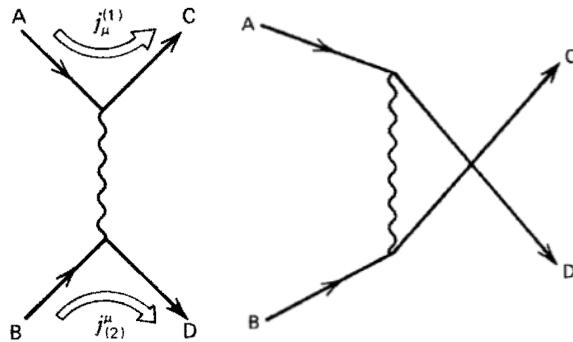


Figure 8.11: Tree-level Feynman diagrams for the  $e^+e^- \rightarrow e^+e^-$  process. This differs from the  $e^+e^- \rightarrow \mu^+\mu^-$  process because it includes both an s-channel diagram and a t-channel diagram. This produces an interference term in the scattering cross-section, and it leads to a more interesting random angle generation problem.

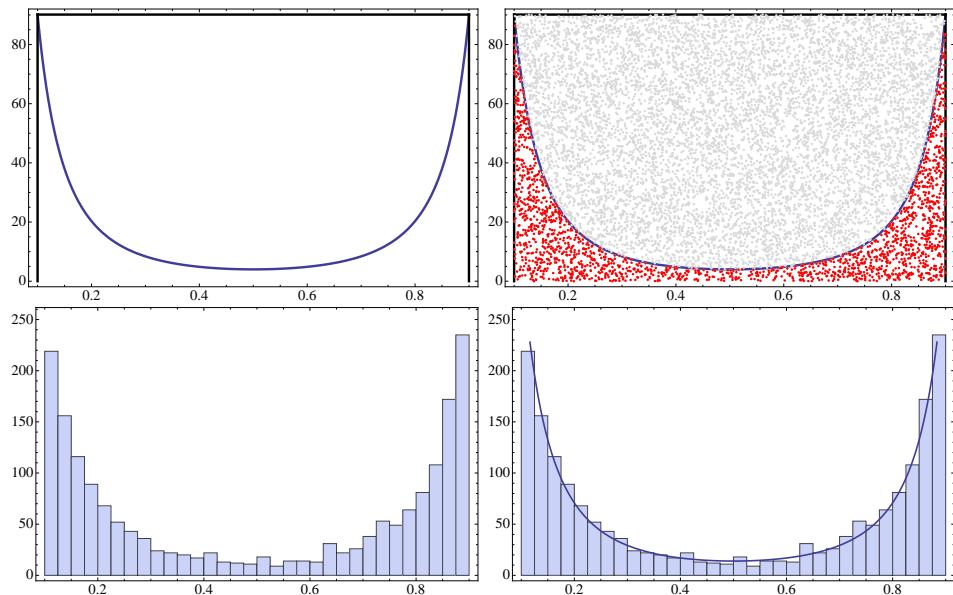


Figure 8.12: Generation of angles for the process  $e^+e^- \rightarrow e^+e^-$ . The panels are as in figure 8.10. Here the acceptance-rejection method can be applied, but only after a provision is made for the divergence of the angular distribution in the forward and in the backward direction. All plots are drawn vs. the angular variable  $x = \sin^2 \theta/2$ . The efficiency of this sampling is very low, rather than a flat  $g(x)$  one should take into account a function that closely matches the target  $f(x)$ .

section is

$$\frac{d\sigma}{d\Omega} = \frac{m_e^2 \alpha^2}{16p^4} \left( \frac{1}{\sin^4 \frac{\theta}{2}} + \frac{1}{\cos^4 \frac{\theta}{2}} - \frac{1}{\sin^2 \frac{\theta}{2} \cos^2 \frac{\theta}{2}} \right) \quad (8.30)$$

(see, e.g., F. Halzen and A. D. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics* (Wiley, 1984). ) If we let  $x = \sin^2 \theta/2$ , the angular pdf becomes

$$f(x) \propto 2 \left( \frac{1}{x^2} + \frac{1}{(1-x)^2} - \frac{1}{x(1-x)} \right) \quad (8.31)$$

Now the angular distribution function has a divergence both in the forward and in the backward direction. This problem is circumvented observing that it is impossible to perform measurements along these directions because of the circulating beam, and that there must be a minimum angle beyond which measurement becomes impossible. This cut in the distribution allows normalization and generation of random angles with the acceptance-rejection method (try it if you know how to code). Figure 8.12 shows the MC simulation results that parallel those of figure 8.10. Notice also that the tree-level diagrams are an incomplete description of the scattering process with two electrons in the final state, and that other processes such as  $e^+e^- \rightarrow e^+e^-\gamma$  must be taken into account.

## 8.6 Cross-sections and the transformation method

In the previous section we have studied the angular distribution of the final state in the center-of-mass (CM) frame, where most cross-sections are defined. However, in many cases we are not so fortunate: it often happens that the laboratory frame does not coincide with the CM frame. Moreover the CM frame often moves at high speed with respect to the laboratory frame, and the two frames are related by a relativistic Lorentz transformation. In such cases we can utilize the transformation method.

The pdf in momentum space is specified by the differential cross-section  $d\sigma/dp_x dp_y dp_z$ , however, from the transformation of volume elements in Cartesian and polar coordinates, we know that

$$dp_x dp_y dp_z = \left| \frac{\partial(p_x, p_y, p_z)}{\partial(p, \theta, \varphi)} \right| dp d\theta d\varphi = p^2 \sin \theta \, dp d\theta d\varphi = p^2 dp \, d\Omega \quad (8.32)$$

Therefore

$$p^2 \frac{\partial^3 \sigma(p_x, p_y, p_z)}{\partial p_x \partial p_y \partial p_z} = \frac{\partial^2 \sigma(p, \Omega)}{\partial p \partial \Omega} \quad (8.33)$$

It often happens that the initial particle momenta are not balanced and that the CM frame does not coincide with the laboratory frame, or that we

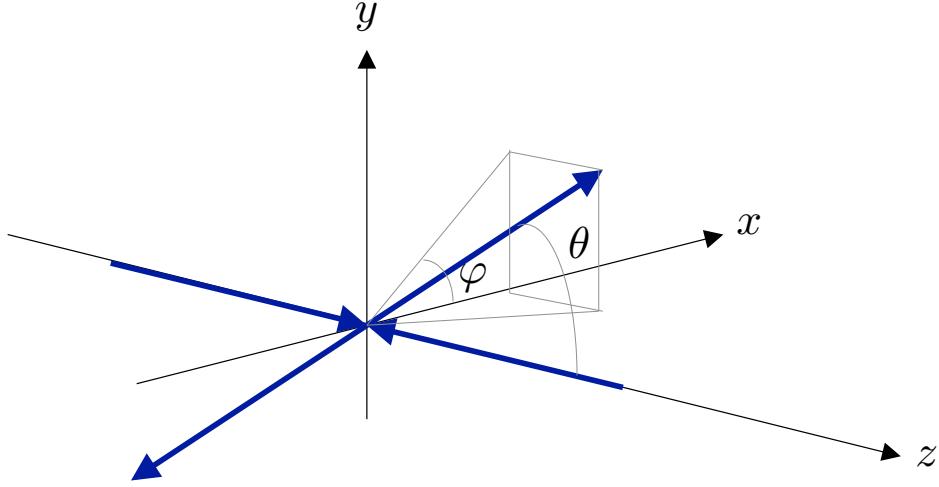


Figure 8.13: Geometry of scattering. The initial particles move along the  $z$ -axis, and collide head-on. Since the initial particle momenta can be different, the  $z$ -axis is also the direction of the relativistic boost.

wish to transform the cross-section to a different moving frame. We consider in general Lorentz transformations along the  $z$ -axis, so that

$$\begin{aligned} p_z &= \gamma(p'_z + \beta E'/c) & p'_z &= \gamma(p_z - \beta E/c) \\ p_y &= p'_y & p'_y &= p_y \\ p_x &= p'_x & p'_x &= p_x \\ E &= \gamma(E' + \beta c p'_z) & E' &= \gamma(E - \beta c p_z) \end{aligned} \quad (8.34)$$

If we introduce polar coordinates such that

$$\begin{aligned} p_x &= p \sin \theta \cos \varphi \\ p_y &= p \sin \theta \sin \varphi \\ p_z &= p \cos \theta \end{aligned} \quad (8.35)$$

we see at once – from the equalities  $p_x = p'_x$  and  $p_y = p'_y$  – that  $\varphi' = \varphi$ , and the Lorentz transformations given above become

$$\begin{aligned} p \cos \theta &= \gamma(p' \cos \theta' + \beta E'/c) & p' \cos \theta' &= \gamma(p \cos \theta - \beta E/c) \\ p \sin \theta &= p' \sin \theta' & p' \sin \theta' &= p \sin \theta \\ \varphi &= \varphi' & \varphi' &= \varphi \\ E &= \gamma(E' + \beta c p'_z) & E' &= \gamma(E - \beta c p_z) \end{aligned} \quad (8.36)$$

In order to apply the transformation method, we note that the equality of probability mass in the usual application of the method, is replaced here

by the equality of the total number of events in corresponding transformed volumes, i.e.,

$$\frac{\partial^2\sigma(p,\Omega)}{\partial p\partial\Omega}dpd\Omega = \frac{\partial^2\sigma(p',\Omega')}{\partial p'\partial\Omega'}dp'd\Omega' = \frac{\partial^2\sigma(p',\Omega')}{\partial p'\partial\Omega'} \left| \frac{\partial(p',\Omega')}{\partial(p,\Omega)} \right| dpd\Omega \quad (8.37)$$

therefore

$$\frac{\partial^2\sigma(p,\Omega)}{\partial p\partial\Omega} = \frac{\partial^2\sigma(p',\Omega')}{\partial p'\partial\Omega'} \left| \frac{\partial(p',\Omega')}{\partial(p,\Omega)} \right| \quad (8.38a)$$

$$\frac{\partial^2\sigma(p',\Omega')}{\partial p'\partial\Omega'} = \frac{\partial^2\sigma(p,\Omega)}{\partial p\partial\Omega} \left| \frac{\partial(p,\Omega)}{\partial(p',\Omega')} \right| \quad (8.38b)$$

By applying the chain rule, we note that

$$\frac{\partial(p,\Omega)}{\partial(p',\Omega')} = \frac{\partial(p,\Omega)}{\partial(p,\theta,\varphi)} \frac{\partial(p,\theta,\varphi)}{\partial(p_x,p_y,p_z)} \frac{\partial(p_x,p_y,p_z)}{\partial(p'_x,p'_y,p'_z)} \frac{\partial(p'_x,p'_y,p'_z)}{\partial(p',\theta',\varphi')} \frac{\partial(p',\theta',\varphi')}{\partial(p',\Omega')} \quad (8.39)$$

where the Lorentz transformation operates only in the Jacobian in the middle. Since

$$\begin{aligned} \frac{\partial(p,\Omega)}{\partial(p,\theta,\varphi)} &= \sin\theta & \frac{\partial(p',\theta',\varphi')}{\partial(p',\Omega')} &= \frac{1}{\sin\theta'} \\ \frac{\partial(p,\theta,\varphi)}{\partial(p_x,p_y,p_z)} &= \frac{1}{p^2 \sin\theta} & \frac{\partial(p'_x,p'_y,p'_z)}{\partial(p',\theta',\varphi')} &= p'^2 \sin\theta' \end{aligned} \quad (8.40)$$

and

$$\frac{\partial(p_x,p_y,p_z)}{\partial(p'_x,p'_y,p'_z)} = \frac{\partial p_z}{\partial p'_z} = \gamma \left( 1 + \frac{\beta}{c} \frac{\partial E'}{\partial p'_z} \right) = \gamma \left( 1 + \frac{\beta}{c} \frac{p'_z c^2}{E'} \right) = \frac{E}{E'} \quad (8.41)$$

we find

$$\frac{\partial(p,\Omega)}{\partial(p',\Omega')} = \frac{p'^2}{p^2} \frac{E}{E'} \quad (8.42)$$

Therefore

$$\frac{\partial^2\sigma(p',\Omega')}{\partial p'\partial\Omega'} = \frac{\partial^2\sigma(p,\Omega)}{\partial p\partial\Omega} \frac{p'^2}{p^2} \frac{E}{E'} \quad (8.43)$$

## 8.7 Monte Carlo simulation of non relativistic electron scattering in matter

Here we examine a simple particle transport Monte Carlo for nonrelativistic electrons with kinetic energy below 50 keV: the simulated process has great interest in several fields, like human radiation physics and semiconductor processing, and at the same time it does not lead to the more complex schemes that must be implemented at higher energies<sup>3</sup>.

---

<sup>3</sup>see also W. Williamson and G. C. Duncan, Am. J. Phys. **54** (1986) 262

### Physical processes:

**Ionization energy loss:** we use a version of the Bethe formula for the mean energy loss for electrons<sup>4</sup>:

$$\frac{dE}{dx} = -\frac{4\pi N_A r_e^2 m_e c^2}{\beta^2} \frac{Z}{A} \rho \left[ \ln \frac{\gamma m_e c^2 \beta \sqrt{\gamma - 1}}{\sqrt{2I}} + \frac{1}{2}(1 - \beta^2) - \frac{2\gamma - 1}{2\gamma^2} \ln 2 + \frac{1}{16} \left( \frac{\gamma - 1}{\gamma} \right)^2 \right] \quad (8.44)$$

where  $E$  is the total electron energy,  $x$  is the path length,  $r_e = e^2/(4\pi\epsilon_0 m_e c^2)$  is the classical electron radius,  $N_A$  is Avogadro's constant (expressed as number of atoms/gram-mole), and where the target material has density  $\rho$ , atomic number  $Z$ , mean ionization potential  $I$ . The mean ionization potential  $I$  (also called mean excitation energy) is approximately  $I \approx 16Z^{0.9}$  (eV) (see also figure 8.14). Figures 8.15 and 8.16 show plots of the mean energy loss vs. energy. It is also useful to evaluate the constant

$$4\pi N_A r_e^2 m_e c^2 \approx 3.07 \cdot 10^{-5} \text{ MeV m}^2 \text{g}^{-1} = 0.307 \frac{\text{MeV}}{\text{g/cm}^2}$$

For non relativistic electrons, the formula for the mean energy loss can be greatly simplified

$$\frac{dE}{dx} = -\frac{4\pi N_A r_e^2 m_e c^2}{\beta^2} \frac{Z}{A} \rho \left( \ln \frac{m_e c^2 \beta^2}{2I} + \frac{1}{2} \right) \quad (8.45)$$

**Scattering process:** the continuous energy loss is the result of a large number of soft scattering processes. However, from time to time the electron approaches a nucleus and undergoes a hard scattering. The scattering process off a naked nucleus is well described by the Rutherford scattering cross-section

$$\frac{d\sigma_R}{d\Omega} = \frac{\alpha^2 \hbar^2 c^2}{p^2 v^2} Z^2 \frac{1}{(1 - \cos \theta)^2} \quad (8.46)$$

where  $p = m_e v$  is the non relativistic electron momentum. The electrons that surround nuclei modify this cross-section: if we assume that the screened Coulomb potential energy decays exponentially

$$V(r) = -\frac{Ze^2}{4\pi\epsilon_0 r} e^{-\mu r} \quad (8.47)$$

---

<sup>4</sup>this is equation (1.28) in C. Grupen and B. Shwartz, *Particle Detectors*, Cambridge Univ. Press, Cambridge 2008. See also E. Fermi, *Nuclear Physics*, The University of Chicago Press, Chicago, 1950, for an introduction to the physics behind this equation.

then it can be shown<sup>5</sup> that the new differential cross-section is

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2 \hbar^2 c^2}{p^2 v^2} Z^2 \frac{1}{(1 - \cos \theta + \hbar^2 \mu^2 / 4p^2)^2} \quad (8.48)$$

The screening effect was included in the theory of multiple scattering developed by Molière, Bethe and others, with a screening radius estimated from the Thomas-Fermi model of the atom

$$\frac{1}{\mu} = 0.885 Z^{-1/3} r_B \quad (8.49)$$

where  $r_B = 4\pi\epsilon_0\hbar^2/m_e e^2 \approx 0.529 \text{ } 10^{-10}\text{m}$  is Bohr's radius.

**Mean free path between scatterings:** now we let  $\eta = \hbar^2 \mu^2 / 4p^2$ , so that the cross-section writes

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2 \hbar^2 c^2}{p^2 v^2} Z^2 \frac{1}{(1 - \cos \theta + 2\eta)^2} \quad (8.50)$$

then, integrating this differential cross-section, we find

$$\sigma = \int_0^{2\pi} d\phi \int_{-1}^{+1} d(\cos \theta) \frac{d\sigma}{d\Omega} \quad (8.51a)$$

$$= 2\pi \frac{\alpha^2 \hbar^2 c^2}{p^2 v^2} Z^2 \int_{-1}^{+1} \frac{dt}{(1 - t + 2\eta)^2} \quad (8.51b)$$

$$= 2\pi \frac{\alpha^2 \hbar^2 c^2}{p^2 v^2} Z^2 \left. \frac{1}{(1 - t + 2\eta)} \right|_{-1}^{+1} \quad (8.51c)$$

$$= 2\pi \frac{\alpha^2 \hbar^2 c^2}{p^2 v^2} Z^2 \left( \frac{1}{2\eta} - \frac{1}{2 + 2\eta} \right) \quad (8.51d)$$

$$= 2\pi \frac{\alpha^2 \hbar^2 c^2}{p^2 v^2} Z^2 \frac{1}{2\eta(1 + \eta)} \quad (8.51e)$$

$$= 2\pi \frac{\alpha^2 \hbar^2 c^2}{(m_e c^2)^2 \beta^4} Z^2 \frac{1}{2\eta(1 + \eta)} \quad (8.51f)$$

The mean free path can be derived from this total cross-section<sup>6</sup>:

$$\begin{aligned} \lambda &= \frac{A}{N_A \rho \sigma} = \frac{A}{N_A \rho} \frac{p^2 v^2}{2\pi \alpha^2 \hbar^2 c^2 Z^2} 2\eta(1 + \eta) \\ &= \frac{A}{Z^2 N_A \rho} \frac{(m_e c^2)^2}{2\pi \alpha^2 \hbar^2 c^2} \beta^4 [2\eta(1 + \eta)] \end{aligned} \quad (8.52)$$

---

<sup>5</sup>see, e.g., F. Mandl, *Quantum Mechanics*, Wiley, Chichester 1992

<sup>6</sup>recall that the number of scatterings per unit length is  $\sigma N$ , where  $N$  is number of atoms per unit volume, and therefore the attenuation of an electron beam is given by  $n(x + dx) = n(x) - N\sigma dx$ , i.e.,  $dn/dx = -N\sigma$ , i.e.,  $\lambda = 1/(N\sigma) = A/(N_A \rho \sigma)$ .

Notice that in these equations Avogadro's constant is once again expressed as number of atoms/gram-mole.

It is useful to evaluate some of the constants in these equations

$$\begin{aligned}\hbar c &\approx 197 \text{ MeV fm} \\ \hbar^2 c^2 &\approx 3.89 \cdot 10^{-26} \text{ MeV}^2 \text{ m}^2 \\ \frac{\hbar^2 c^2}{(m_e c^2)^2} &\approx 1.49 \cdot 10^{-13} \mu\text{m}^2 \\ \frac{\alpha^2 \hbar^2 c^2}{(m_e c^2)^2} &\approx 7.9 \cdot 10^{-18} \mu\text{m}^2 \\ \frac{(m_e c^2)^2}{2\pi\alpha^2\hbar^2 c^2 N_A} &\approx 3.35 \text{ g cm}^{-2}\end{aligned}$$

and note that

$$\begin{aligned}\eta &= \frac{\hbar^2 \mu^2}{4p^2} = \frac{\hbar^2 c^2 \mu^2}{4p^2 c^2} = \frac{\hbar^2 c^2}{4(m_e c^2)^2} \frac{\mu^2}{\beta^2} \\ &= \frac{\hbar^2 c^2}{4(m_e c^2)^2} \frac{Z^{2/3}}{(0.885 r_B)^2 \beta^2} \\ &\approx 1.7 \cdot 10^{-5} \frac{Z^{2/3}}{\beta^2}\end{aligned}$$

and that

$$\lambda \text{ (cm)} \approx 3.35 \frac{A}{Z^2 \rho \text{ (g/cm}^3\text{)}} \beta^4 [2\eta(1 + \eta)]$$

**Simulation of the sequence of scatterings:** for each initial electron with initial momentum  $p_0 = m_e v_0 = (m_e c^2) \beta_0 / c$ , we execute the following loop to simulate a sequence of scatterings (identified by the index  $n$ ):

1. we compute the path length  $\Delta x$  to the next scattering using the exponential pdf  $p(\Delta x) = e^{-\Delta x/\lambda} / \lambda$ ;
2. we use  $\Delta x$  and the Bethe formula to find the mean energy loss  $\Delta E \approx |dE/dx| \Delta x$ ;
3. we update the electron kinetic energy:  $T_{n+1} = T_n - \Delta E$ ;
4. if the updated kinetic energy is  $\leq 0$ , or if the electron exits the material slab, or if the number of steps exceeds a predefined upper limit (this condition is unphysical however a high enough limit has no physical consequences and avoids dangerous loop conditions), then we set  $T_{n+1} = 0$  and we exit the loop;
5. we update the electron momentum:  $p_{n+1}c = \sqrt{2(m_e c^2)T_{n+1}}$ , and  $\beta_{n+1} = p_{n+1}c / (m_e c^2)$ ;
6. we update  $\eta_{n+1} \approx 1.7 \cdot 10^{-5} Z^{2/3} / \beta_{n+1}^2$ ;

7. we compute the azimuthal scattering angle  $\phi$  using a uniform distribution on  $(0, 2\pi)$ ;
8. we compute the scattering angle  $\theta$  using the differential scattering cross-section (see below);
9. using the length  $\Delta x$  and the scattering angles we update the electron position (see below);
10. we return to step 1.

**Generation of the scattering angle:** here we use the transformation method.

Integrating the differential cross-section over the azimuthal angle  $\phi$  we find the probability density of the scattering angle  $\theta$ :

$$p_\theta(\theta) = \frac{1}{\sigma} \frac{d\sigma}{d\theta} = 2\eta(1+\eta) \frac{\sin \theta}{(1 - \cos \theta + 2\eta)^2} \quad (8.53)$$

while the distribution function is

$$P_\theta(\theta) = \int_0^\theta p_\theta(\theta') d\theta' = \int_0^\theta 2\eta(1+\eta) \frac{\sin \theta'}{(1 - \cos \theta' + 2\eta)^2} d\theta' \quad (8.54a)$$

$$= - \frac{2\eta(1+\eta)}{(1 - \cos \theta' + 2\eta)} \Big|_0^\theta \quad (8.54b)$$

$$= - \frac{2\eta(1+\eta)}{(1 - \cos \theta + 2\eta)} + (1+\eta) \quad (8.54c)$$

$$= \frac{(1+\eta)(1 - \cos \theta)}{(1 - \cos \theta + 2\eta)} \quad (8.54d)$$

We know that if we take a random variate  $r_\theta$ , uniformly distributed in  $(0, 1)$  and we set

$$r_\theta = P(\theta) \quad (8.55)$$

then  $\theta$  is distributed according to the pdf  $p_\theta(\theta)$ . Thus we set

$$r_\theta = \frac{(1+\eta)(1 - \cos \theta)}{(1 - \cos \theta + 2\eta)} \quad (8.56)$$

and we find

$$\cos \theta = \frac{(1+\eta) - (1+2\eta)r_\theta}{1 + \eta - r_\theta} \quad (8.57)$$

and finally

$$\theta = \arccos \left[ \frac{(1+\eta) - (1+2\eta)r_\theta}{1 + \eta - r_\theta} \right] \quad (8.58)$$

is the random scattering angle.

**Geometry:** the scattering angle  $\theta$  in the differential cross-section is the deviation from the original direction of the scattered electron. This means that we have to apply a coordinate transformation to find the deviation in the laboratory reference frame. First we note that before the  $(n + 1)$ -th scattering, the momentum vector of the electron points in the direction identified by the angles  $\theta_n$ , and  $\phi_n$  in the laboratory reference frame, and therefore we obtain the corresponding momentum vector simply by rotating the vector  $\mathbf{p}'_n = (0, 0, p_n)^T$  (the momentum vector of the incoming electron in the laboratory frame where the  $z$ -axis is aligned with the direction of the incoming moment vector) first with the matrix

$$R_{xz}(\theta_n) = \begin{pmatrix} \cos \theta_n & 0 & \sin \theta_n \\ 0 & 1 & 0 \\ -\sin \theta_n & 0 & \cos \theta_n \end{pmatrix} \quad (8.59)$$

and next with the matrix

$$R_{xy}(\phi_n) = \begin{pmatrix} \cos \phi_n & -\sin \phi_n & 0 \\ \sin \phi_n & \cos \phi_n & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (8.60)$$

The angles  $\theta_n$  and  $\phi_n$  can be computed from the equality

$$\mathbf{p}_n = (p_x, p_y, p_z) = (p_n \sin \theta_n \cos \phi_n, p_n \sin \theta_n \sin \phi_n, p_n \cos \theta_n)^T \quad (8.61)$$

i.e.,

$$\theta_n = \arctan \frac{p_z}{\sqrt{p_x^2 + p_y^2}} \quad (8.62a)$$

$$\phi_n = \arctan \frac{p_y}{p_x} \quad (8.62b)$$

Since the momentum vector in the primed reference frame after scattering is  $\mathbf{p}'_{n+1} = (p'_n \sin \theta \cos \phi, p'_n \sin \theta \sin \phi, p'_n \cos \theta)^T$ , we find that the new momentum vector in the laboratory frame is

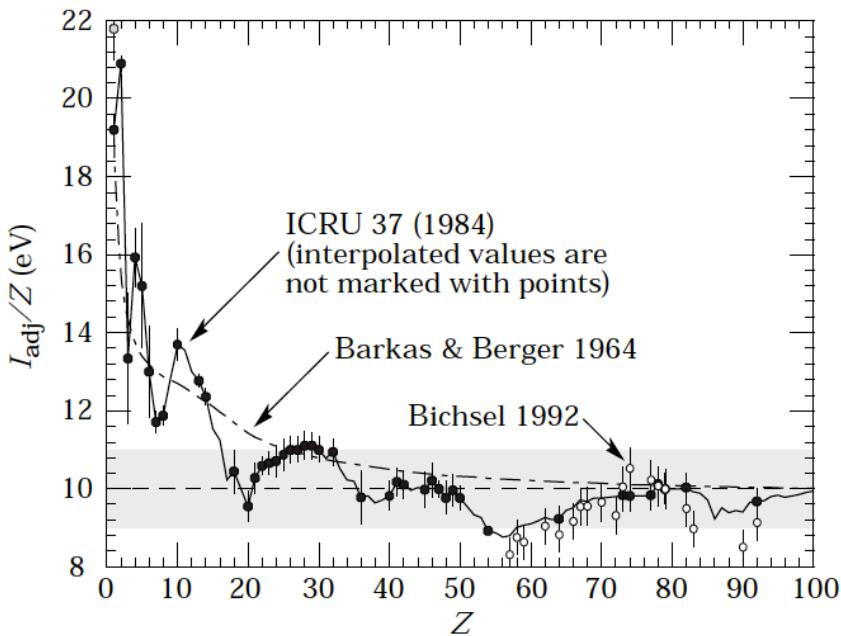
$$\mathbf{p}_{n+1} = R_{xy}(\phi_n)R_{xz}(\theta_n)\mathbf{p}'_{n+1} \quad (8.63)$$

**Limitations and possible refinements:** in this small Monte Carlo there are several approximations and many refinements are possible. Consider, e.g., figure 8.17, which shows the behavior of several formulas for the low-energy differential cross-section: clearly there is some theoretical uncertainty as to the best model to apply at low-energy. Additional uncertainty comes from the material composition: molecular composition turns out to be quite important at very low energy, and in this

MC program should be modified to incorporate mixtures of different atomic species. Other refinements could be introduced at higher energies, e.g., Bethe suggested the replacement  $Z^2 \rightarrow Z(Z + 1)$  in the cross-section formulas, and there are indications that this factor actually compares better with experimental data. Bethe also used a more refined screening model, and thus a better screening parameter  $\eta$ . A prominent feature of this MC program is that it replaces many point interactions with the continuous energy loss of the Bethe formula, while other, more advanced programs use the point interactions. On the other hand, the program cannot easily incorporate another types of continuous interactions, i.e., those with external electric and magnetic fields, because it steps from one scattering to the next, instead of going through a series of fixed steps.

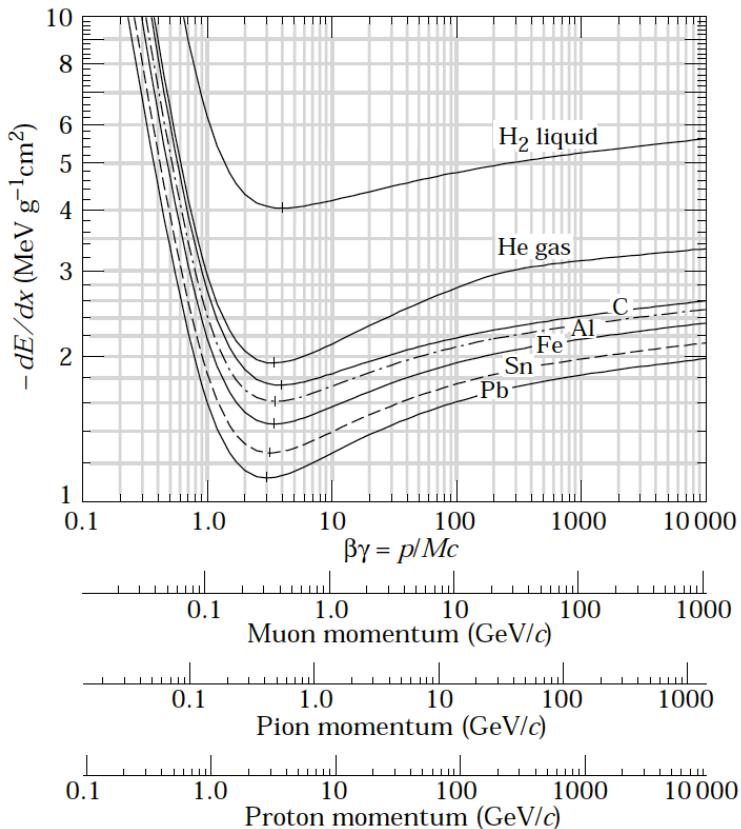
Many other refinements are possible, like full relativistic formulas, etc., however we do not explore this subject further.

Figure 8.18 shows the result of a simulation run with 100 incident 50 keV-electrons on an Al target.



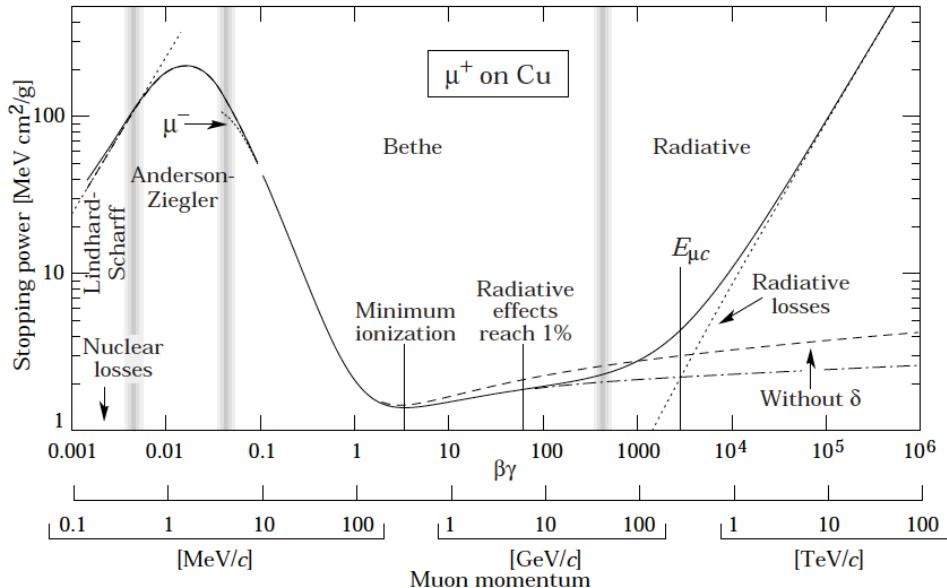
**Figure 27.5:** Mean excitation energies (divided by  $Z$ ) as adopted by the ICRU [10]. Those based on experimental measurements are shown by symbols with error flags; the interpolated values are simply joined. The grey point is for liquid  $\text{H}_2$ ; the black point at 19.2 eV is for  $\text{H}_2$  gas. The open circles show more recent determinations by Bichsel [12]. The dotted curve is from the approximate formula of Barkas [13] used in early editions of this *Review*.

Figure 8.14: Plot of  $I/Z$  vs.  $Z$  (taken from 2010 Review of Particle Physics, K. Nakamura et al. (Particle Data Group), Journal of Physics G **37** (2010) 075021).



**Figure 27.2:** Mean energy loss rate in liquid (bubble chamber) hydrogen, gaseous helium, carbon, aluminum, iron, tin, and lead. Radiative effects, relevant for muons and pions, are not included. These become significant for muons in iron for  $\beta\gamma \gtrsim 1000$ , and at lower momenta for muons in higher- $Z$  absorbers. See Fig. 27.21.

Figure 8.15: Plot of the mean energy loss for muons, pions and protons, according to the Bethe formula (taken from 2010 Review of Particle Physics, K. Nakamura et al. (Particle Data Group), Journal of Physics G **37** (2010) 075021).



**Fig. 27.1:** Stopping power ( $= \langle -dE/dx \rangle$ ) for positive muons in copper as a function of  $\beta\gamma = p/Mc$  over nine orders of magnitude in momentum (12 orders of magnitude in kinetic energy). Solid curves indicate the total stopping power. Data below the break at  $\beta\gamma \approx 0.1$  are taken from ICRU 49 [4], and data at higher energies are from Ref. 5. Vertical bands indicate boundaries between different approximations discussed in the text. The short dotted lines labeled “ $\mu^-$ ” illustrate the “Barkas effect,” the dependence of stopping power on projectile charge at very low energies [6].

Figure 8.16: Plot of the mean energy loss for muons in an extended energy region (taken from 2010 Review of Particle Physics, K. Nakamura et al. (Particle Data Group), Journal of Physics G **37** (2010) 075021. This figure shows that the Bethe formula holds for intermediate energies only, and that one could improve the simulation program developed here by including the detailed description of the low-energy behavior of the mean energy loss.

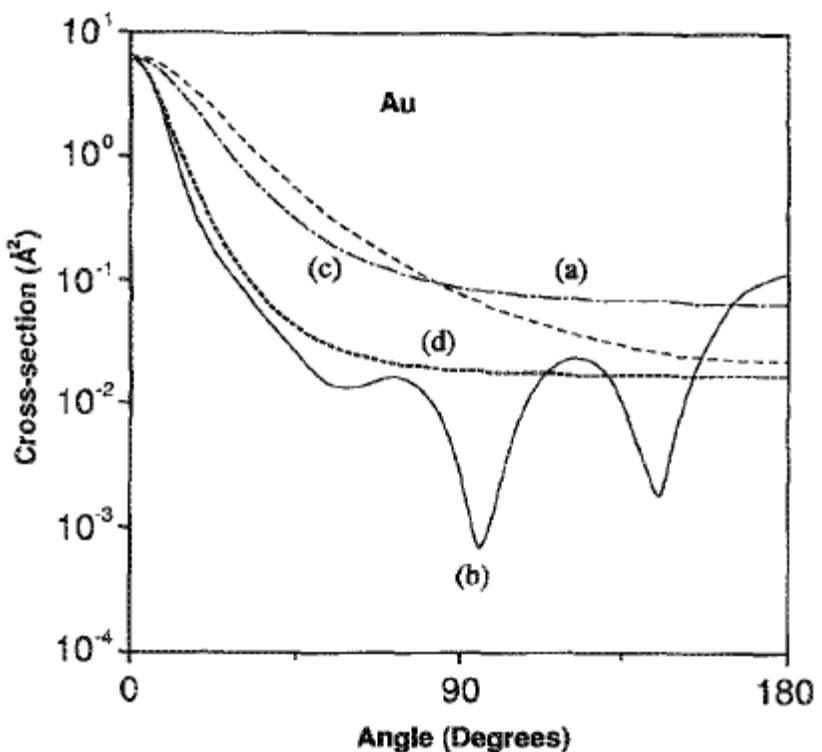


FIG. 4. Differential scattering cross section as a function of angle for 1 keV electrons scattering from Au. (a) Screened Rutherford. (b) Partial wave (Mott) (Ref. 2). (c) Screened Rutherford using the screening parameter Eq. (6). (d) The screened Rutherford plus isotropic distribution ratioed using Eq. (8).

Figure 8.17: Plot of several formulas for the differential cross-section at 1 keV (from R. Browning et al, J. Vac. Sci. Technol. **9** (1991) 3578).

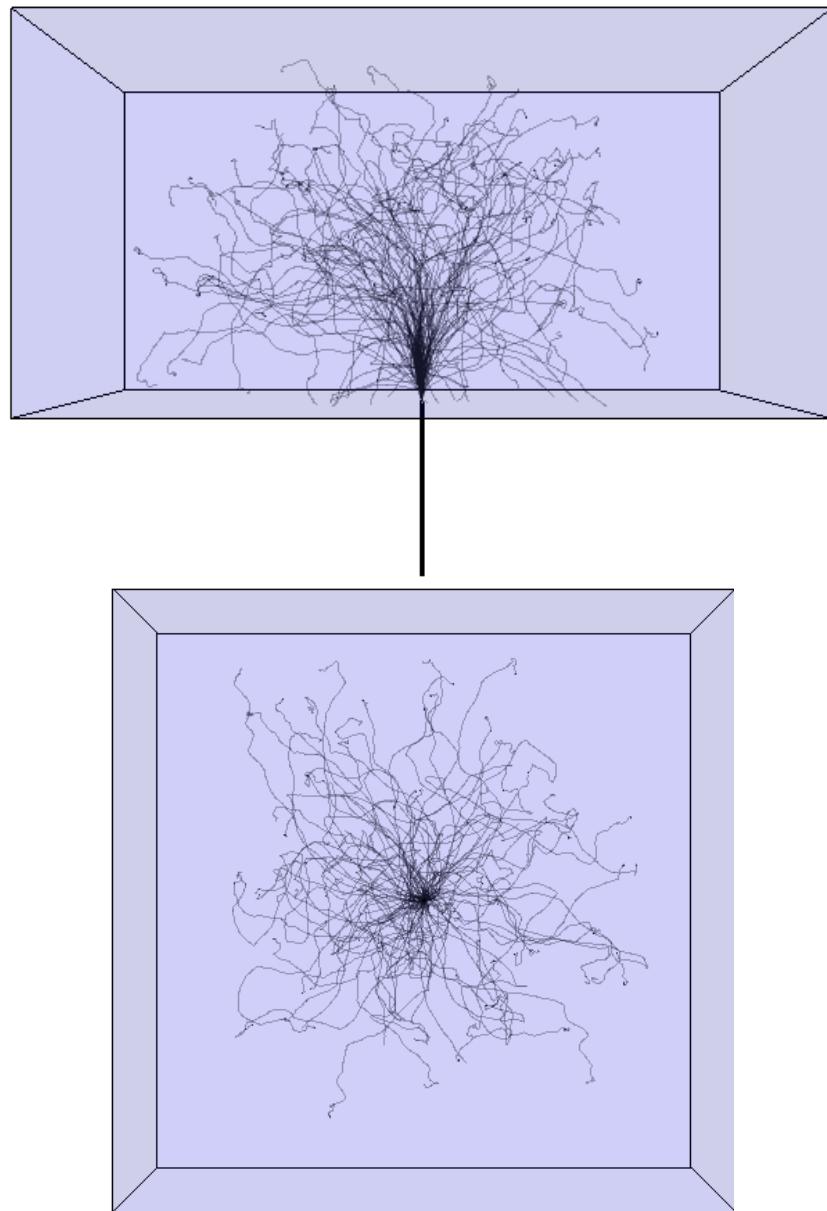


Figure 8.18: Tracks obtained in a simulation with 100 incident electrons, with 50 keV initial kinetic energy, on an Al target. The thick straight line represents the incoming electrons. Upper panel, side view of the block, incident beam hits the block from below. Lower panel, bottom view (as seen by the incoming electrons).

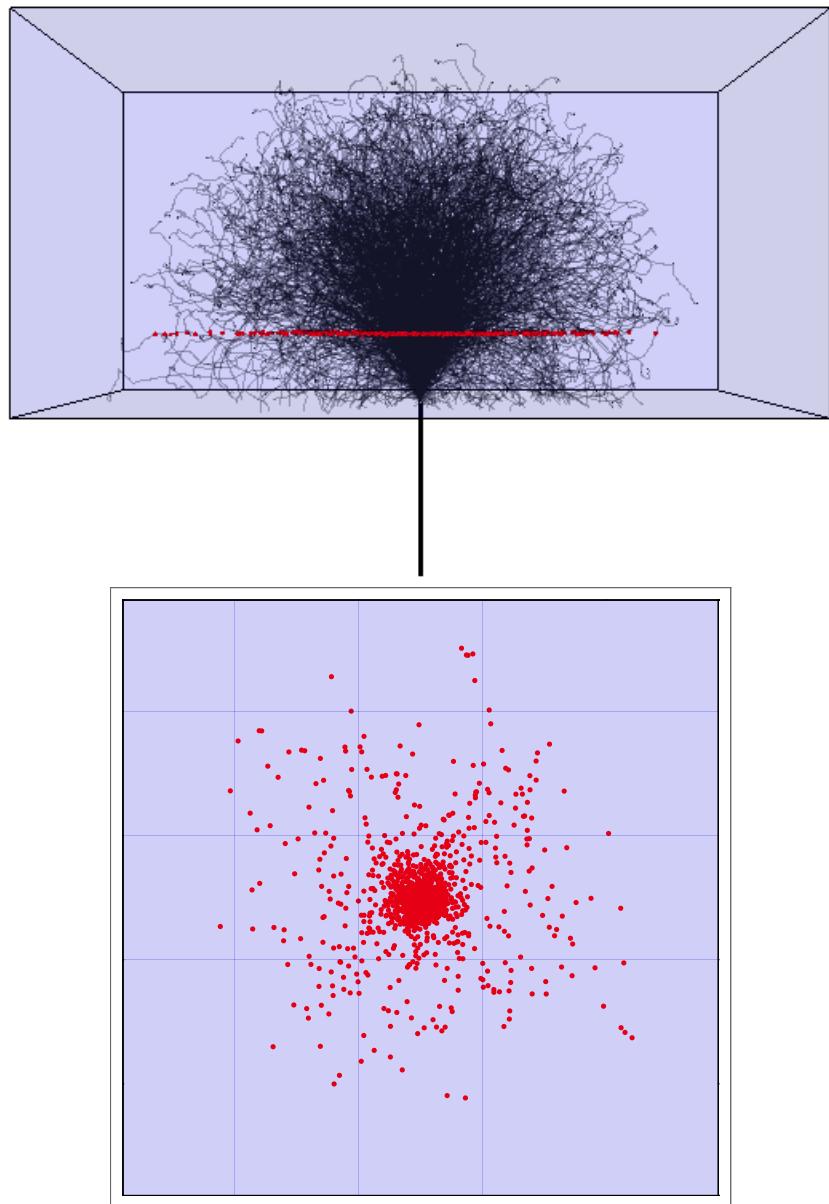


Figure 8.19: Tracks obtained in a simulation with 1000 incident electrons, with 50 keV initial kinetic energy, on the same Al target as in figure 8.18. The red dots correspond to the response of a very thin detector. Upper panel: 3D side view. Lower panel: 2D bottom view of the detector response.

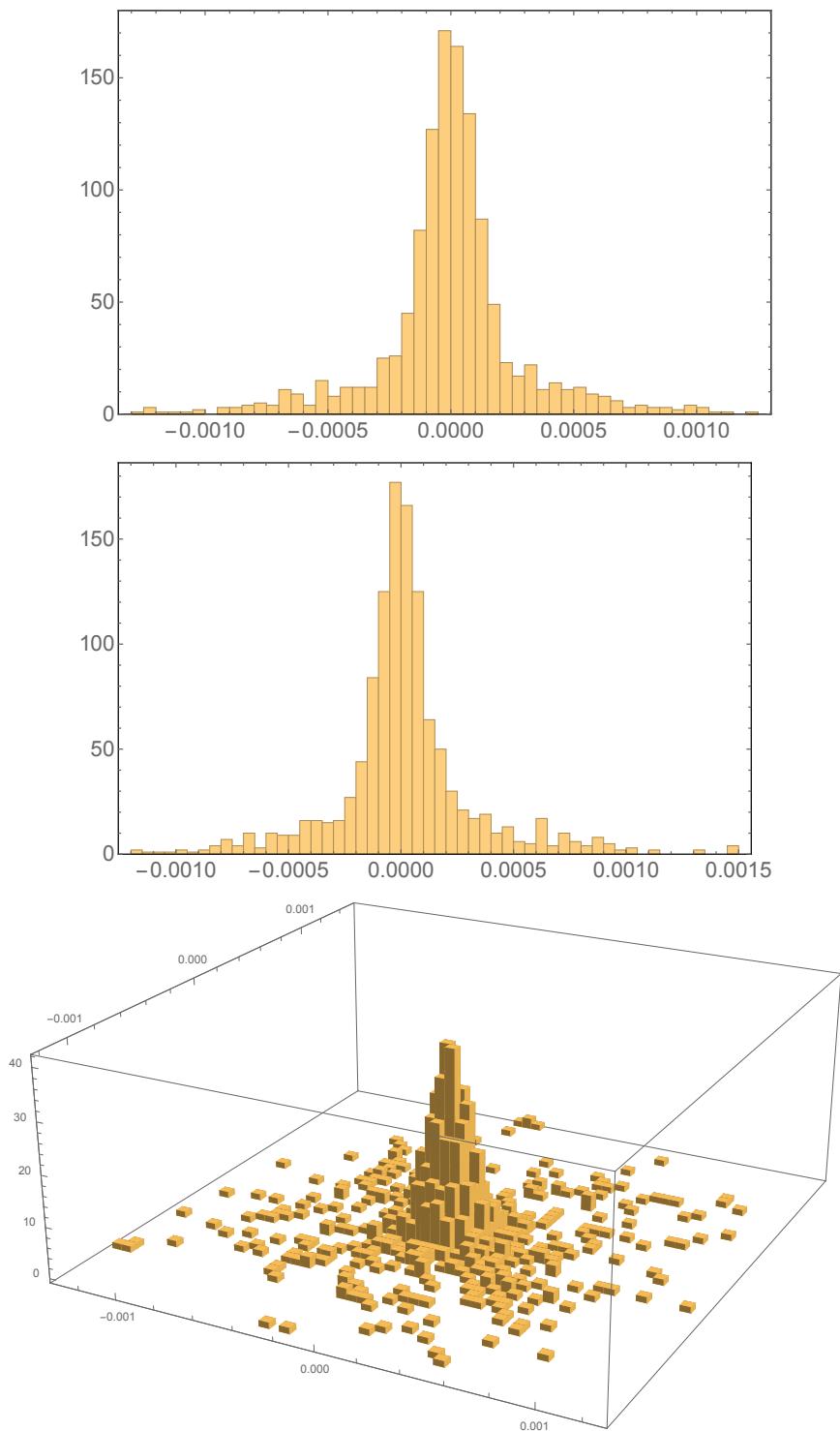


Figure 8.20: Distribution of track positions on the thin detector of figure 8.19. The upper left and upper right plots show the histograms of the positions projected on the  $x$ - and  $y$ -axis, respectively. The lower 3D plot is the histogram (*lego plot*) of the 2D distribution of track positions.

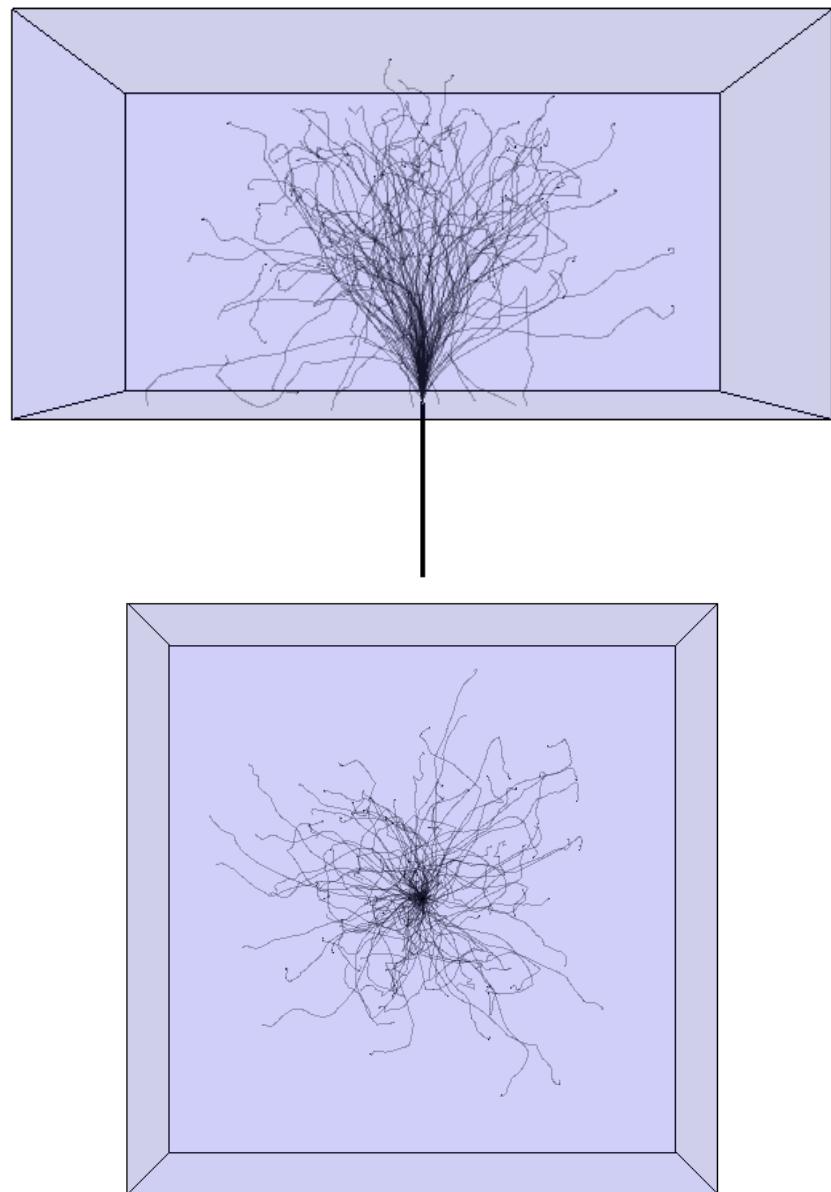


Figure 8.21: Tracks obtained in a simulation with 100 incident electrons, with 50 keV initial kinetic energy, on an C target. The thick straight line represents the incoming electrons. Upper panel, side view of the block, incident beam hits the block from below. Lower panel, bottom view (as seen by the incoming electrons). Compare this picture with figure 8.18, which shows an Al target.

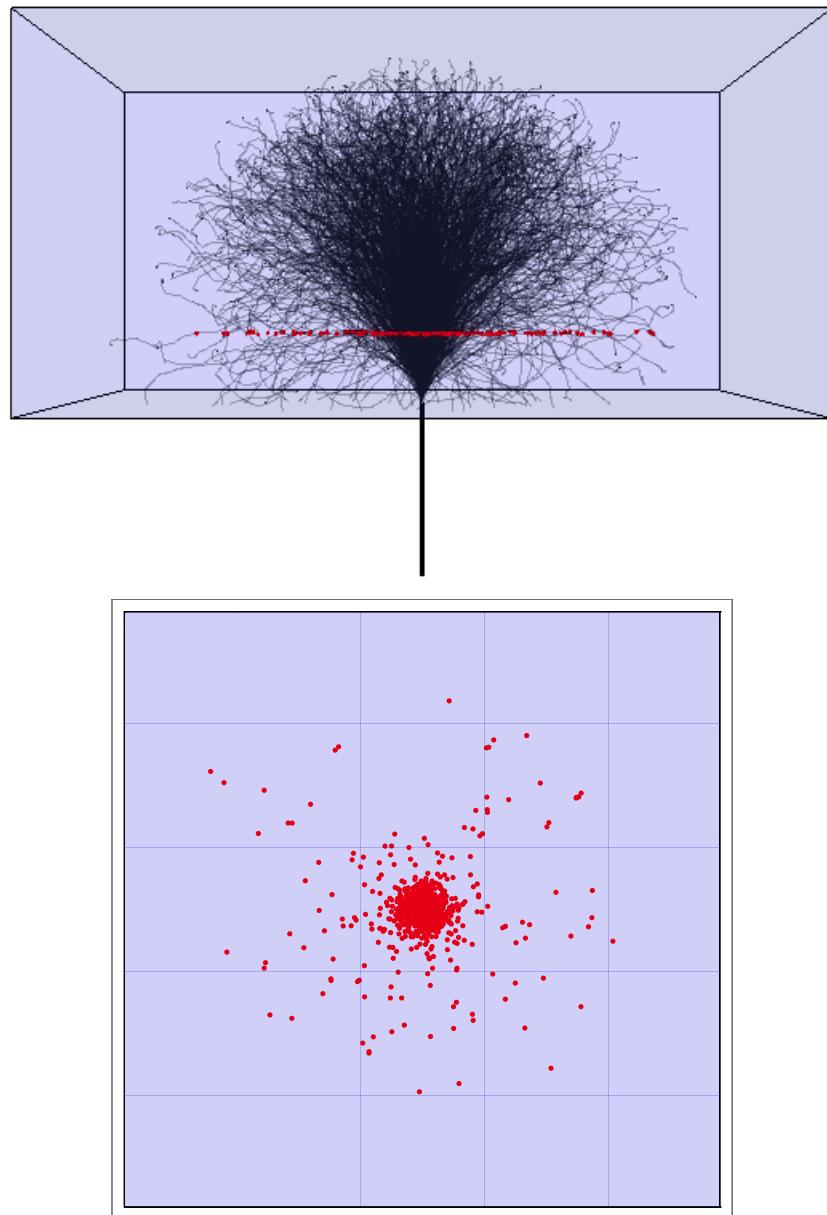


Figure 8.22: Tracks obtained in a simulation with 1000 incident electrons, with 50 keV initial kinetic energy, on the same Al target as in figure 8.18. The red dots correspond to the response of a very thin detector. Upper panel: 3D side view. Lower panel: 2D bottom view of the detector response. Compare this picture with figure 8.19, which shows an Al target.

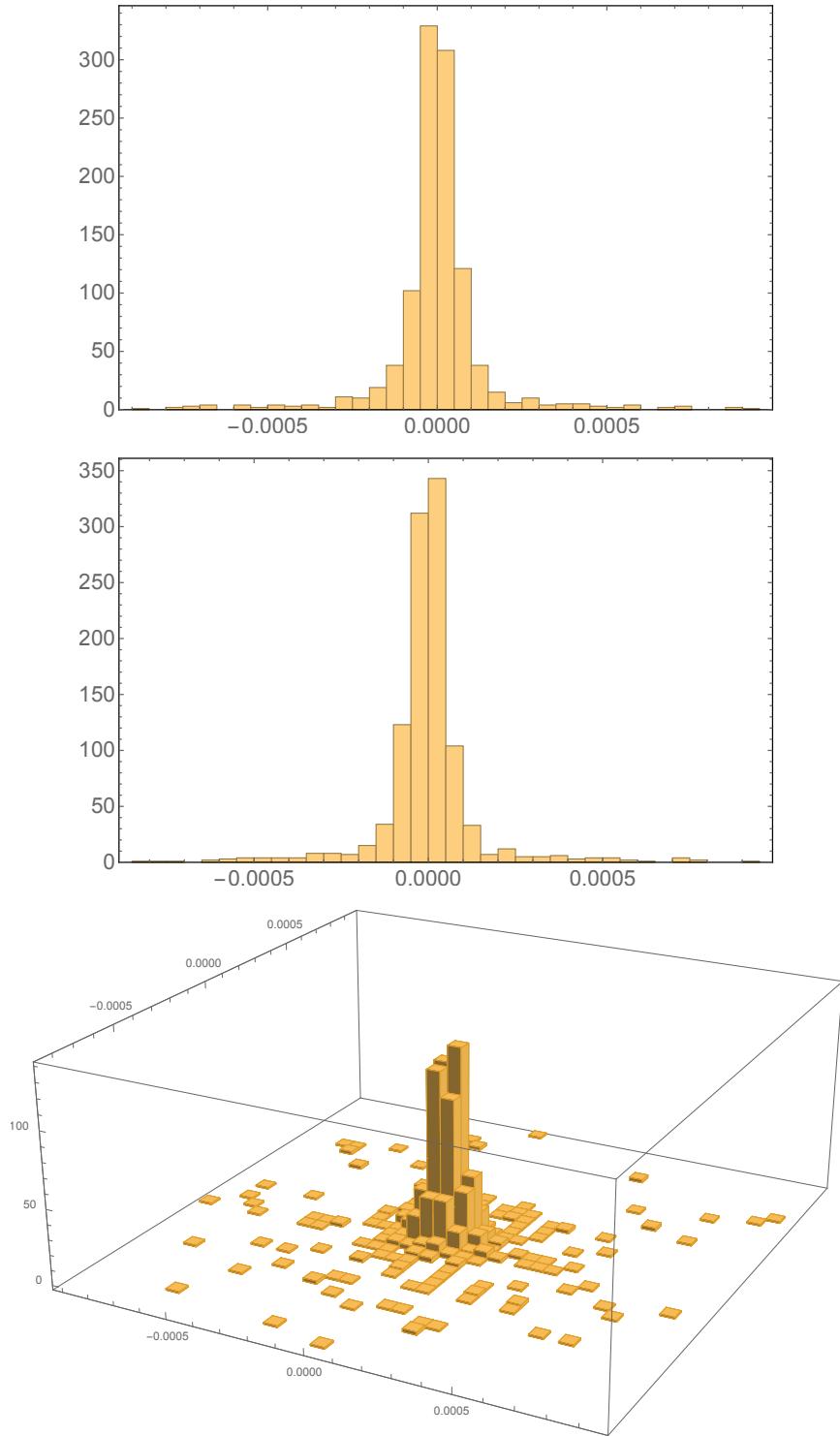


Figure 8.23: Distribution of track positions on the thin detector of figure 8.19. The upper left and upper right plots show the histograms of the positions projected on the  $x$ - and  $y$ -axis, respectively. The lower 3D plot is the histogram (*lego plot*) of the 2D distribution of track positions. Compare this picture with figure 8.20, which shows an Al target.

## 8.8 What's the use of all this?

Mathematically the Monte Carlo method (MC) is a stochastic technique to compute integrals, and thus averages: very often it is exceedingly hard to evaluate integrals with analytic or numerical methods, and in such cases we resort to MC to generate populations of simulated samples and we use them to compute statistically relevant quantities.

The example with the transport Monte Carlo for low-energy electrons, shows how much larger and capable Monte Carlo codes are actually built. EGS is one such simulator of electromagnetic showers: it is available for download in several versions at the web addresses

<http://irs.inms.nrc.ca/software/egsnrc/>

and

<http://rcwww.kek.jp/research/egs/>

Penelope is yet another code for the simulation of electron and photon transport and can be requested at the following web address

[http://www.oecd-nea.org/tools/abstract/detail/nea-1525.](http://www.oecd-nea.org/tools/abstract/detail/nea-1525)

Transport of high-energy hadrons as well as leptons and photons is well simulated by GEANT, available at the web address

[http://geant4.cern.ch/](http://geant4.cern.ch) as well as by FLUKA, available at the web address

<http://www.fluka.org/fluka.php>

In the medical field, the simulation of the behavior of low-energy particles, and several MC codes exist, like GEANT-DNA, PARTRAC, NOREC.

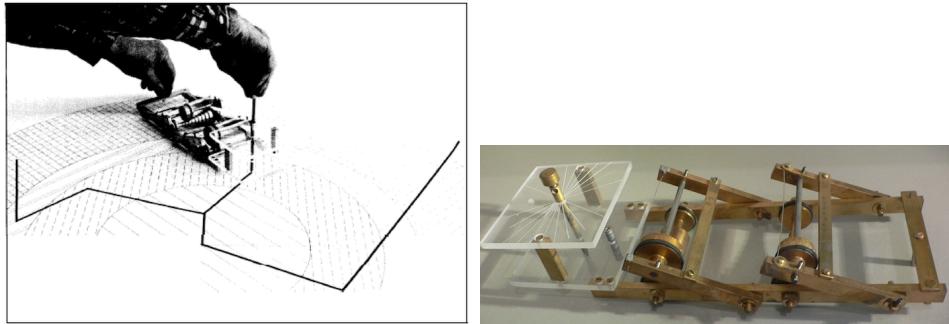


Figure 8.24: (Left panel) The Fermiac in action; (right panel) the Fermiac at the Boston Computer Museum.

## Appendix A: Early history of MC methods

The MC method is so widely used today that it makes little sense trying to describe the latest development in this handout. However it is interesting to mark a few early important developments.

- random sampling has been known in mathematics for a very long time: the Buffon's needle is an example of this kind of sampling. Indeed the great statistician Karl Pearson himself encouraged the effort to produce a table of random numbers<sup>7</sup>, and later used these numbers to obtain complex distributions.
- the nuclear research effort during the Second World War involved extremely complex problems, and Fermi hit on the idea of using a simulation device to follow the course of individual neutrons in a layered structure. The device that he designed and used was jokingly called "Fermiac"<sup>8</sup>, and it is shown in figure 8.24.
- the real turning point took place in 1946, when it occurred to Stanislaw Ulam that random sampling could be used in conjunction with the newly developed computer. Here is how Ulam recalls those early days<sup>9</sup>: "The first thoughts and attempts I made to practice [the Monte Carlo method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether

<sup>7</sup>L. H. C. Tippett, *Random Sampling Numbers*, Cambridge University Press (1927)

<sup>8</sup>N. Metropolis, *The Beginning of the Monte Carlo Method*, Los Alamos Science, Special Issue, 1987

<sup>9</sup>as quoted in R. Eckardt, *Stan Ulam, John von Neumann, and the Monte Carlo Method*, Los Alamos Science, Special Issue, 1987

a more practical method than “abstract thinking” might not be to lay it out say one hundred times and simply observe and count the number of successful plays. This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later... [in 1946, I] described the idea to John von Neumann and we began to plan actual calculations.”



Figure 8.25: A picture of Stan Ulam, Dick Feynman and John Von Neumann at Los Alamos.

- N. Metropolis (see figure 8.26) describes this new beginning very vividly in his 1987 paper:  
“During his wartime stint at the Laboratory, Stan [Ulam] had become aware of the electromechanical computers used for implosion studies, so he was duly impressed, along with many other scientists, by the speed and versatility of the ENIAC. In addition, however, Stan’s extensive mathematical background made him aware that statistical sampling techniques had fallen into desuetude because of the length and tediousness of the calculations. But with this miraculous development of the ENIAC – along with the applications Stan must have been pondering – it occurred to him that statistical techniques should be resuscitated, and he discussed this idea with von Neumann. Thus was triggered the spark that led to the Monte Carlo method. . . . John von

Neumann saw the relevance of Ulam's suggestion and, on March 11, 1947, sent a handwritten letter to Robert Richtmyer, the Theoretical Division leader [at Los Alamos]. His letter included a detailed outline of a possible statistical approach to solving the problem of neutron diffusion in fissionable material.

Johnny's interest in the method was contagious and inspiring. His seemingly relaxed attitude belied an intense interest and a well-disguised impatient drive. His talents were so obvious and his cooperative spirit so stimulating that he garnered the interest of many of us. It was at that time that I suggested an obvious name for the statistical method – a suggestion not unrelated to the fact that Stan had an uncle who would borrow money from relatives because he "just had to go to Monte Carlo." The name seems to have endured."

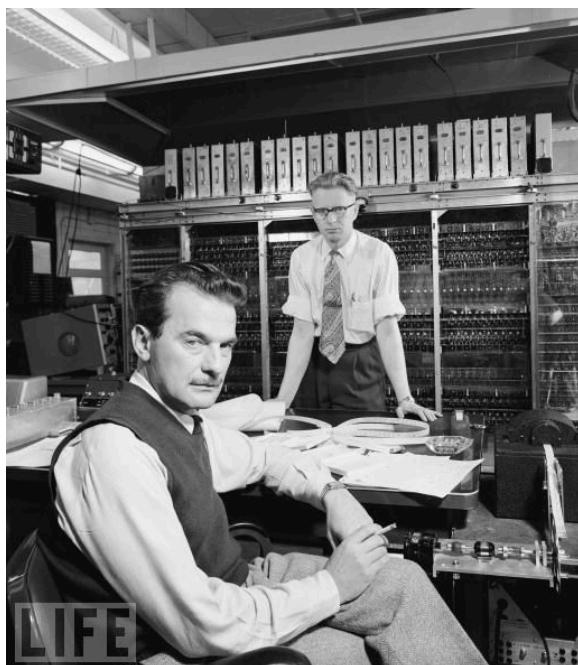


Figure 8.26: Portrait of American computer scientists Nicholas Metropolis (1915 - 1999) (seated) and James Henry Richardson (1918 - 1996) at Los Alamos National Laboratory, Los Alamos, New Mexico, November 1953 (from <http://www.life.com>).

- Again, in the same 1987 paper, Nick Metropolis gives credit to Enrico Fermi as follows:

"Enrico Fermi helped create modern physics. Here, we focus on his interest in neutron diffusion during those exciting times in Rome in the early thirties. According to Emilio Segrè, Fermi's student and col-

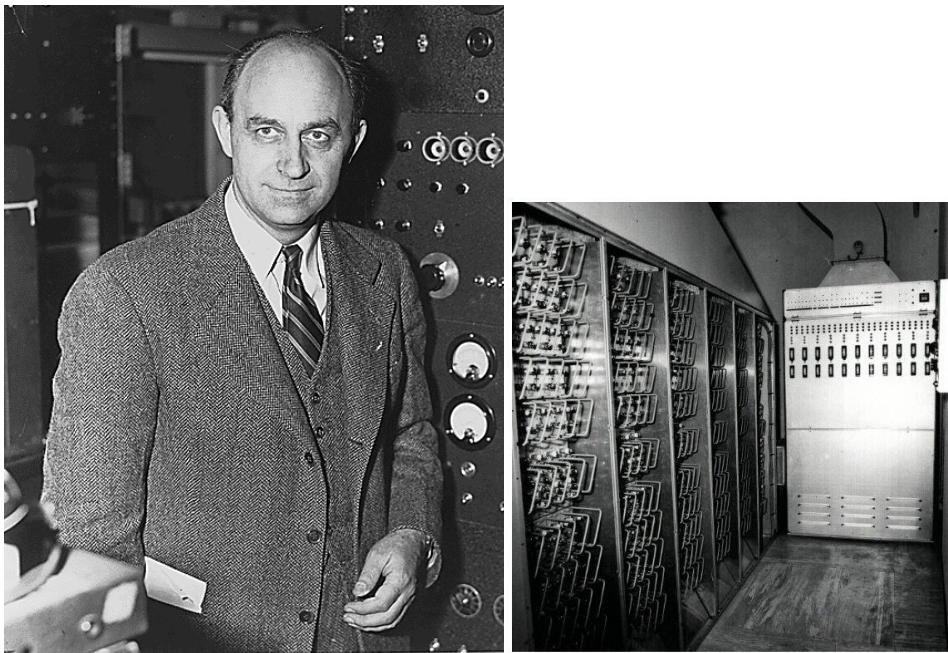


Figure 8.27: (Left panel) Portrait of Enrico Fermi. (Right panel) The first Italian computer “Calcolatrice Elettronica Pisana”, suggested – among others – by Enrico Fermi.

laborator, “Fermi had invented, but of course not named, the present Monte Carlo method when he was studying the moderation of neutrons in Rome. He did not publish anything on the subject, but he used the method to solve many problems with whatever calculating facilities he had, chiefly a small mechanical adding machine.” In a recent conversation with Segrè, I learned that Fermi took great delight in astonishing his Roman colleagues with his remarkably accurate, “too-good-to-believe” predictions of experimental results. After indulging himself, he revealed that his “guesses” were really derived from the statistical sampling techniques that he used to calculate with whenever insomnia struck in the wee morning hours! And so it was that nearly fifteen years earlier, Fermi had independently developed the Monte Carlo method.”

## Appendix B: listing of the Mathematica program for the MC estimation of lemniscate area

This is the Mathematica program for the lemniscate example. It is not the best or most efficient implementation, it has been written in this way for clarity. Students might try their hand at Mathematica and write a faster

code.

```

/* MC area estimate */

/* initialization */
n = 10000; /* total number of points in a single MC realization */
nloop = 1000; /* number of realizations */
area = Table[0, nloop]; /* list of area estimates */

/* loop over MC realizations */
For[l = 1, l <= nloop,
    x = y = Table[0, n]; /* clear the list of point coordinates */
    where = Table[0, n]; /* = 0 if point is out, = 1 if point is
in */

/* start a single MC realization */
For[k = 1, k <= n,
    x[[k]] = RandomReal[-1, 1]; /* random point in square */
    y[[k]] = RandomReal[-1, 1];
    theta = ArcTan[x[[k]], y[[k]]]; /* find polar coordinates
*/
    r = Sqrt[x[[k]]^2 + y[[k]]^2];
    /* test if inside and in this case set where=1 */
    If[Cos[2*theta] >= 0,
        If[r <= a*Sqrt[Cos[2*theta]], where[[k]] = 1]];
    k++];

area[[l]] = 4.*Apply[Plus, where]/n; /* estimate area */

l++];

/* histogram area estimates */
Histogram[area, Frame -> True, FrameStyle -> 18, ImageSize -> 8*72]

```

# Chapter 9

## Random sampling real data: the statistical bootstrap

### 9.1 Introduction

The concept of probability distribution is very powerful, and the associated analytical methods lead to a deep understanding of the problems at hand. However it is not always possible to tackle problems analytically, and we must resort to numerical methods, like the Monte Carlo simulations of the previous chapter. In the Monte Carlo method we use event populations that replace the analytical distributions and we utilize the descriptive sample statistics to produce estimators.

Replacing a distribution with a population of events thus helps carry out calculations that would otherwise be very difficult or impossible, and this concept has spread to other statistical methods, and it is so powerful that it must be fully grasped to appreciate some recent advances in statistics.

In this chapter we consider the case where actual data are drawn from an unknown distribution, and we must squeeze as much information as possible from them. In particular, is there any way to infer the probability distribution (or at least a few of its moments) from data? We already know that the conventional answer to this question involves some form of analytic approach, and, indeed, we shall introduce analytical methods further on in this course. However, there is a remarkable numerical method, that relies on random sampling, and uses the brute force of the computer for statistical inference: it is called “the bootstrap”, and it was invented by Bradley Efron in 1977 (B. Efron, Ann. of Statistics **7** (1979) 1).

This chapter is partly based on a paper by H. Varian, The Mathematica Journal **9** (2005) 768.

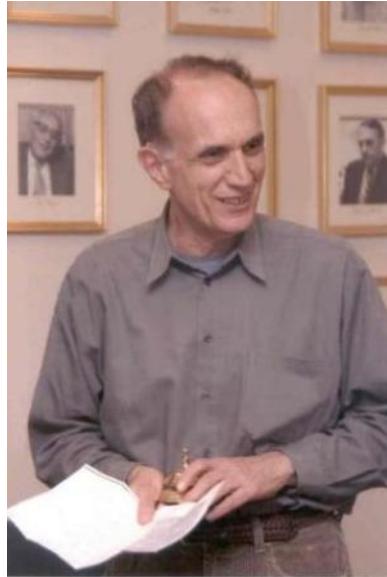


Figure 9.1: Bradley Efron (1938-), inventor of the statistical bootstrap. (Photograph from B. Efron's homepage <http://www-stat.stanford.edu/~ckirby/brad/>)

## 9.2 Statistical bootstrap

The basic idea behind the statistical bootstrap is that the histogram of the collected data is itself an estimate of the distribution of data, and this means that, given  $n$  data points, we assign a probability mass  $1/n$  to each data point. This empirical distribution can be used to estimate all averages and all statistical moments. Unfortunately calculations based on such an approach are almost invariably analytically unfeasible. However we can combine this idea with Monte Carlo, to generate values distributed according to the empirical distribution and calculate numerically. Since the data points themselves define the probability distribution and have equal probability mass, we only need to extract random sets from the whole data set.

Consider the following simple example, where data are drawn from the following pdf:

$$p(x) = 0.3N(0.5, 0.2) + 0.7N(-0.5, 0.4)$$

(recall that  $N(\mu, \sigma)$  denotes a Gaussian pdf with mean  $\mu$  and standard deviation  $\sigma$ ). This distribution is shown in figure 9.2.

The *exact* mean value of the distribution is

$$\int_{-\infty}^{+\infty} xp(x)dx = 0.3 \times 0.5 - 0.7 \times 0.5 = -0.2$$

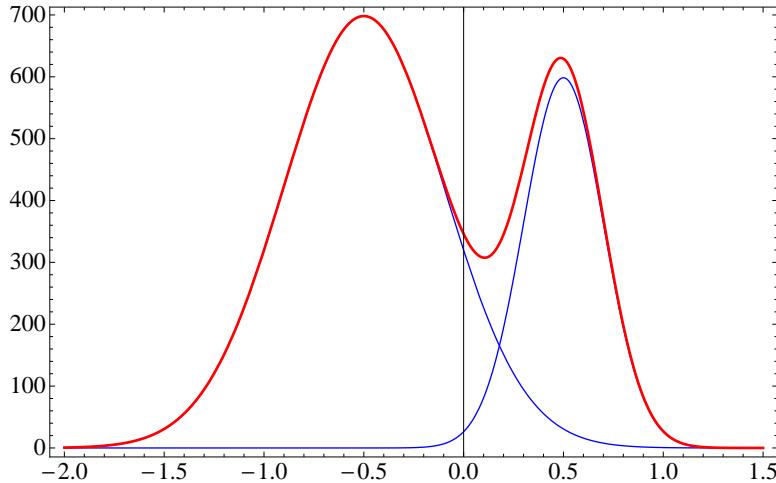


Figure 9.2: Plot of the pdf  $0.3 N(0.5, 0.2) + 0.7 N(-0.5, 0.4)$  (red curve). This pdf is an example of a *mixture model*, i.e., a pdf  $\sum_k a_k p_k(x)$  where the  $p_k(x)$ 's are simple pdf's (often Gaussian pdf's) and  $\sum_k a_k = 1$ . The individual Gaussian components are drawn in blue.

(exercise: find the exact standard deviation of this distribution).

We can evaluate the same mean value numerically, carrying out a standard Monte Carlo simulation. This is shown in figure 9.3

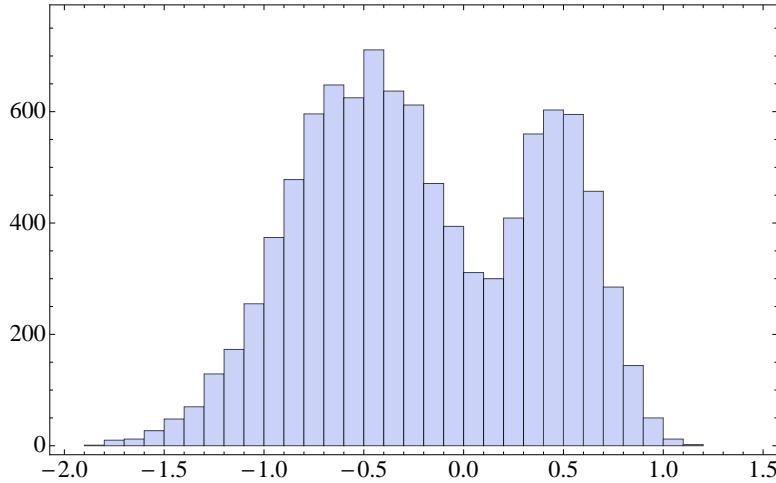


Figure 9.3: Histogram of 10000 random values extracted from the pdf shown in figure 9.2.

Now we can use the sample mean and the sample variance as estimates

of the mean and variance of the pdf  $p(x)$ . Figure 9.4 shows the position of the sample mean.

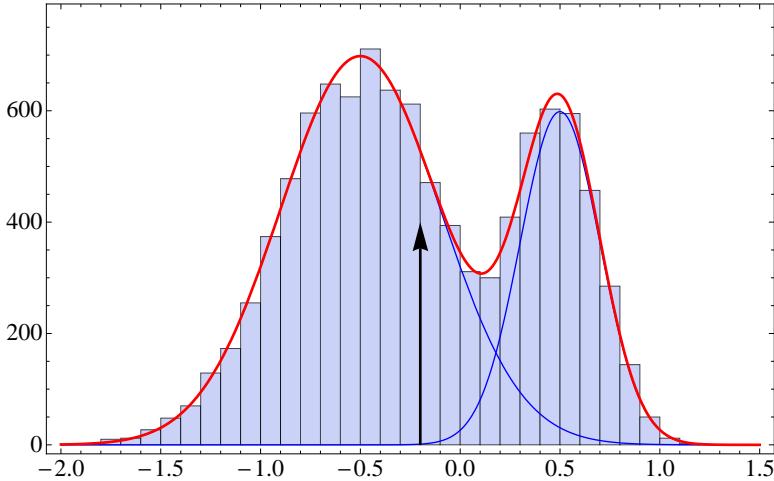


Figure 9.4: The true pdf superposed on the histogram of figure 9.3. The arrow marks the position of the mean value. The exact mean value is -0.2, the sample mean is -0.199, and the two values are indistinguishable from each other in this plot. The sample standard deviation is 0.579 (very close to the exact value 0.577927).

In this example we have drawn 10000 values of the random variable  $x$ , but in a real experiment we may be limited to a much smaller number: in this case the sample mean fluctuates considerably, and we estimate the amplitude of this fluctuation using the sample variance.

The Monte Carlo procedure lets us estimate the amplitude of the fluctuations of the sample variance as well. Figure 9.5 shows the distributions of mean and variance in a numerical experiment where sample mean and sample variance of a 100-sample dataset have been computed for 5000 different datasets.

Here we note that the sample standard deviation of the mean in this MC simulation is quite close to the expected result  $0.577927/\sqrt{100} = 0.0577927$ . But in addition to the sample standard deviation this procedure also gives a sample standard deviation of the sample standard deviation itself, obviating to the difficulty of finding a closed formula for this estimator (in this particular instance the sample standard deviation has a 5% relative error).

Unfortunately, quite often in actual practice we cannot rely on such massive amounts of data, and we must make do with much smaller datasets. We might actually have only one of the 100 sample datasets that we used to produce figure 9.5, and find an experimental histogram like that shown

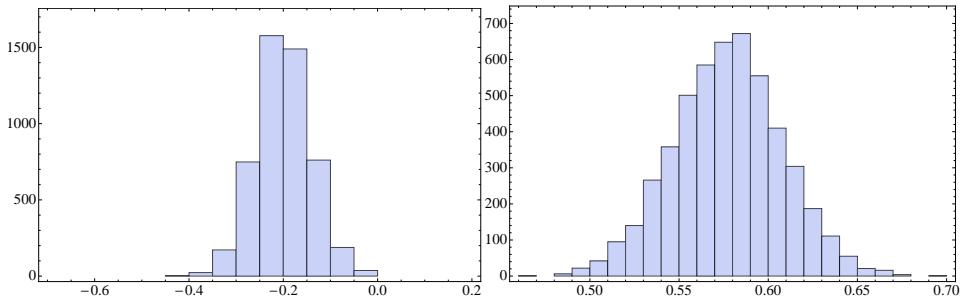


Figure 9.5: Distribution of sample mean (left panel) and sample standard deviation (right panel) for 5000 datasets of 100 independent samples each. The sample mean and sample standard deviation of the first histogram are -0.199809, and 0.0577796, respectively. Similarly, the sample mean and sample standard deviation of the second histogram are 0.577147 and 0.0304807, respectively.

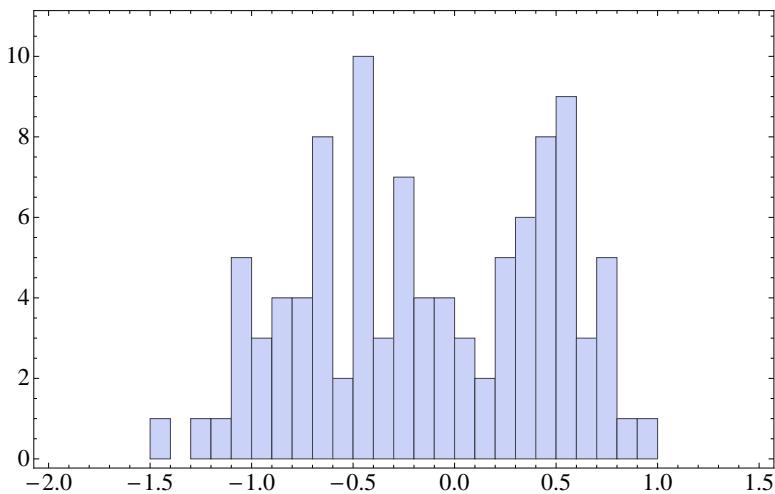


Figure 9.6: Histogram of 100 values of  $x$ . This is how a real experimental distribution might look, when only few data are available.

in figure 9.6

This is where the statistical bootstrap comes into play: we use this distribution as if it were the actual distribution, and we resample from it, drawing each time 100 samples *with replacement*. For instance, drawing with replacement from a small dataset of 10 experimental values like  $\{x_0, x_1, \dots, x_9\}$ , we might find the resampled dataset  $\{x_2, x_7, x_1, x_5, x_2, x_3, x_9, x_5, x_4, x_2\}$ . Just as before we can use the sample means and standard deviations from

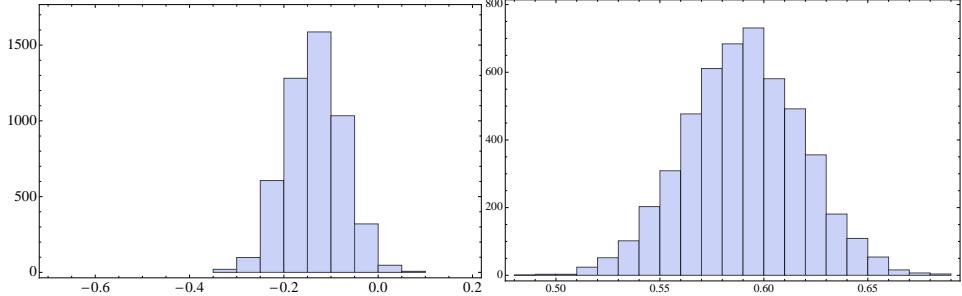


Figure 9.7: Distribution of sample mean (left panel) and sample standard deviation (right panel) for 5000 datasets of 100 samples each, resampled from the empirical distribution shown in figure 9.6. The sample mean and sample standard deviation of the first histogram are -0.134963, and 0.059739, respectively. Similarly, the sample mean and sample standard deviation of the second histogram are 0.590239 and 0.0279584, respectively. Notice that these values are very close to those found with the previous Monte Carlo calculation.

each (resampled) dataset to fill histograms like those of figure 9.5. Figure 9.7 shows the results of a numerical experiment performed with the data in figure 9.6. These result are quite remarkable: they are in excellent agreement with those of the much larger dataset considered above, and they suggest that this resampling procedure is reliable even with scarce data.

### 9.3 Variance of bootstrap estimates

The bootstrap technique has the obvious drawback of modifying the variance of the estimates, because of the correlations intrinsic to the method. Consider for instance the sample mean of a set of  $N$  samples  $\{x_i\}_{i=1,N}$ :

$$x_S = \frac{1}{N} \sum_{i=1}^N x_i$$

which is replaced by

$$x_S^{(R)} = \frac{1}{N} \sum_{i=1}^N n_i x_i$$

where  $R$  indicates a bootstrap resampling, and  $n_i$  is the number of times a given sample appears in the resampled sequence of data, with the condition that  $\sum_i n_i = N$ . Clearly, the probability of a given sequence of  $n_i$ 's is given by a multinomial probability

$$p(n_1, \dots, n_N) = \frac{N!}{\prod_{j=1}^N n_j!} \frac{1}{N^N} \quad (9.1)$$

so that the mean value of  $x_S^{(R)}$  is

$$\begin{aligned}\langle x_S^{(R)} \rangle &= \sum_{\sum_i n_i = N} p(n_1, \dots, n_N) x_S^{(R)} \\ &= \sum_{\sum_k n_k = N} \left( \frac{N!}{\prod_{j=1}^N n_j!} \frac{1}{N^N} \right) \frac{1}{N} \sum_{i=1}^N n_i x_i \\ &= \frac{1}{N^{N+1}} \sum_{i=1}^N \sum_{\sum_k n_k = N} \left( \frac{N!}{\prod_{j=1}^N n_j!} \right) n_i x_i \quad (9.2)\end{aligned}$$

Now let

$$n_i x_i = \left. \frac{d}{dy} e^{n_i x_i y} \right|_{y=0}$$

then we can rewrite equation (9.2) as follows

$$\langle x_S^{(R)} \rangle = \left. \frac{d}{dy} \left[ \frac{1}{N^{N+1}} \sum_{i=1}^N \sum_{\sum_k n_k = N} \left( \frac{N!}{\prod_{j=1}^N n_j!} \right) (e^{x_i y})^{n_i} \right] \right|_{y=0} \quad (9.3a)$$

$$= \left. \frac{d}{dy} \left[ \frac{1}{N^{N+1}} \sum_{i=1}^N (1 + \dots + e^{x_i y} + \dots + 1)^N \right] \right|_{y=0} \quad (9.3b)$$

$$= \frac{1}{N} \sum_{i=1}^N x_i \quad (9.3c)$$

and we find that the mean of the resampled means coincides with the usual sample mean. Similarly, we find for double products

$$\sum_{\sum_k n_k = N} p(n_1, \dots, n_N) n_i x_i n_j x_j = \sum_{\sum_k n_k = N} \left( \frac{N!}{\prod_{\ell=1}^N n_\ell!} \frac{1}{N^N} \right) n_i x_i n_j x_j \quad (9.4a)$$

$$= \left. \frac{\partial^2}{\partial y \partial z} \frac{1}{N^N} \sum_{\sum_k n_k = N} \left( \frac{N!}{\prod_{\ell=1}^N n_\ell!} \right) (e^{x_i y})^{n_i} (e^{x_j z})^{n_j} \right|_{y=0, z=0} \quad (9.4b)$$

$$= \left. \frac{\partial^2}{\partial y \partial z} \frac{1}{N^N} (1 + \dots + e^{x_i y} + \dots + e^{x_j z} + \dots + 1)^N \right|_{y=0, z=0} \quad (9.4c)$$

$$= \frac{N-1}{N} x_i x_j \quad (9.4d)$$

and for triple products

$$\sum_{\sum_\ell n_\ell = N} p(n_1, \dots, n_N) n_i x_i n_j x_j n_k x_k = \frac{(N-1)(N-2)}{N^2} x_i x_j x_k \quad (9.5a)$$

Using these equalities we find the mean of the variance estimate obtained with the bootstrap method

$$\sigma_R^2 = \sum_{\sum_\ell n_\ell = N} p(n_1, \dots, n_N) \frac{1}{N-1} \sum_{i=1}^N n_i \left( x_i - \bar{x}_S^{(R)} \right)^2 \quad (9.6a)$$

$$= \frac{1}{N-1} \sum_{\sum_\ell n_\ell = N} p(n_1, \dots, n_N) \sum_{i=1}^N n_i \left( x_i^2 - \frac{2}{N} \sum_j n_j x_i x_j + \frac{1}{N^2} \sum_{j,k} n_j n_k x_i x_j x_k \right) \quad (9.6b)$$

$$= \frac{1}{N-1} \sum_{i=1}^N \left( x_i^2 - \frac{2}{N} \sum_j \frac{N-1}{N} x_i x_j + \frac{1}{N^2} \sum_{j,k} \frac{(N-1)(N-2)}{N^2} x_i x_j x_k \right) \quad (9.6c)$$

$$= \frac{1}{N-1} \sum_{i=1}^N \left( x_i^2 - \frac{2(N-1)}{N^2} \sum_j x_i x_j + \frac{(N-1)(N-2)}{N^4} \sum_{j,k} x_i x_j x_k \right) \quad (9.6d)$$

which differs from the usual formula of sample variance, as it includes correlations in the bootstrapped data sets.

## 9.4 Example with real data (from Efron and Gong)

The example in the previous section illustrates the power of the resampling method, however it may seem a little artificial. So here is an example with real data (taken from B. Efron and G. Gong, *The American Statistician* **37** (1983) 36). The data are 15 pairs of values {LSAT, GPA}, where

- $\text{LSAT}_i$  = average LSAT score for law school  $i$  in 1973 (LSAT = Law School Admission Test);
- $\text{GPA}_i$  = average LSAT score of entering students at law school  $i$  in 1973 (GPA = Grade Point Average).

Figure 9.8 shows the plot of law school data in the paper by Efron and Gong: it is clear that there is a marked correlation between the LSAT and the GPA scores, and it is natural to try to find an estimator for the correlation coefficient. This can be done with the bootstrap method, and the result is shown in figure 9.9 (also from Efron and Gong), together with the analytical result from a Gaussian-based model: the bootstrap estimate is not very far from the analytical model, however it is much easier to state and to apply (exercises: a) search the literature to find how to build the Gaussian-based model and try to apply it to this case; b) use the data in figure 9.8 and reproduce the histogram in figure 9.9).

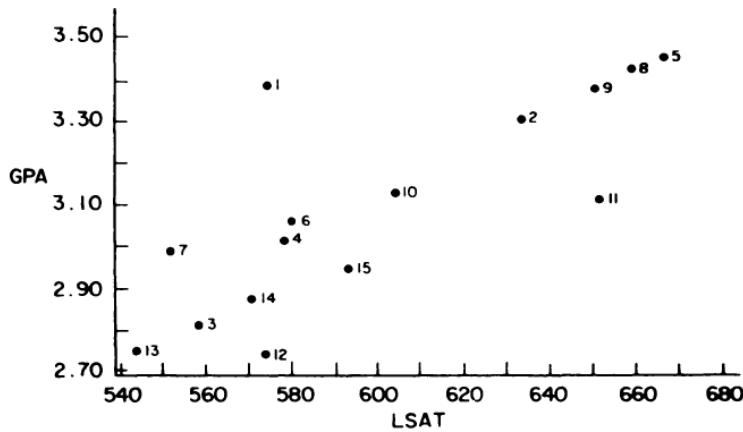


Figure 1. The law school data (Efron 1979B). The data points, beginning with School #1, are (576, 3.39), (635, 3.30), (558, 2.81), (578, 3.03), (666, 3.44), (580, 3.07), (555, 3.00), (661, 3.43), (651, 3.36), (605, 3.13), (653, 3.12), (575, 2.74), (545, 2.76), (572, 2.88), (594, 2.96).

Figure 9.8: Distribution of law school data, from Efron and Gong (see text).

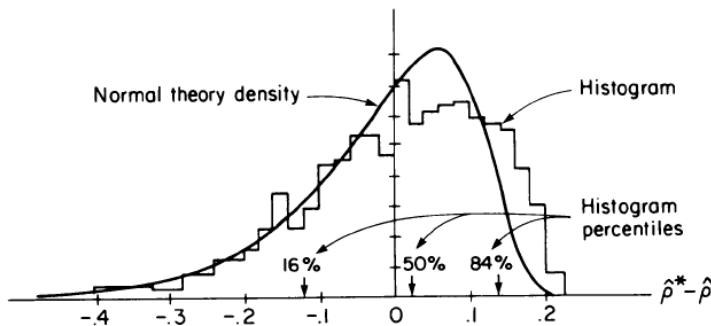


Figure 2. Histogram of  $B = 1000$  bootstrap replications  $\hat{\rho}^*$  for the law school data. The normal theory density curve has a similar shape, but falls off more quickly at the upper tail.

Figure 9.9: Histogram of correlation coefficient from bootstrap resampling the data shown in figure 9.8, again from Efron and Gong (see text).

Exercises:

- search the literature to find how to build the Gaussian-based model and try to apply it to this case
- use the data in figure 9.8 and reproduce the histogram in figure 9.9).

- recall that the estimator of the correlation coefficient is

$$\rho_S = \frac{\text{cov}_S(x, y)}{\sigma_{x,S}\sigma_{y,S}} = \frac{\sum_{k=1}^n (x_k - \mu_{x,S})(y_k - \mu_{y,S})}{\sqrt{\sum_{i=1}^n (x_i - \mu_{x,S})^2 \sum_{j=1}^n (y_j - \mu_{y,S})^2}}$$

and use this estimator to find the correlation coefficient and its standard deviation in the example of the LSAT and GPA scores shown in figure 9.8.

## 9.5 Robust statistics

One important strength of the statistical bootstrap is that it can be used to find the standard deviation of relevant statistical estimators that are difficult to treat by other methods.

Consider the Cauchy distribution

$$p(x) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2} \quad (9.7)$$

we can use the transformation method to generate numbers distributed according to this distribution. The cumulative distribution is<sup>1</sup>

$$F(y) = \int_{-\infty}^y \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (y' - y_0)^2} dy' = \frac{1}{\pi} \left( \arctan \frac{y - y_0}{\Gamma/2} + \frac{\pi}{2} \right) \quad (9.8)$$

and thus we obtain Cauchy-distributed numbers from

$$y = F^{-1}(x) = y_0 + \frac{\Gamma}{2} \tan \left[ \pi \left( x - \frac{1}{2} \right) \right] \quad (9.9)$$

where  $x$  is a uniform random variable in the interval (0,1). Figure 9.10 shows the histogram obtained from a simulation of Cauchy distributed values.

When we introduced the Cauchy distribution, we found that the mean value is undefined. Even the sample mean has a somewhat erratic behavior, because of the outliers: this is shown in figure 9.11 (running sample mean<sup>2</sup>

---

<sup>1</sup>Here we use the indefinite integral

$$\int \frac{1}{a^2 + b^2 x^2} dx = \frac{1}{ab} \arctan \frac{bx}{a}$$

so that

$$\int_{-\infty}^{+\infty} \frac{1}{a^2 + b^2 x^2} dx = \frac{\pi}{ab}$$

<sup>2</sup>the running sample mean  $\mu_S(n)$  of a dataset  $\{x_k\}_{k=1,\dots,N}$  is a function of sample size  $n$ :

$$\mu_S(n) = \frac{1}{n} \sum_{k=1}^n x_k$$

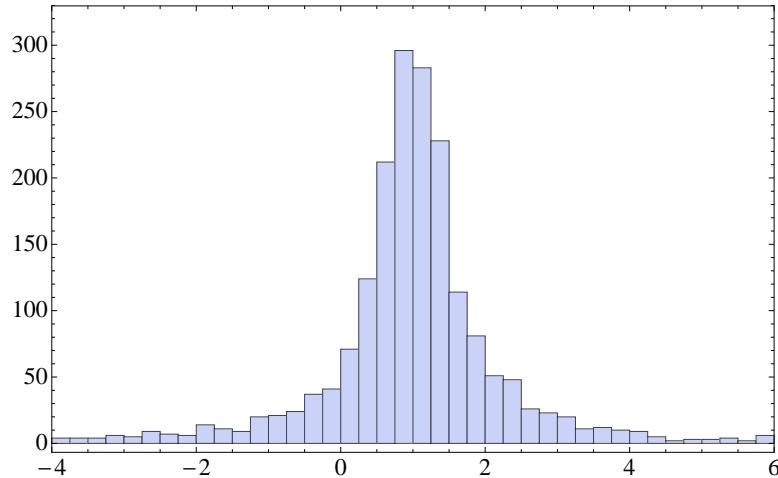


Figure 9.10: Histogram obtained drawing 2000 Cauchy-distributed random numbers; the distribution has the peak value at 1 and  $\Gamma = 1$ . The histogram covers a very small range and includes 1866 entries, many outliers are left out.

vs. sample size). On the other hand the median is much less sensitive to outliers, and quickly converges to the correct value 1: this is shown in figure 9.12 (running sample mean vs. sample size). We say that the median is a *robust* statistic.

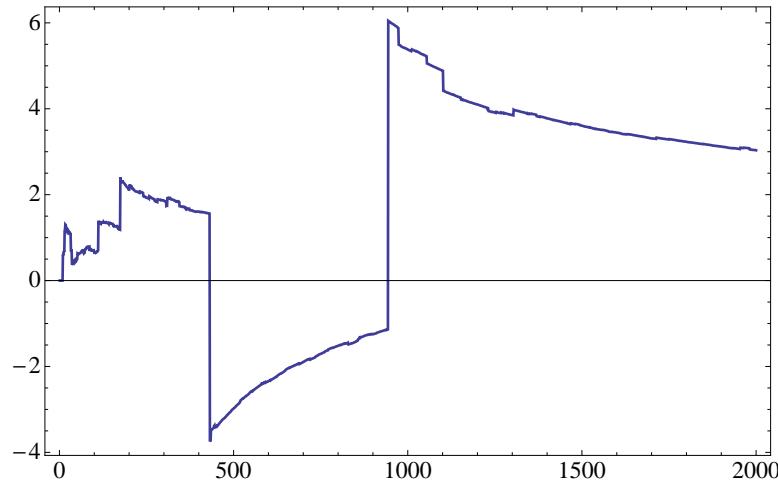


Figure 9.11: Running sample mean vs. sample size for the Cauchy distributed random numbers of figure 9.10. As soon as a large outlier is included, the sample mean fluctuates wildly. In any case the sample mean misses the median value 1.

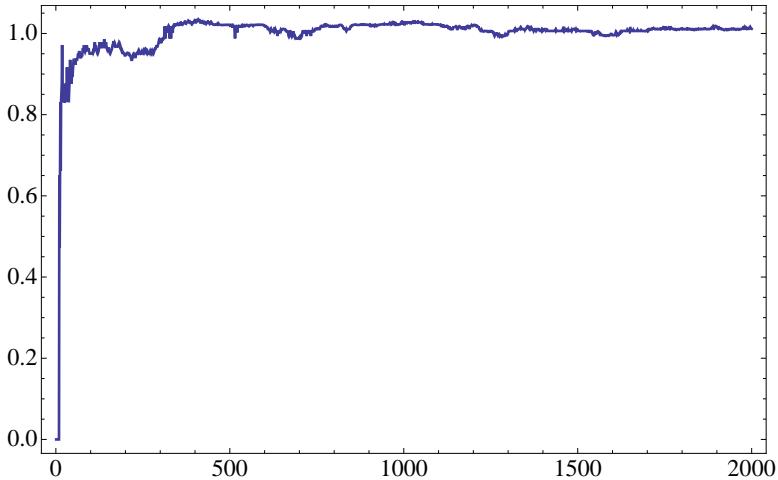


Figure 9.12: Running sample median vs. sample size for the Cauchy distributed random numbers of figure 9.10. The median is much less sensitive to outliers.

While it is rather easy to estimate analytically the fluctuations of the sample mean, it is quite hard to estimate the fluctuations of the sample median. However the problem is easily solved with the statistical bootstrap. Figure 9.13 shows a histogram obtained drawing just 100 numbers from a the same Cauchy distribution of figure 9.10, and figure 9.14 shows the distribution of the sample median. From histogram 9.14 we can calculate the sample mean of the median (in this example 0.998), and its standard deviation (in this example 0.075).

## 9.6 What's the use of all this?

The statistical bootstrap is an important numerical method with many applications. It is an example of a *nonparametric* method, and does not rely on underlying models. It has become an important complement to the traditional analytical methods that we shall study further on in the course.

The bootstrap is also a method with a strong frequentist flavor (it improves as the original sample size increases), which also appeals to Bayesians.

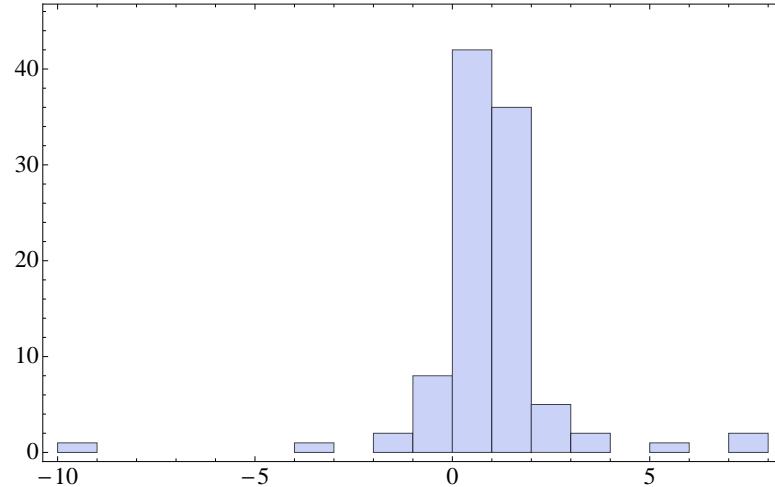


Figure 9.13: Histogram of 100 samples from the same Cauchy distribution of figure 9.10, used for a bootstrap estimate of the sample median.

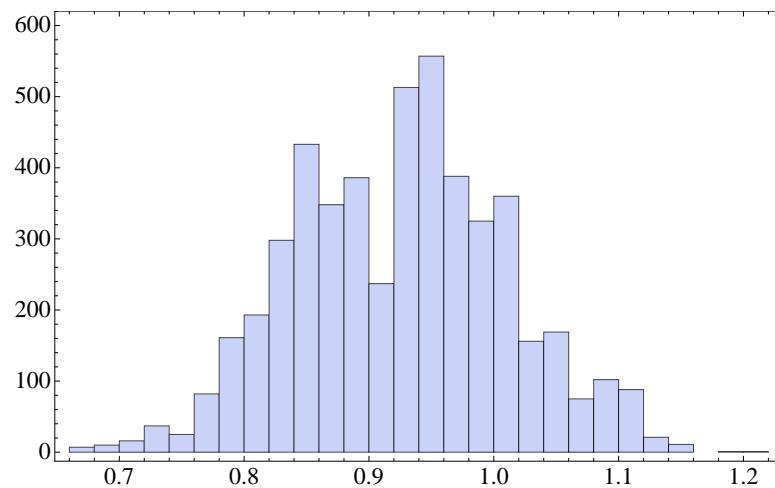


Figure 9.14: Distribution of the sample median obtained bootstrapping the distribution of figure 9.13. In this case the sample mean of the median is 0.998 and the sample standard deviation of the median is 0.075.

## Appendix: listing of the Mathematica program for the Cauchy distribution

This section lists the Mathematica program for the Cauchy example. The program is divided in two parts: the first produces a large dataset of Cauchy-distributed numbers, and plots the running sample mean and the running sample median. The second part produces a small dataset and implements the statistical bootstrap to find the distribution of the median.

```
\* Cauchy example *\

\* set the random seed *\n
SeedRandom[1];

nran = 2000; \* set the dataset size *\n
y0 = 1; \* median value of the Cauchy distribution *\n
gam = 1; \* width  $\Gamma$  of the Cauchy distribution *\n

halfrange = 5.; \* set the histogram range *\n
nbins = 20; \* set the number of bins *\n
dy = halfrange/nbins; \* find the bin size *\n

\* generate the Cauchy distributed random numbers *\n
y = Table[y0 + 0.5*gam*Tan[Pi*(RandomReal[] - 0.5)], {nran}];

\* fill and plot the histogram *\n
Histogram[y, {-halfrange + y0, halfrange + y0, dy},
  PlotRange -> {{-halfrange + y0, halfrange + y0}, All}, Frame
-> True,
  FrameStyle -> 18, ImageSize -> 8*72]

\* select the numbers that fall in the histogram range *\n
sel = Select[y, (# > -halfrange + y0 && # < halfrange + y0) &];
Print["numbers in histogram range: ", Length[sel]];

\* running sample mean and sample median *\n
mv = med = Table[0, {nran}];
For[k = 10, k <= nran,
  mv[[k]] = Mean[Take[y, {1, k}]];
  med[[k]] = Median[Take[y, {1, k}]];
  k++];
\* plot sample mean *\n
ListPlot[mv, Joined -> True, PlotStyle -> AbsoluteThickness[2],
```

```
Frame -> True,
  FrameStyle -> 18, ImageSize -> 8*72];
(* plot sample median *)
Frame -> True, ListPlot[med, Joined -> True, PlotStyle -> AbsoluteThickness[2],
  Frame -> True, FrameStyle -> 18, ImageSize -> 8*72, PlotRange
-> All];

(* statistical bootstrap *)

(* this function defines resampling *)
Clear[Resample]; Resample[list_] := list[[Table[RandomInteger[{1,
Length[list]}], {Length[list]}]]];

nr = 100; (* sample size *)
sample = Table[y0 + 0.5*gam*Tan[Pi*(RandomReal[] - 0.5)], {nr}];
(* data *)

(* fill and plot histogram to show dataset *)
Histogram[sample, Frame -> True, ImageSize -> 8*72, FrameStyle ->
18];

(* resample and find median for each resampled dataset *)
medb = Table[0, {nrep}];
For[n = 1, n <= nrep,
  sam = Resample[sample];
  medb[[n]] = Median[sam];
  n++];
(* fill and plot histogram of median distribution *)
Histogram[medb, Frame -> True, ImageSize -> 8*72, FrameStyle ->
18];

(* compute mean and sd of median distribution *)
Mean[med];
StandardDeviation[med];
```



# Chapter 10

## Parametric statistics: maximum likelihood and confidence intervals

In this chapter we develop some core concepts of statistics, which greatly extend the simple descriptive approach of chapter 7. Very often we are not content with a simple description of data, as we wish to estimate the parameters of some theoretical model. The estimate of the “best values” of the parameters is an example of *point estimate* (we estimate a single value or one vector of values). We are also concerned with the uncertainty of our estimate, and we want to assign a probability to a range that contains the parameter estimate: this is an example of an *interval estimate*. In this chapter we study the *Maximum Likelihood* (ML) method, which is a powerful approach to the systematic construction of estimators and to parameter estimation. It was developed and popularized by R. A. Fisher between 1912 and 1922 [1]. We begin our study with the Bayesian approach to parameter estimation and then we move on to the ML method.

### 10.1 Bayes’ Theorem and the Principle of Maximum Likelihood

Bayes’ theorem in the case of a discrete set of mutually exclusive  $H_n$ ’s is

$$P(H_n|B, I) = \frac{P(B|H_n, I)P(H_n, I)}{\sum_k P(B|H_k, I)P(H_k, I)} \quad (10.1)$$

where  $I$  is some *a priori* (or *prior*) information that may be available from the start. Now let the  $H_n$ ’s be competing, mutually exclusive hypotheses, and let  $B = D$  be a set of data, then we can rewrite the theorem in the form

$$P(H_n|D, I) = \frac{P(D|H_n, I)}{\sum_k P(D|H_k, I)P(H_k, I)} P(H_n, I) \quad (10.2)$$

which shows that the probability of hypothesis  $H_n$  given the data  $D$  is the old probability times a factor that includes the data. This is an important expression:  $P(H_n, I)$  on the r.h.s. is the *a priori* (or *prior*) probability of hypothesis  $H_n$ ,  $L = P(D|H_n, I)$  is the *likelihood*, the normalization factor in the denominator  $\sum_k P(D|H_k, I)P(H_k, I)$  is the *evidence*, and finally the probability  $P(H_n|D, I)$  on the l.h.s is the *a posteriori* (or *posterior*) probability of hypothesis  $H_n$ .

This form of Bayes' theorem can be used to discriminate between different hypotheses (a topic that we shall discuss later). It can also be modified to include parametric forms for the hypotheses: if we have a model that depends on a set of parameters  $\boldsymbol{\theta}$ , then different sets correspond to different – mutually exclusive – hypotheses. If, as often happens, parameters can change continuously, then hypotheses are identified by parameters, probabilities become pdf's, and the sum becomes an integral

$$p(\boldsymbol{\theta}|D, I) = \frac{L(D, \boldsymbol{\theta})}{\int_{\Theta} L(D, \boldsymbol{\theta}) p(\boldsymbol{\theta}|I) d\boldsymbol{\theta}} p(\boldsymbol{\theta}|I) \quad (10.3)$$

where  $\Theta$  is the range of the parameters  $\boldsymbol{\theta}$ .

In the Bayesian approach we select the parameter values that maximize the posterior pdf, and use the posterior pdf to set the confidence intervals as well. However the procedure requires a decision on the prior pdf. Then, when enough data are available the information carried by the prior pdf is simply “washed away” by the information in the data. On the other hand, when data are scarce, the prior information can significantly bias decisions based on Bayes' theorem. This requires a careful selection of the prior pdf, that must be as uninformative as possible. One common – and usually well-justified – choice, is a uniform distribution. If we take a uniform prior, the expression for the posterior probability simplifies considerably

$$p(\boldsymbol{\theta}|D, I) = \frac{L(D, \boldsymbol{\theta})}{\int_{\Theta} L(D, \boldsymbol{\theta}) d\boldsymbol{\theta}} \quad (10.4)$$

and since the normalizing integral (the evidence) does not influence the functional shape of the posterior probability, we see that maximizing the posterior probability is equivalent to maximizing the likelihood. This is the probabilistic basis for the *principle of maximum likelihood*.

Obviously, this approach to the method of maximum likelihood is not entirely acceptable to frequentists. In a frequentist framework, the method is justified by its a posteriori effectiveness, and we shall see that this is measured by statistical indicators like the *mean square error*.

When dealing with likelihoods, we can almost invariably assume that single measurements  $d_k$  are i.i.d. random variables with pdf  $p(d, \boldsymbol{\theta})$ , so that

likelihoods have a multiplicative structure:

$$L(D, \boldsymbol{\theta}) = \prod_k p(d_k, \boldsymbol{\theta}) \quad (10.5)$$

Maximizing a product may not be practical, but the logarithm is a monotonically increasing function of its argument, so that maximizing the logarithm of the likelihood is the same as maximizing the likelihood itself, and therefore we usually maximize

$$\ln L(D, \boldsymbol{\theta}) = \sum_k \ln p(d_k, \boldsymbol{\theta}) \quad (10.6)$$

## 10.2 An example with exponentially distributed data

Now we consider  $n$  exponentially distributed data, with the pdf

$$p(t) = \frac{1}{\tau} e^{-t/\tau} \quad (10.7)$$

so that there is just one parameter  $\theta = \tau$ , and

$$\ln L(D, \tau) = \sum_{k=1}^n \left( -\ln \tau - \frac{t_k}{\tau} \right) \quad (10.8)$$

Taking the derivative of  $\ln L(D, \tau)$  we find

$$\frac{\partial \ln L}{\partial \tau} = \sum_{k=1}^n \left( -\frac{1}{\tau} + \frac{t_k}{\tau^2} \right) = -\frac{1}{\tau^2} \sum_{k=1}^n (\tau - t_k) = -\frac{1}{\tau^2} \left( n\tau - \sum_{k=1}^n t_k \right) = 0 \quad (10.9)$$

therefore the *ML estimator* (MLE) for  $\tau$  is

$$\hat{\tau} = \frac{1}{n} \sum_{k=1}^n t_k \quad (10.10)$$

We use the decoration  $\hat{\cdot}$  to denote the ML estimator. Notice that in this case the sample mean coincides with the ML estimator.

The method can easily be adapted to more complex situations. Consider the case where the detector cannot measure times smaller than  $t_{\min}$  or larger than  $t_{\max}$ , then

$$p(t) = C e^{-t/\tau} \quad (10.11)$$

must be normalized so that

$$1 = \int_{t_{\min}}^{t_{\max}} p(t) dt = C \int_{t_{\min}}^{t_{\max}} e^{-t/\tau} dt = C\tau \left[ e^{-t_{\min}/\tau} - e^{-t_{\max}/\tau} \right] \quad (10.12)$$

thus

$$p(t) = \frac{1}{\tau} \frac{e^{-t/\tau}}{e^{-t_{\min}/\tau} - e^{-t_{\max}/\tau}} \quad (10.13)$$

Therefore we find

$$\ln L(D, \tau) = \sum_{k=1}^n \left\{ -\ln \tau - \frac{t_k}{\tau} - \ln \left[ e^{-t_{\min}/\tau} - e^{-t_{\max}/\tau} \right] \right\} \quad (10.14)$$

and

$$\begin{aligned} \frac{\partial \ln L}{\partial \tau} &= \sum_{k=1}^n \left[ -\frac{1}{\tau} - \frac{t_k}{\tau^2} + \frac{(t_{\min}/\tau^2)e^{-t_{\min}/\tau} - (t_{\max}/\tau^2)e^{-t_{\max}/\tau}}{e^{-t_{\min}/\tau} - e^{-t_{\max}/\tau}} \right] \\ &= -\frac{n}{\tau} + \frac{n}{\tau^2} \frac{t_{\min}e^{-t_{\min}/\tau} - t_{\max}e^{-t_{\max}/\tau}}{e^{-t_{\min}/\tau} - e^{-t_{\max}/\tau}} - \frac{1}{\tau^2} \sum_{k=1}^n t_k = 0 \end{aligned} \quad (10.15)$$

The solution of this equation cannot be found with analytical methods, and we must use a numerical method: the most common is the Newton-Raphson method.

The Newton-Raphson (NR) algorithm for the generic nonlinear equation  $F(\theta) = 0$  works as follows: we assume a reasonable starting value  $\theta_0$ , we take the linear approximation of  $F$  about  $\theta_0$ , and set it to 0

$$F(\theta_1) = F(\theta_0) + F'(\theta_0)(\theta_1 - \theta_0) = 0 \quad (10.16)$$

then we use this equation to evaluate the approximate solution  $\theta_1$ :

$$\theta_1 = \theta_0 - [F'(\theta_0)]^{-1} F(\theta_0) \quad (10.17)$$

We iterate this scheme until the value  $F(\theta_n)$  is sufficiently close to zero, or until the step size  $\delta_n = \theta_n - \theta_{n-1}$  is sufficiently small.

Now note that measurements are limited by our ability to measure very short times, rather than by very long waits, so that we can confidently set  $t_{\max} = \infty$  in (10.14), and we obtain

$$\ln L(D, \tau) = \sum_{k=1}^n \left( -\ln \tau - \frac{t_k}{\tau} + \frac{t_{\min}}{\tau} \right) = -n \ln \tau - \frac{1}{\tau} \sum_{k=1}^n t_k + n \frac{t_{\min}}{\tau} \quad (10.18)$$

This log-likelihood yields the following ML estimator

$$\hat{\tau} = \frac{1}{n} \sum_{k=1}^n t_k - t_{\min} \quad (10.19)$$

which does not coincide with the true value  $\tau$  even in the limit of large  $n$ .

### 10.3 Estimators and their properties

The ML method provides a consistent way to construct estimators. The problem is, how good are they? Can we tell whether there are better estimators? Can some form of optimality be established?

We consider first the generic properties of *point estimators*, i.e., estimates of individual parameters (one can also estimate intervals, distributions, etc.).

An estimator  $\hat{\theta}(D)$  of a parameter  $\theta$  is a function of the dataset  $D$ , and thus it is random variable as well. A good estimator must have an expectation value that coincides with the parameter, i.e.,

$$\mathbf{E}[\hat{\theta}(D)] = \theta \quad (10.20)$$

where the average is performed over the possible realizations of the dataset  $D$ , and in such a case the estimator is *unbiased*. This is not always the case, and if

$$\mathbf{E}[\hat{\theta}(D)] = \theta + b_n \quad (10.21)$$

the estimator is *biased*, with a *bias*  $b_n$  which usually depends on the dataset size  $n$ .

When we observe a particular dataset  $D$  we find the estimator  $\hat{\theta}(D)$ , and the *estimation error* is  $\hat{\theta}(D) - \theta$ , which includes a possible bias. This is the error of an individual estimate, however when characterizing an estimator one is more interested in the *mean square error*

$$\text{MSE}[\hat{\theta}(D)] = \mathbf{E}[(\hat{\theta}(D) - \theta)^2] \quad (10.22)$$

The variance of the estimator is also important:

$$\text{var}[\hat{\theta}(D)] = \mathbf{E}(\{\hat{\theta}(D) - \mathbf{E}[\hat{\theta}(D)]\}^2) \quad (10.23)$$

and because of bias, the variance can differ significantly from the MSE:

$$\text{var}[\hat{\theta}(D)] = \mathbf{E}(\{\hat{\theta}(D) - \mathbf{E}[\hat{\theta}(D)]\}^2) \quad (10.24a)$$

$$= \mathbf{E}\{[\hat{\theta}(D) - (\theta + b_n)]^2\} \quad (10.24b)$$

$$= \mathbf{E}[(\hat{\theta}(D) - \theta)^2] - 2b_n \mathbf{E}[\hat{\theta}(D) - \theta] + b_n^2 \quad (10.24c)$$

$$= \text{MSE}[\hat{\theta}(D)] - b_n^2 \quad (10.24d)$$

An estimator may be biased, but ideally it should converge in probability to the *true value* of the parameter for a large enough dataset. If it does so, then it is *consistent*, and

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}(D) - \theta| > \epsilon) = 0 \quad (10.25)$$

for all  $\epsilon > 0$ .

Finally, many estimators have pdf's that converge to Gaussian pdf's for large  $n$ : in this case estimators are *asymptotically normal*.

## 10.4 Invariance of ML estimates and bias

Consider a parameter transformation like  $a = a(\theta)$ , and assume that  $\hat{\theta}$  is the ML estimator. The corresponding value of  $a$  is  $\hat{a} = a(\hat{\theta})$  and the derivative of the likelihood is

$$\frac{\partial L}{\partial a}\Big|_{a=\hat{a}} = \frac{\partial L}{\partial \theta}\Big|_{\theta=\hat{\theta}} \frac{\partial \theta}{\partial a}\Big|_{a=\hat{a}} \quad (10.26)$$

and we conclude that if  $\partial L/\partial \theta|_{\theta=\hat{\theta}} = 0$  then  $\partial L/\partial a|_{a=\hat{a}} = 0$ , and this means that  $\hat{a}$  is an ML estimate as well.

This invariance property comes at a price. As an example consider the exponentially distributed data of section 10.2: we might be interested in the event rate  $\lambda = 1/\tau$  rather than the mean time between events  $\tau$ . From the previous results we find that

$$\hat{\lambda} = \frac{1}{\hat{\tau}} = \frac{n}{\sum_{k=1}^n d_k} \quad (10.27)$$

is the ML estimate of  $\lambda$ .

We already know that the sample mean

$$\mu_S = \frac{1}{n} \sum_{k=1}^n d_k \quad (10.28)$$

is an unbiased estimator of  $\tau$  (i.e., the ML estimate  $\hat{\tau}$ , which coincides with the sample mean, is an unbiased estimator). However, we also found in section 7.2 that  $\mu_S$  has a gamma distribution

$$p_{\mu_S}(\mu_S) = \frac{n^n}{(n-1)!} \left( \frac{\mu_S^{n-1}}{\tau^n} \right) e^{-n\mu_S/\tau} \quad (10.29)$$

so that

$$\begin{aligned} \mathbf{E}\hat{\lambda} &= \int_0^\infty \frac{1}{t} p_{\mu_S}(t) dt = \int_0^\infty \frac{n^n}{(n-1)!} \left( \frac{t^{n-2}}{\tau^n} \right) e^{-nt/\tau} dt = \frac{n}{n-1} \cdot \frac{1}{\tau} \\ &= \frac{n}{n-1} \lambda \end{aligned} \quad (10.30)$$

From this we see that the MLE of  $\lambda$  is not an unbiased estimator of  $\lambda$ , and that using the unbiased ML estimator for  $\tau$  does not yield an unbiased ML estimator for  $\lambda$ .

To sum it up, we conclude that the maximization procedure is invariant with respect to parameter changes, however it does not guarantee that the resulting estimator is unbiased.

## 10.5 Normally distributed data

Here we consider the very common case of data with a Gaussian distribution, where neither the mean nor the standard deviation are known, and we want to estimate them from data. The distribution of individual data is

$$p(d_k|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(d_k - \mu)^2}{2\sigma^2}\right] \quad (10.31)$$

so that

$$\ln L(D; \mu, \sigma) = \sum_{k=1}^n \left[ -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma^2 - \frac{(d_k - \mu)^2}{2\sigma^2} \right] \quad (10.32)$$

Then,

$$\frac{\partial \ln L}{\partial \mu} = \sum_{k=1}^n \frac{(d_k - \hat{\mu})}{\hat{\sigma}^2} = 0 \quad (10.33a)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \sum_{k=1}^n \left[ -\frac{1}{2\hat{\sigma}^2} + \frac{(d_k - \hat{\mu})^2}{2(\hat{\sigma}^2)^2} \right] = 0 \quad (10.33b)$$

and we find

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n d_k \quad (10.34a)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (d_k - \hat{\mu})^2 \quad (10.34b)$$

The ML estimator of the mean value  $\mu$  coincides with the sample mean, while the ML estimator of the variance has the  $n$  factor instead of  $n - 1$ : this ML estimator is biased, although asymptotically it coincides with the sample variance, i.e., it is consistent.

## 10.6 Consistency of ML estimators

The ML estimators are often consistent – i.e., asymptotically unbiased – and we can see this as follows. Define first the mean log-likelihood of i.i.d. data:

$$\ell(D, \theta) = \frac{1}{n} \ln L(D, \theta) = \frac{1}{n} \sum_{k=1}^n \ln p(x_k, \theta) \quad (10.35)$$

which is a sort of sample mean, and is itself a random variable. Notice that by maximizing the likelihood we also maximize  $\ell(D, \theta)$ . Notice also that – because of the law of large numbers – *the mean log-likelihood approaches a*

limit that does not depend on the data set, i.e.,  $\ell(D, \theta) \rightarrow \ell(\theta)$ , for large  $n$ , and in particular

$$\ell(D, \theta) \rightarrow \ell(\theta) = \mathbf{E}[\ln p(x, \theta)] = \int_{\{x\}} [\ln p(x, \theta)] p(x, \theta_0) dx \quad (10.36)$$

because the values of  $\ell(D, \theta)$  follow a pdf  $p(x, \theta_0)$  that depends on the actual distribution of the data values, and therefore on the *true value*  $\theta_0$ .

Now notice that<sup>1</sup>:

$$\ell(\theta) - \ell(\theta_0) = \mathbf{E}[\ell(D, \theta) - \ell(D, \theta_0)] \quad (10.37a)$$

$$= \int_{\{x\}} [\ln p(x, \theta) - \ln p(x, \theta_0)] p(x, \theta_0) dx \quad (10.37b)$$

$$= \int_{\{x\}} \ln \frac{p(x, \theta)}{p(x, \theta_0)} p(x, \theta_0) dx \quad (10.37c)$$

$$\leq \int_{\{x\}} \left[ \frac{p(x, \theta)}{p(x, \theta_0)} - 1 \right] p(x, \theta_0) dx \quad (10.37d)$$

$$= \int_{\{x\}} [p(x, \theta) - p(x, \theta_0)] dx = 1 - 1 = 0 \quad (10.37e)$$

Therefore, for large  $n$ ,  $\ell(D, \theta_0)$ , and equivalently  $\ln L(D, \theta_0)$  correspond to a maximum, i.e.,  $\theta_0$  is equal to the value  $\hat{\theta}$  that maximizes the likelihood, and we conclude that  $\hat{\theta}$  is a consistent estimator.

While this proof holds for many common realizations of the ML formalism, the underlying assumptions are not always true, and some ML estimators are not consistent [12, 22].

## 10.7 Estimator efficiency and the Cramér-Rao-Fisher bound

Efficiency is one further important feature that distinguishes estimators: an efficient estimator has minimal MSE. Here we study an important inequality that establishes the absolute minimum of estimator variance, and is thus related to efficiency, the Cramér-Rao-Fisher bound. We start by establishing two important identities, the *Bartlett identities*.

### 10.7.1 Bartlett identities

Consider a family of normalized pdf's  $p(x, \theta)$  which depend on a parameter  $\theta$ , so that

$$\int_{\{x\}} p(x, \theta) dx = 1 \quad (10.38)$$

---

<sup>1</sup>Here we use the inequality  $\ln x \leq x - 1$

where  $\{x\}$  denotes the range of  $x$ , and which holds for all  $\theta$ 's. Then

$$\frac{\partial}{\partial \theta} \int_{\{x\}} p(x, \theta) dx = \int_{\{x\}} \frac{\partial p(x, \theta)}{\partial \theta} dx = 0 \quad (10.39)$$

We divide and multiply by  $p(x, \theta)$  inside the integral:

$$0 = \int_{\{x\}} \frac{\partial p(x, \theta)}{\partial \theta} dx = \int_{\{x\}} \frac{1}{p(x, \theta)} \frac{\partial p(x, \theta)}{\partial \theta} p(x, \theta) dx \quad (10.40a)$$

$$= \int_{\{x\}} \frac{\partial \ln p(x, \theta)}{\partial \theta} p(x, \theta) dx \quad (10.40b)$$

$$= \mathbf{E} \left[ \frac{\partial \ln p(x, \theta)}{\partial \theta} \right] \quad (10.40c)$$

(first Bartlett identity). We can repeat the process taking the derivative of

$$\int_{\{x\}} \frac{\partial \ln p(x, \theta)}{\partial \theta} p(x, \theta) dx = 0 \quad (10.41)$$

and find

$$0 = \frac{\partial}{\partial \theta} \int_{\{x\}} \frac{\partial \ln p(x, \theta)}{\partial \theta} p(x, \theta) dx \quad (10.42a)$$

$$= \int_{\{x\}} \left[ \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} p(x, \theta) + \frac{\partial \ln p(x, \theta)}{\partial \theta} \frac{\partial p(x, \theta)}{\partial \theta} \right] dx \quad (10.42b)$$

$$= \int_{\{x\}} \left[ \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} + \frac{1}{p(x, \theta)} \frac{\partial \ln p(x, \theta)}{\partial \theta} \frac{\partial p(x, \theta)}{\partial \theta} \right] p(x, \theta) dx \quad (10.42c)$$

$$= \int_{\{x\}} \left\{ \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} + \left[ \frac{\partial \ln p(x, \theta)}{\partial \theta} \right]^2 \right\} p(x, \theta) dx \quad (10.42d)$$

$$= \mathbf{E} \left[ \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} \right] + \mathbf{E} \left[ \left( \frac{\partial \ln p(x, \theta)}{\partial \theta} \right)^2 \right] \quad (10.42e)$$

(second Bartlett identity).

Using the first Bartlett identity, we can also write the second Bartlett identity in the form

$$\text{var} \left[ \frac{\partial \ln p(x, \theta)}{\partial \theta} \right] = \mathbf{E} \left[ \left( \frac{\partial \ln p(x, \theta)}{\partial \theta} \right)^2 \right] = -\mathbf{E} \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} \quad (10.43)$$

As a side remark, I note that it is easy to generalize the second Bartlett identity to multiparameter distributions, and find

$$\mathbf{E} \left[ \left( \frac{\partial \ln p(x, \boldsymbol{\theta})}{\partial \theta_i} \right) \left( \frac{\partial \ln p(x, \boldsymbol{\theta})}{\partial \theta_j} \right) \right] = -\mathbf{E} \left[ \frac{\partial^2 \ln p(x, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] \quad (10.44)$$

where  $\boldsymbol{\theta} = \{\theta_i\}_{i=1,\dots,n}$ .

### 10.7.2 The Cramér-Rao-Fisher bound

The Bartlett identities hold for any parametric family of pdf's, and in particular they hold for the likelihood, which is just the pdf of data which are assumed to be distributed according to some well-defined probability model. Here we show an important statement which gives a lower bound to the variance attainable by *all possible estimators* of the “true” parameter  $\theta_0$ , *assuming the model of data embodied in the likelihood function*.

Here we consider an estimator  $\theta$  – obtained with some unspecified method – which has the bias  $b_n$  (the bias depends on the dataset size  $n$ ), i.e.,

$$\mathbf{E}[\theta(D)] = \int_D \theta(x)L(x, \theta_0)dx = \theta_0 + b_n \quad (10.45)$$

and we take the derivative with respect to  $\theta$  of this expression, noting that  $\theta$  does not depend on the “true” value  $\theta$  of the parameter, i.e.,

$$\frac{\partial}{\partial\theta_0}\mathbf{E}[\theta(D)] = 1 + \frac{\partial b_n}{\partial\theta_0} \quad (10.46a)$$

$$= \int_D \theta(x)\frac{\partial L(x, \theta_0)}{\partial\theta_0}dx \quad (10.46b)$$

$$= \int_D \theta(x)\frac{\partial \ln L(x, \theta_0)}{\partial\theta_0}L(x, \theta_0)dx \quad (10.46c)$$

$$= \mathbf{E}\left[\theta(D)\frac{\partial \ln L(D, \theta_0)}{\partial\theta_0}\right] \quad (10.46d)$$

Now we can rearrange this expression using the definition of covariance and the first Bartlett identity:

$$1 + \frac{\partial b_n}{\partial\theta_0} = \mathbf{E}\left[\theta(D)\frac{\partial \ln L(D, \theta_0)}{\partial\theta_0}\right] \quad (10.47a)$$

$$= \text{cov}\left[\theta(D), \frac{\partial \ln L(D, \theta_0)}{\partial\theta_0}\right] + \mathbf{E}[\theta(D)]\mathbf{E}\left[\frac{\partial \ln L(D, \theta_0)}{\partial\theta_0}\right] \quad (10.47b)$$

$$= \text{cov}\left[\theta(D), \frac{\partial \ln L(D, \theta_0)}{\partial\theta_0}\right] \quad (10.47c)$$

Using Schwartz's inequality for covariance (proved in the appendix of chapter 7)

$$[\text{cov}(x, y)]^2 \leq \sigma_x^2 \sigma_y^2 \quad (10.48)$$

we find

$$\begin{aligned} \left(1 + \frac{\partial b_n}{\partial\theta_0}\right)^2 &= \left\{\text{cov}\left[\theta(D), \frac{\partial \ln L(D, \theta_0)}{\partial\theta_0}\right]\right\}^2 \\ &\leq \text{var}[\theta(D)]\text{var}\left[\frac{\partial \ln L(D, \theta_0)}{\partial\theta_0}\right] \end{aligned} \quad (10.49)$$

and finally, using again Bartlett's identities, we prove the inequality

$$\begin{aligned} \text{var}[\theta(D)] &\geq \\ &\geq \frac{\left(1 + \frac{\partial b_n}{\partial \theta_0}\right)^2}{\text{var}\left[\frac{\partial \ln L(D, \theta_0)}{\partial \theta_0}\right]} = \frac{\left(1 + \frac{\partial b_n}{\partial \theta_0}\right)^2}{\mathbf{E}\left[\left(\frac{\partial \ln L(D, \theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{\left(1 + \frac{\partial b_n}{\partial \theta_0}\right)^2}{-\mathbf{E}\frac{\partial^2 \ln L(D, \theta_0)}{\partial \theta_0^2}} \end{aligned} \quad (10.50)$$

which is the Cramér-Rao-Fisher bound. For unbiased estimators this gives

$$\begin{aligned} \text{var}[\hat{\theta}(D)] &\geq \\ &\geq \frac{1}{\text{var}\left[\frac{\partial \ln L(D, \theta_0)}{\partial \theta_0}\right]} = \frac{1}{\mathbf{E}\left[\left(\frac{\partial \ln L(D, \theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{1}{-\mathbf{E}\frac{\partial^2 \ln L(D, \theta_0)}{\partial \theta_0^2}} \end{aligned} \quad (10.51)$$

### 10.7.3 Example with exponentially distributed data

As an example, consider the (unbiased) estimator of the mean time of the exponential distribution

$$\hat{\tau} = \frac{1}{n} \sum_{k=1}^n t_k \quad (10.52)$$

The estimator is unbiased, therefore  $b_n = 0$ . We found earlier that

$$\frac{\partial \ln L}{\partial \tau} = \sum_{k=1}^n \left( -\frac{1}{\tau} + \frac{t_k}{\tau^2} \right) \quad (10.53)$$

therefore

$$\frac{\partial^2 \ln L}{\partial \tau^2} = \sum_{k=1}^n \left( \frac{1}{\tau^2} - \frac{2t_k}{\tau^3} \right) = \frac{1}{\tau^3} \left( n\tau - 2 \sum_{k=1}^n t_k \right) \quad (10.54)$$

and

$$-\mathbf{E}\frac{\partial^2 \ln L}{\partial \tau^2} = -\frac{1}{\tau^3} \left( n\tau - 2\mathbf{E}\sum_{k=1}^n t_k \right) = -\frac{1}{\tau^3} (n\tau - 2n\tau) = \frac{n}{\tau^2} \quad (10.55)$$

and finally, using the Cramér-Rao-Fisher bound, we find

$$\text{var}(\hat{\tau}) \geq \frac{\tau^2}{n} \quad (10.56)$$

However we know that the variance of the estimator  $\hat{\tau}$  is just  $\tau^2/n$ , therefore the ML estimator satisfies the equality in the Cramér-Rao-Fisher bound, and we conclude that the ML estimator is an estimator of  $\tau$  with maximal efficiency. This result is actually a generic property of ML estimators, as we shall see below.

## 10.8 Optimality and Gaussianity of consistent ML estimators at large $n$

We have already found that the Cramér-Rao-Fisher bound (10.50) or (10.51) relates the variance of the estimator to the variance of the first derivative of the log-likelihood.

Recall that using the Bartlett identities we find equivalent expressions for the variance of the derivative of the log-likelihood – this is equation (10.43) above:

$$\begin{aligned} \text{var} \left( \frac{\partial \ln L(D, \theta)}{\partial \theta} \right) \\ = \mathbf{E} \left[ \left( \frac{\partial \ln L(D, \theta)}{\partial \theta} \right)^2 \right] = -\mathbf{E} \left( \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \right) \quad (10.57) \end{aligned}$$

which hold for any value of  $\theta$ , and in particular they hold for the “true” value  $\theta_0$

$$\begin{aligned} \text{var} \left( \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right) \\ = \mathbf{E} \left[ \left( \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right)^2 \right] = -\mathbf{E} \left( \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right) \quad (10.58) \end{aligned}$$

Now we expand the derivative of the log-likelihood up to first order about the “true” value  $\theta_0$

$$\frac{\partial \ln L(D, \theta)}{\partial \theta} \approx \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} + \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} (\theta - \theta_0) \quad (10.59)$$

Then for large  $n$  – so that, because of the consistency of ML estimators  $\hat{\theta} \approx \theta_0$  – the following relation holds

$$0 = \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \approx \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} + \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} (\hat{\theta} - \theta_0) \quad (10.60)$$

and therefore

$$\frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \approx - \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} (\hat{\theta} - \theta_0) \quad (10.61)$$

Now we make the additional assumption that  $\partial^2 \ln L(D, \theta)/\partial \theta^2$  and  $\hat{\theta}$  are uncorrelated – this means that the peak position and the width of the likelihood function are uncorrelated – we use equation (10.61), and the Bartlett

identities, and we find

$$\text{var} \left( \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right) = \mathbf{E} \left[ \left( \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right)^2 \right] \quad (10.62a)$$

$$= \mathbf{E} \left[ \left( \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} (\hat{\theta} - \theta_0) \right)^2 \right] \quad (10.62b)$$

$$= \mathbf{E} \left[ \left( \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right)^2 \right] \mathbf{E}[(\hat{\theta} - \theta_0)^2] \quad (10.62c)$$

$$= \mathbf{E} \left[ \left( \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right)^2 \right] \text{var}(\hat{\theta}) \quad (10.62d)$$

Then, using equation (10.58), we find

$$\text{var}(\hat{\theta}) = \frac{\mathbf{E} \left( - \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right)}{\mathbf{E} \left[ \left( \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right)^2 \right]} \quad (10.63)$$

However we are evaluating this expression for large  $n$ , and from the law of large numbers, the derivatives in the expressions converge in probability to the expectation values, and thus we find that asymptotically

$$\text{var}(\hat{\theta}) \approx \left[ \mathbf{E} \left( - \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right) \right]^{-1} \approx \left[ \mathbf{E} \left( - \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} \right) \right]^{-1} \quad (10.64)$$

i.e., ML estimators are asymptotically optimal (they realize the Cramér-Rao-Fisher bound).

In the case of ML estimators, the derivative of the log-likelihood satisfies the equation

$$\frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = 0 \quad (10.65)$$

where  $\hat{\theta}$  is the ML estimator, it depends on the dataset  $D$  and is itself a random variable, and since  $\partial \ln L(D, \theta)/\partial \theta$  is a sum of  $n$  i.i.d. terms,

$$\frac{\partial \ln L(D, \theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_{k=1}^n \ln p(x_k, \theta) = \sum_{k=1}^n \frac{\partial \ln p(x_k, \theta)}{\partial \theta} \quad (10.66)$$

where  $p(x, \theta)$  is the pdf of each individual data point  $x$ , then for large  $n$  the derivative of the log-likelihood is approximately Gaussian (assuming that the conditions for the validity of the CLT are satisfied). From the first Bartlett identity we also know that its mean value is 0, and from the proof above we know that its variance is the inverse of the variance of the ML estimator.

Can we infer the Gaussianity of MLE from the Gaussianity of the log-likelihood? It turns out that we can. Indeed, from the mean value theorem, we know that

$$0 = \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} + \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_1} (\hat{\theta} - \theta_0) \quad (10.67)$$

where  $\theta_1 \in (\theta_0, \hat{\theta})$ ; and because of consistency  $\hat{\theta} = \theta_0$  for large  $n$  – we also have  $\theta_1 = \theta_0$  for large  $n$ . Then

$$\begin{aligned} \hat{\theta} - \theta_0 &= -\frac{\frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0}}{\left( \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_1} \right)} \approx \frac{\frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0}}{\mathbf{E} \left( -\frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \Big|_{\theta=\theta_0} \right)} \\ &= \text{var}(\hat{\theta}) \frac{\partial \ln L(D, \theta)}{\partial \theta} \Big|_{\theta=\theta_0} \end{aligned} \quad (10.68)$$

Since, for large  $n$ , the derivative of the log-likelihood is Gaussian, then the estimator  $\hat{\theta}$  is also Gaussian, with distribution  $N(\theta_0, 1/\sqrt{\text{var}(\hat{\theta})})$ .

**Exercise:** Extend these conclusions to the multiparameter case.

## 10.9 Interlude on information measures

When dealing with estimator efficiency and with the Cramér-Rao-Fisher bound, it is customary to introduce the quantity the *Fisher information*, which is not quite what we mean when we talk about information today, but is closely related to some important measures of information. For this reason we briefly discuss information measures before moving on to confidence intervals.

### 10.9.1 Boltzmann entropy

The Boltzmann entropy is a measure of the number of thermodynamical configurations, and it is defined by the equation

$$S_B = k_B \ln \Omega \quad (10.69)$$

where  $\Omega$  is the number of configurations, subject to the physical constraints imposed by the system under study. This expression can be further elaborated in simple situations like the case where there are  $M$  available states and a total of  $N = N_1 + N_2 + \dots + N_M$  “particles”, and where  $N_k$  is the occupancy number of the  $k$ -th state. Then, the number of states that correspond to these occupancy numbers is

$$\Omega = \frac{N!}{N_1! N_2! \dots N_M!} \quad (10.70)$$

and

$$\begin{aligned} \ln \Omega &= \ln N! - \sum_{k=1}^M \ln N_k! \approx N \ln N - N - \sum_{k=1}^M (N_k \ln N_k - N_k) \\ &= - \sum_{k=1}^M (N_k \ln N_k - N_k \ln N) = -N \sum_{k=1}^M \frac{N_k}{N} \ln \frac{N_k}{N} \end{aligned} \quad (10.71)$$

We denote with  $p_k = N_k/N$  the fraction of “particles” in state  $k$ , so that

$$S_B = -k_B N \sum_{k=1}^M p_k \ln p_k = k_B N \sum_{k=1}^M p_k \ln \frac{1}{p_k} \quad (10.72)$$

The Boltzmann entropy has a central role in statistical mechanics, and the expression (10.72) goes beyond the simple example examined here. The particular form of Boltzmann entropy (10.72) is also called *Gibbs entropy* or *statistical entropy*.

### 10.9.2 Shannon information

In his landmark paper ”A Mathematical Theory of Communication” [30], Claude E. Shannon introduced an extremely important measure of information. He considered a source of  $M$  symbols, that outputs these symbols with probability  $p_1, \dots, p_M$ , and defined the information carried by the  $k$ -th symbol as  $I_k = -\log_2 p_k = \log_2 1/p_k$  bits. Then, the average information output is

$$S_S = - \sum_{k=1}^M p_k \log_2 p_k = \sum_{k=1}^M p_k \log_2 \frac{1}{p_k} \quad (10.73)$$

which is clearly proportional to the formal expression of the Boltzmann entropy.

It is natural to ask how the maximum of the Shannon entropy is attained: to find it we must maximize the expression (10.73) with the normalization

constraint  $\sum_k p_k = 1$ , i.e., we must introduce one Lagrange multiplier  $\lambda$ . Thus we maximize the expression

$$S_S - \lambda \sum_{k=1}^M p_k = - \sum_{k=1}^M p_k \log_2 p_k - \lambda \sum_{k=1}^M p_k = - \sum_{k=1}^M p_k \frac{\ln p_k}{\ln 2} - \lambda \sum_{k=1}^M p_k \quad (10.74)$$

i.e., we must solve the equations

$$\ln p_k + 1 = -\lambda \ln 2 \quad (10.75)$$

The solution is easy to find:

$$p_k = \exp(-\lambda \ln 2 - 1) = \frac{1}{M} \quad (10.76)$$

i.e., the Shannon entropy is maximized by the (discrete) uniform distribution, and the maximum value is

$$S_S = \log_2 M \quad (10.77)$$

Any other discrete distribution has a lower entropy. Thus we can also view the Shannon entropy as a sort of *distance* of a given distribution from the uniform distribution.

### 10.9.3 Kullback-Leibler divergence

The “obvious” extension of the Shannon entropy to continuous distributions is

$$S_C = \int_{-\infty}^{+\infty} p(x) dx \log_2 \frac{1}{p(x) dx} \quad (10.78)$$

because  $p(x)dx$  is the probability that  $x$  lies in the interval  $(x, x + dx)$ . However this expression diverges, and it must be modified to yield a suitable information measure.

A way out is suggested by the Boltzmann entropy for a model that differs slightly from the one that led us to equation (10.71). Now we attach a degeneracy number  $g_k$  to each state, so that the number of states is

$$\Omega = \frac{N!}{N_1! N_2! \dots N_M!} g_1^{N_1} g_2^{N_2} \dots g_M^{N_M} \quad (10.79)$$

and therefore

$$\ln \Omega = \ln N! - \sum_{k=1}^M \ln N_k! + \sum_{k=1}^M N_k \ln g_k = -N \sum_{k=1}^M (N_k/N) \ln \frac{(N_k/N)}{g_k} \quad (10.80)$$

And finally, by denoting again with  $p_k = N_k/N$  the fraction of “particles” in state  $k$ , we find a new expression for the Boltzmann entropy in this special case

$$S_B = -k_B N \sum_{k=1}^M p_k \ln \frac{p_k}{g_k} \quad (10.81)$$

This suggests the introduction of a “relative entropy”, where the information is relative to a reference (unnormalized) distribution  $g_k$ . This quantity has indeed been introduced in the mathematical literature by Kullback and Leibler [23] and is called the *Kullback-Leibler divergence*:

$$I_{KL} = \sum_{k=1}^M p_k \ln \frac{p_k}{g_k} \quad (10.82)$$

The extremal value of  $I_{KL}$  can be found just as in the case of the Shannon entropy, we introduce a Lagrange multiplier and we find the extremum of  $I_{KL} + \lambda \sum_k p_k$ . Thus we obtain the equations

$$\ln p_k + 1 - \ln g_k - \lambda = 0 \quad (10.83)$$

the solution is  $p_k = g_k e^{\lambda-1}$ , and the extreme value is obtained when the two normalized distributions are equal. Moreover, the extreme value is equal to 0.

In the continuous case there are no problems now, and the Kullback-Leibler divergence is

$$I_{KL} = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{g(x)} dx \quad (10.84)$$

It is easy to show that  $I_{KL} \geq 0$ : to this end we define the auxiliary function  $\phi(t) = t \ln t$ , and we note that

$$\begin{aligned} \phi(t) &= \phi(1) + \phi'(1)(t-1) + \frac{1}{2}\phi''(h)(t-1)^2 \\ &= (t-1) + \frac{1}{2h}(t-1)^2 \end{aligned} \quad (10.85)$$

where  $t < h < 1$ . Therefore

$$I_{KL} = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{g(x)} dx = - \int_{-\infty}^{+\infty} \frac{p(x)}{g(x)} \ln \frac{p(x)}{g(x)} g(x) dx \quad (10.86a)$$

$$= \int_{-\infty}^{+\infty} \phi\left(\frac{p(x)}{g(x)}\right) g(x) dx \quad (10.86b)$$

$$= \int_{-\infty}^{+\infty} \left[ \left( \frac{p(x)}{g(x)} - 1 \right) + \frac{1}{2h} \left( \frac{p(x)}{g(x)} - 1 \right)^2 \right] g(x) dx \quad (10.86c)$$

$$= \int_{-\infty}^{+\infty} \frac{1}{2h} \left( \frac{p(x)}{g(x)} - 1 \right)^2 g(x) dx \quad (10.86d)$$

$$= \int_{-\infty}^{+\infty} \frac{1}{2h} \frac{(p(x) - g(x))^2}{g(x)} dx \geq 0 \quad (10.86e)$$

This shows that the Kullback-Leibler divergence is minimum when  $p(x) = g(x)$  and that it is actually a sort of distance between the distributions  $p(x)$  and  $g(x)$ .

However the Kullback-Leibler divergence is not symmetrical and therefore it is not a true distance. This can be obviated, e.g., the Jeffreys distance between distributions is  $J(p(x), g(x)) = I_{KL}(p(x), g(x)) + I_{KL}(g(x), p(x))$ .

It is also important to note that the Kullback-Leibler divergence does not depend on representation (it is invariant with respect to variable transformations – prove this as an exercise).

### Exercises

- Find the KL divergence for two normal distributions with different mean values and variances;
- Find the KL divergence for two exponential distributions with different mean;
- Which distribution is closer – in the KL sense – to a gamma distribution with given parameters? A normal distribution or a log-normal distribution, both with the same mean and variance?

#### 10.9.4 Fisher information

Now we are ready to go back to the Cramér-Rao-Fisher bound, where we met the quantity

$$I(\theta) = \mathbf{E} \left[ \left( \frac{\partial \ln p(x, \theta)}{\partial \theta} \right)^2 \right] = -\mathbf{E} \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} \quad (10.87)$$

which is usually called *Fisher information*. Just as it happens for the Bartlett identities, it is straightforward to extend this quantity to the multi parameter case, where we define a (symmetric) Fisher information matrix with elements

$$I_{i,j} = \mathbf{E} \left[ \left( \frac{\partial \ln p(x, \theta)}{\partial \theta_i} \right) \left( \frac{\partial \ln p(x, \theta)}{\partial \theta_j} \right) \right] = -\mathbf{E} \left[ \frac{\partial^2 \ln p(x, \theta)}{\partial \theta_i \partial \theta_j} \right] \quad (10.88)$$

It turns out that the Fisher information is closely related to the Kullback-Leibler divergence. Consider a parametric family of distributions  $p(x, \theta)$ , and the KL divergence of  $p(x, \theta)$  and  $p(x, \theta + \epsilon)$ , where  $\epsilon$  is an infinitesimal parameter change:

$$\begin{aligned} I_{KL}(p(x, \theta), p(x, \theta + \epsilon)) &= \int_{-\infty}^{+\infty} p(x, \theta) \ln \frac{p(x, \theta)}{p(x, \theta + \epsilon)} dx \\ &= \mathbf{E}(\ln p(x, \theta) - \ln p(x, \theta + \epsilon)) \end{aligned} \quad (10.89)$$

The Taylor expansion of  $\ln p(x, \theta + \epsilon)$  up to second order is

$$\ln p(x, \theta + \epsilon) \approx \ln p(x, \theta) + \frac{\partial \ln p(x, \theta)}{\partial \theta} \epsilon + \frac{1}{2} \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} \epsilon^2 \quad (10.90)$$

therefore, using the first Bartlett identity to get rid of the first derivative, we find

$$\begin{aligned} I_{KL}(p(x, \theta), p(x, \theta + \epsilon)) &= \\ \mathbf{E}(\ln p(x, \theta) - \ln p(x, \theta + \epsilon)) &= -\frac{1}{2} \mathbf{E} \left[ \frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} \right] \epsilon^2 = \frac{1}{2} I(\theta) \epsilon^2 \end{aligned} \quad (10.91)$$

We conclude that the Fisher information is a local version of the KL divergence. However, unlike the KL divergence, it is symmetric (it does not depend on the sign of  $\epsilon$ ). Indeed, it can be shown that in a multi parameter space,  $I_{ij}$  behaves as a metric tensor. From equation (10.64) we also see that the Fisher information is the inverse of the ML parameter variance (with respect to the posterior distribution of the parameter, which corresponds to the likelihood), and in this sense the variance (and for the multi parameter case, the covariance matrix) also sets the metric properties of the parameter space.

### **Example: Fisher information for a family of zero-mean normal distributions**

The generic pdf of the family of zero-mean normal distributions is

$$p(x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{x^2}{2\sigma^2} \right) \quad (10.92)$$

then

$$\ln p(x, \theta) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{x^2}{2\sigma^2} \quad (10.93)$$

We must evaluate the derivatives

$$\frac{\partial \ln p(x, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{x^2}{2(\sigma^2)^2} \quad (10.94)$$

$$-\frac{\partial^2 \ln p(x, \sigma^2)}{\partial (\sigma^2)^2} = -\frac{1}{2(\sigma^2)^2} + \frac{x^2}{(\sigma^2)^3} \quad (10.95)$$

so that the Fisher information is

$$I(\theta) = \mathbf{E} \left( -\frac{1}{2(\sigma^2)^2} + \frac{x^2}{(\sigma^2)^3} \right) = -\frac{1}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^2} = \frac{1}{2(\sigma^2)^2} \quad (10.96)$$

The last equation shows very clearly that a smaller variance yields a larger Fisher information, i.e., a concentrated distribution is more “informative” than a spread-out distribution.

### **Example: Fisher information for a family of Bernoulli distributions**

Consider

$$p(x, \theta) = \theta^x (1 - \theta)^{1-x} \quad (10.97)$$

where  $x$  is an integer which is either 0 or 1. Then

$$\ln p(x, \theta) = x \ln \theta + (1 - x) \ln(1 - \theta) \quad (10.98)$$

and

$$\frac{\partial \ln p(x, \theta)}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta} \quad (10.99)$$

$$-\frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} = \frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2} \quad (10.100)$$

so that the Fisher information is

$$I(\theta) = \mathbf{E} \left( \frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2} \right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)} \quad (10.101)$$

### **Example: Fisher information for a family of exponential distributions**

Now let

$$p(x, \theta) = \theta e^{-x\theta} \quad (10.102)$$

then

$$\ln p(x, \theta) = \ln \theta - x\theta \quad (10.103)$$

and

$$\frac{\partial \ln p(x, \theta)}{\partial \theta} = \frac{1}{\theta} - x \quad (10.104)$$

$$-\frac{\partial^2 \ln p(x, \theta)}{\partial \theta^2} = \frac{1}{\theta^2} \quad (10.105)$$

so that the Fisher information is

$$I(\theta) = \mathbf{E} \left( \frac{1}{\theta^2} \right) = \frac{1}{\theta^2} \quad (10.106)$$

## 10.10 Confidence intervals

When we estimate a parameter (a *point estimate*) we are still missing something: we need an estimate of the uncertainty. In practice this means that we must determine an interval around the parameter estimate with an attached probability which quantifies the uncertainty (this is an example of an *interval estimate*).

Here we start with the Bayesian construction, which is much simpler than the frequentist construction, although it is not acceptable to frequentists.

### 10.10.1 Bayesian credible intervals

We start from Bayes' theorem, as in section 10.1:

$$p(\theta|D, I) = \frac{p(D|\theta, I)}{\int_{\theta} p(D|\theta, I)p(\theta, I)d\theta} p(\theta, I) \quad (10.107)$$

where  $D$  represents the data set,  $\theta$  is a parameter that we wish to estimate from data, and  $I$  is all the prior information that helps us determine the prior distribution and the physical model for the distribution of data (the likelihood function). This expression involves a *prior* pdf of the parameter  $\theta$ ,  $p(\theta, I)$ , and a *posterior* pdf,  $p(\theta|D)$ , and therefore it is unacceptable to frequentists, like F. James [20], who insist that we must uncover the underlying truth of nature and that there is no such thing as a pdf of a “true” parameter. Bayesians do not share this view, and they stress the “subjective” character of knowledge – in practice they do not deny the existence of an underlying reality, but they dismiss the possibility of ever reaching an absolute knowledge of this reality, and for this reason they advocate the use of probability distributions of theoretical parameters as a measure of our “degree of belief” in their value. All in all, using the posterior pdf it is easy to define and construct the Bayesian equivalent of the confidence intervals, i.e., the *credible intervals*, even though one gives up the practical concept of *coverage* (see below)<sup>2,3</sup>.

---

<sup>2</sup>The author of these notes has a personal preference for the Bayesian view.

<sup>3</sup>Some of James' criticism is clearly ill-directed. For instance in [20] he dismisses the Bayesian approach to the uniform prior noting that a variable transformation can change

According to the Bayesian view we can easily define an interval with a given probability  $C$  that the value of the parameter falls into that interval. The value  $C$  is called the *confidence level*, and there is an infinite number of ways to choose the endpoints of this interval. Figure 10.1 shows an example pdf and the corresponding distribution function, and figure 10.2 shows three possible credible intervals: two of them *do not* cover the peak value. There are several rules for choosing the credible interval (or the confidence interval), one of them is the choice of the shortest interval; figure 10.3 shows a plot of the interval length vs. the position of its left endpoint. Common choices for the interval are

- the endpoints are equidistant from the parameter estimate;
- the interval is shortest;
- the interval is central, i.e., the probability mass of the excluded tails is the same on the left and on the right (this is the most common choice, see figure 10.4).

---

a uniform prior into a strongly nonuniform prior, and concluding that the usual choice – when completely ignorant we pick a uniform prior – makes little or no sense. We already know from the discussion of Bertrand’s paradox that the choice of a uniform pdf may be ambiguous, however we have also learned from Jaynes’ solution of that paradox, that physical problems often (always?) do have symmetry properties that ensure that a *unique* choice of the uniform prior is at hand.

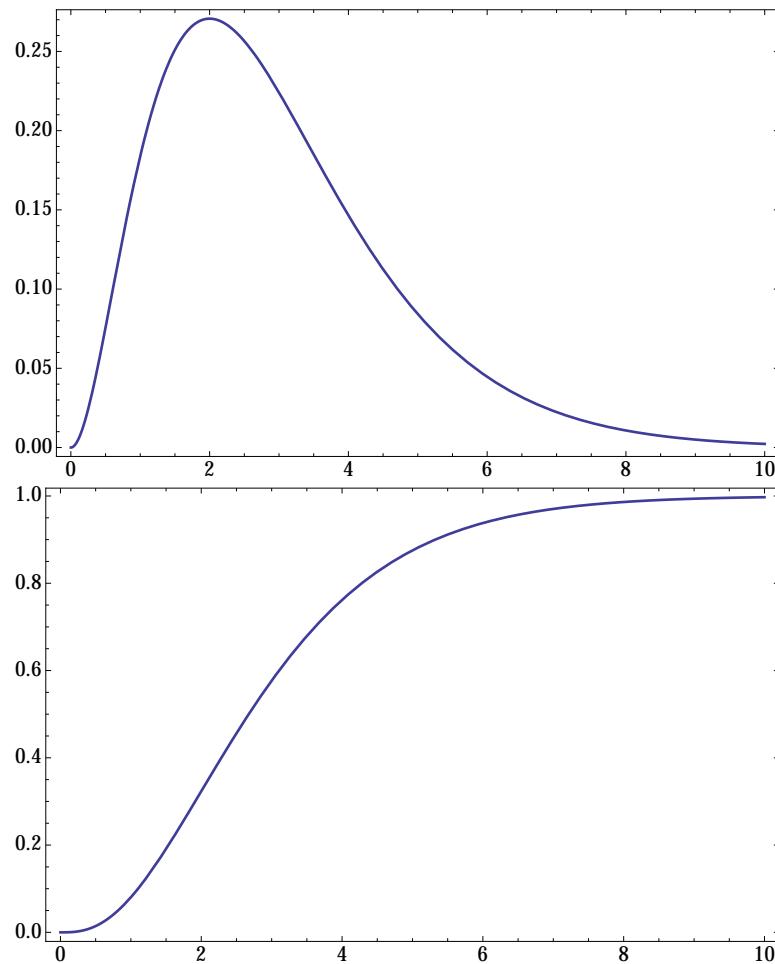


Figure 10.1: The left panel shows an example  $p(x)$  vs.  $x$ , where the posterior pdf  $p(x)$  is a gamma distribution (here  $k = 3$ ,  $\theta = 1$ , with the notation of section 4.11). In this case the pdf peaks at  $x = 2$ . The right panel shows the corresponding cumulative distribution function.

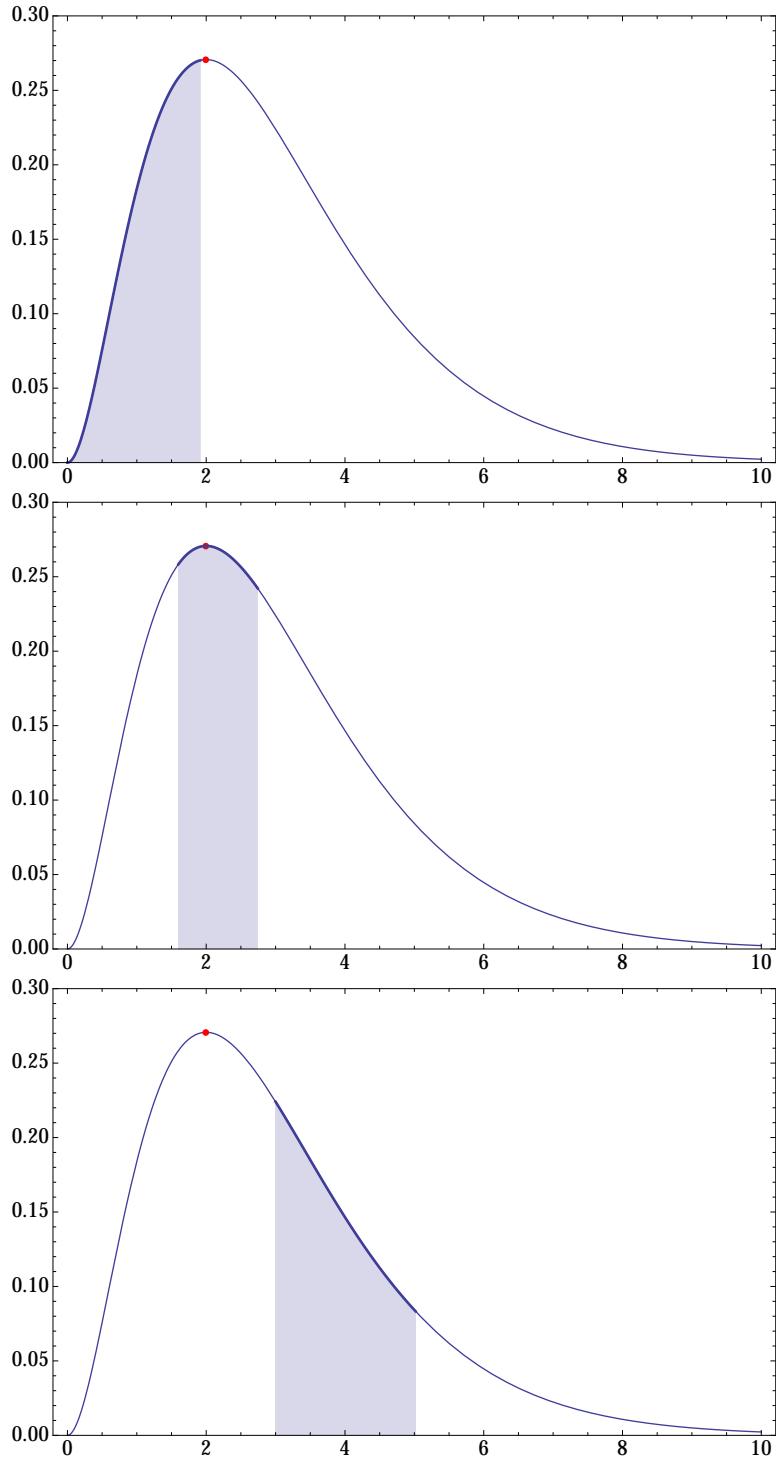


Figure 10.2: The gray regions mark the equal probability masses corresponding to three different intervals with confidence level  $C = 0.3$ . This is a very low level, which is used here only as an example. The intervals in the top and the bottom panels do not cover the peak value (marked by a red dot). All the intervals are asymmetric about the peak value.

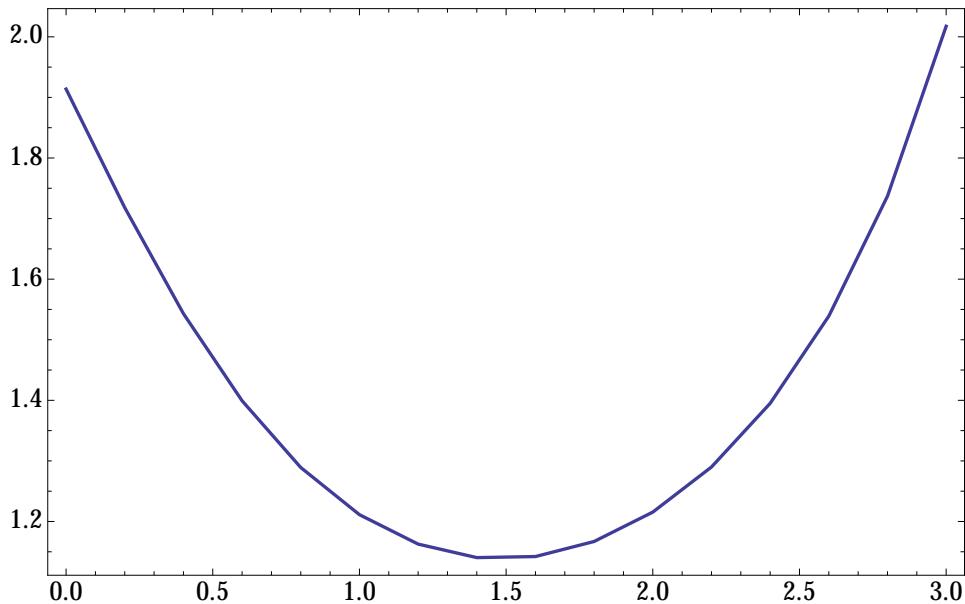


Figure 10.3: Interval length vs. the position of the left endpoint in the example pdf of figure 10.1.

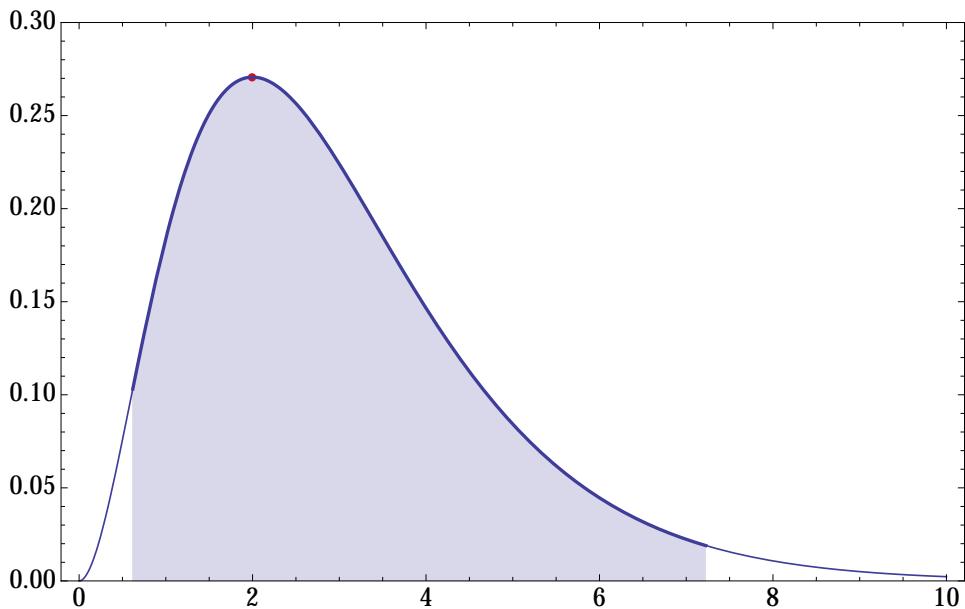


Figure 10.4: Central interval with confidence level  $C = 0.95$ . Each tail corresponds to 2.5% probability. Note that in this case the central interval is *not* symmetric about the estimate. Here the peak value corresponds to  $x = 2$  and the interval endpoints are  $x = 0.619$  and  $x = 7.225$ .

### 10.10.2 Frequentist confidence intervals

The standard frequentist interpretation is that data are distributed according to a well-defined underlying distribution, and that with enough data we should be able to pinpoint the “true” value of a parameter  $\theta$  with increasing precision, according to the weak law of large numbers. This assumes that our model of the distribution of data is correct, that the “true” value is absolutely fixed, and therefore that the pdf  $p_\theta(\theta)$  has no meaning. This view also assumes that the same regularity observed in the past shall hold in the future, so that it shall actually be possible to utilize the weak law of large number, and in this sense it assumes that the underlying theory shall never be falsified.

In this framework we can still use the empirical pdf’s obtained from data to estimate the unknown position of the “true” parameters, specifying the probability that a certain interval “covers” the “true” value  $\theta$ . The classical construction of confidence intervals is originally due to J. Neyman [25].

To understand the Neyman construction, consider first the following example, the sample mean of  $n$  exponentially distributed samples: we know from section 10.2 that the sample mean is also the ML estimator  $\hat{\theta}$ . We assume that the “true” value of the decay time is  $\tau$ , and we also know the pdf of  $\hat{\theta}$  from the results of section 7.2, which is a Gamma distribution

$$p(\hat{\theta}) = \frac{\hat{\theta}^{n-1} e^{-\hat{\theta}/\theta}}{\theta^n \Gamma(n)} \quad (10.108)$$

with  $\theta = \tau/n$ . Then it is possible to construct an interval with a given probability of covering the “true” value of the parameter (see figure 10.5).

The construction shown in figure 10.5 can be repeated for different values of the “true” parameter. Start by choosing a value for  $\theta$  (the “true” parameter): this determines the distribution of the estimator  $\hat{\theta}$ , and we can determine an interval  $(\theta_-, \theta_+)$  such that

$$P(\theta_- \leq \hat{\theta} \leq \theta_+) = C$$

where  $C$  is a constant called *confidence level*, where the endpoints depend in general on the actual value of the parameter,  $\theta_- = \theta_-(\theta)$ ,  $\theta_+ = \theta_+(\theta)$ , and where we fix the interval so that it covers the “true” value  $\theta$ . In this context the interval is itself a random variable, and  $C$  is the probability that it covers the “true” value  $\theta$ . Again, there are many different ways to choose the interval, and we are free to specify the kind of coverage.

The classical construction of Neyman is shown in figure 10.6, where we see a band which is called the *confidence belt*. Using the confidence belt we can construct the *confidence interval* for  $\theta$ , as explained in the caption of figure 10.6.

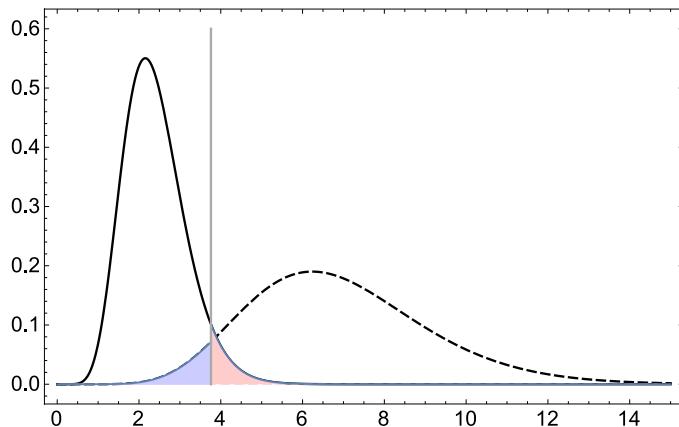


Figure 10.5: Neyman construction for a specific “true” value  $\theta = 0.376$  of the decay time in the example with  $n = 10$  exponentially distributed samples (i.e.,  $\tau = 3.76$ ). The figure shows the gamma pdf’s (10.108) of the estimator  $\hat{\theta}$  for different values of  $\theta$ . Vertical bar: position of the “true” value; solid curve: pdf with the assumption  $\theta = 0.239$  (i.e.,  $\tau = 2.39$ ), 5% probability that the estimator is larger than the “true” value; dashed curve: pdf with the assumption  $\theta = 0.693$  (i.e.,  $\tau = 6.93$ ), 5% probability that the estimator is smaller than the “true” value. Therefore, when we define the interval  $(2.39, 6.93)$  the probability of missing the “true” value – i.e., the “true” value is outside the interval – because it is either too large or too small is 10%. In other words the interval  $(2.39, 6.93)$  has a 90% probability of containing the “true” value. This construction is repeated for all other possible values of the “true” parameter and this produces a band in the  $(\hat{\theta}, \theta)$  plane (see figure 10.6).

The Neyman construction is simplest in the case of a Gaussian density

$$p(\mu_S, \mu) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left[-\frac{(\mu_S - \mu)^2}{2\sigma^2}\right] \quad (10.109)$$

where  $\mu_S$  is a sample mean, and where  $\mathbf{E}\mu_S = \mu$  (this defines the central line of the confidence belt). In this case it is easy to see that the confidence belt is delimited by straight lines at  $45^\circ$ , and has the simple shape shown in figure 10.7.

If the underlying distribution of the estimator is Gaussian, it is common to characterize the confidence interval by means of distance from the mean measured in units of standard deviation. In this case we find the confidence levels listed in table 10.1.

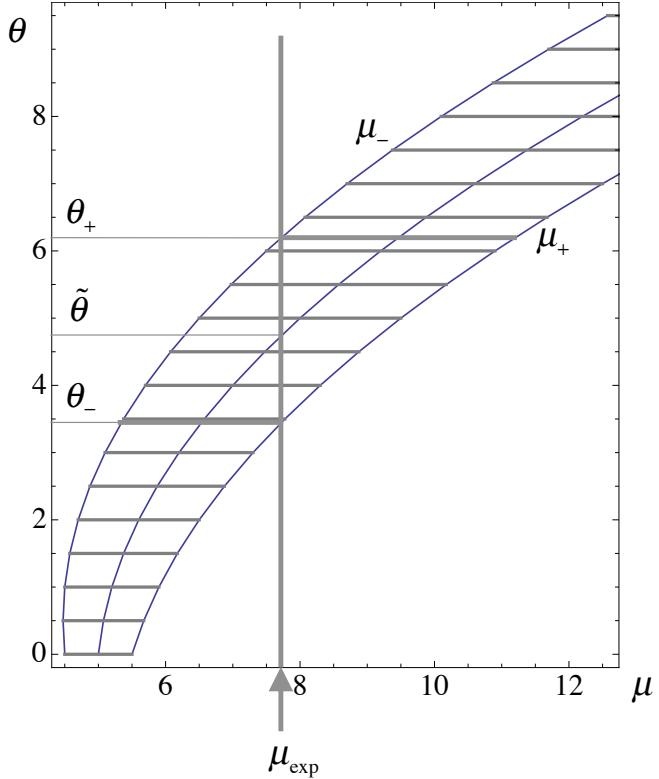


Figure 10.6: Generic confidence belt for a parameter  $\theta$  which is estimated using a measured mean value  $\mu$ , which is related to the parameter by the equation  $\mathbf{E}\mu = f(\theta)$ . The central line represents the function  $\theta = f^{-1}(\mathbf{E}\mu)$ , and the  $\mu_-(\theta)$  and  $\mu_+(\theta)$  curves represent the lower and upper limits of an interval that contains the measured  $\mu_S$  with probability  $C$ . In the Neyman construction, for a given value of  $\theta$ , the interval is itself a random variable (in this case, a two dimensional random variable, defined by the  $(\mu_-, \mu_+)$  pair), and the limits are chosen such that  $\theta \in (\mu_-, \mu_+)$ , and then we say that  $(\mu_-, \mu_+)$  covers the “true” value  $\theta$  with probability  $C$ . Now notice that a measured value  $\mu_{\text{exp}}$  belongs one of these intervals if and only if it belongs to one of the intervals in the range  $\theta_- \leq \theta \leq \theta_+$ , therefore for any  $\theta$  such that  $\theta_- \leq \theta \leq \theta_+$ , the corresponding interval defined by  $\mu_-$  and  $\mu_+$  covers  $\theta$  with probability  $C$ . The limits  $\theta_-$  and  $\theta_+$  define the (frequentist) *confidence interval* of  $\theta$ , and they correspond to the collection of intervals with a given confidence level – i.e., all the intervals of this set have the property that they cover the “true” value with a given probability.

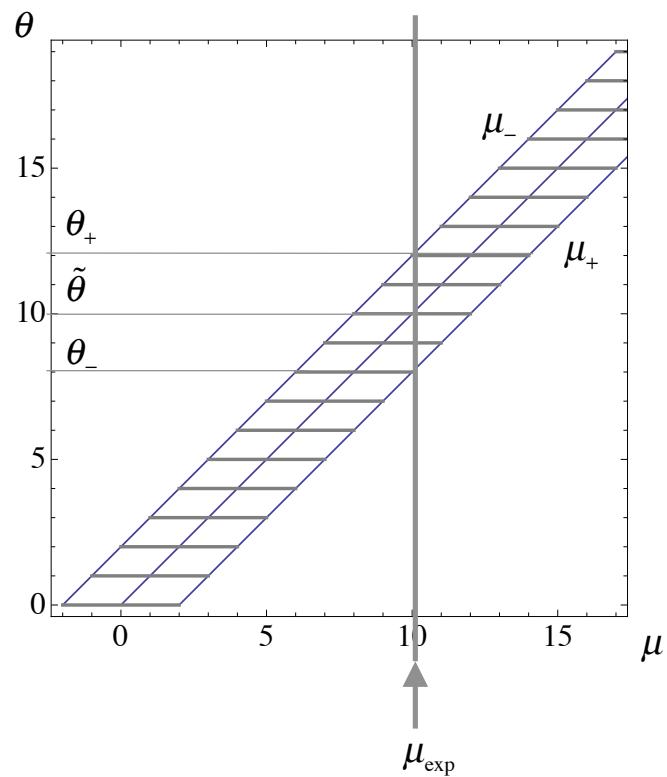


Figure 10.7: Confidence belt for a Gaussian pdf.

Table 10.1: Confidence level for a Gaussian distribution, and intervals at  $\pm k\sigma$  about the mean.

$k$	confidence level
1	0.682689
2	0.954500
3	0.997300
4	0.999937

### 10.10.3 Example: mean of exponentially distributed data

We have seen above that the mean of exponentially distributed data follows a gamma distribution and that it is an unbiased ML estimator of the parameter  $\tau$ . In this case the standard deviation of the sample mean is  $\tau^2/n$ , therefore the confidence belt widens as  $\tau$  increases (for fixed  $n$ ) (see figure 10.8, left panel). Figure 10.8 (right panel) shows the lower and upper values of the central 95% confidence intervals for the sample mean of the exponential distribution, compared with the corresponding intervals obtained with the Gaussian approximation of the gamma distribution. Notice that the Gaussian confidence interval approximates quite well the interval from the gamma distribution, however for small  $n$  the lower bound can extend to negative – and unphysical – values.

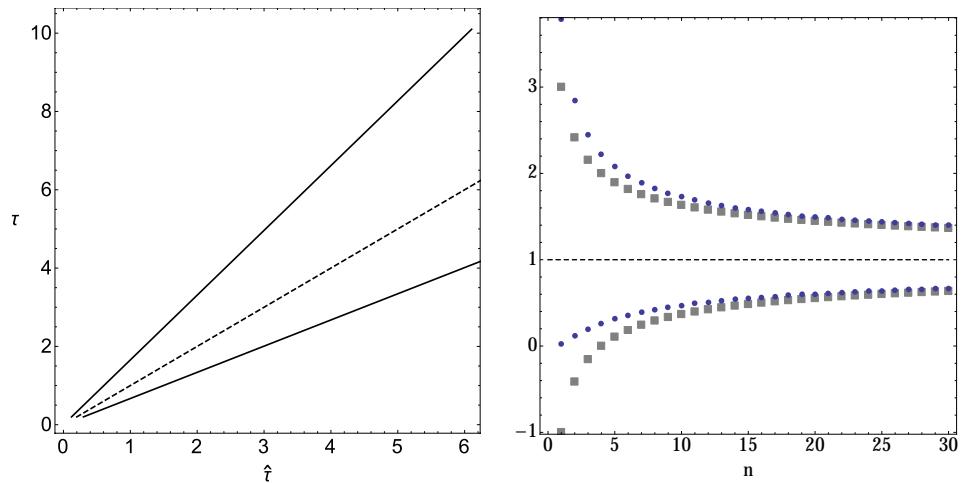


Figure 10.8: Left panel: confidence belt for the mean of exponentially distributed data ( $\tau$  vs.  $\hat{\tau}$ ). The solid lines show the lower and upper limits of the central interval with a confidence level  $C = 0.95$ , while the dashed line shows the position of the “true” mean  $\tau$ . Right panel: plot of lower and upper values of the central 95% confidence intervals for the sample mean of sets of exponentially distributed data vs. the sample size  $n$  (circles). Here  $C = 0.95$ . The corresponding values for the Gaussian approximation ( $2\sigma$  intervals) are also shown (squares).

#### 10.10.4 Confidence intervals for a more complex estimator: the correlation coefficient

Even in the simple case of a bivariate Gaussian distribution, the pdf of the sample estimator of the correlation coefficient – for a given estimate  $r_S$  obtained from data – is quite complex, and definitely non-Gaussian<sup>4</sup>:

$$p(r, r_S) = \frac{1}{\pi} (n-2)(1-r^2)^{(n-4)/2} (1-r_S^2)^{(n-1)/2} \int_0^\infty \frac{d\beta}{(\cosh \beta - r r_S)^{n-1}} \quad (10.110)$$

However in cases such as this (and more complex) we can determine the confidence interval from a MC simulation. Figure 10.9 shows 10000 points generated according to a bivariate Gaussian distribution: the distributions of the  $x$  and  $y$  coordinates are shown in figure 10.10. The correlation coefficient of each data set can be estimated from the formula

$$r_S = \frac{\text{cov}_S(x, y)}{\sigma_{x,S}\sigma_{y,S}} = \frac{\sum_{k=1}^n (x_k - \mu_{x,S})(y_k - \mu_{y,S})}{\sqrt{\sum_{i=1}^n (x_i - \mu_{x,S})^2 \sum_{j=1}^n (y_j - \mu_{y,S})^2}} \quad (10.111)$$

derived in section 7.1, and this is shown in figure 10.11. The MC distribution allows the determination of the limits of the central confidence interval, as shown in figure 10.12.

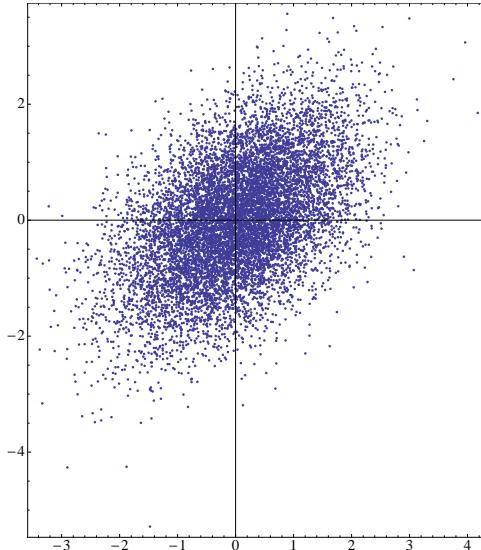


Figure 10.9: 10000 points generated according to a bivariate Gaussian distribution with  $r = 0.5$ .

---

<sup>4</sup>see, e.g., <http://mathworld.wolfram.com/CorrelationCoefficientBivariateNormalDistribution.html>

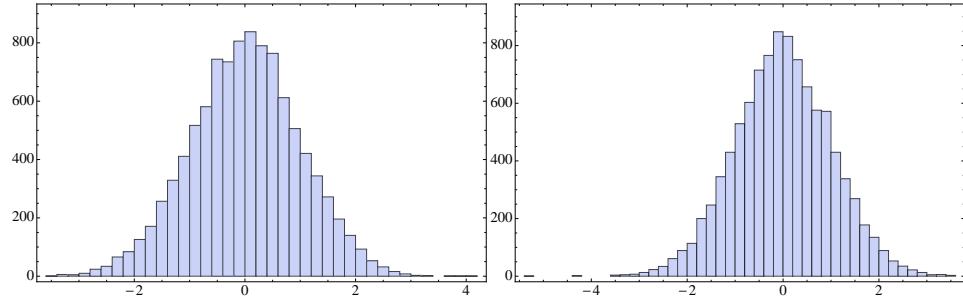


Figure 10.10: Individual histograms of the  $x$  and  $y$  coordinates of the points in figure 10.9.

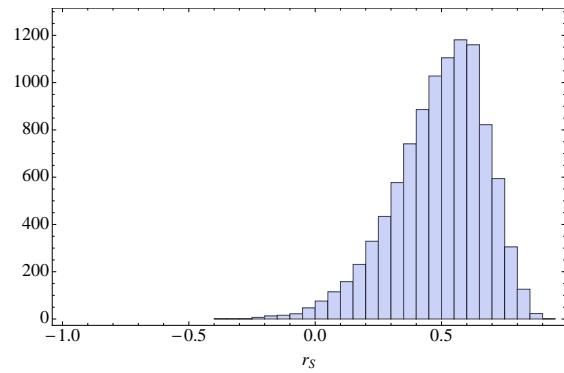


Figure 10.11: Histogram of the sample correlation coefficient for 10000 datasets of 20 samples each.

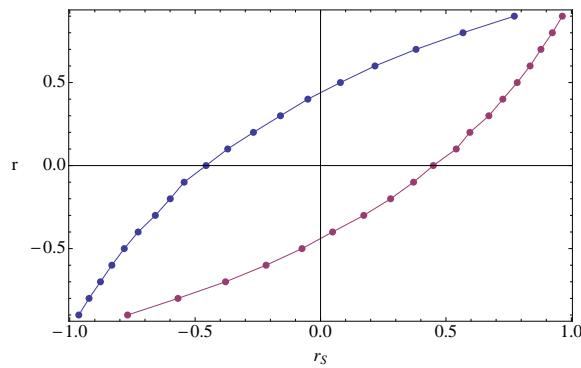


Figure 10.12: 95% confidence belt for the correlation coefficient of 20 samples, from the MC calculation.

### 10.10.5 Graphical method to evaluate estimator variance in the ML method

If we expand the log-likelihood about the maximum, we find

$$\ln L(D, \theta) = \ln L(D, \hat{\theta}) + \frac{1}{2} \sum_{i,j} \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\theta=\hat{\theta}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) + \dots \quad (10.112)$$

In particular, in the case of a single parameter,

$$\ln L(D, \theta) = \ln L(D, \hat{\theta}) + \frac{1}{2} \left. \frac{\partial^2 \ln L(D, \hat{\theta})}{\partial \theta^2} \right|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \dots \quad (10.113)$$

and if we take

$$\sigma_{\hat{\theta}}^2 \approx \left( - \left. \frac{\partial^2 \ln L(D, \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \right)^{-1} \quad (10.114)$$

as an estimate of the variance of the parameter, we see that

$$L(D, \theta) \propto \exp \left[ \frac{1}{2} \left. \frac{\partial^2 \ln L(D, \hat{\theta})}{\partial \theta^2} \right|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 \right] = \exp \left[ - \frac{(\theta - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2} \right] \quad (10.115)$$

and this defines a Gaussian distribution with variance  $\sigma_{\hat{\theta}}^2$ .

Thus, taking  $(\theta - \hat{\theta})^2 = \sigma_{\hat{\theta}}^2$ , we find

$$\ln L = \ln L_{\max} - \frac{1}{2} \quad (10.116)$$

and conversely, if we find the values  $\theta$  that correspond to a decrease of the log-likelihood of 1/2 with respect to its maximum, we define the  $\pm 1 \sigma$  interval about the mean. This -1/2 rule lets us define the confidence interval for the estimator  $\hat{\theta}$ . Although this strictly holds only in the asymptotic (Gaussian) limit, it is commonly used to determine confidence intervals for non-Gaussian likelihoods as well. This can be easily extended to more than one standard deviation: e.g., for a  $2\sigma_{\hat{\theta}}$  interval, the likelihood change is -2.

**Exercise:** Extend these conclusions to the multiparameter case.

### 10.10.6 The Feldman-Cousins construction

The Neyman construction fails in some important cases. For instance, it is often troublesome to determine confidence intervals for non-negative quantities (like masses) that are very small, like the neutrino masses. The Bayesian method does not suffer from these problems, because we include the prior knowledge on non-negativity in the prior distribution itself (we choose a prior

that vanishes for negative values). However frequentists reject the Bayesian construction because they do not accept 1) the perceived arbitrariness in the choice of the prior distribution, 2) the notion that the value to be determined can be regarded as a random variable. On the frequentist side, several methods have been proposed to solve the problems of confidence intervals close to physical boundaries. One of them has gained widespread popularity in the last few years, the Feldman-Cousins method [11]. Basically, it uses the residual freedom for the construction of the confidence intervals to obtain a smooth transition from standard confidence intervals to upper (or lower) confidence limits, thereby also solving the problem of smoothing the transition from two-sided intervals (the usual intervals with two excluded tails) to one-sided intervals (those where one tail only is excluded, and which are used to set upper or lower bounds).

The idea of the Feldman-Cousins construction is to introduce an “ordering principle” in the construction of confidence belt. The order is based on the likelihood ratio of “signal’ + background” vs. “background only”.

## 10.11 Some interesting examples

### 10.11.1 Angular estimate

In the examples and in some proofs on ML estimators we have assumed i.i.d. data, however this is not always the case. Consider for instance two measurement where we independently measure a sine  $s$  and a cosine  $c$ : both measurements are normally distributed with standard deviation  $\sigma$ , so that

$$L(s, c; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(s - \sin \theta)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(c - \cos \theta)^2}{2\sigma^2}\right) \quad (10.117)$$

i.e.,

$$\ln L(s, c; \theta) = -\ln(2\pi\sigma^2) - \frac{(s - \sin \theta)^2}{2\sigma^2} - \frac{(c - \cos \theta)^2}{2\sigma^2} \quad (10.118)$$

and

$$\begin{aligned} \frac{\partial \ln L(s, c; \theta)}{\partial \theta} &= \cos \theta \frac{(s - \sin \theta)}{\sigma^2} - \sin \theta \frac{(c - \cos \theta)}{\sigma^2} \\ &= \frac{1}{\sigma^2} (s \cos \theta - c \sin \theta) = 0 \end{aligned} \quad (10.119)$$

and finally we obtain the ML estimate of  $\tan \theta$ , or  $\theta$ :

$$\tan \hat{\theta} = \frac{s}{c} \quad (10.120)$$

$$\hat{\theta} = \arctan \frac{s}{c} \quad (10.121)$$

Now notice that

$$\mathbf{E} \left( -\frac{\partial^2 \ln L(s, c; \theta)}{\partial \theta^2} \Big|_{\theta_0} \right) = \mathbf{E} \left[ \frac{1}{\sigma^2} (s \sin \theta_0 + c \cos \theta_0) \right] = \frac{1}{\sigma^2} \quad (10.122)$$

so that the estimator  $\hat{\theta}$  has the minimum variance

$$\text{var}(\hat{\theta}) = \sigma_\theta^2 = \sigma^2 \quad (10.123)$$

### 10.11.2 Cauchy distribution

Now consider data that are distributed according to a Cauchy distribution<sup>5</sup>

$$p(x) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - a)^2} \quad (10.124)$$

so that the log-likelihood is

$$\ln L(D; \Gamma, a) = \sum_k \left\{ -\ln 2\pi + \ln \Gamma - \ln[\Gamma^2/4 + (x_k - a)^2] \right\} \quad (10.125)$$

Then we must solve the following equations to maximize the likelihood and estimate  $a$  and  $\Gamma$ :

$$\frac{\partial \ln L(D; \Gamma, a)}{\partial a} = \sum_k \frac{2(x_k - a)}{\Gamma^2/4 + (x_k - a)^2} = 0 \quad (10.126a)$$

$$\frac{\partial \ln L(D; \Gamma, a)}{\partial \Gamma} = \frac{n}{\Gamma} - \sum_k \frac{\Gamma/2}{\Gamma^2/4 + (x_k - a)^2} = 0 \quad (10.126b)$$

In general, these equations can only be solved numerically. Notice, however, that although the ML formalism leads to a rather complex numerical problem, it still gives a definite answer for a problem where the standard descriptive statistics performs badly.

## 10.12 Extended ML

### 10.12.1 A simple example: determination of a forward-backward asymmetry

In this example<sup>6</sup> we want to determine a forward-backward asymmetry, and we need to count the total number of events  $F$  in the forward direction, and

---

<sup>5</sup>this could be the case where we measure the mass of a resonance with a Breit-Wigner mass distribution.

<sup>6</sup>borrowed from L. Lyons, W. W. M. Allison, and J. Pañella Comellas, Nucl. Instr. Meth. **A245** (1986) 530.

$B$  in the backward direction. In such cases the asymmetry is defined by the ratio

$$A = \frac{F - B}{F + B} \quad (10.127)$$

Traditionally, the search for asymmetries has always been associated to “New Physics”, and recently there have been claims of significant asymmetries in  $t\bar{t}$  production from the experiments D0 ad CDF at the Tevatron, the collider at Fermilab<sup>7</sup>: although a small forward-backward asymmetry is predicted by the Standard Model in  $t\bar{t}$  production, both experiment detect much higher asymmetries, although their results are not compatible.

Now let  $f$  be the probability that a specific particle is observed in the forward direction, while  $1 - f$  is the probability that it is observed in the backward direction (in the case of tau production, we expect the  $t$  to be aligned with the incoming quark and the  $\bar{t}$  to be aligned with the incoming antiquark: this defines the “forward” direction). Then the probability of observing  $F$  events in the forward direction, and  $B$  events in the backward direction is described by a binomial model

$$L(F, B; f) = \binom{N}{F} f^F (1-f)^B = \frac{N!}{F! B!} f^F (1-f)^B \quad (10.128)$$

where  $N = F + B$ , so that

$$\ln L(F, B; f) = \ln N! - \ln F! - \ln B! + F \ln f + B \ln(1-f) \quad (10.129)$$

and the maximum likelihood condition is

$$\frac{\partial \ln L(F, B; f)}{\partial f} = \frac{F}{f} - \frac{B}{1-f} = 0 \quad (10.130)$$

Finally we find the ML estimator of  $f$

$$\hat{f} = \frac{F}{N} \quad (10.131)$$

and its variance

$$\sigma_{\hat{f}}^2 = \text{var} \hat{f} \approx \left( -\frac{\partial^2 \ln L(F, B; f)}{\partial f^2} \right)^{-1} = \left( \frac{F}{\hat{f}^2} + \frac{B}{(1-\hat{f})^2} \right)^{-1} = \frac{1}{N} \hat{f}(1-\hat{f}) \quad (10.132)$$

Correspondingly, the estimated rates of forward and backward events are

$$\hat{F} = N \hat{f} = F \quad \hat{B} = N(1 - \hat{f}) = B \quad (10.133)$$

with the variances

$$\text{var} \hat{F} = \text{var} \hat{B} = N^2 \sigma_{\hat{f}}^2 = \frac{FB}{N} \quad (10.134)$$

---

<sup>7</sup>see, e.g., arXiv:hep-ex/1107.4995

Clearly, in this approach  $\hat{F}$  and  $\hat{B}$  are completely correlated; the covariance is

$$\text{cov}(\hat{F}, \hat{B}) \approx \left[ -\left( \frac{\partial^2 \ln L}{\partial F \partial B} \right) \right]^{-1} = \left[ \frac{1}{N^2} \left( \frac{\partial^2 \ln L}{\partial f^2} \right) \right]^{-1} = -N^2 \sigma_{\hat{f}}^2 \quad (10.135)$$

so that the correlation coefficient is

$$r = \frac{\text{cov}(\hat{F}, \hat{B})}{\sqrt{\text{var} \hat{F} \text{ var} \hat{B}}} = -1 \quad (10.136)$$

In the previous discussion we have assumed a fixed total number of events,  $N = F + B$ , however we can let this number fluctuate, and include its probability distribution in the definition of the likelihood. We assume a Poisson distribution for  $N$ , and the new likelihood is

$$L(N, F, B; f, \nu) = \frac{\nu^N e^{-\nu}}{N!} \binom{N}{F} f^F (1-f)^B = \frac{\nu^N e^{-\nu}}{F! B!} f^F (1-f)^B \quad (10.137)$$

This new likelihood where the dataset size is not fixed, but is itself a random variable, is called *Extended Maximum Likelihood* (EML), and it is very similar to the standard likelihood, although there are subtle differences<sup>8</sup>. We find

$$\ln L(N, F, B; f, \nu) = N \ln \nu - \nu + F \ln f + B \ln(1-f) - \ln F! - \ln B! \quad (10.138)$$

and the equations

$$\frac{\partial \ln L}{\partial f} = \frac{F}{f} - \frac{B}{(1-f)} = 0 \quad (10.139a)$$

$$\frac{\partial \ln L}{\partial \nu} = \frac{N}{\nu} - 1 = 0 \quad (10.139b)$$

from which we find the ML estimators

$$\hat{f} = \frac{F}{N}; \quad \hat{\nu} = N \quad (10.140)$$

and the corresponding variances

$$\sigma_{\hat{f}}^2 = \text{var} \hat{f} \approx \left[ -\frac{\partial^2 \ln L}{\partial f^2} \right]^{-1} = \frac{1}{\hat{\nu}} \hat{f} (1 - \hat{f}); \quad \sigma_{\hat{N}}^2 = \text{var} \hat{\nu} \approx \left[ -\frac{\partial^2 \ln L}{\partial \nu^2} \right]^{-1} = \hat{\nu} \quad (10.141)$$

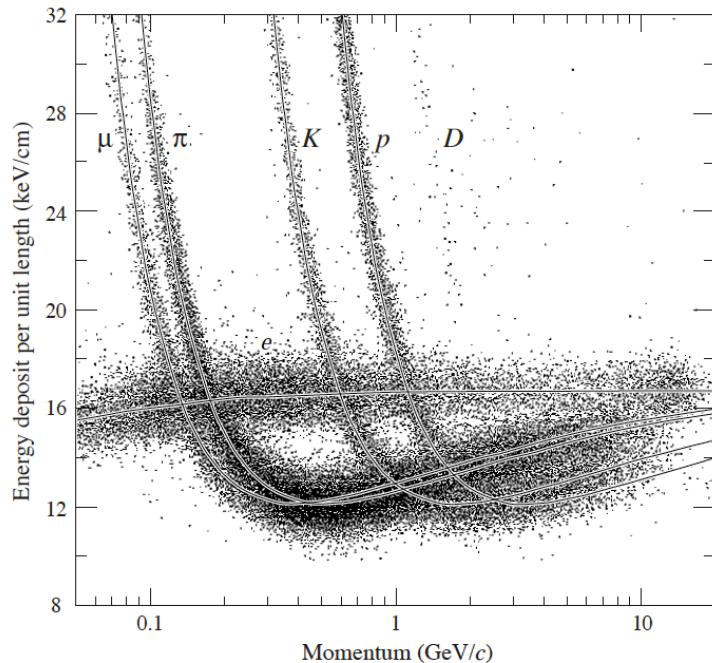
and no correlation.

---

<sup>8</sup>for an extensive discussion of the theory of EML, see R. Barlow, Nucl. Instr. Meth. **A297** (1990) 496

### Example: pion-kaon separation

Here we consider the  $dE/dx$  ionization loss in a detector. This loss has roughly a  $1/\beta^2$  dependence, and thus it is different for particles with the same charge and energy but different mass. This is clearly displayed in figures 10.13 and 10.14, which show the energy losses in two different experiments: we see that a simultaneous measurement of  $dE/dx$  and momentum helps in particle identification.



**Figure 28.15:** The PEP4/9-TPC energy deposit measurements (185 samples, 8.5 atm Ar-CH<sub>4</sub> 80:20). The ionization rate at the Fermi plateau (at high  $\beta$ ) is 1.4 times that for the minimum at lower  $\beta$ . This ratio increases to 1.6 at atmospheric pressure.

Figure 10.13: Energy deposition from  $dE/dx$  in the PEP4 TPC detector (taken from 2010 Review of Particle Physics, K. Nakamura et al. (Particle Data Group), Journal of Physics G **37** (2010) 075021).

The identification is uncertain when the  $dE/dx$  curves are very close, and at a given momentum we obtain  $dE/dx$  histograms like figure 10.15. This can be modeled with a Gaussian mixture model

$$p(x; \boldsymbol{\mu}, \sigma) = \sum_j \frac{f_j}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu_j)^2}{2\sigma_j^2} \right] \quad (10.142)$$

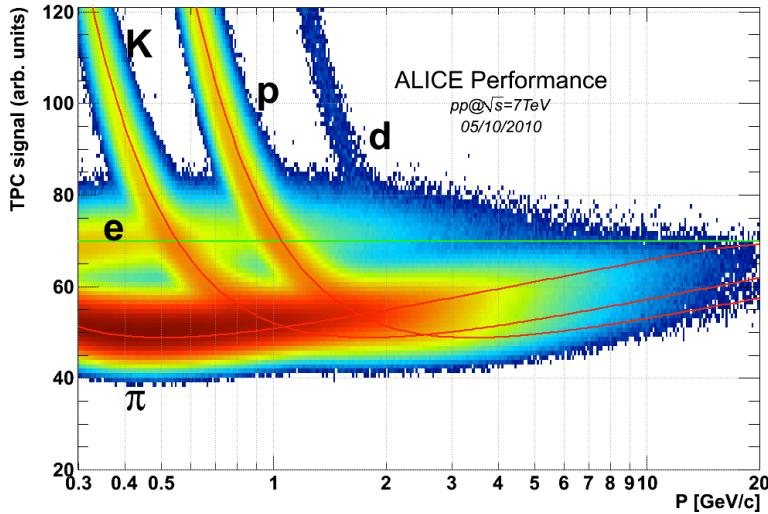


Figure 10.14:  $dE/dx$  vs. particle momentum in the ALICE experiment at the LHC (ALICE Collaboration, Acta Phys. Pol. **B42** (2011) 1401). This is a typical  $dE/dx$  plot, where it is clear that it is possible to identify particles measuring both the  $dE/dx$  energy loss and momentum.

so that the likelihood is

$$L(D; \boldsymbol{\mu}, \sigma) = \prod_k \left\{ \sum_j \frac{f_j}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_k - \mu_j)^2}{2\sigma^2} \right] \right\} \quad (10.143)$$

where  $N$  is the total number of events,  $n_j = N f_j$  is the number of particles of the  $j$ -th species with the constraint  $\sum_j n_j = N$ , and where we assume different mean energy losses  $\mu_j$ , but the same standard deviation  $\sigma$ . Now the log-likelihood is

$$\ln L(D; \boldsymbol{\mu}, \sigma) = \sum_k \ln \left\{ \sum_j \frac{n_j}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_k - \mu_j)^2}{2\sigma^2} \right] \right\} \quad (10.144)$$

Maximizing the log-likelihood of a mixture model is not very practical, and several strategies have been developed to ease this task: one possibility is releasing the constraint and let the total number of counts fluctuate as well: then the extended likelihood is

$$L_E(D; \boldsymbol{\mu}, \sigma) = \frac{\nu^N}{N!} e^{-\nu} \prod_k \left\{ \sum_j \frac{f_j}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_k - \mu_j)^2}{2\sigma^2} \right] \right\} \quad (10.145)$$

In this case computational considerations mix with the actual statistical problem. Although the extended maximum likelihood is a viable alternative,

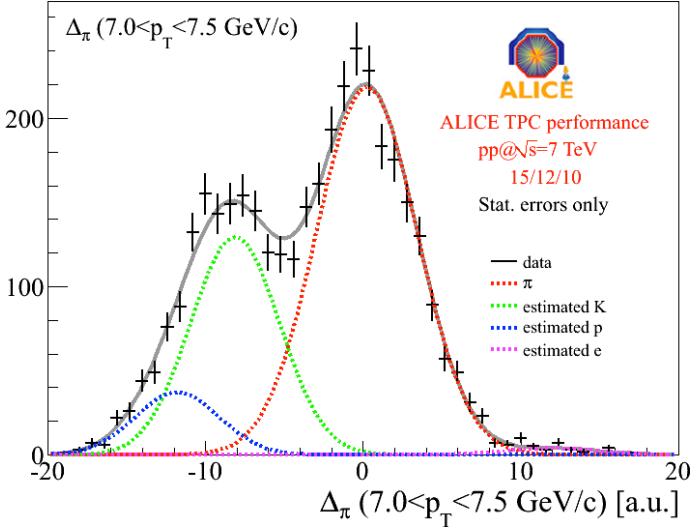


Figure 10.15: Histogram of  $dE/dx$  at fixed particle momentum in the ALICE experiment at the LHC (ALICE Collaboration, Acta Phys. Pol. **B42** (2011) 1401). This is a section of the plot shown in figure 10.14, and the contributions of the different particles have been determined on the basis of a Gaussian distribution of the events about the mean energy loss.

at the moment the method of choice is the Expectation-Maximization (EM) algorithm<sup>9</sup>.

## 10.13 ML with binned data

When there are plenty of data, and likelihood maximization must be performed numerically, the computational load can be very high. In such cases it is more practical to work with binned data. This involves the use of Poisson distributions as in the case of extended ML.

### 10.13.1 Signal enhancement

This is a simple example taken from the course on Modern Methods in Data Analysis held in Heidelberg, which illustrates the use of Poisson statistics in the case of just two channels.

We measure some number of events in two channels, and we compare these results with our MC simulations. We find that the signal is enhanced with respect to the MC simulation (see table), and we wish to find the enhancement factor  $f$ .

---

<sup>9</sup>A. P. Dempster, N. M. Laird, and D. B. Rubin, J. Royal Stat. Soc. **B39** (1977) 1.

Table 10.2: Channel counts

Channel	Measured $n_i$	Expected $n_i$	Signal $s_i$ (from MC)	Background $b_i$ (from MC)
1	6	$1.1 \pm 0.3$	$0.9 \pm 0.3$	$0.2 \pm 0.1$
2	24	$28.0 \pm 6.0$	$4 \pm 0.6$	$24.0 \pm 6.0$

The model is

$$\mu_i = fs_i + b_i \quad (10.146)$$

where  $\mu_i$  is the mean number of counts in channel  $i$ . The corresponding likelihood is

$$L = \frac{\mu_1^{n_1}}{n_1!} e^{-\mu_1} \cdot \frac{\mu_2^{n_2}}{n_2!} e^{-\mu_2} \quad (10.147)$$

and the log-likelihood is

$$\ln L = n_1 \ln \mu_1 - \mu_1 + n_2 \ln \mu_2 - \mu_2 + \text{const.} \quad (10.148)$$

(the constant term includes the factorials that do not depend on  $f$ ). Since

$$\frac{\partial \mu_i}{\partial f} = s_i \quad (10.149)$$

we obtain

$$\frac{\partial \ln L}{\partial f} = n_1 \frac{s_1}{\mu_1} - s_1 + n_2 \frac{s_2}{\mu_2} - s_2 = 0 \quad (10.150)$$

from which we find

$$\hat{f} = 2.62 \begin{array}{l} +1.05 \\ -0.88 \end{array} \quad (10.151)$$

using the values in the table, and the confidence interval determined by the  $-1/2$  rule (see figure 10.16).

### 10.13.2 ML with binned data

Now consider a histogram with  $N$  bins, where we expect to find an average number of  $\nu_i(\theta)$  counts. For instance, in the case of a pdf  $f(x, \theta)$ , and a bin that spans the interval  $(x_{min}^{(i)}, x_{max}^{(i)})$ , and a total of  $n_{tot}$  counts,

$$\nu_i = n_{tot} \int_{x_{min}^{(i)}}^{x_{max}^{(i)}} f(x, \theta) dx \quad (10.152)$$

with the normalization

$$\sum_i \nu_i = n_{tot} \quad (10.153)$$

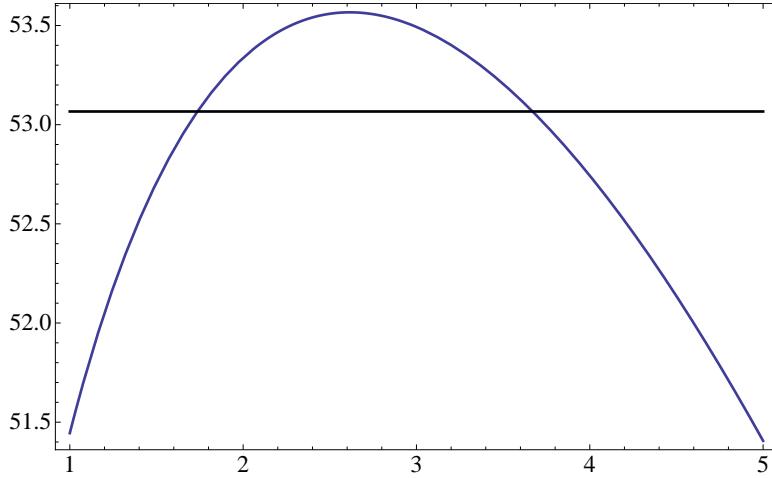


Figure 10.16: Plot of the log-likelihood function  $\ln L = n_1 \ln \mu_1 - \mu_1 + n_2 \ln \mu_2 - \mu_2$  vs.  $f$ . The horizontal line defines the position  $\ln L - 1/2$  for the determination of the confidence interval.

If we consider the whole histogram as a single measurement where data are distributed in individual bins, the joint probability of the histogram is given by the multinomial probability

$$p(D, \boldsymbol{\nu}) = \frac{n_{tot}!}{n_1! \dots n_N!} \left( \frac{\nu_1}{n_{tot}} \right)^{n_1} \dots \left( \frac{\nu_N}{n_{tot}} \right)^{n_N} = L(D, \boldsymbol{\nu}) \quad (10.154)$$

Then the log-likelihood is

$$\ln L(D, \boldsymbol{\nu}) = \sum_i n_i \ln \nu_i + \text{const.} \quad (10.155)$$

Now we assume that the total observed count  $n_{tot}$  is a random variable as well, and it is reasonable to assume that it has a Poisson distribution. Then, the *extended likelihood function* is

$$L(D, \boldsymbol{\nu}) = \frac{\nu_{tot}^{n_{tot}} e^{-\nu_{tot}}}{n_{tot}!} \frac{n_{tot}!}{n_1! \dots n_N!} \left( \frac{\nu_1}{\nu_{tot}} \right)^{n_1} \dots \left( \frac{\nu_N}{\nu_{tot}} \right)^{n_N} = \prod_i \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i} \quad (10.156)$$

where  $\sum_i \nu_i = \nu_{tot}$  and  $\sum_i n_i = n_{tot}$  (note the subtle change of meaning of the symbols), and we see that this likelihood is the product of  $N$  Poisson distributions<sup>10</sup>.

---

<sup>10</sup>Note that this is also the basis for the common recipe that "events have a Poisson distribution in each bin of the experimental histogram".

### 10.13.3 Example: radioactive decay

In this case we consider an exponential pdf

$$p(t, \lambda) = \lambda e^{-\lambda t} \quad (10.157)$$

an initial number  $N$  of radioactive atoms, and a time resolution  $\Delta t$ , that forces us to introduce a binned likelihood (we can only count the number of decays in the time interval  $\Delta t$ , we cannot directly measure the event-to-event time intervals; we also neglect dead time). Now let  $T_j = (2j+1) \Delta t/2$  for  $j = 0, 1, \dots$ , be the central time value of the  $j$ -th bin, so that the average number of events in the  $j$ -th bin is

$$\begin{aligned} \nu_j(\lambda) &= N \int_{T_j - \Delta t/2}^{T_j + \Delta t/2} \lambda e^{-\lambda t} dt = N \left( e^{-\lambda(T_j - \Delta t/2)} - e^{-\lambda(T_j + \Delta t/2)} \right) \\ &= Ne^{-\lambda T_j} \left( e^{\lambda \Delta t/2} - e^{-\lambda \Delta t/2} \right) \\ &\approx N\lambda\Delta t e^{-\lambda T_j} \end{aligned}$$

where the approximated result holds for small  $\Delta t$ .

The corresponding extended likelihood of the binned data is

$$L_E(D, \lambda) = \prod_i \frac{[\nu_i(\lambda)]^{n_i}}{n_i!} e^{-\nu_i(\lambda)} \quad (10.158)$$

and the log-likelihood is

$$\ln L_E(D, \lambda) = \sum_i \{ n_i \ln [\nu_i(\lambda)] - \ln n_i! - \nu_i(\lambda) \} \quad (10.159)$$

and therefore the binned EML estimate of  $\lambda$  is the solution of the equation

$$\frac{\partial \ln L_E(D, \lambda)}{\partial \lambda} = \sum_i \left\{ \frac{n_i}{\nu_i(\lambda)} \frac{\partial \nu_i(\lambda)}{\partial \lambda} - \frac{\partial \nu_i(\lambda)}{\partial \lambda} \right\} = 0 \quad (10.160)$$

Using the expression for  $\nu_j(\lambda)$  we find

$$\frac{\partial \nu_i(\lambda)}{\partial \lambda} = -NT_i e^{-\lambda T_i} \left( e^{\lambda \Delta t/2} - e^{-\lambda \Delta t/2} \right) + \frac{N\Delta t}{2} e^{-\lambda T_i} \left( e^{\lambda \Delta t/2} + e^{-\lambda \Delta t/2} \right) \quad (10.161)$$

and

$$\frac{1}{\nu_i(\lambda)} \frac{\partial \nu_i(\lambda)}{\partial \lambda} = \frac{\partial \ln \nu_i(\lambda)}{\partial \lambda} = -T_i + \frac{\Delta t}{2} \frac{e^{\lambda \Delta t/2} + e^{-\lambda \Delta t/2}}{e^{\lambda \Delta t/2} - e^{-\lambda \Delta t/2}} \quad (10.162)$$

The resulting equation is

$$\begin{aligned}
 & - \sum_i n_i T_i + \frac{N \Delta t}{2} \frac{e^{\lambda \Delta t / 2} + e^{-\lambda \Delta t / 2}}{e^{\lambda \Delta t / 2} - e^{-\lambda \Delta t / 2}} \\
 & = - \frac{N}{\Delta t} \left( \sum_i T_i e^{-\lambda T_i} \Delta t \right) \left( e^{\lambda \Delta t / 2} - e^{-\lambda \Delta t / 2} \right) \\
 & \quad + \frac{N}{2} \left( \sum_i e^{-\lambda T_i} \Delta t \right) \left( e^{\lambda \Delta t / 2} + e^{-\lambda \Delta t / 2} \right) \quad (10.163)
 \end{aligned}$$

If  $\lambda \Delta t \ll 1$ , we can approximate this equation as follows

$$- \sum_i n_i T_i + \frac{N}{\lambda} = 0 \quad (10.164)$$

i.e.,

$$\hat{\lambda}^{-1} = \frac{1}{N} \sum_i n_i T_i \quad (10.165)$$

which is equivalent to the result found in the unbinned case.

#### 10.13.4 Example: estimate of the exponent of a power-law distribution

Consider the following power-law pdf<sup>11</sup>:

$$p(x, \gamma) = \frac{C}{x^\gamma} \quad (10.166)$$

in the range  $(x_{\min}, x_{\max})$  so that the normalization constant can be derived from the condition

$$1 = \int_{x_{\min}}^{x_{\max}} \frac{C}{x^\gamma} dx = \frac{C}{\gamma - 1} (x_{\min}^{-\gamma+1} - x_{\max}^{-\gamma+1}) \quad (10.167)$$

i.e.,

$$C = \frac{\gamma - 1}{x_{\min}^{-\gamma+1} - x_{\max}^{-\gamma+1}} \quad (10.168)$$

The distribution function follows from the integral

$$F(x) = \int_{x_{\min}}^x \frac{C}{s^\gamma} ds = \frac{x_{\min}^{-\gamma+1} - x^{-\gamma+1}}{x_{\min}^{-\gamma+1} - x_{\max}^{-\gamma+1}} \quad (10.169)$$

---

<sup>11</sup>power-law pdf's are quite common in many fields of physics; for a recent comparison of different methods of estimation of the exponent, see H. Bauke, Eur. Phys. J. B58 (2007) 167.

and we can generate random numbers distributed according to this distribution with the transformation method, solving the equation

$$F(x) = \frac{x_{\min}^{-\gamma+1} - x^{-\gamma+1}}{x_{\min}^{-\gamma+1} - x_{\max}^{-\gamma+1}} = r \quad (10.170)$$

where  $r$  is a uniform variate on  $(0, 1)$ , i.e.,

$$x = \left[ x_{\min}^{-\gamma+1} - r(x_{\min}^{-\gamma+1} - x_{\max}^{-\gamma+1}) \right]^{1/(1-\gamma)} \quad (10.171)$$

We can use the transformation method to set up a MC procedure, and produce a simulated dataset, as shown in figure 10.17. In this case we can apply the ML formalism to the simulated data to retrieve the value of the exponent  $\gamma$ , assuming that the range  $(x_{\min}, x_{\max})$  is known beforehand.

Whether data are simulated or not, the likelihood is

$$L(D, \gamma) = \prod_k \frac{C(\gamma)}{x_k^\gamma} \quad (10.172)$$

and therefore the log-likelihood is

$$\ln L(D, \gamma) = \sum_k (\ln C(\gamma) - \gamma \ln x_k) \quad (10.173)$$

and finally we have to solve the equation

$$\frac{\partial L(D, \gamma)}{\partial \gamma} = N \frac{\partial \ln C}{\partial \gamma} - \sum_k \ln x_k \quad (10.174)$$

where  $N$  is the size of the data set. Since

$$\ln C = \ln(\gamma - 1) - \ln[e^{(1-\gamma) \ln x_{\min}} - e^{(1-\gamma) \ln x_{\max}}] \quad (10.175)$$

we find

$$\frac{\partial \ln C}{\partial \gamma} = \frac{1}{(\gamma - 1)} + \frac{\ln x_{\min} x_{\min}^{(1-\gamma)} - \ln x_{\max} x_{\max}^{(1-\gamma)}}{x_{\min}^{(1-\gamma)} - x_{\max}^{(1-\gamma)}} \quad (10.176)$$

and the ML estimate of  $\gamma$  is the solution of

$$\frac{1}{n} \frac{\partial L(D, \gamma)}{\partial \gamma} = \frac{1}{(\gamma - 1)} + \frac{\ln x_{\min} x_{\min}^{(1-\gamma)} - \ln x_{\max} x_{\max}^{(1-\gamma)}}{x_{\min}^{(1-\gamma)} - x_{\max}^{(1-\gamma)}} - \frac{1}{N} \sum_k \ln x_k = 0 \quad (10.177)$$

This is a nonlinear equation and we must resort to a numerical method of solution, like the Newton-Raphson algorithm: it is actually a rather simple

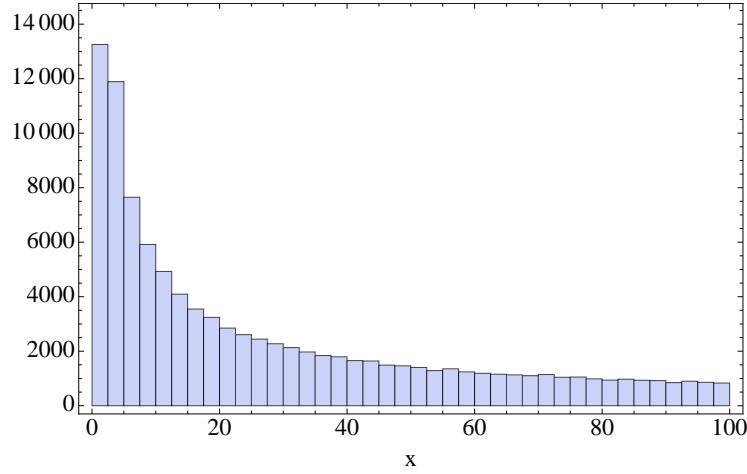


Figure 10.17: Histogram of 100000 power-law distributed numbers from a computer simulation with  $\gamma = 0.8$ ,  $x_{\min} = 1$ , and  $x_{\max} = 100$ .

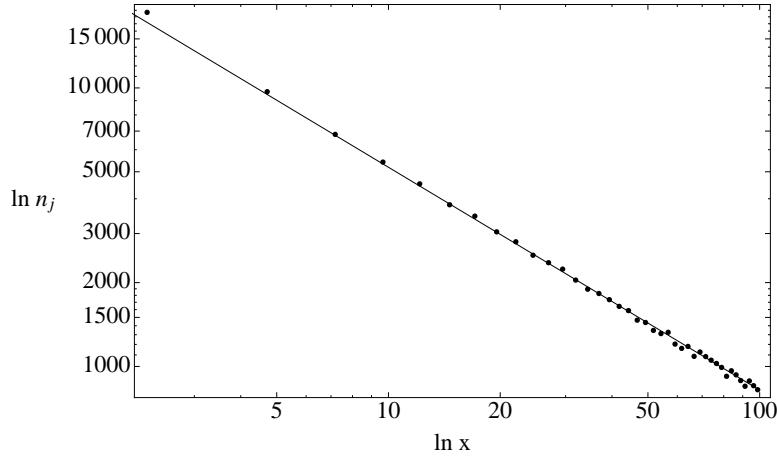


Figure 10.18: Log-log plot of the power-law distributed data of figure 10.17. The solid line is a plot of the approximate mean count  $\nu_j(\gamma) \approx C(\gamma)X_j^{-\gamma}\Delta x$ .

problem, because the sample mean of the logarithms of our data must be evaluated only once. However, in actual practice, we might not know the exact values of  $x_k$  because of the limited resolution of the experimental apparatus. Then the actual data would be the number of counts in each bin of a histogram, like figure 10.17.

Now let  $X_j$  be the central value of the  $j$ -th bin, and  $\Delta x$  be the bin width,

then the average number of events in the  $j$ -th bin is

$$\nu_j(\gamma) = N \int_{X_j - \Delta x/2}^{X_j + \Delta x/2} \frac{C(\gamma)}{x^\gamma} dx = \frac{C(\gamma)}{\gamma - 1} [(X_j - \Delta x/2)^{-\gamma+1} - (X_j + \Delta x/2)^{-\gamma+1}] \quad (10.178)$$

If  $\Delta x \ll x_{\min}$ , we can approximate this expression as follows

$$\begin{aligned} \nu_j(\gamma) &\approx \frac{C(\gamma)}{\gamma - 1} \left\{ \left[ X_j^{-\gamma+1} + (\gamma - 1) X_j^{-\gamma} \Delta x / 2 \right] - \left[ X_j^{-\gamma+1} - (\gamma - 1) X_j^{-\gamma} \Delta x / 2 \right] \right\} \\ &= C(\gamma) X_j^{-\gamma} \Delta x \end{aligned} \quad (10.179)$$

The corresponding extended likelihood of the binned data is

$$L_E(D, \gamma) = \prod_i \frac{[\nu_i(\gamma)]^{n_i}}{n_i!} e^{-\nu_i(\gamma)} \quad (10.180)$$

and the log-likelihood is

$$\ln L_E(D, \gamma) = \sum_i \{ n_i \ln [\nu_i(\gamma)] - \ln n_i! - \nu_i(\gamma) \} \quad (10.181)$$

and therefore the EML estimate of  $\gamma$  is the solution of the equation

$$\frac{\partial \ln L_E(D, \gamma)}{\partial \gamma} = \sum_i \left\{ \frac{n_i}{\nu_i(\gamma)} \frac{\partial \nu_i(\gamma)}{\partial \gamma} - \frac{\partial \nu_i(\gamma)}{\partial \gamma} \right\} = 0 \quad (10.182)$$

i.e.,

$$\sum_i \frac{n_i}{\nu_i(\gamma)} \frac{\partial \nu_i(\gamma)}{\partial \gamma} = \sum_i \frac{\partial \nu_i(\gamma)}{\partial \gamma} \quad (10.183)$$

Using the approximate expression given above, we find

$$\frac{\partial \nu_i(\gamma)}{\partial \gamma} \approx \frac{\partial C(\gamma)}{\partial \gamma} X_j^{-\gamma} \Delta x - \gamma C(\gamma) X_j^{-\gamma-1} \Delta x \quad (10.184)$$

and therefore

$$\sum_i \frac{n_i}{X_i} \left[ \frac{\partial \ln C}{\partial \gamma} - \gamma X_i \right] = \sum_i \left[ \frac{\partial C(\gamma)}{\partial \gamma} X_j^{-\gamma} \Delta x - \gamma C(\gamma) X_j^{-\gamma-1} \Delta x \right] \quad (10.185)$$

or also

$$\frac{\partial \ln C}{\partial \gamma} \sum_i \frac{n_i}{X_i} - N\gamma = C(\gamma) \Delta x \left[ \frac{\partial \ln C(\gamma)}{\partial \gamma} \sum_i X_j^{-\gamma} - \gamma \sum_i X_j^{-\gamma-1} \right] \quad (10.186)$$

Thus, we have to solve a new nonlinear equations, and we resort again to the Newton-Raphson algorithm (or another algorithm for the solution of nonlinear equations).

**Exercise:** now we take  $M$  intervals  $\Delta x_j$  such that  $\Delta x_j = \alpha \Delta x_{j-1}$ , where the constant  $\alpha > 1$ , with the condition  $\sum \Delta x_j = x_{\max} - x_{\min}$ , i.e.,

$$\Delta x_0 \frac{1 - \alpha^M}{1 - \alpha} = x_{\max} - x_{\min} \quad (10.187)$$

and finally

$$\Delta x_0 = \frac{\alpha - 1}{\alpha^M - 1} (x_{\max} - x_{\min}) \quad (10.188)$$

(this fixes  $\Delta x_0$ , given  $\alpha$  and the  $(x_{\min}, x_{\max})$  range), and

$$\Delta x_j = \alpha^j \frac{\alpha - 1}{\alpha^M - 1} (x_{\max} - x_{\min}) \quad (10.189)$$

How does the binned likelihood change<sup>12</sup>?

## 10.14 What's the use of all this?

The maximum likelihood method is central to frequentist statistics and it finds wide ranging applications in many problems. The method is also the theoretical foundation on which other methods are built, as we shall see in the next chapter.

---

<sup>12</sup>since  $\ln \Delta x_j - \ln \Delta x_{j-1} = \ln \alpha = \text{constant}$  this is also called *logarithmic binning*.

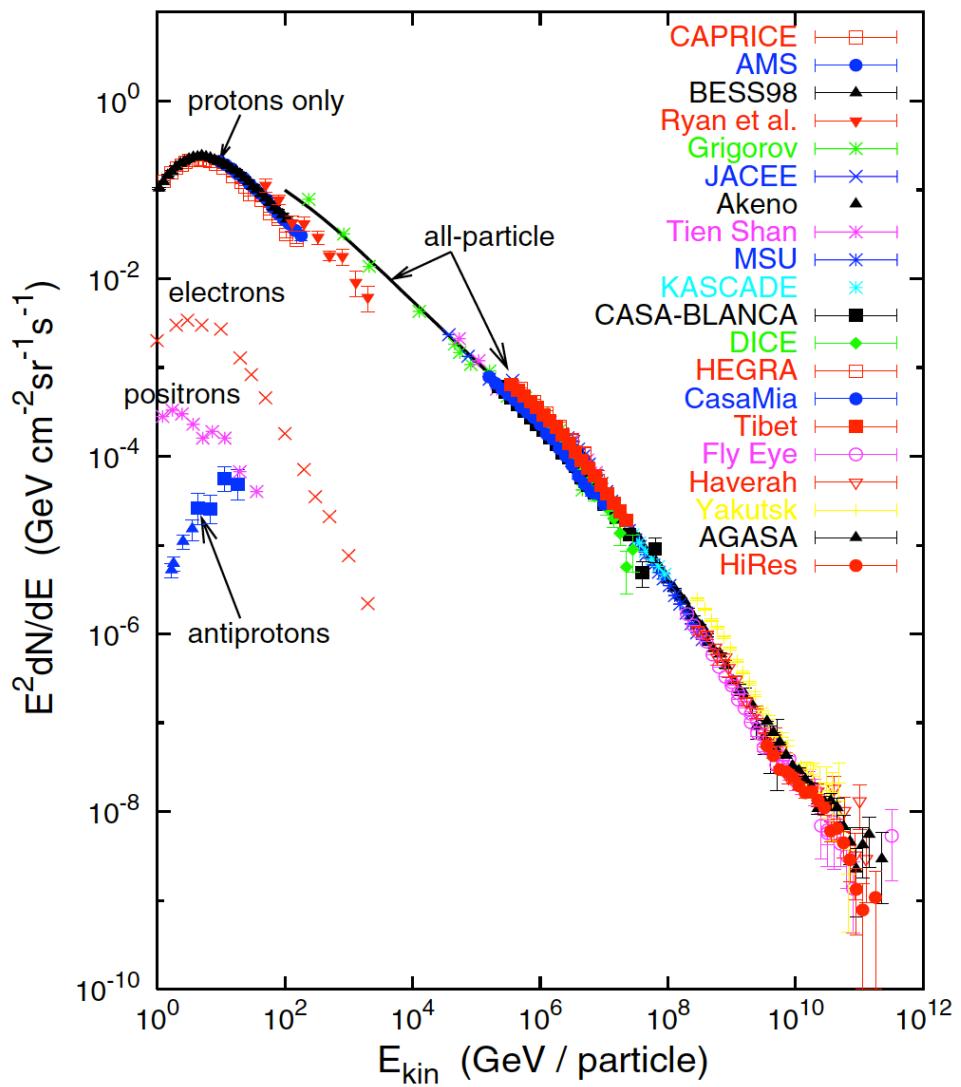


Figure 10.19: The energy spectrum of cosmic rays follows a near-perfect power-law at high energy (from B. Baret and V. Van Elewyck, Rep. Prog. Phys. **74** (2011) 046902).



# Chapter 11

## Fitting data

### 11.1 Parametric models and normally distributed data: the method of least squares

In the case of normally distributed scalar data that fluctuate about a mean value and that depend on some parametric model  $\mu_k = \mu_k(\theta)$ , and where the individual precision  $\sigma_k$  of each data point is known, we have

$$p(d_k|\theta) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{[d_k - \mu_k(\theta)]^2}{2\sigma_k^2}\right\} \quad (11.1)$$

therefore we find

$$\ln L(D, \theta) = \sum_k \left\{ -\frac{\ln(2\pi\sigma_k^2)}{2} - \frac{[d_k - \mu_k(\theta)]^2}{2\sigma_k^2} \right\} \quad (11.2)$$

The first term does not depend on the parameters  $\theta$ , therefore maximizing the likelihood is equivalent to minimizing the *chi-square*  $\chi^2$ :

$$\chi^2 = \sum_k \frac{[d_k - \mu_k(\theta)]^2}{\sigma_k^2} \quad (11.3)$$

Since minimizing the chi-square means minimizing a sum of squares, this approach is called the *method of least squares*.

#### 11.1.1 Correlated data

Until now we have always assumed independent measurements, however here we see that the chi-square can easily be extended to correlated, normally distributed data. If  $V$  is the (known) correlation matrix, then the joint

distribution of data is

$$p(D; \theta) = \frac{1}{(2\pi)^n |V|^{n/2}} \exp \left[ -\frac{1}{2} \sum_{i,j} (d_i - \mu_i(\theta))(V^{-1})_{ij}(d_j - \mu_j(\theta)) \right] = L(D, \theta) \quad (11.4)$$

therefore the log-likelihood is

$$\ln L(D, \theta) = -\frac{1}{2} \sum_{i,j} (d_i - \mu_i(\theta))(V^{-1})_{ij}(d_j - \mu_j(\theta)) + \text{const.} \quad (11.5)$$

and

$$\chi^2 = -2 \ln L(D, \theta) = \sum_{i,j} (d_i - \mu_i(\theta))(V^{-1})_{ij}(d_j - \mu_j(\theta)) + \text{const.} \quad (11.6)$$

### 11.1.2 Expansion about the minimum

When we expand the log-likelihood about the estimator  $\hat{\theta}$ , we find

$$\ln L(D, \theta) = \ln L(D, \hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^T U^{-1}(\theta - \hat{\theta}) \quad (11.7)$$

and from the asymptotic normality of the likelihood we know that  $U$  is the covariance matrix of the parameters

$$U_{ij} = \text{cov}(\theta_i, \theta_j) \quad (11.8)$$

In the same way, we can expand the  $\chi^2$  about the minimum

$$\chi^2(\theta) = \chi^2(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H(\theta - \hat{\theta}) \quad (11.9)$$

where

$$H_{ij} = \left. \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right|_{\hat{\theta}} \quad (11.10)$$

is the Hessian matrix. Therefore

$$U_{ij}^{-1} = \left. \frac{\partial^2 \ln L(D, \theta)}{\partial \theta_i \partial \theta_j} \right|_{\hat{\theta}} = -\frac{1}{2} \left. \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right|_{\hat{\theta}} = -\frac{1}{2} H_{ij} \quad (11.11)$$

### 11.1.3 Criticism of the ML approach to least squares

We have derived the principle of chi-square minimization from the maximization likelihood, however it is found that minimizing the chi-square yields satisfactory results even in the case of non-Gaussian data. For this reason

the derivation above has been criticized, and the efficacy of chi-square minimization has been considered itself a sufficient justification for the method. For example Yule and Kendall write:

“It was formerly the custom, and is still so in works on the theory of observations, to derive the method of least squares from certain theoretical considerations, the assumed normality of the errors of observation being one such. It is, however, more than doubtful whether the conditions for the theoretical validity of the method are realised in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has stood the test of experience”<sup>1</sup>.

## 11.2 Chi-square and confidence intervals

When we are close enough to the peak of the likelihood function, we find that

$$\ln p(\theta|D) = \ln L(D, \theta) = \ln L(D, \hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^T \left( \frac{H}{2} \right) (\theta - \hat{\theta}) \quad (11.12)$$

and because of the relationship

$$\chi^2 = -2 \ln L(D, \theta) \quad (11.13)$$

the  $-\Delta(1/2)$  rule for the likelihood becomes a  $+\Delta 1$  rule for the chi-square. The quadratic expansion of the chi-square about the minimum defines a paraboloid, and from this a Gaussian likelihood. This means that we can compute confidence regions for these multivariate Gaussians. However we must be aware that the probability content is different. For instance for a binormal (the contour plot of an example pdf like of this kind is shown in figure 11.1) we find the probabilities in table 11.1.

## 11.3 The $\chi^2$ distribution

The  $\chi^2$  is the sum of the squares of many random variables

$$\frac{d_k - \mu_k(\theta)}{\sigma_k} \quad (11.14)$$

with Gaussian pdf  $N(0, 1)$ : here we find the pdf of the  $\chi^2$  random variable. If a single random variable  $x$  has pdf  $p_x(x) = N(0, 1)$ , then  $y = x^2$  has pdf

$$p_y(y) = 2p_x(x(y)) \left| \frac{dx}{dy} \right| = 2 \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} \quad (11.15)$$

---

<sup>1</sup>G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, 14th edition (1958), as quoted in R. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley (1989)

Table 11.1: Cumulative probabilities for the regions defined by  $r = \sqrt{x^2 + y^2} < k\sigma$  for the binormal distribution  $p(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right)$ .

$k$	$P(r < k\sigma)$
0.5	0.117503
1.	0.393469
1.5	0.675348
2.	0.864665
2.5	0.956063
3.	0.988891
3.5	0.997813
4.	0.999665
4.5	0.99996
5.	0.999996

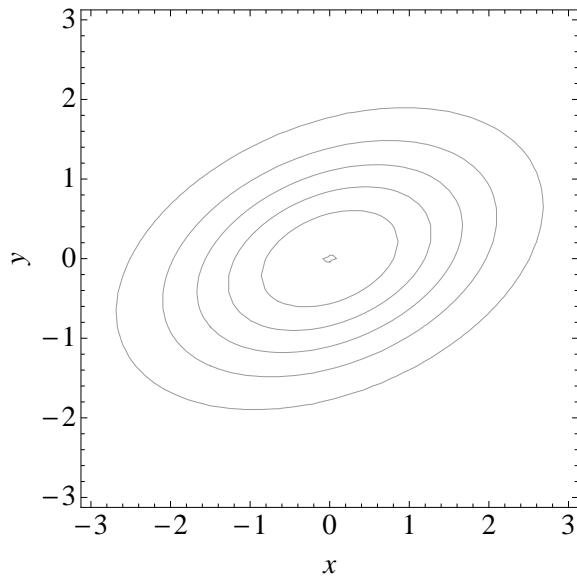


Figure 11.1: Contour plot of the pdf of a binormal distribution; this is an instance of the behavior of a two-parameter likelihood function close to the peak. The regions delimited by the lines of constant likelihood are used to define two-dimensional confidence regions (the confidence regions are multidimensional generalizations of the confidence intervals).

for  $y \in (0, \infty)$ . The characteristic function of this pdf is

$$f(t) = \int_0^{+\infty} p_y(y) e^{ity} dy = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} e^{ity} dy \quad (11.16)$$

and using the transformation  $y = x^2$  the integral becomes

$$\begin{aligned} f(t) &= \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} e^{-x^2/2} e^{itx^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-x^2 \frac{(1-2it)}{2}\right) dx = \frac{1}{(1-2it)^{1/2}} \end{aligned} \quad (11.17)$$

Since the cf of a sum of independent variates is the product of the individual cf's, we find the cf of the  $\chi^2$  pdf with  $n$  terms:

$$f_{\chi^2}(t) = \frac{1}{(1-2it)^{n/2}} \quad (11.18)$$

The Fourier inversion of this cf yields the  $\chi^2$  pdf:

$$p_{\chi^2}(\chi^2) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{1}{(1-2it)^{n/2}} e^{-it\chi^2} dt \quad (11.19)$$

The integrand has a pole at  $t = -i/2$  (i.e., in the lower half-plane), and to simplify matters, we assume that  $n = 2m$ , so that the residual of the integrand is

$$\text{Res} = \frac{1}{(m-1)!} \frac{d^{m-1}}{dt^{m-1}} \left( t + \frac{i}{2} \right)^m \left. \frac{e^{-it\chi^2}}{(1-2it)^m} \right|_{t=-i/2} = \frac{(-i\chi^2)^{m-1}}{(-2i)^m \Gamma(m)} e^{-\chi^2/2} \quad (11.20)$$

and

$$\begin{aligned} p_{\chi^2}(\chi^2) &= -\frac{1}{2\pi} 2\pi i \frac{(-i\chi^2)^{m-1}}{(-2i)^m \Gamma(m)} e^{-\chi^2/2} = \frac{(\chi^2)^{m-1}}{2^m \Gamma(m)} e^{-\chi^2/2} \\ &= \frac{(\chi^2)^{n/2-1}}{2^{n/2} \Gamma(n/2)} e^{-\chi^2/2} \end{aligned} \quad (11.21)$$

We can reach the same result by a different route. Consider  $n$  uncorrelated Gaussian variates  $x_i$  with pdf  $N(0, 1)$ , then the vector  $D = (x_1, \dots, x_n)$  has a multivariate Gaussian pdf

$$p_D(D) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2}\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\chi^2}{2}\right) \quad (11.22)$$

Since the probability of finding  $D$  in a shell with radius  $\chi$  and thickness  $d\chi$  is equal to

$$p_\chi(\chi)d\chi = \int \dots \int_{\chi \leq (x_1^2 + \dots + x_n^2)^{1/2} \leq \chi + d\chi} p_D(D) dx_1 \dots dx_n \quad (11.23a)$$

$$= p_D(D(\chi)) \int \dots \int_{\chi \leq (x_1^2 + \dots + x_n^2)^{1/2} \leq \chi + d\chi} dx_1 \dots dx_n \quad (11.23b)$$

$$= \frac{2\pi^{n/2}}{\Gamma(n/2)} \chi^{n-1} p_D(D(\chi)) d\chi \quad (11.23c)$$

$$= \frac{2\pi^{n/2}}{\Gamma(n/2)} \chi^{n-1} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\chi^2}{2}\right) d\chi \quad (11.23d)$$

$$= \frac{1}{2^{n/2-1} \Gamma(n/2)} \chi^{n-1} \exp\left(-\frac{\chi^2}{2}\right) d\chi \quad (11.23e)$$

(where we use the expression for the surface of the  $n$ -dimensional hyperspherical shell derived in the appendix) and from the transformation of variables which relates  $\chi$  to  $\chi^2$  we find

$$p_{\chi^2}(\chi^2) = p_\chi(\chi) \frac{1}{2\chi} = \frac{(\chi^2)^{n/2-1}}{2^{n/2} \Gamma(n/2)} \exp\left(-\frac{\chi^2}{2}\right) \quad (11.24)$$

which coincides with the previous result.

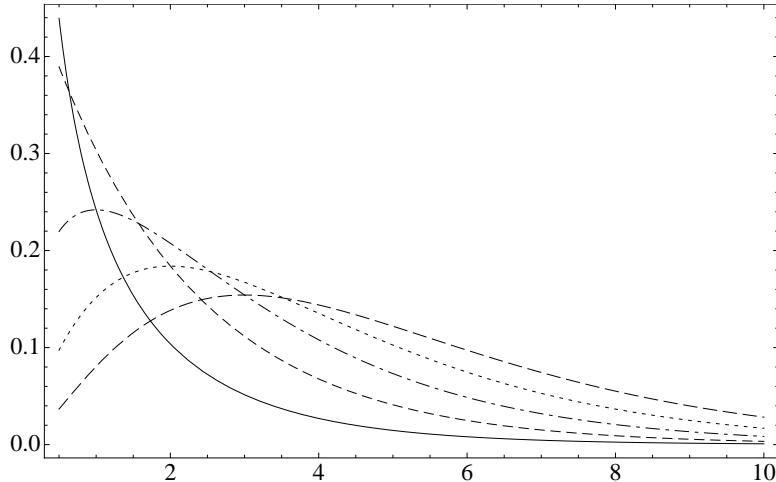


Figure 11.2: Plot of the chi-square pdf ( $p_{\chi^2}(\chi^2)$  vs.  $\chi^2$ ) for  $n = 1$  (solid line);  $n = 2$  (dashed line);  $n = 3$  (dashed-dotted line);  $n = 4$  (dotted line);  $n = 5$  (long dashes).

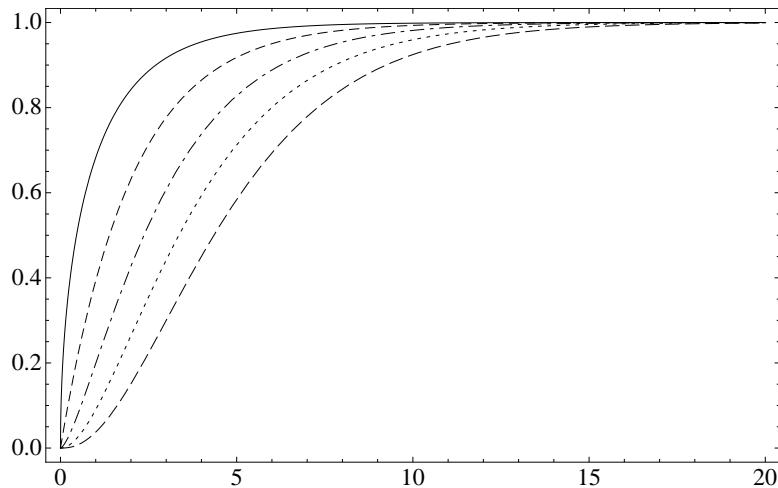


Figure 11.3: Plot of the chi-square cdf ( $\int_0^{\chi^2} p_{\chi^2}(x)dx$  vs.  $\chi^2$ ) for  $n = 1$  (solid line);  $n = 2$  (dashed line);  $n = 3$  (dashed-dotted line);  $n = 4$  (dotted line);  $n = 5$  (long dashes).

The chi-square distribution is a special case of the gamma distribution

$$p(x) = \frac{x^{\alpha-1} e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)} \quad (11.25)$$

with shape parameter  $\alpha = n/2$  and scale parameter  $\theta = 2$ , therefore the mean value is

$$\alpha\theta = n \quad (11.26)$$

and the variance is

$$\alpha\theta^2 = 2n \quad (11.27)$$

Figure 11.2 displays the chi-square pdf for  $n = 1, \dots, 5$ , and figure 11.3 shows the corresponding cdf's. The parameter  $n$  is also called the *number of degrees of freedom* of the distribution: each degree of freedom gives a unit contribution to the mean value of  $\chi^2$ . We shall soon see that this distribution is quite important in defining the goodness of fits to data.

## 11.4 Straight line fit

Now we go back to the formula of the  $\chi^2$  for uncorrelated data

$$\chi^2 = \sum_k \frac{[y_k - \mu_k(\theta)]^2}{\sigma_k^2} \quad (11.28)$$

and assume that scalar data  $y_k$  are associated to a scalar, *exact*, input variable  $x_k$ , and that they can be described by a linear model

$$\mu_k(a, b) = ax_k + b \quad (11.29)$$

i.e.,

$$\chi^2 = \sum_k \frac{[y_k - (ax_k + b)]^2}{\sigma_k^2} \quad (11.30)$$

Then, to minimize the  $\chi^2$  we must solve the system

$$\frac{\partial \chi^2}{\partial a} \Big|_{\hat{a}, \hat{b}} = -2 \sum_k \frac{x_k[y_k - (\hat{a}x_k + \hat{b})]}{\sigma_k^2} = 0 \quad (11.31a)$$

$$\frac{\partial \chi^2}{\partial b} \Big|_{\hat{a}, \hat{b}} = -2 \sum_k \frac{[y_k - (\hat{a}x_k + \hat{b})]}{\sigma_k^2} = 0 \quad (11.31b)$$

i.e.,

$$\hat{a} \sum_k \frac{x_k^2}{\sigma_k^2} + \hat{b} \sum_k \frac{x_k}{\sigma_k^2} = \sum_k \frac{x_k y_k}{\sigma_k^2} \quad (11.32a)$$

$$\hat{a} \sum_k \frac{x_k}{\sigma_k^2} + \hat{b} \sum_k \frac{1}{\sigma_k^2} = \sum_k \frac{y_k}{\sigma_k^2} \quad (11.32b)$$

This can be recast in matrix form, setting

$$S_0 = \sum_k \frac{1}{\sigma_k^2}; \quad S_x = \sum_k \frac{x_k}{\sigma_k^2}; \quad S_y = \sum_k \frac{y_k}{\sigma_k^2}; \quad S_{xx} = \sum_k \frac{x_k^2}{\sigma_k^2}; \quad S_{xy} = \sum_k \frac{x_k y_k}{\sigma_k^2};$$

then

$$\begin{pmatrix} S_{xx} & S_x \\ S_x & S_0 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix} \quad (11.33)$$

and

$$\begin{aligned} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} &= \begin{pmatrix} S_{xx} & S_x \\ S_x & S_0 \end{pmatrix}^{-1} \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix} \\ &= \frac{1}{S_{xx}S_0 - S_x^2} \begin{pmatrix} S_0 & -S_x \\ -S_x & S_{xx} \end{pmatrix} \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix} = \begin{pmatrix} \frac{S_0S_{xy} - S_xS_y}{S_{xx}S_0 - S_x^2} \\ \frac{S_{xx}S_y - S_xS_{xy}}{S_{xx}S_0 - S_x^2} \end{pmatrix} \end{aligned} \quad (11.34)$$

The errors on the parameter estimates can be computed with error propagation. Since

$$\frac{\partial S_y}{\partial y_j} = \frac{1}{\sigma_j^2}; \quad \frac{\partial S_{xy}}{\partial y_j} = \frac{x_j}{\sigma_j^2}; \quad (11.35)$$

$$\sigma_{\hat{a}}^2 = \sum_j \left( \frac{S_0 x_j / \sigma_j^2 - S_x / \sigma_j^2}{S_{xx} S_0 - S_x^2} \right)^2 \sigma_j^2 = \frac{S_0^2 S_{xx} - S_0 S_x^2}{(S_{xx} S_0 - S_x^2)^2} = \frac{S_0}{S_{xx} S_0 - S_x^2} \quad (11.36a)$$

$$\sigma_{\hat{b}}^2 = \sum_j \left( \frac{S_{xx} / \sigma_j^2 - S_x x_j / \sigma_j^2}{S_{xx} S_0 - S_x^2} \right)^2 \sigma_j^2 = \frac{S_0 S_{xx}^2 - S_{xx} S_x^2}{(S_{xx} S_0 - S_x^2)^2} = \frac{S_{xx}}{S_{xx} S_0 - S_x^2} \quad (11.36b)$$

$$\text{cov}(\hat{a}, \hat{b}) = \sum_j \left( \frac{S_0 x_j / \sigma_j^2 - S_x / \sigma_j^2}{S_{xx} S_0 - S_x^2} \right) \left( \frac{S_{xx} / \sigma_j^2 - S_x x_j / \sigma_j^2}{S_{xx} S_0 - S_x^2} \right) \sigma_j^2 = -\frac{S_x}{S_{xx} S_0 - S_x^2} \quad (11.36c)$$

$$\rho = \frac{\text{cov}(\hat{a}, \hat{b})}{\sigma_{\hat{a}} \sigma_{\hat{b}}} = -\frac{S_x}{\sqrt{S_0 S_{xx}}} \quad (11.36d)$$

This kind of straight line fit is an example of the *Least Squares method*.

#### 11.4.1 Numerical example

Consider the linear model  $y = 0.5x + 1.1$ , and assume that measurement errors are normally distributed, with standard deviations in the range (0.25, 0.75). Figure 11.4 shows a dataset with 20 datapoints.

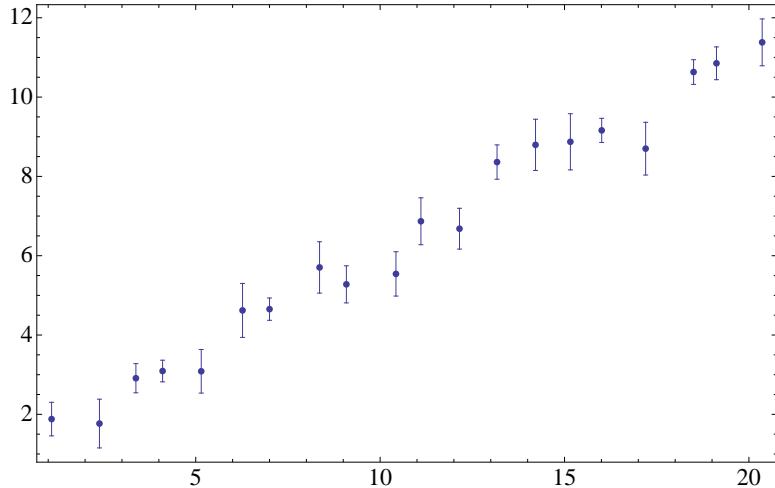


Figure 11.4: Plot of a MC-generated dataset with 20 datapoints  $(x, y)$ . The underlying linear model is  $y = 0.5x + 1.1$ . The datapoints have a Gaussian pdf about the mean value defined by the linear model. The distribution of the standard deviations of the individual errors is unimportant.

Using the formulas for  $a$ ,  $b$  and their variances, and assuming that their pdf

is Gaussian, we find the estimates

$$\hat{a} = 0.511 \pm 0.016; \quad \hat{b} = 1.020 \pm 0.193$$

(the errors are indicated at 1 standard deviation); the corresponding straight line is shown in figure 11.5, and the 2 standard deviation confidence intervals are

$$\hat{a} \in (0.478, 0.543); \quad \hat{b} \in (0.634, 1.406)$$

In the case of general least squares, the pdf's of  $\hat{a}$  and  $\hat{b}$  are difficult to

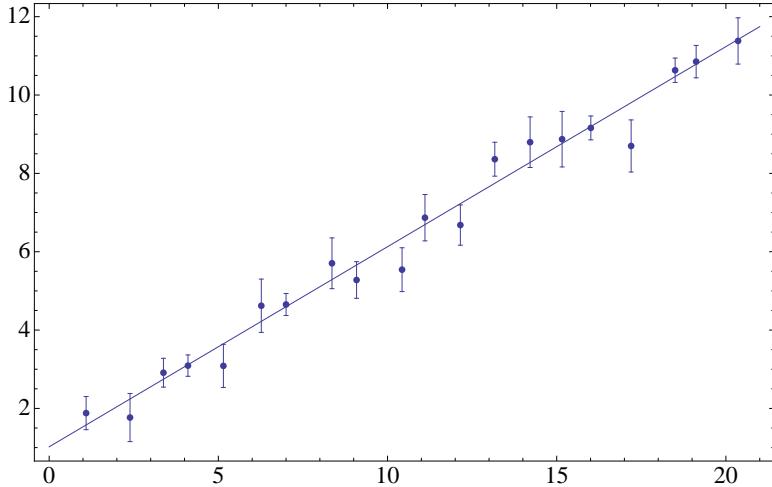


Figure 11.5: Straight line fit of the datapoints shown in figure 11.4.

find, however here we know that they are the sums of Gaussian random variables, and therefore their pdf's are also Gaussian. In other cases, where the distribution of the measurements  $y_k$  is not known, we can attack the problem with the statistical bootstrap. In this example, we carry out 10000 MC resamplings to find the empirical distributions of  $\hat{a}$  and  $\hat{b}$ , shown if figure 11.6 and 11.7: from these histograms we find the bootstrap estimates

$$\hat{a}_b = 0.511 \pm 0.011; \quad \hat{b}_b = 1.013 \pm 0.125$$

The standard deviations are lower than those estimated from error propagation. Why the difference? This can be understood from a plot of the empirical pdf obtained from the histogram in figure 11.6, shown in figure 11.8: the plot has a vertical log scale, and in such a representation a Gaussian pdf becomes a parabola; two Gaussian pdfs are shown for comparison. We see that the tails of the empirical pdf are markedly non-Gaussian, and these long tails produce the discrepancy between the estimates. This is not a problem of the fit procedure but reflects the biased distribution of data points induced by the bootstrap method.

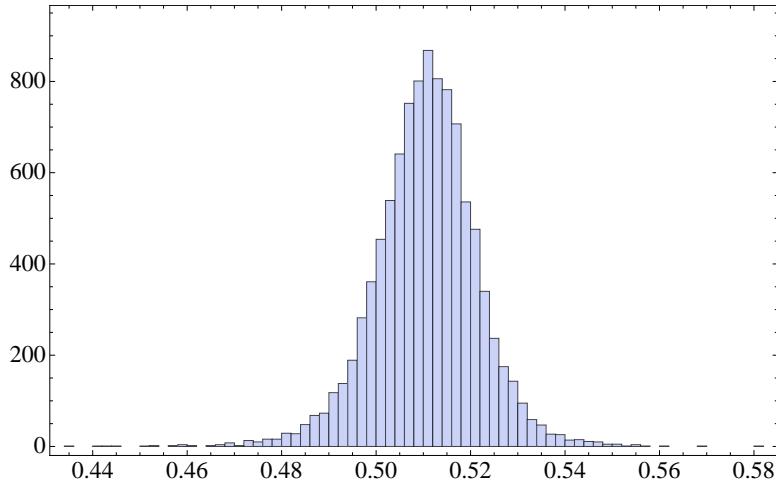


Figure 11.6: Histogram of  $\hat{a}_b$  in 10000 bootstrap iterations. The sample mean is 0.511 and the sample standard deviation is 0.011.

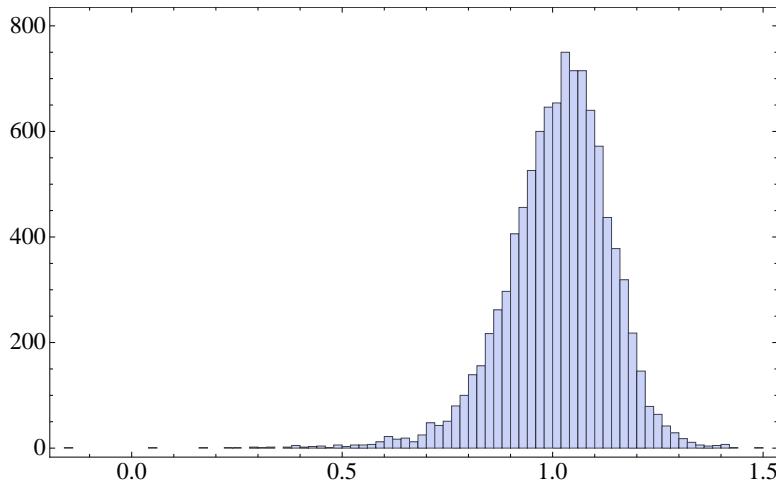


Figure 11.7: Histogram of  $\hat{b}_b$  in the same 10000 bootstrap iterations as in figure 11.6. The sample mean is 1.013 and the sample standard deviation is 0.125.

Figure 11.9 shows a linear fit for a larger dataset. We see that the fit line adapts nicely to the data points. In such a case we expect the residuals (the differences  $y_k - (\hat{a}x_k + \hat{b})$ ) to have a pdf that mirrors that of measurement errors: this is shown in figure 11.10, which confirms that the distribution of residuals is indeed like that of the original (MC) errors.

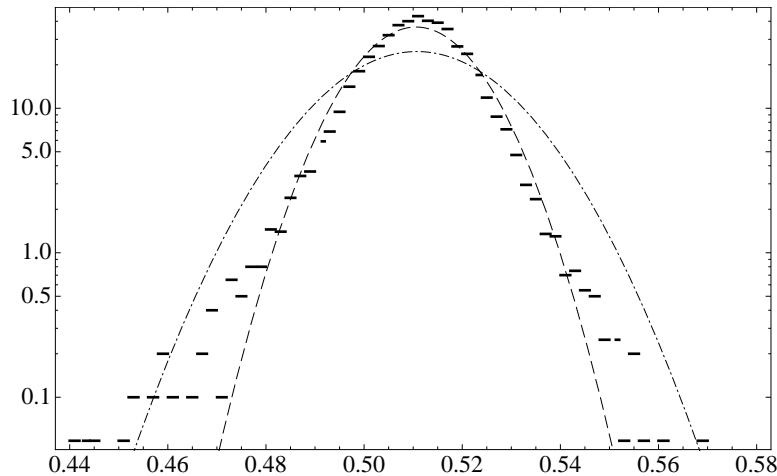


Figure 11.8: The empirical pdf of  $\hat{a}_b$  obtained with the statistical bootstrap is compared with two Gaussian pdf's in this log-scale plot. One Gaussian has the same mean and standard deviation obtained from the bootstrap histogram in figure 11.6 (dashed line); the other Gaussian has the parameters estimated from error propagation (dashed-dotted line). The first Gaussian fits the peak well, but fails to fit the tails. The deviation from normality is due to the (non-Gaussian) distribution of individual measurement errors in the bootstrap procedure.

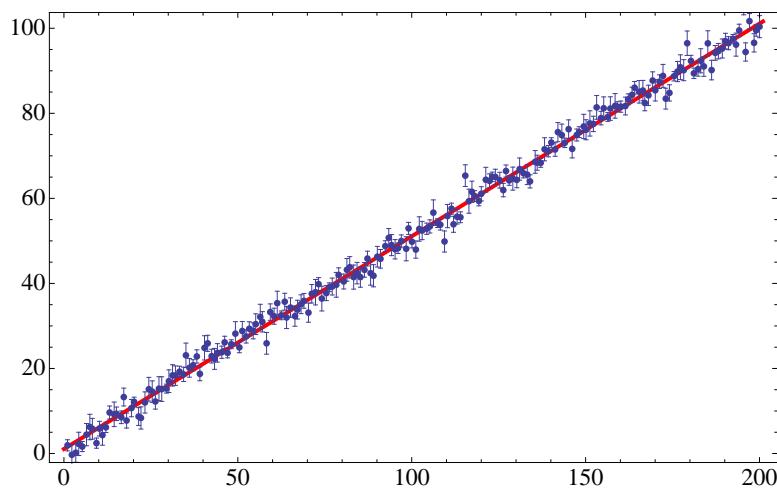


Figure 11.9: Example of a straight line fit for a larger dataset (200 data-points).

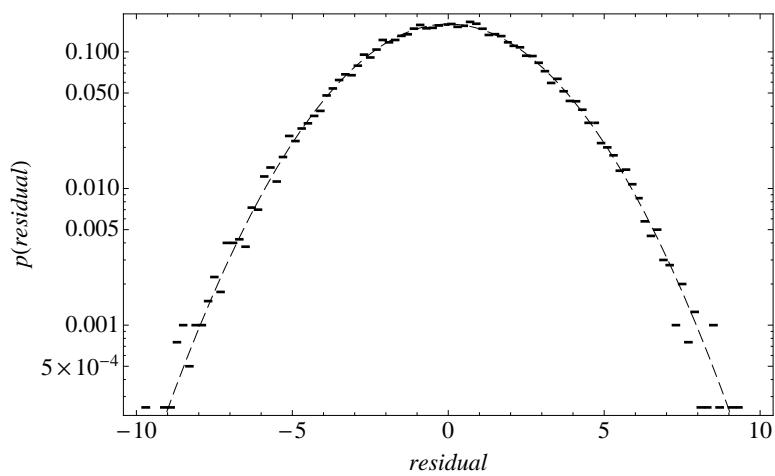


Figure 11.10: Distribution of residuals for a large dataset (20000 datapoints). The MC simulation produced data with a Gaussian distribution. Residuals mirror this original distribution: the dashed line is a Gaussian pdf with the same standard deviation as the distribution of measurement errors in the MC simulation.

## 11.5 General linear model

Now let's return to the expression

$$\chi^2 = \chi_{min}^2 + \sum_{i,k} (d_i - \mu(x_i, \boldsymbol{\theta})) (V^{-1})_{ik} (d_k - \mu(x_k, \boldsymbol{\theta})) \quad (11.37)$$

where we assume the more general form of linear model

$$\mu(x, \boldsymbol{\theta}) = \sum_{j=1}^m a_j(x) \theta_j \quad (11.38)$$

and where the  $a_j$  are linearly independent functions of  $x$ . Notice that  $\mu$  is linear in  $\boldsymbol{\theta}$  and has no need to be linear in  $x$ , and that  $\boldsymbol{\theta}$  is a parameter vector with  $m$  components. Now, for the individual measurements we can write

$$\mu(x_i, \boldsymbol{\theta}) = \sum_{j=1}^m a_j(x_i) \theta_j = \sum_{j=1}^m A_{ij} \theta_j \quad (11.39)$$

where the matrix  $A_{ij} = a_j(x_i)$  has  $n$  rows and  $m$  columns. Therefore, in vector form

$$\chi^2 = \chi_{min}^2 + (\mathbf{d} - A\boldsymbol{\theta})^T V^{-1} (\mathbf{d} - A\boldsymbol{\theta}) \quad (11.40)$$

and setting derivatives to zero, we find

$$\nabla \chi^2 = - \left\{ A^T V^{-1} (\mathbf{d} - A\hat{\boldsymbol{\theta}}) + \left[ (\mathbf{d} - A\hat{\boldsymbol{\theta}})^T V^{-1} A \right]^T \right\} = 0 \quad (11.41)$$

and finally, thanks to the symmetry of the correlation matrix, this reduces to

$$\nabla \chi^2 = -2[A^T V^{-1} \mathbf{d} - A^T V^{-1} A \hat{\boldsymbol{\theta}}] = 0 \quad (11.42)$$

the corresponding  $m$  scalar equations are called *normal equations*. Then, if  $A^T V^{-1} A$  is not singular,

$$\hat{\boldsymbol{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \mathbf{d} = B\mathbf{d} \quad (11.43)$$

or, in scalar form

$$\hat{\theta}_i = \sum_j B_{ij} d_j \quad (11.44)$$

Therefore we find

$$\frac{\partial \hat{\theta}_i}{\partial d_j} = B_{ij} \quad (11.45)$$

and thus

$$\text{cov}(\hat{\theta}_i, \hat{\theta}_k) = \sum_{j,l} B_{ij} B_{kl} \text{cov}(d_j, d_l) = \sum_{j,l} B_{ij} V_{jl} (B^T)_{lk} = (B V B^T)_{ik} \quad (11.46)$$

or also

$$\begin{aligned} U^{-1} &= BVB^T \\ &= (A^T V^{-1} A)^{-1} A^T V^{-1} V [(A^T V^{-1} A)^{-1} A^T V^{-1}]^T = (A^T V^{-1} A)^{-1} \end{aligned} \quad (11.47)$$

if we use inverse of the covariance matrix  $U_{ik}^{-1} = \text{cov}(\hat{\theta}_i, \hat{\theta}_k)$ .

Finally we remark that since the best-fit parameter estimates  $\hat{\boldsymbol{\theta}}$  are linear combinations of the data, and we assume that the data are normally distributed, then the parameter estimates are Gaussian random variables as well.

### 11.5.1 Chi-square distribution

When we perform a least squares fit, the resulting, minimum chi-square, obviously has a chi-square distribution. However now data are constrained by the fit. If there are  $m$  parameters, this means that there are  $m$  constraints that can be used to fix  $m$   $y$ -values if the other  $n - m$  are given. In other words, only  $n - m$  variates are actually free to fluctuate, and we say that there are  $n - m$  *degrees of freedom*

## 11.6 Nonlinear least squares

Now let's return to the original  $\chi^2$  expression

$$\chi^2 = \sum_k \frac{[d_k - \mu_k(\boldsymbol{\theta})]^2}{\sigma_k^2} \quad (11.48)$$

and assume that the model  $\mu$  is nonlinear. Then the normal equations

$$(\nabla \chi^2)_j = -2 \sum_k \frac{\partial \mu_k}{\partial \theta_j} \frac{[d_k - \mu_k(\boldsymbol{\theta})]}{\sigma_k^2} = 0 \quad (11.49)$$

are nonlinear as well, and in general there is no easy analytical way to solve them. This means that we must resort to numerical methods of solution, like the Newton-Raphson method.

In the procedure outlined above there are in general  $m$  parameters, and  $m$  normal (scalar) equations, so that  $F$  is vector valued and  $F'$  is an  $m \times m$  square matrix. This procedure can be difficult to implement in practical cases, and there are many variants. For instance, the basic equation

$$F(\boldsymbol{\theta}_1) = F(\boldsymbol{\theta}_0) + F'(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) = 0 \quad (11.50)$$

tells us that the gradient sets the direction of maximum change, therefore we can also decide to minimize directly the chi-square, following the direction of steepest descent (opposite to the gradient). There are many variants of these methods and we do not pursue them further here.

## 11.7 Least squares with binned data

We have seen earlier that we can treat counts in histogram bins as Poisson variates. Since a Poisson variate with mean value  $\nu$  has variance  $\nu$ , and since Poisson variates closely approximate Gaussian variates for large  $\nu$ , we can write the chi-square for binned data (counts in histogram bins) as follows:

$$\chi^2 = \sum_k \frac{[n_k - \nu_k(\theta)]^2}{\nu_k(\theta)} \quad (11.51)$$

where  $n_k$  is the number of counts in the  $k$ -th bin

$$\nu_k(\theta) = n_{tot} \int_{x_k}^{x_{k+1}} f(x, \theta) dx \quad (11.52)$$

$n_{tot}$  is the total number of counts, the  $k$ -th bin corresponds to the interval  $(x_k, x_{k+1})$ ,  $f(x, \theta)$  is the model pdf that we want to fit, and  $\sum_k n_k = n_{tot}$ .

This chi-square may be difficult to minimize, because of the nonlinear dependence on  $\theta$  due to the  $\nu(\theta)$  at the denominator. We can simplify the expression assuming that  $n_k$  is actually close to the predicted mean value and variance, and using this assumption to modify the denominator as follows:

$$\chi^2 = \sum_k \frac{[n_k - \nu_k(\theta)]^2}{n_k} \quad (11.53)$$

this is the Modified Least Squares (MLS) method. The rule of thumb for the applicability of the MLS method is that bins have at least 5 counts each.

## 11.8 Polynomial fits

The polynomial model,

$$\mu(x, a) = \sum_{j=0}^m a_j x^j \quad (11.54)$$

is linear with respect to the parameters  $a$  – although it is a polynomial of degree  $m$  with respect to  $x$  – and thus falls into the category of least squares fits that we discussed above. Now we use this polynomial model to fit the set of data points  $(x, y)$  shown in figure 11.11, ignoring the underlying linear model, much like in a real experiment.

This means that we must try different polynomial forms: since we have 10 data points and we find one equation per parameter, we cannot go beyond a polynomial of 9th degree. The results of the polynomial fits with degrees 0, ..., 9, are shown in figure 11.12. Polynomials of higher degree adapt better and better to the data points, until the degree 9 polynomial

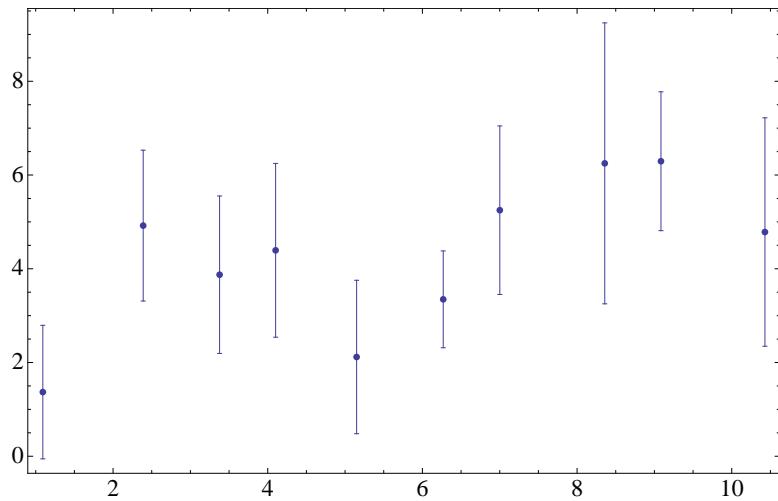


Figure 11.11: A set of point that we want to fit with a polynomial model. These points are the result of a small MC simulation with an underlying linear model.

goes exactly through all of them (in this case 10 equations are used to determine – exactly – 10 parameters).

The chi-square values corresponding to these fits are shown in table 11.2, and we see that the  $\chi^2$  decreases as  $m$  increases. Moreover the chi-square per degree of freedom ( $\chi^2_{min}/(9 - m)$ ) is about 1 for low  $m$ , and decreases sharply for higher  $m$  values: we know that on average each degree of freedom contributes 1 to the total chi-square, therefore we expect the chi-square per degree of freedom to be about 1. This hints at an important role of the chi-square statistics, an ability to discriminate between competing hypotheses (in this case the different polynomial degrees). For this reason we turn now to a systematic study of hypothesis testing, to uncover the role of chi-square and of other statistics.

## 11.9 What's the use of all this?

The method of least squares is widely used, in many different variants. Grasping the concepts in this chapter is essential to a large part of statistics. Fits of linear and nonlinear models are ubiquitous in physics, although the mathematics is often cumbersome and uninspiring.

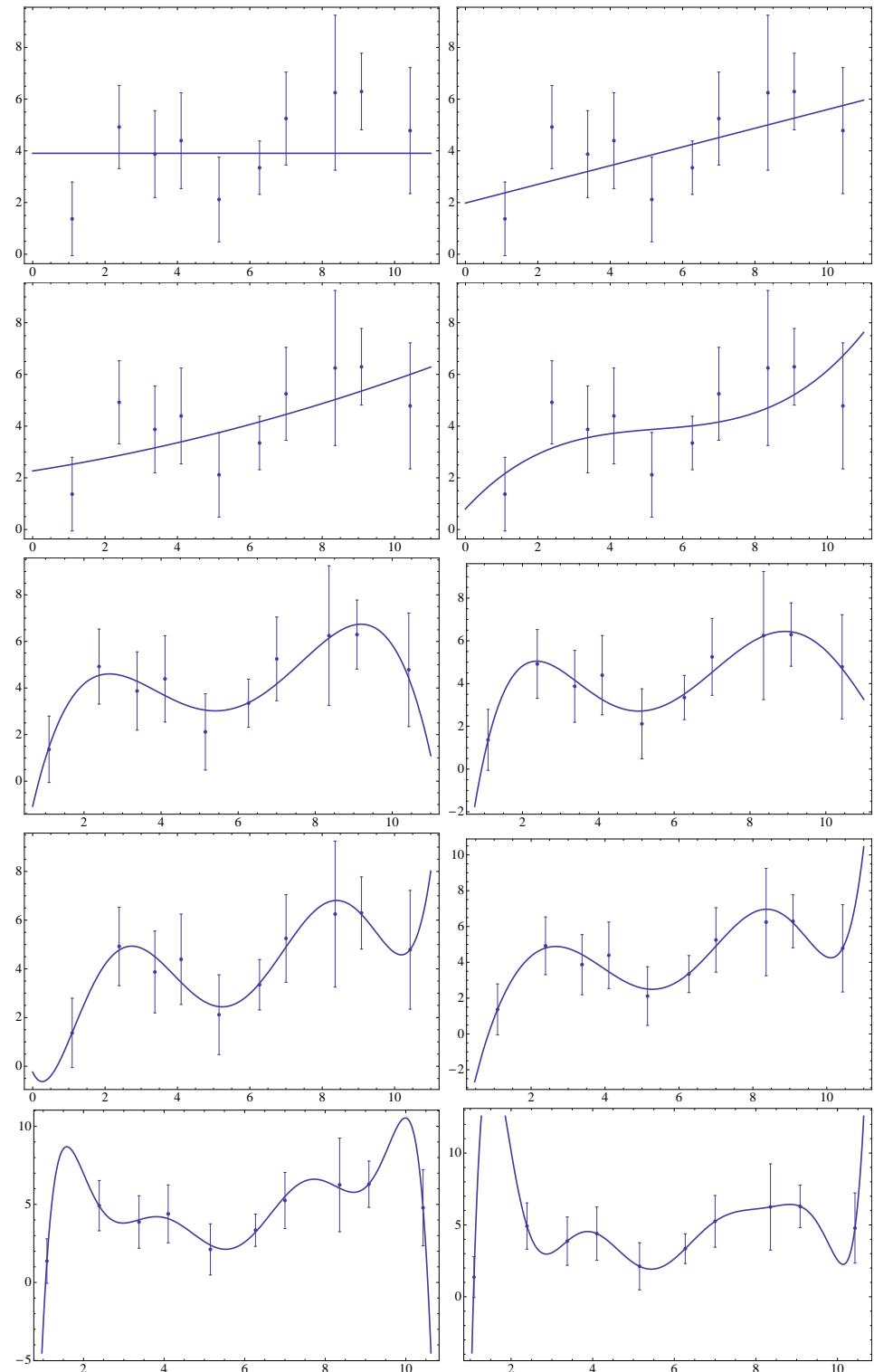


Figure 11.12: These figures show fits of the data points of figure 11.11 with polynomials of different degrees: the first one is a 0-degree polynomial (upper left corner), the last one is a 9-th degree polynomial (lower right corner).

Table 11.2:  $\chi^2$  values in the polynomial fits of figure 11.12; since there are 10 data points the fit with a polynomial of degree  $m$  has  $9 - m$  degrees of freedom.

degree $m$	$\chi^2_{min}$	$\chi^2_{min}/(9 - m)$
0	9.02602	1.00289
1	5.42686	0.678357
2	5.38898	0.769854
3	5.04511	0.840851
4	1.06332	0.212665
5	0.743402	0.18585
6	0.526778	0.175593
7	0.519602	0.259801
8	0.103418	0.103418
9	0	undefined

## Appendix: volume of a sphere in $n$ dimensions

We can easily find the volume  $V_n$  of an  $n$ -dimensional sphere, starting from its definition

$$V_n(R) = \int \dots \int_{x_1^2 + \dots + x_n^2 \leq R^2} dx_1 \dots dx_n \quad (11.55)$$

where  $R$  is the radius of the sphere. A change of variables  $x = Rx'$  leads to the expression

$$V_n(R) = R^n \int \dots \int_{x_1^2 + \dots + x_n^2 \leq 1} dx_1 \dots dx_n = R^n v_n \quad (11.56)$$

where  $v_n = V_n(1)$ . Now we notice that the inequality

$$x_1^2 + \dots + x_n^2 \leq 1 \quad (11.57)$$

can be substituted by the pair of inequalities

$$x_1^2 + x_2^2 \leq 1 \quad (11.58a)$$

$$x_3^2 + \dots + x_n^2 \leq 1 - x_1^2 - x_2^2 \quad (11.58b)$$

and therefore the integral can be broken in two parts (for  $n \geq 3$ ):

$$v_n = \int \dots \int_{x_1^2 + \dots + x_n^2 \leq 1} dx_1 \dots dx_n \quad (11.59a)$$

$$= \int \int_{x_1^2 + x_2^2 \leq 1} dx_1 dx_2 \int \dots \int_{x_3^2 + \dots + x_n^2 \leq 1 - x_1^2 - x_2^2} dx_3 \dots dx_n \quad (11.59b)$$

$$= v_{n-2} \int \int_{x_1^2 + x_2^2 \leq 1} (1 - x_1^2 - x_2^2)^{(n-2)/2} dx_1 dx_2 \quad (11.59c)$$

$$= 2\pi v_{n-2} \int_0^1 (1 - r^2)^{(n-2)/2} r dr \quad (11.59d)$$

$$= \pi v_{n-2} \int_0^1 (1 - s)^{n/2-1} ds \quad (11.59e)$$

$$= \frac{2\pi}{n} v_{n-2} \quad (11.59f)$$

Since  $v_1 = 2$ ,  $v_2 = \pi$ ,  $v_3 = 4\pi/3$ , we find from the recurrence formula

$$v_n = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} \quad (11.60)$$

and finally

$$V_n(R) = R^n \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} \quad (11.61)$$

Clearly, the volume of a thin  $n$ -dimensional spherical shell can be found from

$$\Delta V_n = \frac{dV_n}{dR} \Delta R = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} n R^{n-1} \Delta R \quad (11.62)$$

i.e., the surface area of the  $n$ -dimensional sphere is

$$S_n = \frac{dV_n}{dR} = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} n R^{n-1} = \frac{2\pi^{n/2}}{\Gamma(n/2)} R^{n-1} \quad (11.63)$$



# Chapter 12

## Taking decisions

We often have to take decisions on the basis of our data. When can we rule out an hypothesis? When can we claim a discovery? When is a fit good enough for our purposes? Here I try to summarize the current thinking on these topics: notice that the following sections deal with frequentist arguments only, and that Bayesians base all decisions on posterior distributions.

### 12.1 Definitions and basic concepts

The topic of statistical hypothesis test was first introduced by R. A. Fisher, and later it was further developed by J. Neyman and E. Pearson. The two approaches are different and led to harsh personal attacks between the two frequentist camps, and to a decades-long debate. The Bayesian approach is different, but was rejected by all frequentists (Fisher repeatedly attacked the Bayesian approach as “non-scientific”).

In this section I introduce the scheme of Neyman and Pearson.

- hypotheses can be *simple* or *composite*. A composite hypothesis represents a whole array of possible hypotheses, e.g., a pdf that depends on one or more parameters (a given set of parameter values fixes a member of the pdf family and selects a simple hypothesis); for instance “the data are drawn from a Poisson distribution with mean 4.4” is a simple hypothesis, while “the data are drawn from a Poisson distribution with mean greater than 4.4” is composite;
- the *null hypothesis*  $H_0$  is the conventional, minimal hypothesis that would be true if there were no signal in the data. The choice of the null hypothesis is not obvious, and it is no less arbitrary than the prior distribution in Bayesian statistics;
- the *alternative hypotheses*  $H_1, H_2, \dots$ , are the specific hypotheses (alternative to  $H_0$ ) that we make, and that we wish to test;

- a *test statistic*  $t(D)$  is a function of data  $D$ , such that we know its conditional pdf  $p(t|H_k)$ ; e.g., in the section on polynomial fit we have guessed that the  $\chi^2$  can be such a statistic, and in each case we know its distribution  $p_{\chi^2}(\chi^2|m)$  (a chi-square distribution with  $m$  degrees of freedom). As another example we could also take the case of repeated throws of a coin where we try to infer whether the coin is fair or not: here we could take as test statistic the number of heads (or tails) in the set of throws;
- the *critical region* is that region  $C$  of the  $t$  space where  $H_0$  is rejected, i.e., if  $t(D)$  falls into  $C$  then  $H_0$  is rejected; its complement,  $C^*$  is the *acceptance region* (see figure 12.1);
- the probability that  $t(D)$  falls into the critical region

$$\alpha = \int_C p(t|H_0) dt \quad (12.1)$$

is called *significance level* or *size* of the test;

- the probability that  $t(D)$  falls into the acceptance region is

$$1 - \alpha = \int_{C^*} p(t|H_0) dt \quad (12.2)$$

and this is sometimes called *specificity* of the test;

- the probability of rejecting  $H_0$  if  $H_0$  is true is equal to the significance level  $\alpha$  of the test;
- if  $t$  is (as often happens) a 1D statistic, then the regions  $C$  and  $C^*$  are usually separated by a single value  $t_{cut}$ , called the *cut* or *decision boundary* (see figure 12.1); figure 12.2 shows a 2D example in a real experiment [2];
- if we reject a true hypothesis we commit an *error of the first kind*;
- if we accept a false hypothesis we commit an *error of the second kind*;
- if the true hypothesis is not the null hypothesis, but an alternative hypothesis  $H_1$ , then the true conditional distribution of the test statistic  $t$  is  $p(t|H_1)$ , the probability of finding  $t(D)$  in the acceptance region of  $H_0$  is

$$\beta = \int_{C^*} p(t|H_1) dt \quad (12.3)$$

and in this case we would accept  $H_0$  and commit an error of the second kind; the probability of correctly finding  $t(D)$  in the critical region of  $H_0$  is

$$1 - \beta = \int_C p(t|H_1) dt \quad (12.4)$$

and is called *power* of the test;

	$H_0$ true	$H_0$ false
accept $H_0$	right decision	wrong decision, type II error
reject $H_0$	wrong decision, type I error	right decision

Table 12.1: Possible cases in the decision to accept or reject a null hypothesis  $H_0$ . When we must decide between two competing hypotheses  $H_0$  and  $H_1$ , the significance level  $\alpha$  is the probability of wrongly rejecting  $H_0$  when it is true, and  $\beta$  is the probability of wrongly accepting  $H_0$ , when it is false.

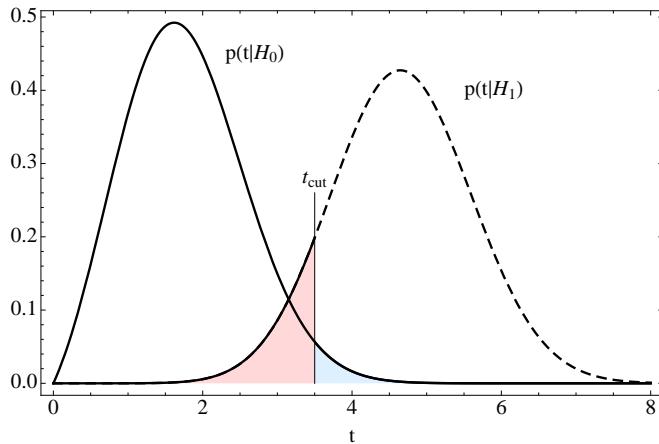


Figure 12.1: Comparison of hypotheses  $H_0$  and  $H_1$  with a 1D statistic  $t$ . The statistic has different distributions when  $H_0$  is true or  $H_1$  is true. The region to the left of  $t_{\text{cut}}$  is the acceptance region for  $H_0$ , while the region to the right of  $t_{\text{cut}}$  is the critical region for  $H_0$ . The probability  $\alpha$  that the test statistic  $t$  falls into the critical region of  $H_0$  is the area under  $p(t|H_0)$  (marked in light blue in this figure):  $\alpha = \int_C p(t|H_0)dt$ .  $\alpha$  is the probability of rejecting  $H_0$  when it is true, and is called the *significance level* of the test. If the true hypothesis is not the null hypothesis, but an alternative hypothesis  $H_1$ , then the true conditional distribution of the test statistic  $t$  is  $p(t|H_1)$ , and the probability of finding  $t$  in the acceptance region of  $H_0$  – i.e., the probability of rejecting  $H_1$  when it is true – is the area under  $p(t|H_1)$  (marked in light red in this figure):  $\beta = \int_{C^*} p(t|H_1)dt$ . The complementary quantity  $1 - \beta = \int_C p(t|H_1)dt$  is the *power* of the test.

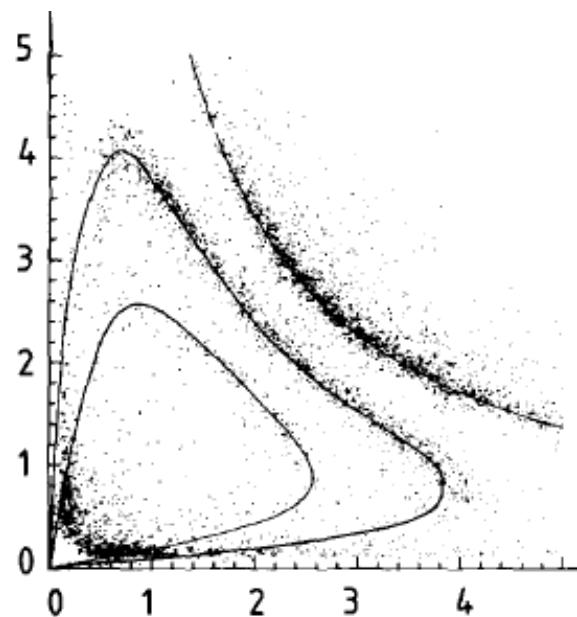


Figure 12.2: Kinematic distributions of electron, pion and muon pairs in experiment NA7 at CERN (distributions of scattering angles  $\theta_2$  vs.  $\theta_1$ , mrad) [2]. The experiment was designed to measure the electromagnetic form factor of charged pions, and detected electron, muon and pion pairs. This two-body plot shows the scattering angle of each particle in the pair, with respect to the beam direction. The outer curve corresponds to electrons, and the most internal curve to pions. The experimental distribution of the scattering angles allows the definition of empirical 2D pdf's, and of the critical and acceptance regions for the three hypotheses of electron, muon or pion pair.

## 12.2 P-values

The concept of *P-value* was introduced by Fisher. A p-value is attached to a given measurement, and is the probability of obtaining a more extreme value than the measured one (see figure 12.3).

The p-value is very controversial, and the debate on its usefulness is heated, especially in the social sciences. See, e.g., [31] for a good summary of the present debate.

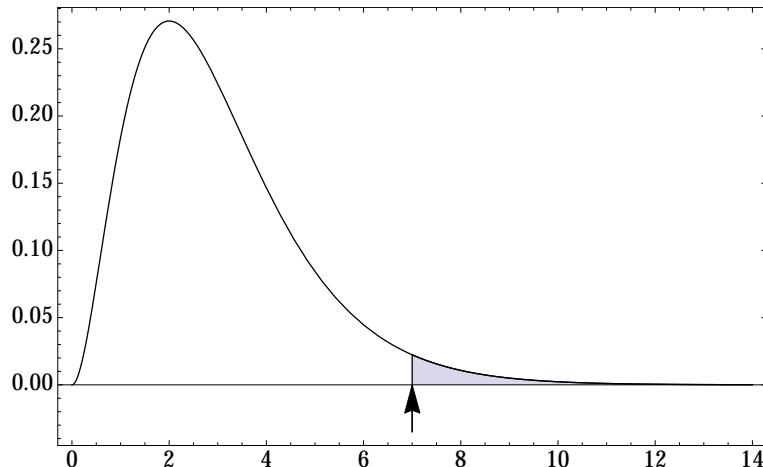


Figure 12.3: Illustration of p-value: pdf of the test statistic for the null hypothesis. The position of the test statistic in a given measurement is marked by the arrow, and the p-value is the probability of exceeding this measurement, given the distribution of the test statistic. In the usual practice, if the p-value is smaller than some predefined level (once 5%, according to the now-obsolete Fisher prescription), the null statistics is rejected.

## 12.3 The Neyman-Pearson lemma

Hypothesis testing deals not only with observed data, but also with the design of experiments: it is important to design experiments that test the null and the alternative hypothesis minimizing both kinds of errors.

When we consider the framework of Neyman and Pearson, and we go back to figure 12.1, we notice that as we decrease the significance level  $\alpha$  (and thus the size of the critical region) increasing  $t_{cut}$ , we also increase  $\beta$ , i.e., we decrease the power  $1 - \beta$ , and vice versa. In particular, if  $t_{cut} \rightarrow -\infty$ , then  $\alpha \rightarrow 1$  and  $\beta \rightarrow 0$ , while if  $t_{cut} \rightarrow +\infty$ , then  $\alpha \rightarrow 0$  and  $\beta \rightarrow 1$ , and we guess that there must be some tradeoff that optimizes the test statistic, so that both  $\alpha$  and  $\beta$  are small.

The Neyman-Pearson (NP) lemma is an answer to this problem. The lemma basically states that the likelihood ratio is the most powerful test of size  $\alpha$  in deciding which of two hypotheses is true (i.e., with given significance level  $\alpha$  for the rejection of the null hypothesis, compared to another hypothesis  $H_1$ ). In symbols, if

$$P(t(D) \leq \eta | H_0) = \alpha \quad (12.5)$$

which defines the critical region as the region where the test statistic is below a certain threshold  $\eta$ , then the likelihood ratio

$$t(D) = r(D, H_0, H_1) = \frac{L(D, H_0)}{L(D, H_1)} \quad (12.6)$$

is the most powerful test, i.e., it has the greatest power  $1 - \beta$ , and the least probability  $\beta$  of an error of type II.

Now we prove the NP lemma. Let  $C_{NP}$  be the critical region associated to the likelihood ratio

$$C_{NP} = \left\{ D : r(D, H_0, H_1) = \frac{L(D, H_0)}{L(D, H_1)} \leq \eta \right\} \quad (12.7)$$

where  $D$  is any data set, and consider also another test  $A$  with significance level  $\alpha$ , and a different critical region  $C_A$  in parameter space: Notice first that for any hypothesis  $H$

$$P(C_{NP}|H) = P(C_{NP} \cap C_A|H) + P(C_{NP} \cap C_A^*|H) \quad (12.8)$$

and

$$P(C_A|H) = P(C_{NP} \cap C_A|H) + P(C_{NP}^* \cap C_A|H) \quad (12.9)$$

Since, by assumption both tests (NP and A) have the same significance level  $\alpha$  for the null hypothesis

$$P(C_{NP}|H_0) = P(C_A|H_0) = \alpha \quad (12.10)$$

we see that the above pair of equalities implies

$$P(C_{NP} \cap C_A^*|H_0) = P(C_{NP}^* \cap C_A|H_0) \quad (12.11)$$

Moreover the power of the NP test is  $1 - \beta_{NP} = P(C_{NP}|H_1)$ , and that of the A test is  $1 - \beta_A = P(C_A|H_1)$ , therefore, from the same two equalities – now for  $H_1$  – we obtain

$$1 - \beta_{NP} = P(C_{NP}|H_1) = P(C_{NP} \cap C_A|H_1) + P(C_{NP} \cap C_A^*|H_1) \quad (12.12)$$

and

$$1 - \beta_A = P(C_A|H_1) = P(C_{NP} \cap C_A|H_1) + P(C_{NP}^* \cap C_A|H_1) \quad (12.13)$$

Next, we note that

$$P(C_{NP} \cap C_A^* | H_1) = \int_{C_{NP} \cap C_A^*} L(D|H_1) dD \quad (12.14a)$$

$$\geq \frac{1}{\eta} \int_{C_{NP} \cap C_A^*} L(D|H_0) dD \quad (12.14b)$$

$$= \frac{1}{\eta} P(C_{NP} \cap C_A^* | H_0) \quad (12.14c)$$

$$= \frac{1}{\eta} P(C_{NP}^* \cap C_A | H_0) \quad (12.14d)$$

$$= \frac{1}{\eta} \int_{C_{NP}^* \cap C_A} L(D|H_0) dD \quad (12.14e)$$

$$\geq \int_{C_{NP}^* \cap C_A} L(D|H_1) dD \quad (12.14f)$$

$$= P(C_{NP}^* \cap C_A | H_1) \quad (12.14g)$$

where we have used repeatedly  $L(D|H)$ , the likelihood function, which is the pdf of data assuming the hypothesis  $H$ . Therefore

$$P(C_{NP} \cap C_A^* | H_1) \geq P(C_{NP}^* \cap C_A | H_1) \quad (12.15)$$

Now we go back to the equations for power, and find

$$(1 - \beta_{NP}) - (1 - \beta_A) = P(C_{NP} \cap C_A^* | H_1) - P(C_{NP}^* \cap C_A | H_1) \geq 0 \quad (12.16)$$

i.e.,

$$(1 - \beta_{NP}) \geq (1 - \beta_A) \quad (12.17)$$

which means that no test can have a power higher than the NP test for a given significance level – if we know the correct likelihood function. Notice also that the test statistic

$$\ln \frac{L(D, H_0)}{L(D, H_1)}$$

is fully equivalent to the likelihood ratio, since the logarithm is a monotonic function.

## 12.4 Using $\chi^2$ to test goodness-of-fit

We have noted above that on average each degree of freedom has a unitary contribution to the total  $\chi^2$  value. We also know the  $\chi^2$  distribution for  $n$  degrees of freedom, therefore we can evaluate the probability of finding a given  $\chi^2$  value (see table 12.2 which lists the quantiles<sup>1</sup> for selected cumulative probabilities, and up to 30 degrees of freedom).

---

<sup>1</sup>The quantile  $q$  associated to a given pdf  $p(x)$  and to a probability  $P$  is given by the solution of the equation  $\int_{-\infty}^q p(x) dx = P$ .

Table 12.2: Quantile values of the  $\chi^2$  distribution with  $n$  degrees of freedom.

$n$	1%	5%	10%	50%	90%	95%	98%	99%
1	0.0001571	0.003932	0.01579	0.4549	2.706	3.841	5.412	6.635
2	0.02010	0.1026	0.2107	1.386	4.605	5.991	7.824	9.210
3	0.1148	0.3518	0.5844	2.366	6.251	7.815	9.837	11.34
4	0.2971	0.7107	1.064	3.357	7.779	9.488	11.67	13.28
5	0.5543	1.145	1.610	4.351	9.236	11.07	13.39	15.09
6	0.8721	1.635	2.204	5.348	10.64	12.59	15.03	16.81
7	1.239	2.167	2.833	6.346	12.02	14.07	16.62	18.48
8	1.646	2.733	3.490	7.344	13.36	15.51	18.17	20.09
9	2.088	3.325	4.168	8.343	14.68	16.92	19.68	21.67
10	2.558	3.940	4.865	9.342	15.99	18.31	21.16	23.21
11	3.053	4.575	5.578	10.34	17.28	19.68	22.62	24.72
12	3.571	5.226	6.304	11.34	18.55	21.03	24.05	26.22
13	4.107	5.892	7.042	12.34	19.81	22.36	25.47	27.69
14	4.660	6.571	7.790	13.34	21.06	23.68	26.87	29.14
15	5.229	7.261	8.547	14.34	22.31	25.00	28.26	30.58
16	5.812	7.962	9.312	15.34	23.54	26.30	29.63	32.00
17	6.408	8.672	10.09	16.34	24.77	27.59	31.00	33.41
18	7.015	9.390	10.86	17.34	25.99	28.87	32.35	34.81
19	7.633	10.12	11.65	18.34	27.20	30.14	33.69	36.19
20	8.260	10.85	12.44	19.34	28.41	31.41	35.02	37.57
21	8.897	11.59	13.24	20.34	29.62	32.67	36.34	38.93
22	9.542	12.34	14.04	21.34	30.81	33.92	37.66	40.29
23	10.20	13.09	14.85	22.34	32.01	35.17	38.97	41.64
24	10.86	13.85	15.66	23.34	33.20	36.42	40.27	42.98
25	11.52	14.61	16.47	24.34	34.38	37.65	41.57	44.31
26	12.20	15.38	17.29	25.34	35.56	38.89	42.86	45.64
27	12.88	16.15	18.11	26.34	36.74	40.11	44.14	46.96
28	13.56	16.93	18.94	27.34	37.92	41.34	45.42	48.28
29	14.26	17.71	19.77	28.34	39.09	42.56	46.69	49.59
30	14.95	18.49	20.60	29.34	40.26	43.77	47.96	50.89

Table 12.3:  $\chi^2$  values in the polynomial fits of figure 11.12; since there are 10 data points the fit with a polynomial of degree  $m$  has  $9 - m$  degrees of freedom. Here the p-value is the probability of finding a  $\chi^2$  smaller than that reported in the fit. A 1% p-value (either in the upper or the lower tail) would exclude the polynomial degrees 0 and 5, ..., 9.

degree $m$	$\chi^2_{min}$	$\chi^2_{min}/(9 - m)$	p-value (%)
0	9.02602	1.00289	99.73
1	5.42686	0.678357	98.02
2	5.38898	0.769854	85.46
3	5.04511	0.840851	71.73
4	1.06332	0.212665	4.270
5	0.743402	0.18585	0.6494
6	0.526778	0.175593	0.06576
7	0.519602	0.259801	0.01543
8	0.103418	0.103418	$2.977 \cdot 10^{-6}$
9	0	undefined	undefined

Going back to table 11.2 we can also compute the corresponding *p-values* for the  $\chi^2$  values that we found in the fit.

Finally, notice that for a Gaussian likelihood, the logarithm of the likelihood ratio is equal to the difference of the corresponding  $\chi^2$ 's, and therefore in this case the  $\chi^2$  corresponds to the most powerful statistic according to Neyman-Pearson.

## 12.5 Significance of a signal

Consider a histogram with a small peak over a fluctuating background. Is that a signal? How do we decide? In this case we can proceed as follows:

1. find the average number of counts  $\nu_B$  in the fluctuating close to the peak;
2. assume the null hypothesis that the peak is just a fluctuation of the background; this implicitly defines the alternative hypothesis that the signal is not a fluctuation;
3. make a reasonable assumption for the statistics of the fluctuating background, i.e., in the case of the histogram take a Poisson statistics;
4. if the peak corresponds to  $n_P$  counts, compute the corresponding p-

value; in the case of the histogram, this is

$$P(n > n_P) = \sum_{n_P}^{\infty} \frac{\nu_B^n}{n!} e^{-\nu_B} = 1 - \sum_0^{n_P} \frac{\nu_B^n}{n!} e^{-\nu_B} \quad (12.18)$$

5. use the p-value to accept or to reject the null hypothesis; always remember that the threshold p-value that sets acceptance or rejection is somewhat arbitrary.

Now we examine a specific example: we observe a time series like that shown in figure 12.4, which represents a voltage vs. time. The figure shows a structureless signal: we can histogram the recorded amplitudes and we find the result shown in figure 12.5, which indicates that this is a noise with Gaussian statistics. We can also compute the periodogram of the signal with the Fast Fourier Transform (FFT), and find the result shown in figure 12.6, which shows that this is white noise: thus we are dealing with a Gaussian white noise. In a different measurement we observe the time series shown in figure 12.7. We glimpse a hint of nonrandomness, but is not apparent in the histogram of the sample amplitudes 12.8. The periodogram, however, helps us detect a clear peak which stands well above the noise background (see figure 12.9). In yet another series of measurement we observe the time series shown in figure 12.10, and we find the corresponding periodogram shown in figure 12.11: now there might again be a peak in the same position of figure 12.9, however it is not at all clear. What can we do?

First, we find the pdf of a white noise periodogram, to establish our baseline, the null hypothesis. If we use the FFT for  $N$  voltage samples  $V_n$

$$\begin{aligned} F_k &= a_k + i b_k = \sum_{n=0}^{N-1} V_n e^{-2\pi i n k / N} \\ &= \sum_{n=0}^{N-1} V_n \cos\left(\frac{2\pi n k}{N}\right) - i \sum_{n=0}^{N-1} V_n \sin\left(\frac{2\pi n k}{N}\right) \end{aligned} \quad (12.19)$$

we obtain the periodogram estimate

$$S_k = \frac{|F_k|^2}{N^2} = \frac{a_k^2 + b_k^2}{N^2} \quad (12.20)$$

The samples have the same kind of fluctuation and have a Gaussian distribution as shown by the histograms of the sample amplitudes, thus the histograms let us estimate the variance  $\sigma^2$  of the samples. Since the samples are independent, have zero mean, and have the same variance, and the real and the imaginary parts of the FFT are linear combinations of the samples, then  $a_k$  and  $b_k$  are normally distributed as well, they have zero mean,

and – using error propagation – their variances are

$$\sigma_{a,k}^2 = \sigma^2 \sum_{n=0}^{N-1} \cos^2 \left( \frac{2\pi nk}{N} \right) = \frac{N\sigma^2}{2} (1 + \delta_{k,0}) \quad (12.21a)$$

$$\sigma_{b,k}^2 = \sigma^2 \sum_{n=0}^{N-1} \sin^2 \left( \frac{2\pi nk}{N} \right) = \frac{N\sigma^2}{2} (1 + \delta_{k,0}) \quad (12.21b)$$

i.e., for  $k > 0$ ,

$$\sigma_{a,k}^2 = \sigma_{b,k}^2 = \frac{N\sigma^2}{2} \quad (12.22)$$

Since  $a_k$  and  $b_k$  are Gaussian random variables, then  $s_k = 2N S_k / \sigma^2$  has a chi-square distribution with two degrees of freedom:

$$p_s(s_k) = \frac{1}{2} e^{-s_k/2} \quad (12.23)$$

therefore

$$p_S(S_k) = p_s(S_k) \frac{ds_k}{dS_k} = \frac{N}{\sigma^2} e^{-NS_k/\sigma^2} \quad (12.24)$$

This means that the p-value (probability of finding an amplitude larger than that observed) associated to a given spectral amplitude  $S_k^{(0)}$  is

$$\int_{S_k^{(0)}}^{\infty} \frac{N}{\sigma^2} e^{-NS_k/\sigma^2} dS_k = e^{-NS_k^{(0)}/\sigma^2} \quad (12.25)$$

If the p-value is smaller than a given, predefined threshold<sup>2</sup>, then the null hypothesis must be rejected and we can claim the detection of a spectral peak.

---

<sup>2</sup>to avoid bias it is important to define the threshold before performing the experiment.

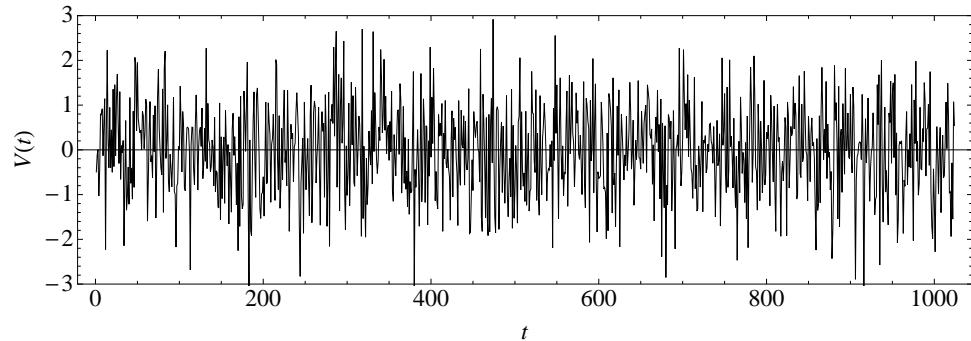


Figure 12.4: A structureless time series showing a voltage vs. time: a visual inspection shows white noise only. In this figure there are  $2^{10} = 1024$  samples, which is common in laboratory instrumentation.

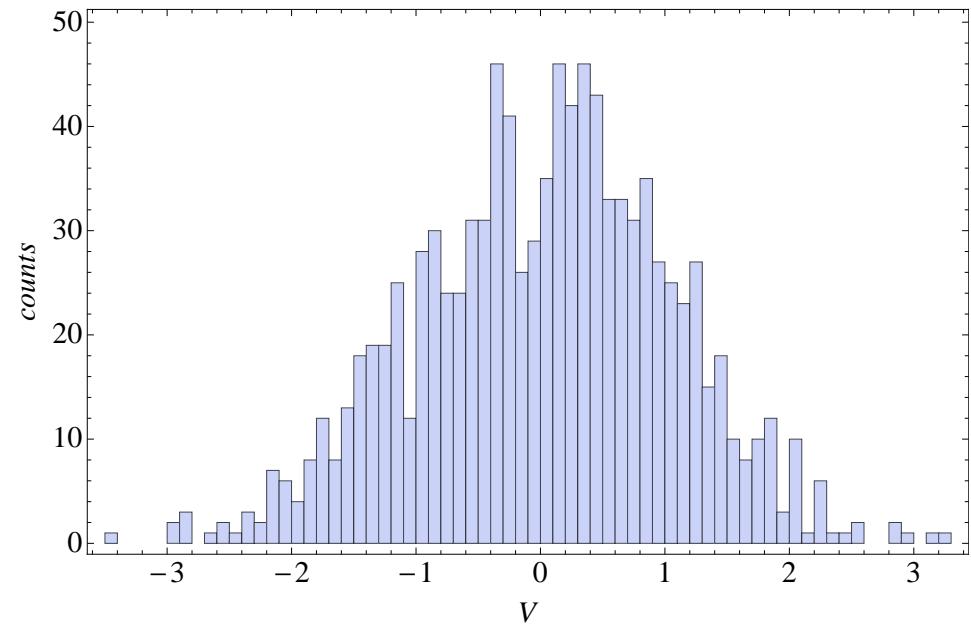


Figure 12.5: Histogram of the sample amplitudes of the time series in figure 12.4. This is an indication that the signal has a Gaussian pdf (a careful analysis would prove it conclusively).

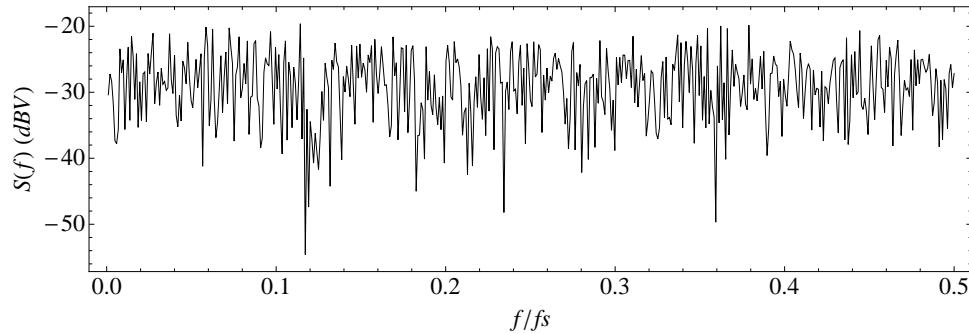


Figure 12.6: Estimate of the spectral density (periodogram) of the time series shown in figure 12.4 (dBV vs. frequency; frequency is measured in units of sampling frequency  $fs$ ). This single realization of the periodogram is flat, and this indicates that we are dealing with a white noise.

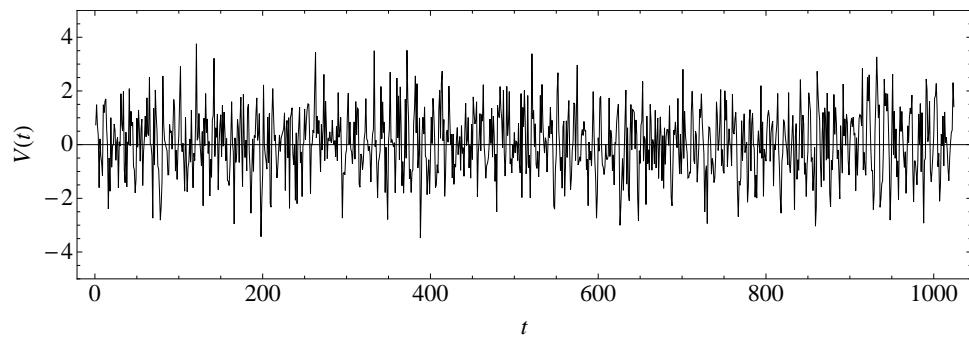


Figure 12.7: Another time series showing a voltage vs. time: a visual inspection shows some nonrandom correlation. Just as in figure 12.4 there are  $2^{10} = 1024$  samples.

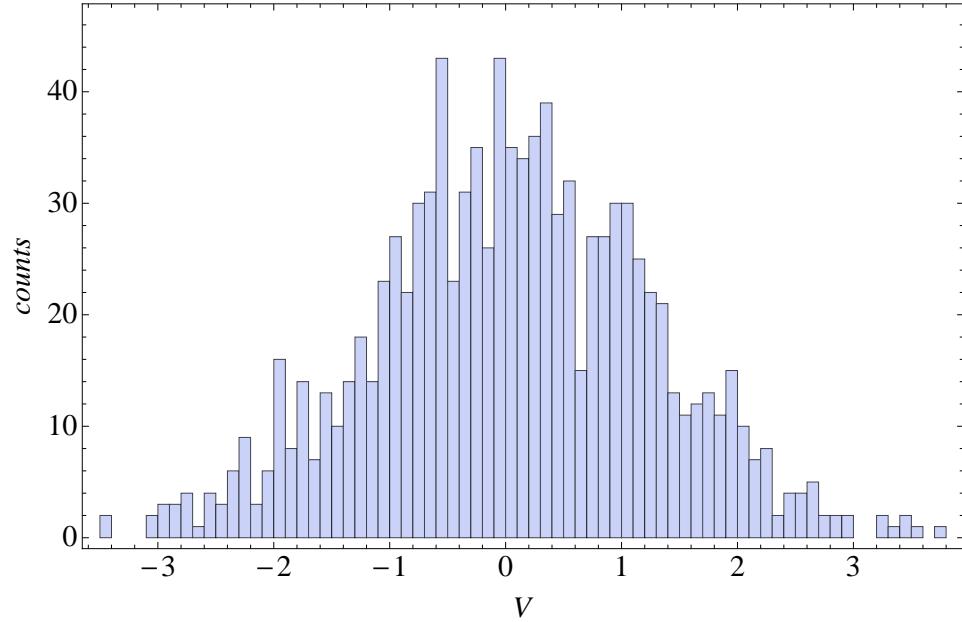


Figure 12.8: Histogram of the sample amplitudes of the time series in figure 12.7. The histogram does not seem significantly different from that in figure 12.5, and indeed in this case it is not useful to uncover a hidden signal.

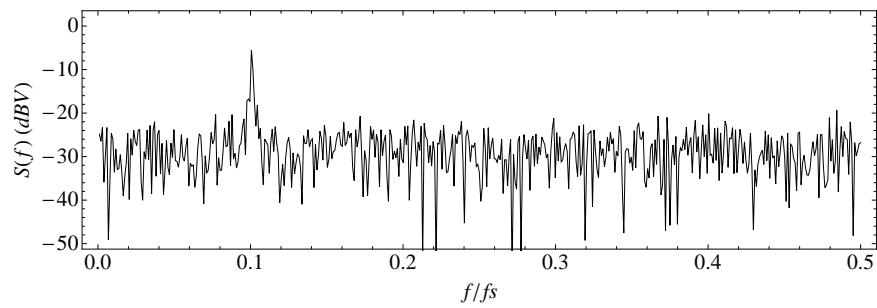


Figure 12.9: Estimate of the spectral density (periodogram) of the time series shown in figure 12.7. This single realization of the periodogram shows white noise plus a clear peak.

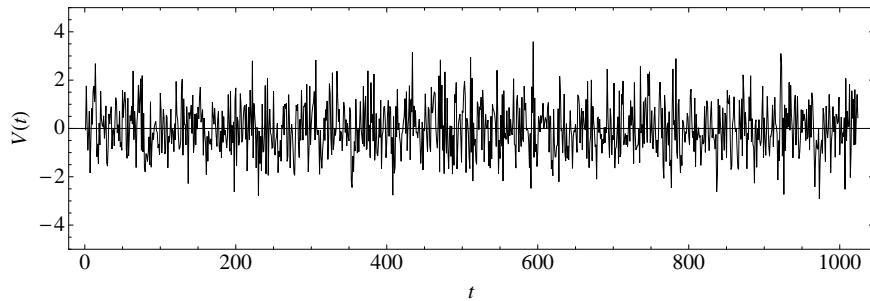


Figure 12.10: Yet another time series showing a voltage vs. time. Just as in figure 12.4 there are  $2^{10} = 1024$  samples. Now is there a signal hidden in the noise background?

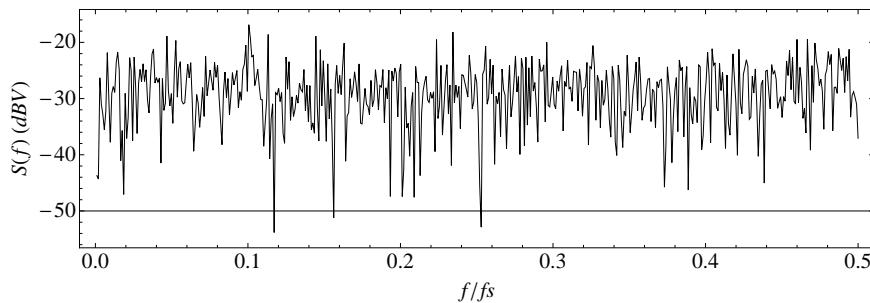


Figure 12.11: Estimate of the spectral density (periodogram) of the time series shown in figure 12.10. This single realization of the periodogram shows white noise plus a possible peak in the same position as figure 12.9. Accepting this as a true peak or not depends on our definition of the threshold for the p-value. In this case, given the large visible fluctuations of the background, it is clear that the peak cannot be very significant.

## 12.6 A decision problem: compatibility of measurements

It often happens that measurements must be compared to test their compatibility, such as in quality tests in manufacturing, or in many cases in physics, where you want to assign a measurement to a certain class. Here I consider two cases with Gaussian random variables: a) where the standard deviations  $\sigma_x, \sigma_y$  are known, and b) where they are not known and must be estimated from the dataset.

### 12.6.1 Standard deviations known

This case is easy: here the null hypothesis is that measurements do not differ, and testing for compatibility of two measurements  $x$  and  $y$  means essentially that the difference  $x - y$  is indistinguishable from 0, i.e.,  $|x - y| < k\sqrt{\sigma_x^2 + \sigma_y^2}$  where  $k$  is the number of standard deviations that corresponds to the required significance level.

### 12.6.2 Standard deviations unknown: Student's $t$

In this case we consider datasets of  $n$  independent, normally distributed measurements  $x_k$ , and we compare the compatibility of different datasets. Since we assume normality, then  $\mu_S$  is a Gaussian variate. To estimate the variance of a given data set we use the sample variance  $\sigma_S^2$

$$\sigma_S^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \mu_S)^2 \quad (12.26)$$

We also know that in this case, sample mean and sample variance are uncorrelated (see the proof in the Appendix to this chapter). Since both the data  $x_k$  and  $\mu_S$  are normal variates and

$$\mathbf{E} (x_k - \mu_S)^2 = \frac{n-1}{n} \sigma^2 \quad (12.27)$$

then  $\sigma_S$  has a chi-square distribution with  $n - 1$  degrees of freedom.

When the standard deviation is known the random variable

$$\frac{\mu_S - \mu}{\sigma / \sqrt{n}} \quad (12.28)$$

has a normal pdf  $N(0, 1)$ . However – as discussed above – in its place we often have to use the random variable

$$v = \frac{\mu_S - \mu}{\sigma_S / \sqrt{n}} = \frac{(\mu_S - \mu) / (\sigma / \sqrt{n})}{\sigma_S / \sigma} \quad (12.29)$$

and we must find its pdf under general conditions.

We note that  $v$  has the generic functional form  $t = x(s/\sqrt{n})^{-1/2}$ , where  $x$  has a normal distribution and  $s$  has a chi-square distribution with  $n$  degrees of freedom, and when we define the auxiliary variate  $y = x$  we can readily find the pdf of  $t$ . Indeed, the inverse transformation is  $x = y$ ,  $s = ny^2/t^2$ , and therefore the transformation from the  $(x, s)$  pair to the  $(y, t)$  pair has the Jacobian matrix

$$\frac{\partial(x, s)}{\partial(y, t)} = \begin{pmatrix} 1 & 0 \\ 2ny/z^2 & -2ny^2/t^3 \end{pmatrix} \quad (12.30)$$

and therefore

$$\left| \frac{\partial(x, s)}{\partial(y, t)} \right| = \frac{2ny^2}{t^3} \quad (12.31)$$

From the assumed statistical independence of  $x$  and  $s$ , their joint pdf is

$$p_{x,s}(x, s) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \frac{s^{n/2-1}}{2^{n/2}\Gamma(n/2)} \exp\left(-\frac{s}{2}\right) \quad (12.32)$$

and thus

$$\begin{aligned} p_{y,t}(y, t) &= p_{x,s}(x(y, t), s(y, t)) \left| \frac{\partial(x, s)}{\partial(y, t)} \right| \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \frac{(n^{1/2}y/t)^{n-2}}{2^{n/2}\Gamma(n/2)} \exp\left(-\frac{ny^2}{2t^2}\right) \frac{2ny^2}{t^3} \end{aligned} \quad (12.33)$$

The pdf of  $t$  is obtained integrating over the nuisance random variable  $y$ :

$$p_t(t) = \int_y p_{y,t}(y, t) dy \quad (12.34a)$$

$$= \frac{1}{2^{n/2-1}\sqrt{2\pi n}\Gamma(n/2)} \int_y \exp\left[-\frac{(1+t^2/n)}{2}\left(\frac{y}{t/\sqrt{n}}\right)^2\right] \left(\frac{y}{t/\sqrt{n}}\right)^n d\left(\frac{y}{t/\sqrt{n}}\right) \quad (12.34b)$$

$$= \frac{1}{2^{n/2-1}\sqrt{2\pi n}\Gamma(n/2)} \int_z \exp\left[-\frac{(1+t^2/n)}{2}z^2\right] z^n dz \quad (12.34c)$$

$$= \frac{2^{(n+1)/2}(1+t^2/n)^{-(n+1)/2}}{2^{n/2}\sqrt{2\pi n}\Gamma(n/2)} \int_0^\infty e^{-w} w^{(n+1)/2-1} dw \quad (12.34d)$$

$$= \frac{(1+t^2/n)^{-(n+1)/2}}{\sqrt{\pi n}\Gamma(n/2)} \Gamma\left(\frac{n+1}{2}\right) \quad (12.34e)$$

This is called *Student's t distribution*, and it substitutes the Gaussian distribution in our considerations on standard deviations.

### 12.6.3 Another application of Student's $t$ : confidence limits

In chapter 10 we considered the problem of finding confidence limits for normally distributed data, with a known variance. In many cases the variance is not known, and must be estimated from the sample variance. Very often this does not change the construction of confidence limits, because we have plenty of data, and we can determine the variance with a high precision. However this is not always the case, the size  $n$  of the data set is not always large, and we must modify the procedure to take into account the actual distribution of the fluctuations about the “true” mean. The answer is provided by Student's  $t$ , as outlined in the previous section.

## 12.7 What's the use of all this?

Taking decision is important in many fields, and in physics this happens in many cases, e.g., when you claim a discovery. Therefore it is important to grasp the principles of hypothesis testing, while you wait for that important discovery that awaits you in a brilliant research work ...

## Appendix: correlation of sample mean and sample variance for independent, normally distributed samples

In the text I assert the lack of correlation of sample mean and sample variance for normally distributed, independent samples, and here I prove this statement. We recall at once that in chapter 7 we found that

$$\mathbf{E} (x_k - \mu_S)^2 = \frac{n-1}{n} \sigma^2 \quad (12.35)$$

with

$$\mu_S = \frac{1}{n} \sum_k x_k \quad (12.36)$$

Now we must evaluate

$$\mathbf{E} [\mu_S \sigma_S^2] = \frac{1}{n(n-1)} \mathbf{E} \sum_{k,l} [x_k (x_l - \mu_S)^2] \quad (12.37a)$$

$$= \frac{1}{n(n-1)} \sum_{k,l} \mathbf{E} (x_k x_l^2 - 2x_k x_l \mu_S + x_k \mu_S^2) \quad (12.37b)$$

$$= \frac{1}{n(n-1)} \left[ \sum_{k,l} \mathbf{E}(x_k x_l^2) - \frac{2}{n} \sum_{k,l,m} \mathbf{E}(x_k x_l x_m) + \frac{1}{n} \sum_{k,l,m} \mathbf{E}(x_k x_l x_m) \right] \quad (12.37c)$$

$$= \frac{1}{n(n-1)} \left[ \sum_{k,l} \mathbf{E}(x_k x_l^2) - \frac{1}{n} \sum_{k,l,m} \mathbf{E}(x_k x_l x_m) \right] \quad (12.37d)$$

Now we concentrate on the individual terms in this sum:

- $\mathbf{E}(x_k x_l^2)$

$$\mathbf{E}(x_k x_l^2) = \mathbf{E} \{ [(x_k - \mu) + \mu][(x_l - \mu) + \mu]^2 \} \quad (12.38a)$$

$$= \mathbf{E} \{ (x_k - \mu)(x_l - \mu)^2 + 2\mu(x_k - \mu)(x_l - \mu) + \mu^2(x_k - \mu) \} \quad (12.38b)$$

$$+ \mu(x_l - \mu)^2 + 2\mu^2(x_l - \mu) + \mu^3 \quad (12.38c)$$

$$= 2\mu\sigma^2\delta_{k,l} + \mu\sigma^2 + \mu^3 \quad (12.38d)$$

- $\mathbf{E}(x_k x_l x_m)$

$$\mathbf{E}(x_k x_l x_m) = \mathbf{E}\{(x_k - \mu) + \mu][(x_l - \mu) + \mu][(x_m - \mu) + \mu]\} \quad (12.39a)$$

$$= \mathbf{E}\{(x_k - \mu)(x_l - \mu)(x_m - \mu) + \mu(x_l - \mu)(x_m - \mu) \quad (12.39b)$$

$$+ \mu(x_k - \mu)(x_m - \mu) + \mu^2(x_m - \mu)\} \quad (12.39c)$$

$$+ \mu(x_k - \mu)(x_l - \mu) + \mu^2(x_l - \mu) + \mu^2(x_k - \mu) + \mu^3\} \quad (12.39d)$$

$$= \mu\sigma^2\delta_{l,m} + \mu\sigma^2\delta_{k,m} + \mu\sigma^2\delta_{k,l} + \mu^3 \quad (12.39e)$$

Then

$$\mathbf{E}[\mu_S \sigma_S^2] = \frac{1}{n(n-1)} \left[ \sum_{k,l} \mathbf{E}(x_k x_l^2) - \frac{1}{n} \sum_{k,l,m} \mathbf{E}(x_k x_l x_m) \right] \quad (12.40a)$$

$$= \frac{1}{n(n-1)} \left[ \sum_{k,l} (2\mu\sigma^2\delta_{k,l} + \mu\sigma^2 + \mu^3) - \frac{1}{n} \sum_{k,l,m} (\mu\sigma^2\delta_{l,m} + \mu\sigma^2\delta_{k,m} + \mu\sigma^2\delta_{k,l} + \mu^3) \right] \quad (12.40b)$$

$$= \frac{1}{n(n-1)} [(2n\mu\sigma^2 + n^2\mu\sigma^2 + n^2\sigma^3) - (3n\sigma^2 + n^2\mu^3)] \quad (12.40c)$$

$$= \mu\sigma^2 \quad (12.40d)$$

Finally

$$\text{cov}(\mu_S, \sigma_S^2) = \mathbf{E}[\mu_S \sigma_S^2] - \mathbf{E}(\mu_S)\mathbf{E}(\sigma_S) = 0 \quad (12.41)$$

i.e., in the case of normally distributed, independent samples, sample mean and sample variance are uncorrelated.

# Bibliography

- [1] John Aldrich. Ra fisher and the making of maximum likelihood 1912–1922. *Statistical Science*, 12(3):162–176, 1997.
- [2] SR Amendolia, B Badelek, G Batignani, GA Beck, EH Bellamy, E Bertolucci, D Bettoni, H Bilokon, A Bizzeti, G Bologna, et al. Measurement of the pion form factor in the time-like region for  $|q_1^2|$  values between 0.1 (gev/c) $^2$  and 0.18 (gev/c) $^2$ . *Physics Letters B*, 138(5):454–458, 1984.
- [3] T. Apostol. *Calculus*, volume 2. Wiley, 1969.
- [4] Gely P Basharin, Amy N Langville, and Valeriy A Naumov. The life and work of A. A. Markov. *Linear Algebra and its Applications*, 386:3–26, 2004.
- [5] I.A. Cosma and L. Evers. Markov chains and monte carlo methods. *African Institute for Mathematical Sciences, Cape Town*, 2010.
- [6] D.R. Cox and H.D. Miller. *The theory of stochastic processes*, volume 134. Chapman & Hall/CRC, 1977.
- [7] F. N. David. *Games, Gods and Gambling: A History of Probability and Statistical Ideas*. Dover, 1968.
- [8] Z.E. Dell and S.V. Franklin. The Buffon–Laplace needle problem in three dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):P09010, 2009.
- [9] Paul Ehrenfest and Tatiana Ehrenfest. Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. *Physikalische Zeitschrift*, 8(9):311–314, 1907.
- [10] R. Engstrom. Burle Photomultiplier Handbook (a reprint of the RCA photomultiplier manual. Technical report, Burle Industries Inc, Tube Products Division. Lancaster USA, 1980.

- [11] Gary J Feldman and Robert D Cousins. Unified approach to the classical statistical analysis of small signals. *Physical Review D*, 57(7):3873, 1998.
- [12] Thomas S Ferguson. An inconsistent maximum likelihood estimate. *Journal of the American Statistical Association*, 77(380):831–834, 1982.
- [13] Ramon Ferrer-i Cancho and Brita Elvevåg. Random texts do not exhibit the real zipf's law-like rank distribution. *PLoS One*, 5(3):e9411, 2010.
- [14] Bernard D Flury. Acceptance-rejection sampling made easy. *SIAM Review*, 32(3):474–476, 1990.
- [15] B.V. Gnedenko. *Teoria della probabilità*. Ed. Riuniti, 1979.
- [16] C.M. Grinstead and J.L. Snell. *Introduction to probability*. Amer Mathematical Society, 1997.
- [17] T.E. Harris. *The theory of branching processes*. Dover Publications, 2002.
- [18] D. Hawkins and S. Ulam. Theory of multiplicative processes, i. *Los Alamos Scientific Laboratory, LADC-265*, 1944.
- [19] R. Hersh and R. J. Griego. Brownian Motion and Potential Theory. *Sci. Am.*, 220:67–74, 1969.
- [20] Frederick James. *Statistical methods in experimental physics*. World Scientific, 2006.
- [21] Edwin T Jaynes. The well-posed problem. *Foundations of Physics*, 3(4):477–492, 1973.
- [22] C.H. Kraft and L. M LeCam. A remark on the roots of the maximum likelihood equation. *Univ. of California Pub. in Stat.*, 1:277, 1953.
- [23] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [24] R. Michel. On Stirling's Formula. *Am. Math. Monthly*, 109:388–390, 2002.
- [25] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.
- [26] Donald H Perkins. *Introduction to high energy physics*. Cambridge University Press, 4th edition edition, 2000.

- [27] Y.A. Rozanov. *Probability theory: a concise course*. Dover Publications, 1977.
- [28] Richard Ruggles and Henry Brodie. An Empirical Approach to Economic Intelligence in World War II. *Journal of the American Statistical Association*, 42(237):72–91, 1947.
- [29] F. Sauli. *Principles of operation of multiwire proportional and drift chambers*, volume 77. Cern, 1977.
- [30] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27(3):379, 1948.
- [31] Jonathan AC Sterne and George Davey Smith. Sifting the evidence—what's wrong with significance tests? *Physical Therapy*, 81(8):1464–1469, 2001.
- [32] J. Tabak. *Probability and Statistics: The Science of Uncertainty*. FactsOnFile, 2004.
- [33] B. J. West and M. Schlesinger. The noise in natural phenomena. *Am. Sci.*, 78:40, 1990.
- [34] H. Wilf. *Generatingfunctionology*. Academic Press (New York), 1994.





# Index

- acceptance region, 302
- acceptance-rejection method, 184
- addition law for conditional probabilities, 21
- addition law for probabilities, 16
  - mutually exclusive events, 17
  - non-mutually exclusive events, 19
- additive processes, 120
- Allan variance, 166
- average, 42
- Bartlett identities, 236
- Bartlett identity
  - first, 237
  - second, 237
- Bayes' theorem, 23
- Bell's inequality, 95
- Bernoulli variate, 48
- Bernoulli's lemniscate, 171
- Berry-Esseen theorem, 120
- Bertrand's paradox, 26, 90
- Bethe formula, 191
- Binet's formula, 104
- Boltzmann entropy, 242
- Boltzmann's H-theorem, 142
- Boost C++ Library, 176
- bootstrap, 288
- Borel-Cantelli lemma
  - first, 32, 50
  - second, 33
- Box-Müller algorithm, 183
- Buffon's needle, 41
- Cauchy-Schwartz inequality, 167
- cell plating statistics, 66
- central limit theorem, 101, 118
- CERNLIB, 181
- characteristic function
  - sum of random variates, 112
- characteristic function, 112
- binomial distribution, 116
- exponential distribution, 116
- normal distribution, 114, 116
- Poisson distribution, 116
- series expansion, 113
- uniform distribution, 115
- Chebyshev's inequality, 47, 49
  - generalized, 49
- Chernoff bound, 53
- CLT, see central limit theorem, 101
- combinations, 7
- conditional probability, 20, 21
- confidence belt, 254
- confidence intervals, 249
  - classical construction, 254
  - frequentist, 254
  - ML estimators, 261
- confidence level, 254
  - Bayesian, 250
- consistency of maximum likelihood estimators, 235
- correlation coefficient, 83, 152, 259
- covariance, 45
- coverage, 249
- Cramér-Rao-Fisher bound, 238
- credible intervals, 249
- critical region, 302
- De Moivre-Laplace theorem, 69
- dead time, 67
- decision boundary, 302
- descriptive statistics, 149

- detailed balance, 139
- detector
  - nonparalyzable, 68
  - paralyzable, 68
- diehard tests of randomness, 178
- diffusion equation, 24
- dispersion, see variance, 45
- distribution
  - Bernoulli, 58, 248
  - beta, 88
  - binomial, 58, 105, 116
  - Cauchy-Lorentz (Breit-Wigner), 85
  - chi-square, 281, 283, 293, 307
  - exponential, 66, 116, 231, 239, 248, 271
  - gamma, 84, 155, 285
  - Gaussian, see normal distribution, 69
  - Landau, 87
  - Laplace, 89
  - log-normal, 83, 120
  - logistic, 89
  - multinomial, 60
  - multivariate normal, 81
  - normal, 69, 71, 106, 114, 116, 183, 235, 247
  - Poisson, 63, 104, 106, 114, 116
  - Poisson (high-rate limit), 74
  - power-law, 272
  - Rayleigh, 87
  - uniform, 40, 44, 57, 114, 115, 165
- distribution function, 39
- Ehrenfest urn model, 129
- electron scattering, 190
- entropy, 165
- error function, 74
- error propagation, 78
- estimator, 149
  - bias, 233, 234
  - biased, 151
  - consistency, 233, 235
  - efficiency, 236
- estimation error, 233
- mean square error, 233
- MSE, 233
- unbiased, 150
- variance, 233
- event
  - complementary, 16
  - impossible, 15
  - incompatible, 16
  - mutually exclusive, 16
  - sure, 15
- event intersection, 16
- event union, 16
- expectation, 42
- extended maximum likelihood, 263, 265, 268, 270, 271, 275
- Fibonacci sequence, 103
- Fisher information, 246
- fit
  - nonlinear, 293
  - polynomial, 294
  - straight line, 285, 287
- Galton-Watson process, 108
- gambler's ruin, 23
- generating functions, 101
- Gibbs entropy, 243
- GNU Scientific Library, 176
- goodness-of-fit, 307
- H-theorem, 142
- Hessian matrix, 280
- Hidden Markov Models, 143
- HMM, 143
- hypersphere
  - surface, 299
  - volume, 298
- hypothesis
  - alternative, 301
  - composite, 301
  - null, 301
  - simple, 301
- indicator random variable, 58

- information
  - Fisher, 246
  - Kullback-Leibler, 244
  - Shannon, 243
- information measures, 242
- interval estimate, 229
- Kullback-Leibler divergence, 244
- kurtosis, 114, 115
- law of large numbers, 5
  - strong, 50
  - weak, 47
- least squares, 279
  - binned data, 294
  - general linear model, 292
  - nonlinear, 293
- lemniscate, 171
- linear congruential generator, 177
- logarithmic binning, 276
- lottery ticket problem, 66
- Markov chain, 125
  - homogeneous, 125
  - n-step transition kernel, 127
  - state classification, 131
  - stationary distribution, 137
- Markov processes, 125
- Markov property, 125
- Markov's inequality, 49
- mathematical expectation, 42
- mean square value, 45
- mean value, 42
- median, 115
- Mellin convolution, 78
- mixture model, 215
- MLE, 231
- modal value, 115
- mode, 115
- moment generating function, 46
- moments about the mean, 46, 113, 114
- moments of a distribution, 46, 112
- moments of a probability distribution, 113
- multiplicative linear congruential generator, 177
- multiplicative processes, 120
- n-dimensional volume element, 99
- Netlib, 176
- Newton-Raphson algorithm, 232, 273
- Neyman-Pearson lemma, 305
- Numerical Recipes, 175
- order statistics, 161
- p-value, 305, 309
- pdf, see probability density function, 39
- Pearson's correlation coefficient, see correlation coefficient, 153
- permutations, 6
- photomultiplier noise statistics, 107
- point estimate, 229, 233
- power law, 121
- power of a test, 303
- principle of maximum likelihood, 229
- PRN, see pseudo-random numbers, 175
- probability density
  - joint, 39
- probability density function, 39
- probability distribution, 39
- probability generating function, 104, 112
  - binomial distribution, 105
  - multivariate, 117
  - Poisson distribution, 104, 106
  - random sum of random variates, 105
    - sum of random variates, 105
- pseudo-random numbers, 175
- random variable, 37
  - Bernoulli, 58
- random variables
  - orthogonal transformations, 79
  - product, 78
  - sum, 75

transformations, 76  
random walk, 25, 128, 134  
RANDU, 177, 180  
RANLUX, 181  
relative frequency, 5  
return probability, 132

sample covariance, 151  
sample mean, 149  
exponentially distributed samples,  
    154  
    variance of, 151  
sample space, 15  
sample variance, 149  
sampling  
    with replacement, 6  
    without replacement, 7  
Shannon information, 243  
Shannon's entropy, 165  
signal enhancement, 268  
significance level, 302  
significance of a signal, 309  
skewness, 114  
specificity, 302  
stable distribution, 117  
standard deviation, 45  
    inadequacy of, 164  
statistical bootstrap, 214  
statistical entropy, 243  
statistical independence, 22  
statistics  
    Bose-Einstein, 8  
    Fermi-Dirac, 8  
Stirling's formula, 8  
stochastic matrix, 126  
Student's  $t$ , 316

test statistic, 302  
time reversal, 139  
transformation method, 182

variance, 45, 114  
variante, see also random variable, 37

weak convergence theorem, 106, 117

Zipf's law, 121