Lecture Notes in Physics 909

Luca Lista

Statistical Methods for Data Analysis in Particle Physics

Springer

Lecture Notes in Physics

Volume 909

The Lecture Notes in Physics

The series Lecture Notes in Physics (LNP), founded in 1969, reports new developments in physics research and teaching-quickly and informally, but with a high quality and the explicit aim to summarize and communicate current knowledge in an accessible way. Books published in this series are conceived as bridging material between advanced graduate textbooks and the forefront of research and to serve three purposes:

- to be a compact and modern up-to-date source of reference on a well-defined topic
- to serve as an accessible introduction to the field to postgraduate students and nonspecialist researchers from related areas
- to be a source of advanced teaching material for specialized seminars, courses and schools

Both monographs and multi-author volumes will be considered for publication. Edited volumes should, however, consist of a very limited number of contributions only. Proceedings will not be considered for LNP.

Volumes published in LNP are disseminated both in print and in electronic formats, the electronic archive being available at springerlink.com. The series content is indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia.

Proposals should be sent to a member of the Editorial Board, or directly to the managing editor at Springer:

Christian Caron
Springer Heidelberg
Physics Editorial Department I
Tiergartenstrasse 17
69121 Heidelberg/Germany
christian.caron@springer.com

More information about this series at
http://www.springer.com/series/5304

Luca Lista

Statistical Methods for Data Analysis in Particle Physics

Springer

Luca Lista
INFN Sezione di Napoli
Napoli, Italy

ISSN 0075-8450 ISSN 1616-6361 (electronic)
Lecture Notes in Physics
ISBN 978-3-319-20175-7 ISBN 978-3-319-20176-4 (eBook)
DOI 10.1007/978-3-319-20176-4

Library of Congress Control Number: 2015944537

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

The following notes collect material from a series of lectures presented during a course on "Statistical methods for data analysis" I gave for Ph.D. students in physics at the University of Naples "Federico II" from 2009 to 2014. The aim of the course was to elaborate on the main concepts and tools that allow to perform statistical analysis of experimental data.

An introduction to probability theory and basic statistics is provided, which serves as a refresher course for students who did not take a formal course on statistics before starting their Ph.D.

Many of the statistical tools that have been covered have applications in high energy physics (HEP), but their scope could well range outside the boundaries of HEP.

A shorter version of the course was presented at CERN in November 2009 as lectures on statistical methods in LHC data analysis for the ATLAS and CMS experiments. The chapter discussing upper limits was improved after lectures on the subject I gave in Autrans, France, in May 2012 at the IN2P3 School of Statistics. I was also invited to conduct a seminar on statistical methods in Gent University, Belgium, in October 2014, which gave me an opportunity to review some of my material and add new examples.

Napoli, Italy — Luca Lista

Contents

1 Probability Theory 1
- 1.1 The Concept of Probability 1
- 1.2 Classical Probability 2
- 1.3 Issues with the Generalization to the Continuum 4
  - 1.3.1 The Bertrand's Paradox 5
- 1.4 Axiomatic Probability Definition 6
- 1.5 Probability Distributions 6
- 1.6 Conditional Probability and Independent Events 7
- 1.7 Law of Total Probability 8
- 1.8 Average, Variance and Covariance 9
- 1.9 Variables Transformations 12
- 1.10 The Bernoulli Process 13
- 1.11 The Binomial Process 14
  - 1.11.1 Binomial Distribution and Efficiency Estimate 15
- 1.12 The Law of Large Numbers 18

References 19

2 Probability Distribution Functions 21
- 2.1 Definition of Probability Distribution Function 21
- 2.2 Average and Variance in the Continuous Case 22
- 2.3 Cumulative Distribution 23
- 2.4 Continuous Variables Transformation 24
- 2.5 Uniform Distribution 24
- 2.6 Gaussian Distribution 26
- 2.7 Log-Normal Distribution 27
- 2.8 Exponential Distribution 28
- 2.9 Poisson Distribution 31
- 2.10 Other Distributions Useful in Physics 34
  - 2.10.1 Argus Function 34
  - 2.10.2 Crystal Ball Function 35
  - 2.10.3 Landau Distribution 37

2.11 Central Limit Theorem 38
2.12 Convolution of Probability Distribution Functions 40
2.13 Probability Distribution Functions in More than One Dimension ... 41
2.13.1 Marginal Distributions 41
2.13.2 Conditional Distributions 45
2.14 Gaussian Distributions in Two or More Dimensions 46
References 51

3 Bayesian Approach to Probability 53
3.1 Bayes' Theorem 53
3.2 Bayesian Probability Definition 58
3.3 Bayesian Probability and Likelihood Functions 60
3.3.1 Repeated Use of Bayes' Theorem and Learning Process 61
3.4 Bayesian Inference 62
3.5 Bayes Factors 65
3.6 Arbitrariness of the Prior Choice 66
3.7 Jeffreys' Prior 67
3.8 Error Propagation with Bayesian Probability 68
References 68

4 Random Numbers and Monte Carlo Methods 69
4.1 Pseudorandom Numbers 69
4.2 Pseudorandom Generators Properties 69
4.3 Uniform Random Number Generators 71
4.3.1 Remapping Uniform Random Numbers 72
4.4 Non Uniform Random Number Generators 72
4.4.1 Gaussian Generators Using the Central Limit Theorem ... 73
4.4.2 Non-uniform Distribution From Inversion of the Cumulative Distribution 73
4.4.3 Gaussian Numbers Generation 75
4.5 Monte Carlo Sampling 76
4.5.1 Hit-or-Miss Monte Carlo 76
4.5.2 Importance Sampling 78
4.6 Numerical Integration with Monte Carlo Methods 79
References 80

5 Parameter Estimate 81
5.1 Measurements and Their Uncertainties 82
5.2 Nuisance Parameters and Systematic Uncertainties 84
5.3 Estimators 84
5.4 Properties of Estimators 85
5.4.1 Consistency 85
5.4.2 Bias 86
5.4.3 Minimum Variance Bound and Efficiency 86
5.4.4 Robust Estimators 87

5.5 Maximum-Likelihood Method 87
5.5.1 Likelihood Function 88
5.5.2 Extended Likelihood Function 89
5.5.3 Gaussian Likelihood Functions 91
5.6 Errors with the Maximum-Likelihood Method 91
5.6.1 Properties of Maximum-Likelihood Estimators 94
5.7 Minimum χ^2 and Least-Squares Methods 95
5.7.1 Linear Regression 97
5.7.2 Goodness of Fit 98
5.8 Error Propagation 99
5.8.1 Simple Cases of Error Propagation 100
5.9 Issues with Treatment of Asymmetric Errors 101
5.10 Binned Samples 104
5.10.1 Minimum-χ^2 Method for Binned Histograms 105
5.10.2 Binned Poissonian Fits 105
5.11 Combining Measurements 106
5.11.1 Weighted Average 107
5.11.2 χ^2 in n Dimensions 108
5.11.3 The Best Linear Unbiased Estimator 108
References 111

6 Confidence Intervals 113
6.1 Neyman's Confidence Intervals 113
6.1.1 Construction of the Confidence Belt 113
6.1.2 Inversion of the Confidence Belt 115
6.2 Binomial Intervals 116
6.3 The "Flip-Flopping" Problem 117
6.4 The Unified Feldman–Cousins Approach 119
References 121

7 Hypothesis Tests 123
7.1 Introduction to Hypothesis Tests 123
7.2 Fisher's Linear Discriminant 126
7.3 The Neyman–Pearson Lemma 128
7.4 Likelihood Ratio Discriminant 129
7.5 Kolmogorov–Smirnov Test 129
7.6 Wilks' Theorem 131
7.7 Likelihood Ratio in the Search for a New Signal 133
References 135

8 Upper Limits 137
8.1 Searches for New Phenomena: Discovery and Upper Limits 137
8.2 Claiming a Discovery 138
8.2.1 The p-Value 138
8.2.2 Significance 139
8.2.3 Significance and Discovery 140

8.3 Excluding a Signal Hypothesis 141
8.4 Significance and Parameter Estimates Using Likelihood Ratio 141
8.4.1 Significance Evaluation with Toy Monte Carlo 142
8.5 Definitions of Upper Limits 143
8.6 Poissonian Counting Experiments 143
8.6.1 Simplified Significance Evaluation for Counting Experiments 144
8.7 Bayesian Approach 144
8.7.1 Bayesian Upper Limits for Poissonian Counting 145
8.7.2 Limitations of the Bayesian Approach 146
8.8 Frequentist Upper Limits 147
8.8.1 The Counting Experiment Case 148
8.8.2 Upper Limits from Neyman's Confidence Intervals 149
8.8.3 Frequentist Upper Limits on Discrete Variables 149
8.8.4 Feldman–Cousins Upper Limits for Counting Experiments 151
8.9 Can Frequentist and Bayesian Upper Limits Be "Unified"? 153
8.10 Modified Frequentist Approach: The CL_s Method 154
8.11 Incorporating Nuisance Parameters and Systematic Uncertainties 157
8.11.1 Nuisance Parameters with the Bayesian Approach 157
8.11.2 Hybrid Treatment of Nuisance Parameters 158
8.12 Upper Limits Using the Profile Likelihood 159
8.13 Variations of Profile-Likelihood Test Statistics 160
8.13.1 Test Statistic for Positive Signal Strength 161
8.13.2 Test Statistics for Discovery 161
8.13.3 Test Statistics for Upper Limits 161
8.13.4 Higgs Test Statistic 162
8.13.5 Asymptotic Approximations 162
8.14 The Look-Elsewhere Effect 169
8.14.1 Trial Factors 170
References 171

List of Figures

Fig. 1.1 Probability of different outcomes of the sum of two dices 4
Fig. 1.2 Illustration of Betrand's paradox: three different choices of *random extraction* of a chord in the circle lead apparently to probabilities that the cord is longer than the inscribed triangle's side of $\frac{1}{2}$ (*left*), $\frac{1}{3}$ (*center*) and $\frac{1}{4}$ (*right*) respectively 5
Fig. 1.3 If two events, A and B, are represented as sets, the conditional probability $P(A|B)$ is equal to the area of the intersection, $A \cap B$, divided by the area of B 7
Fig. 1.4 Visual example of partition of an event set 9
Fig. 1.5 Visualization of the law of total probability 9
Fig. 1.6 A set of $n = 10$ balls, of which $r = 3$ are *red* determines a Bernoulli process. The probability to randomly extract a *red ball* in the set shown in this figure is $p = r/n = 3/10 = 30\,\%$ 13
Fig. 1.7 Building a binomial process as subsequent random extractions of a single *red or white ball* (Bernoulli process). The "tree" in figure displays all the possible combinations at each extraction step. Each of the "branching" has a corresponding probability equal to p or $1-p$ for *red and white ball* respectively. The number of paths corresponding to each possible combination is shown in parentheses, and is equal to a binomial coefficient 15
Fig. 1.8 The Yang Hui triangle published in 1303 by Zhu Shijie (1260–1320) 16
Fig. 1.9 Binomial distribution for $N = 15$ and for $p = 0.2$, 0.5 and 0.8 17

Fig. 1.10 An illustration of the law of large numbers using a simulation of die rolls. The average of the first N out of 1000 random extraction is reported as a function of N. The 1000 extractions have been repeated twice (*red and blue lines*) 18

Fig. 2.1 Uniform distributions with different extreme values 25
Fig. 2.2 Gaussian distributions with different values of average and standard deviation parameters 26
Fig. 2.3 Lognormal distributions with different parameters μ and σ 28
Fig. 2.4 Exponential distributions with different values of the parameter λ 29
Fig. 2.5 An horizontal axis representing occurrence times (*dots*) of some events. The time origin ($t_0 = 0$) is marked as a *cross* 29
Fig. 2.6 Poisson distributions with different rate parameter 32
Fig. 2.7 A uniform distribution of events in a variable x. Two intervals are shown of sizes x and X, where $x \ll X$ 32
Fig. 2.8 Argus distributions with different parameter values 35
Fig. 2.9 Crystal ball distributions with $\mu = 0$, $\sigma = 1$, $n = 2$ and different values of the tail parameter α 36
Fig. 2.10 Landau distributions with $\mu = 0$, and different values of σ 37
Fig. 2.11 Approximate visual demonstration of the central limit theorem using a Monte Carlo technique. A random variable x_1 is generated uniformly in the interval $[-\sqrt{3}, \sqrt{3}[$, in order to have average value $\mu = 0$ and standard deviation $\sigma = 1$. The *top-left* plot shows the distribution of 10^5 random extractions of x_1; the other plots show 10^5 random extractions of $(x_1 + x_2)/\sqrt{2}$, $(x_1 + x_2 + x_3)/\sqrt{3}$ and $(x_1 + x_2 + x_3 + x_4)/\sqrt{4}$ respectively, where all x_i obey the same uniform distribution as x_1. A Gaussian curve with $\mu = 0$ and $\sigma = 1$, with proper normalization in order to match the sample size, is superimposed to the extracted sample in the four cases. The Gaussian approximation is more and more stringent as a larger number of variables is added 38
Fig. 2.12 Same as Fig. 2.11, using a PDF that is uniformly distributed in two disjoint intervals, $[-\frac{3}{2}, -\frac{1}{2}[$ and $[\frac{1}{2}, \frac{3}{2}[$, in order to have average value $\mu = 0$ and standard deviation $\sigma = 1$. The sum of 1, 2, 3, 4, 6 and 10 independent random extractions of such a variable, divided by $\sqrt{n}$, $n = 1, 2, 3, 4, 6, 10$ respectively, are shown with a Gaussian distribution having $\mu =$ and $\sigma = 1$ superimposed 39

Fig. 2.13 In the two-dimensional plane (x, y), a slice in x corresponds to a probability $\delta P(x) = f_x(x)\delta x$, a slice in y corresponds to a probability $\delta P(y) = f_y(y)\delta y$, and their intersection to a probability $\delta P(x, y) = f(x, y)\delta x\delta y$ 42
Fig. 2.14 Example of a PDF of two variables x and y that are uncorrelated but are not independent 44
Fig. 2.15 Illustration of conditional PDF in two dimensions 46
Fig. 2.16 Plot of the one-sigma contour for a two-dimensional Gaussian PDF. The two ellipse axes are equal to $\sigma_{x'}$ and $\sigma_{y'}$; the x' axis is rotated of an angle ϕ with respect to the x axis and the lines tangent to the ellipse parallel to the x and y axes have a distance with respect to the respective axes equal to σ_y and σ_x respectively 48
Fig. 2.17 Plot of the two-dimensional one- and two-sigma Gaussian contours 49

Fig. 3.1 Visualization of the conditional probabilities, $P(A|B)$ and $P(B|A)$ due to Robert Cousins. The events A and B are represented as subsets of a sample space Ω 54
Fig. 3.2 Visualization of the Bayes' theorem due to Robert Cousins. The areas of events A an B, $P(A)$ and $P(B)$ respectively, simplify as we multiply $P(A|B)P(B)$ and $P(B|A)P(A)$ 55
Fig. 3.3 Visualization of the ill/healthy case considered in the Example 3.7. The *red areas* correspond to the cases of a positive diagnosis for a ill person ($P(+|\text{ill})$, *vertical red area*) and a positive diagnosis for a healthy person ($P(+|\text{healthy})$, *horizontal red area*). The probability of being really ill in the case of a positive diagnosis, $P(\text{ill}|+)$, is equal to the ratio of the *vertical red area* and the *total red area*. In the example it was assumed that $P(-|\text{ill}) \simeq 0$ 57

Fig. 4.1 Logistic map [2] 71
Fig. 4.2 Sketch of hit-or-miss Monte Carlo method 77
Fig. 4.3 Variation of hit-or-miss Monte Carlo using the importance sampling 78

Fig. 5.1 Relation between probability and inference 83
Fig. 5.2 Example of unbinned extended maximum-likelihood fit of a simulated dataset. The fit curve is superimposed on the data points as *solid blue line* 90
Fig. 5.3 Scan of $-2\ln L$ as a function of a parameter θ. the error interval is determined looking at the excursion of $-2\ln L$ around the minimum, at $\hat{\theta}$, finding the values for which $-2\ln L$ increases by one unit 93

Fig. 5.4 Two-dimensional contour plot showing the 1σ error ellipse for the fit parameters s and b (number of signal and background events) corresponding to the fit in Fig. 5.2 93
Fig. 5.5 Example of minimum-χ^2 fit of a simulated dataset. The points with the error bars, assumed resulting from Gaussian PDFs, are used to fit a function model of the type $y = f(x) = Axe^{-Bx}$, where A and B are unknown parameters determined form the fit procedure. The fit curve is superimposed as *solid blue line* 96
Fig. 5.6 Plot of a variable transformation $\eta = \eta(\theta)$, and visualization of the procedure of error propagation using local linear approximation 100
Fig. 5.7 Transformation of a variable x into a variable x' through a piece-wise linear transformation characterized by two different slopes. If x obeys a Gaussian PDF with standard deviation σ, x' obeys a *bifurcated* Gaussian, made of two Gaussian halves having different standard deviation parameters, σ'_+ and σ'_- 103

Fig. 6.1 Graphical illustration of Neyman's belt construction (*left*) and inversion (*right*) 114
Fig. 6.2 Three possible choices of ordering rule: central interval (*top*) and fully asymmetric intervals (*bottom left, right*) 115
Fig. 6.3 Illustration of the *flip-flopping* problem. The plot shows the quoted central value of μ as a function of the measured x (*dashed line*), and the 90 % confidence interval corresponding to a choice to quote a central interval for $x/\sigma \geq 3$ and an upper limit for $x/\sigma < 3$ 118
Fig. 6.4 Ordering rule in the Feldman–Cousins approach, based on the likelihood ratio $\lambda(x|\theta_0) = f(x|\theta_0)/f(x|\theta_{\text{best}}(x))$ 120
Fig. 6.5 Neyman confidence belt constructed using the Feldman–Cousins ordering 121

Fig. 7.1 Probability distribution functions for a discriminating variable $t(x) = x$ which has two different PDFs for the signal (*red*) and background (*yellow*) hypotheses under test 124
Fig. 7.2 Signal efficiency versus background misidentification probability 125
Fig. 7.3 Examples of two-dimensional selections of a signal (*blue dots*) against a background (*red dots*). A linear cut is chosen on the *left plot*, while a box cut is chosen on the *right plot* 126

Fig. 7.4 Graphical representation of the Kolmogorov–Smirnov test 130

Fig. 8.1 Poisson distribution in the case of a null signal and an expected background of $b = 4.5$. The probability corresponding to $n \geq 8$ (*red area*) is 0.087, and gives a p-value assuming the event counting as test statistics 139

Fig. 8.2 Upper limits at the 90 % CL (*top*) and 95 % CL (*bottom*) for Poissonian process using the Bayesian approach as a function of the expected background b and for number of observed events n from $n = 0$ to $n = 10$ 147

Fig. 8.3 Graphical illustration of Neyman's belt construction for upper limits determination 149

Fig. 8.4 Poisson distribution in the case of a signal $s = 4$ and $b = 0$. The *white bins* show the smallest possible fully asymmetric confidence interval ($\{2, 3, 4, \cdots\}$ in this case) that gives at least the coverage of $1 - \alpha = 90\,\%$ 150

Fig. 8.5 90 % Confidence belt for a Poissonian process using Feldman–Cousins ordering, in the case of $b = 3$ 151

Fig. 8.6 Upper limits at 90 % confidence belt for Poissonian process using Feldman–Cousins ordering as a function of the expected background b and for number of observed events n from 0 to 10 152

Fig. 8.7 Example of determination of CL_s from pseudoexperiments. The distribution of the test statistics $-2 \ln \lambda$ is shown in *blue* assuming the signal-plus-background hypothesis and in red assuming the background-only hypothesis. The *black line* shows the value of the test statistics measured in data, and the hatched areas represent CL_{s+b} (*blue*) and $1 - \mathrm{CL}_b$ (*red*) 156

Fig. 8.8 A toy Monte Carlo data sample superimposed to an exponential background model (*left*) and to an exponential background model plus a Gaussian signal (*right*) 164

Fig. 8.9 Distribution of the test statistic q for the background-only hypothesis (*blue*) and for the signal-plus-background hypothesis (*red*). Superimposed is the value determined with the presented data sample (*black arrow*). p-values can be determined from the *shaded areas* of the two PDFs 166

Fig. 8.10 Scan of the test statistic q as a function of the parameter of interest μ 166

Fig. 8.11 Toy data sample superimposed to an exponential background model (*top*) and to an exponential background model plus a Gaussian signal (*bottom*) adding a 10 % uncertainty to the background normalization 168

Fig. 8.12 Scan of the test statistic q as a function of the parameter of interest μ including systematic uncertainty on β (*dark brown*) compared with the case with no uncertainty (*red*) 169
Fig. 8.13 Visual illustration of upcrossing, computed to determine $\langle N_{u_0} \rangle$. In this example, we have $N_u = 3$ 171

List of Examples

Example 1.1 Combination of Detector Efficiencies 8
Example 1.2 Application to Sum of Dice Rolls 12

Example 2.3 Strip Detectors 25
Example 2.4 Relation Between Uniform and Exponential Distributions ... 29
Example 2.5 Relation Between Uniform, Binomial and Poisson Distributions 32
Example 2.6 Uncorrelated Variables that Are Not Independent 44

Example 3.7 An Epidemiology Example 55
Example 3.8 Purity of a Sample with Particle Identification 57
Example 3.9 Extreme Cases of Prior Beliefs 59
Example 3.10 Posterior for a Poisson Distribution 64

Example 4.11 Transition From Regular to "Unpredictable" Sequences 70
Example 4.12 Extraction of an Exponential Random Variable 74
Example 4.13 Extraction of a Uniform Point on a Sphere 75
Example 4.14 Combined Application of Different Monte Carlo Techniques 79

Example 5.15 A Very Simple Estimator in a Gaussian Case 84
Example 5.16 Estimators with Variance Below the Cramér–Rao Bound are Not Consistent 86
Example 5.17 Maximum-Likelihood Estimate for an Exponential Distribution 94
Example 5.18 Bias of the Maximum-Likelihood Estimate of a Gaussian Variance 95
Example 5.19 Reusing Multiple Times the Same Measurement Doesn't Improve a Combination 110

Example 6.20 Application of the Clopper–Pearson Method 117

Example 8.21 p-Value for a Poissonian Counting 139
Example 8.22 A Realistic Example of Bump Hunting 163
Example 8.23 Adding a Systematic Uncertainty..... 167

List of Tables

Table 1.1 Possible combinations of two dice roll (center column) leading a given sum (left column) and the corresponding probability (right column) 3

Table 2.1 Probabilities corresponding to $Z\sigma$ one-dimensional and two-dimensional contours for different values of Z 50

Table 3.1 Assessing evidence with Bayes factors according to the scale proposed in [4] 65

Table 3.2 Jeffreys' priors corresponding to the parameters of some of the most frequently used PDFs 67

Table 8.1 Significances expressed as "$Z\sigma$" and corresponding p-values in a number of typical cases 140

Table 8.2 Upper limits in presence of negligible background evaluated under the Bayesian approach for different number of observed events n 146

Table 8.3 Upper and lower limits in presence of negligible background ($b = 0$) obtained using the Feldman–Cousins approach 152

Chapter 1
Probability Theory

1.1 The Concept of Probability

Many processes in nature have uncertain outcomes. This means that their result cannot be predicted before the process occurs. A *random* process is a process that can be reproduced, to some extent, within some given boundary and initial conditions, but whose outcome is uncertain. This situation may be due to insufficient information about the process intrinsic dynamics which prevents to predict its outcome, or lack of sufficient accuracy in reproducing the initial conditions in order to ensure its exact reproducibility. Some processes like quantum mechanics phenomena have intrinsic randomness. This will lead to possibly different outcomes if the experiment is repeated several times, even if each time the initial conditions are exactly reproduced, within the possibility of control of the experimenter. *Probability* is a measurement of how *favored* one of the possible outcomes of such a random process is compared with any of the other possible outcomes.

There are two main different approaches to the concept of probability which result in two different meanings of probability. These are referred to as *frequentist* and *Bayesian* probabilities.

- *Frequentist* probability is defined as the fraction of the number of occurrences of an event of interest over the total number of possible events in a repeatable experiment, in the *limit* of very large number of experiments. This concept can only be applied to processes that can be repeated over a reasonably long range of time. It is meaningless to apply the frequentist concept of probability to an unknown event, like the possible values of an unknown parameter. For instance, the probability of a possible score in a sport game, or the probability that the mass of an unknown particle is larger than 200 GeV, are meaningless in the frequentist approach.
- *Bayesian* probability measures *someone's degree of belief* that a statement is true (this may also refer to future events). The quantitative definition of

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_1

Bayesian probability makes use of an extension of the Bayes theorem, and will be discussed in Sect. 3.1. Bayesian probability can be applied wherever the frequentist probability is meaningful, as well as on a wider variety of cases in which one wants to determine a probability of unknown events or of unknown quantities. For instance, after some direct and/or indirect experimental measurements, the probability that an unknown particle's mass is larger than 200 GeV could be meaningful in the Bayesian sense. Other examples of valid Bayesian probabilities that have no corresponding meaning under the frequentist approach are the outcome of a future elections, as well as statements about past but unknown events, like about uncertain features of prehistoric extinct species, and so on.

This chapter will start from classical probability theory, as formulated since the eighteenth century, and will mainly discuss frequentist probability. Bayesian probability will be the subject of Chap. 3.

1.2 Classical Probability

In 1814, Pierre-Simon Laplace wrote [1]:

> The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.

This formulation is the basis of the definition of *classical probability*. We intend as *random variable* the outcome of a repeatable experiment whose result is uncertain. An *event*[1] consists of the occurrence of a certain condition on the random variable resulting from an experiment, e.g.: a coin toss gave a head or a dice roll gave an even value. The probability of an *event* defined by the random variable having one of a number of possible favorable cases can be written, according to Laplace, as:

$$\text{Probability} = \frac{\text{Number of favorable cases}}{\text{Number of possible cases}} . \tag{1.1}$$

This approach can be used in practice only for relatively simple problems, since it assumes that all possible cases under consideration are equally probable, which may not always be the case in complex examples. Examples of cases where the classical probability can be applied are coin tossing, where the two faces of a coin

[1]Note that in physics often *event* is intended as an *elementary* event. So, the use of the term event in statistics may sometimes lead to confusion.

Table 1.1 Possible combinations of two dice roll (center column) leading a given sum (left column) and the corresponding probability (right column)

Sum	Favorable cases	Probability
2	(1, 1)	1/36
3	(1, 2), (2, 1)	1/18
4	(1, 3), (2, 2), (3, 1)	1/12
5	(1, 4), (2, 3), (3, 2), (4, 1)	1/9
6	(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)	5/36
7	(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)	1/6
8	(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)	5/36
9	(3, 6), (4, 5), (5, 4), (6, 3)	1/9
10	(4, 6), (5, 5), (6, 4)	1/12
11	(5, 6), (6, 5)	1/18
12	(6, 6)	1/36

are assumed to have equal probability, equal to $^1/_2$ each, or dice roll, where each of the six dice faces[2] has equal probability equal to $^1/_6$, and so on.

Starting from simple cases, like coins or dices, more complex models can be built using combinatorial analysis, in which, in the simplest cases, one may proceed by enumeration of all the finite number of possible cases, and again the probability of an event is given by counting the number of favorable cases and dividing it by the total number of possible cases of the combinatorial problem. An easy case of combinatorial analysis is given by the roll of two dices, taking the sum of the two outcomes as final result. The possible number of outcomes is given by the $6 \times 6 = 36$ different combinations that may lead to a sum ranging from 2 to 12. The possible combinations are enumerated in Table 1.1, and the corresponding probabilities, computed as the number of favorable cases divided by 36, are shown in Fig. 1.1.

Several examples can be treated in this way, decomposing the problem in all the possible *elementary* outcomes, and identifying an event as the occurrence of one of the elementary outcomes from a specific set of possible outcomes. For instance, the event "sum of dices = 4" will correspond to the set of possible outcomes $\{(1, 3), (2, 2), (3, 1)\}$. Other events (e.g.: "sum is an odd number", or "sum is greater than 5") may be associated to different sets of possible combinations. In general, formulating the problem in terms of sets allows to replace logical "and", "or" and "not" in the sentence defining the outcome condition by the intersection, union and complement of the corresponding sets of elementary outcomes.

It is possible to build more complex models, but as soon as the problem increases in complexity, the difficulty to manage the combinatorial analysis often rapidly grows up and may easily become hard to manage. In many realistic cases, like in the case of a particle detector where effects like efficiency, resolution, etc. must be taken into account, it may be hard to determine a number of equally probable cases, and a different approach should be followed. It can also be shown that classical

[2] Or even different from 6, like in many role-playing games that use dices of non-cubic solid shapes.

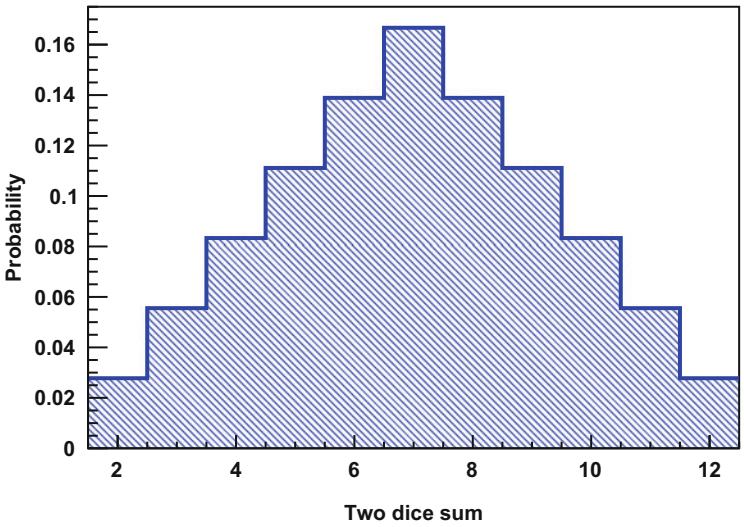

Fig. 1.1 Probability of different outcomes of the sum of two dices

probability is easy to define only for discrete cases, and the generalization to the continuum case is not unambiguously defined, as will be seen in Sect. 1.3.1.

1.3 Issues with the Generalization to the Continuum

The generalization of classical probability to continuous random variables requires to extend the definition of probability introduced in the previous Sect. 1.2. In particular, the concept of *equiprobability* can't be generalized to the continuum case in an unambiguous way. Imagine we have a continuous variable x for which we want to define an extension of the equiprobability concept we applied to the six possible outcomes of a dice. We can think of partitioning the validity range of x (say $[x_1, x_2]$) into intervals $I_1, \cdots, I_N$ all having the same size, for an arbitrarily large number N, i.e. each measuring $(x_2 - x_1)/N$. We can say that the *probability distribution* of the variable x is *uniform* if the probability that the value of x resulting from a random extraction falls into any of the N intervals is the same, i.e. it is equal to $1/N$. This definition of uniform probability of x in the interval $[x_1, x_2]$ clearly changes under reparametrization, i.e. if we transform the variable x into $y = Y(x)$ (assuming for simplicity that Y is a monotonous increasing function of x), and the interval $[x_1, x_2]$ into $[y_1, y_2] = [Y(x_1), Y(x_2)]$. In this case, the transformed intervals $J_1, \cdots, J_N = Y(I_1), \cdots, Y(I_N)$ will not have all the same size, unless the transformation Y is linear. So, if x has a uniform distribution, $y = Y(x)$ in general does not have a uniform distribution.

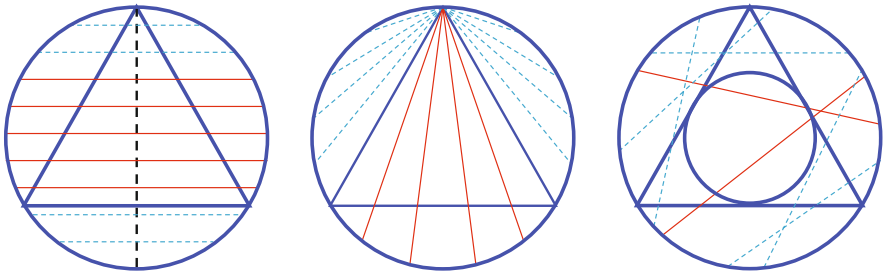

Fig. 1.2 Illustration of Betrand's paradox: three different choices of *random extraction* of a chord in the circle lead apparently to probabilities that the cord is longer than the inscribed triangle's side of $\frac{1}{2}$ (*left*), $\frac{1}{3}$ (*center*) and $\frac{1}{4}$ (*right*) respectively

1.3.1 The Bertrand's Paradox

The arbitrariness in the definition of uniform random extraction becomes evident in a famous paradox, called the Bertrand's paradox[3] that can be formulated as follows:

- Consider an equilateral triangle inscribed in a circle. Extract *uniformly* one of the possible chords of the circle. What is the probability that the length of the extracted chord is larger than the side of the inscribed triangle?

The apparent paradox arises because the *uniform* extraction of a chord on a circle is not a well-defined process. Below we give three possible example of random extractions that give apparently different probabilities.

1. Let's take the circle's diameter passing by one of the vertices of the triangle and let's extract *uniformly* a point on this diameter. Then, let's takes the chord perpendicular to the diameter passing by the extracted point. Evidently from Fig. 1.2 (left plot) one half of the extracted chords will have a length larger than the triangle's side, since the basis of the triangle cuts the vertical diameter at half the radius size. Repeating the example for all possible diameters of the circle, that we assume to extract uniformly with respect to the azimuthal angle, we can spans in this way all possible chords of the circle. Hence, the probability in question would be $P = \frac{1}{2}$.
2. Let's take, instead, one of the chords starting from one of the vertices of the triangle (Fig. 1.2, center plot) extracting *uniformly* an angle with respect to the tangent to the circle passing by that vertex. The chord is longer than the triangle's side when it intersects the basis of the triangle and shorter otherwise. This occurs in one thirds of the cases, since the angles of the equilateral triangle measure

[3]This apparent paradox is due to the French mathematician Joseph Louis François Bertrand (1822–1900).

$\pi/3$, and the chords span an angle of π. Repeating for any possible point on the circumference of the circle, one would derive that $P = \frac{1}{3}$, which is different from $P = \frac{1}{2}$ as we derived in the first case.
3. Let's extract *uniformly* a point in the circle and let's construct the chord passing by that point perpendicular to the radius that passes by the same point (Fig. 1.2, right plot). We can derive in this way that $P = \frac{1}{4}$, since the chords starting from a point contained inside (outside) the circle inscribed in the triangle would have a length longer (shorter) than the triangle's side, and the ratio of the areas of the circle inscribed in the triangle to the area of the circle that inscribes the triangle is equal to $\frac{1}{4}$.

The paradox is clearly only apparent, because the process of uniform random extraction of a chord in a circle is not univocally defined.

1.4 Axiomatic Probability Definition

An axiomatic definition of probability as a theory of measurement that is valid either in the discrete and the continuous case is due to Kolmogorov [2]. Let's consider a *measure space*, $(\Omega, F \subseteq 2^{\Omega}, P)$, where P is a function that maps elements of F, a subset of the power set 2^{Ω} of Ω (i.e.: F contains subsets of Ω), to real numbers. In literature Ω is called *sample space* and F is called *event space*. P is a *probability measure* if the following properties are satisfied:

1. $P(E) \geq 0, \forall E \in F$,
2. $P(\Omega) = 1$,
3. $\forall (E_1, \cdots, E_n) \in F^n : E_i \cup E_j = 0, P(\cup_{I=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$.

This definition allows to generalize the classical probability to the case of continuous variables, as we will see in the following sections. Bayesian probability, defined in Chap. 3, also obeys Kolmogorov's axioms.

1.5 Probability Distributions

Let's assume that a random variable x has possible values (outcomes) $x_1, \cdots, x_N$ which occur each with a probability $P(\{x_i)\}) = P(x_i)$, $i = 1, \cdots N$. We define *probability distribution* the function that associates to a possible value x_i of the *random variable* x its probability $P(x_i)$. The probability of an *event* E corresponding to a set of distinct possible elementary outcomes $\{x_{E_1}, \cdots, x_{E_K}\}$ is, according to the third Kolmogorov's axiom:

$$P(E) = \sum_{j=1}^{K} P(x_{E_j}) \,. \tag{1.2}$$

From the second Kolmogorov's axiom, it is also clear that the probability of the event Ω corresponding to the set of *all* possible outcomes, $x_1, \cdots, x_N$ must be one. Equivalently, the sum of the probabilities of all possible outcomes is equal to one:

$$\sum_{i=1}^{n} P(x_i) = 1\,. \tag{1.3}$$

This property of probability distributions is called *normalization*.

1.6 Conditional Probability and Independent Events

Given an event A and an event B, the *conditional probability*, $P(A|B)$, is defined as the probability of A given the condition that the event B has occurred, and is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,. \tag{1.4}$$

The conditional probability can be visualized in Fig. 1.3. While the probability of A, $P(A)$, corresponds to the area of the set A, relative to the area of the whole sample space Ω, which is equal to one, the conditional probability, $P(A|B)$, corresponds to the area of the *intersection* of A and B, relative to the area of the set B.

An event A is said to be *independent* on event B if the conditional probability of A, given B, is equal to the probability of A, i.e.: the occurrence of B does not change the probability of A:

$$P(A|B) = P(A)\,. \tag{1.5}$$

Two events are independent if, and only if, the probability of their simultaneous occurrence is equal to the product of their probabilities, i.e.:

$$P(\text{“}A \text{ and } B\text{”}) = P(A \cap B) = P(A)P(B)\,. \tag{1.6}$$

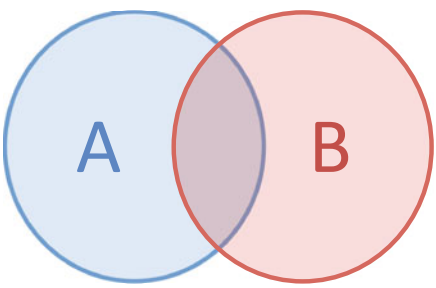

Fig. 1.3 If two events, A and B, are represented as sets, the conditional probability $P(A|B)$ is equal to the area of the intersection, $A \cap B$, divided by the area of B

Given the symmetry of Eq. (1.6), if A is independent on B, then B is also independent on A.

Example 1.1 – Combination of Detector Efficiencies

Consider an experimental apparatus made of two detectors A and B and a particle that traverses both. If the two detectors give independent signals, the probability ε_{AB} that a particle gives a signal in both detector is given by the product of the probability ε_A that the detector A gives a signal times the probability ε_B that the detector B gives a signal:

$$\varepsilon_{AB} = \varepsilon_A \varepsilon_B \,. \tag{1.7}$$

ε_A and ε_B are also called *efficiencies* of the detectors A and B respectively.

This result clearly does not hold if there are causes of simultaneous inefficiency of both detectors, e.g.: fraction of times where the electronics systems for both A and B are simultaneously switched off for short periods, or geometrical overlap of inactive regions.

1.7 Law of Total Probability

Let's consider a number of sets (events) $\{E_1, \cdots, E_n\}$, subsets of another set E_0 included in the sample space Ω, such that the set of the E_i is a *partition* of E_0, i.e.: $E_i \cap E_j = 0$ for all i and j, and:

$$\bigcup_{i=1}^{n} E_i = E_0 \,. \tag{1.8}$$

This case is visualized in Fig. 1.4. It is easy to prove that the probability corresponding to E_0 is equal to the sum of the probabilities of E_i (law of total probability):

$$P(E_0) = \sum_{i=1}^{n} P(E_i) \,. \tag{1.9}$$

Given a partition $\{A_1, \cdots, A_n\}$ of the sample space Ω of disjoint sets ($A_i \cap A_j = 0$ and $\sum_i P(A_i) = 1$), we can build the sets:

$$E_i = E_0 \cap A_i \,, \tag{1.10}$$

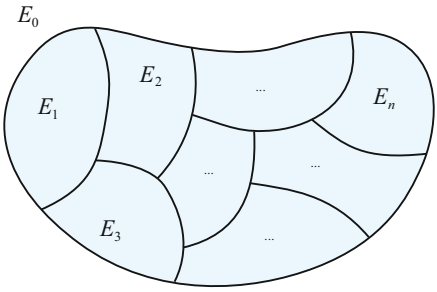

Fig. 1.4 Visual example of partition of an event set

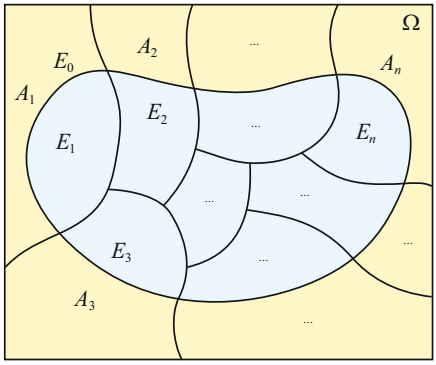

Fig. 1.5 Visualization of the law of total probability

as visualized in Fig. 1.5. This corresponds, from Eq. (1.4), to a probability:

$$P(E_i) = P(E_0 \cap A_i) = P(E_0|A_i)P(A_i)\,. \tag{1.11}$$

In this case, we can rewrite Eq. (1.9) as:

$$P(E_0) = \sum_{i=1}^{n} P(E_0|A_i)P(A_i)\,. \tag{1.12}$$

This decomposition, which can be interpreted as *weighted average* (see Sect. 5.11.1) of the probabilities $P(A_i)$ with weights $w_i = P(E_0|A_i)$, will be later used for Bayesian probability, discussed in Chap. 3.

1.8 Average, Variance and Covariance

In this section a number of useful quantities related to the probability distribution of a random variable are defined. Let's consider a random variable x which can assume the N possible values (outcomes) $x_1, \cdots, x_N$.

The *average value* or *expected value* of a random variable x with probability distribution P is defined as:

$$\boxed{\langle x \rangle = \sum_{i=1}^{N} x_i P(x_i) \, .} \tag{1.13}$$

Sometimes the notation $\mathrm{E}[x]$ or $\bar{x}$ is also used in literature to indicate the average value.

The *variance* of a random variable x is defined as:

$$\boxed{\mathrm{V}[x] = \sum_{i=1}^{N} (x_i - \langle x \rangle)^2 P(x_i) \, ,} \tag{1.14}$$

and the *standard deviation* is the square root of the variance:

$$\boxed{\sigma_x = \sqrt{\mathrm{V}[x]} = \sqrt{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2 P(x_i)} \, .} \tag{1.15}$$

Sometimes the standard deviation is confused, in some physics literature, with the *root mean square*, abbreviated as *r.m.s.*, which instead should be more properly defined as:

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2 P(x_i)} = \sqrt{\langle x^2 \rangle} \, . \tag{1.16}$$

The variance can also be written as:

$$V[x] = \left\langle (x - \langle x \rangle)^2 \right\rangle \, , \tag{1.17}$$

and it's easy to demonstrate that it can be equivalently written as:

$$\boxed{V[x] = \left\langle x^2 \right\rangle - \langle x \rangle^2 \, .} \tag{1.18}$$

It is also easy to demonstrate that, if we have two random variables x and y, the averages can be added linearly, i.e.:

$$\langle x + y \rangle = \langle x \rangle + \langle y \rangle \, , \tag{1.19}$$

Given a constant a, we have:

$$\langle ax \rangle = a \langle x \rangle \tag{1.20}$$

and

$$\mathrm{V}[ax] = a^2 \mathrm{V}[x] \,. \tag{1.21}$$

Given two variables x and y, their *covariance* is defined as:

$$\boxed{\mathrm{cov}(x, y) = \langle xy \rangle - \langle x \rangle \langle y \rangle \ ,} \tag{1.22}$$

and the *correlation coefficient* of x and y is defined as:

$$\boxed{\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} \,.} \tag{1.23}$$

It can be demonstrated that:

$$\mathrm{V}[x + y] = \mathrm{V}[x] + \mathrm{V}[y] + 2\,\mathrm{cov}(x, y) \,. \tag{1.24}$$

So, variances of uncorrelated variables can be added linearly.

Given n random variables, $x_1, \cdots, x_n$, the symmetric matrix $C_{ij} = \mathrm{cov}(x_i, y_j)$ is called *covariance matrix*. The diagonal terms are always positive or null, and correspond to the variances of the n variables.

Another quantity that is sometimes important is the *skewness* of a distribution defined as:

$$\boxed{\gamma_1 = \left\langle \left(\frac{x - \langle x \rangle}{\sigma_x} \right)^3 \right\rangle = \frac{\gamma}{\sigma^3} \,,} \tag{1.25}$$

where the quantity:

$$\boxed{\gamma = \left\langle x^3 \right\rangle - 3 \langle x \rangle \left\langle x^2 \right\rangle + 2 \langle x \rangle^3} \tag{1.26}$$

is called *unnormalized swekness*. Symmetric distribution have skewness equal to zero.

The *kurtosis*, finally, is defined as:

$$\boxed{\beta_2 = \left\langle \left(\frac{x - \langle x \rangle}{\sigma_x} \right)^4 \right\rangle \,.} \tag{1.27}$$

Usually the *kurtosis coefficient*, γ_2 is defined as:

$$\boxed{\gamma_2 = \beta_2 - 3} \tag{1.28}$$

in order to have $\gamma_2 = 0$ for a normal distribution. γ_2 is also defined as *excess*.

1.9 Variables Transformations

Given a random variable x, let's consider the variable y transformed via the function $y = Y(x)$. If x can only assume the values $\{x_1, \cdots, x_N\}$, then y can only assume one of the values $\{y_1, \cdots, y_M\} = \{Y(x_1), \cdots, Y(y_N)\}$. Note that N is equal to M only if all the values $Y(x_i)$ are different from each other. The probability of any value y_j is given by the sum of the probabilities of all values x_i that are transformed into y_j by Y:

$$P(y_j) = \sum_{i:Y(x_i)=y_j} P(x_i)\,. \tag{1.29}$$

Note that there could also be a single possible value of the index i for which $Y(x_i) = y_i$. In this case, we will have just $P(y_j) = P(x_i)$.

The generalization to more variables is straightforward: assume that we have two random variables x and y, and we have a variable z defined as $z = Z(x, y)$; if $\{z_1, \cdots, z_M\}$ is the set of all possible values of z, given all the possible values x_i and y_j, the probability of each possible value of z, z_k, will be given by:

$$P(z_k) = \sum_{i,j:Z(x_i,y_j)=z_k} P(x_i, y_j)\,. \tag{1.30}$$

This expression is consistent with the example of the sum of two dices considered in Sect. 1.2:

Example 1.2 – Application to Sum of Dice Rolls

Let's compute the probabilities that the sum of two dice rolls is even or odd. We can consider all cases in Table 1.1, and add the probabilities for all even and odd values. Even values and their probabilities are: $2 : {}^1/_{36}$, $4 : {}^1/_{12}$, $6 : {}^5/_{36}$, $8 : {}^5/_{36}$, $10 : {}^1/_{12}$, $12 : {}^1/_{36}$. So, the probability of an even result is: $(1 + 3 + 5 + 5 + 3 + 1)/36 = {}^{18}/_{36} = {}^1/_2$. This is consistent with the fact that each dice has probability ${}^1/_2$ of an even or odd result, and an even result as the sum of two dices can occur either as sum of two even results or as sum of two odd results, each case with probability ${}^1/_2 \times {}^1/_2 = {}^1/_4$. So, the probability of either of the two cases is ${}^1/_4 + {}^1/_4 = {}^1/_2$.

1.10 The Bernoulli Process

Let's consider a basket that contains a number n of balls that have two possible colors, say red and white. We know the number r of red balls in the basket, hence the number of white balls is $n - r$. The probability to randomly extract a red ball in that set of n balls is $p = r/n$, according to Eq. (1.1) (Fig. 1.6).

A variable x equal to the number of extracted red balls is called *Bernoulli variable*, and can assume only the values of 0 or 1. The probability distribution of x is just given by $P(1) = p$, $P(0) = 1 - p$. The average of a Bernoulli variable is easy to compute:

$$\langle x \rangle = P(0) \times 0 + P(1) \times 1 = P(1) = p\,. \tag{1.31}$$

Similarly, the average of x^2 is:

$$\left\langle x^2 \right\rangle = P(0) \times 0^2 + P(1) \times 1^2 = P(1) = p\,, \tag{1.32}$$

hence the variance of x, using Eq. (1.18), is:

$$\mathrm{V}[x] = \left\langle x^2 \right\rangle - \langle x \rangle^2 = p(1-p)\,. \tag{1.33}$$

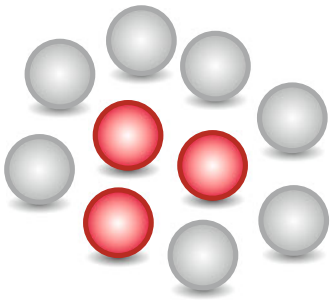

Fig. 1.6 A set of $n = 10$ balls, of which $r = 3$ are *red* determines a Bernoulli process. The probability to randomly extract a *red ball* in the set shown in this figure is $p = r/n = 3/10 = 30\,\%$

1.11 The Binomial Process

A binomial process consists of performing a given number N of independent Bernoulli extractions all having the same probability p. This could be implemented for instance by randomly extracting a (red or white) ball in a set of balls from a basket containing a fraction p of red balls; after each extraction, the extracted ball is placed again in the basket and the extraction is repeated N times. Figure 1.7 shows the possible outcomes of a binomial process as subsequent random extractions of a single ball.

The number n of red balls (positive outcomes) over the total N extraction is called binomial variable. Its probability distribution can be simply determined by looking at Fig. 1.7, considering how many red and white extraction have occurred in each extraction, assigning each extraction a probability p or $1 - p$ respectively, and considering the number of possible paths leading to a given combination of red/white extractions.

The latter term takes the name of *binomial coefficient* and can be demonstrated by recursion to be equal to[4]:

$$\binom{n}{N} = \frac{n!}{N!(N-n)!}\,. \tag{1.34}$$

The probability distribution of the binomial variable n for a given N and p can be obtained considering that the n extractions are independent, hence the corresponding probability terms (p for a red extraction, $1 - p$ for a white extraction) can be multiplied, according to Eq. (1.6). Multiplying finally this product by the binomial coefficient from Eq. (1.34) to take into account the possible extraction paths leading

[4]The coefficients present in the binomial distribution are the same that appear in the expansion of the binomial power $(a + b)^n$. A simple iterative way to compute those coefficients is known in literature as the Pascal's triangle. Different countries quote this triangle according to different authors, e.g.: the Tartaglia's triangle in Italy, Yang Hui's triangle in China, etc. In particular, the following publications of the triangle are present in literature:

- India: published in the tenth century, referring to the work of Pingala, dating back to fifth–second century BC.
- Persia: Al-Karaju (953–1029) and Omar Jayyám (1048–1131), often referred to as Kayyám triangle in modern Iran
- China: Yang Hui (1238–1298), referred to as Yang Hui's triangle by Chinese; see Fig. 1.8
- Germany: Petrus Apianus (1495–1552)
- Italy: Nicolò Fontana Tartaglia (1545), referred to as Tartaglia's triangle in Italy
- France: Blaise Pascal (1655), refereed to as Pascal's triangle.

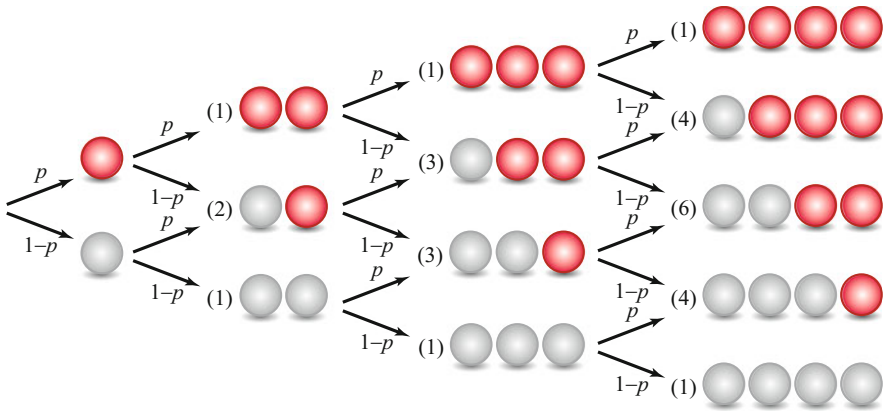

Fig. 1.7 Building a binomial process as subsequent random extractions of a single *red or white ball* (Bernoulli process). The "tree" in figure displays all the possible combinations at each extraction step. Each of the "branching" has a corresponding probability equal to p or $1-p$ for *red and white ball* respectively. The number of paths corresponding to each possible combination is shown in parentheses, and is equal to a binomial coefficient

to the same outcome, the probability to obtain n red and $N-n$ white extractions can be written as:

$$\boxed{P(n;N) = \frac{n!}{N!(N-n)!} p^N (1-p)^{N-n}\,.} \tag{1.35}$$

The Binomial distribution is shown in Fig. 1.9 for $N = 15$ and for $p = 0.2$, 0.5 and 0.8.

Since the binomial variable n is the sum of N independent Bernoulli variables with probability p, the average and variance of n can be computed as N times the average and variance of a Bernoulli variable (Eqs. (1.31) and (1.33)):

$$\langle n \rangle = Np\,, \tag{1.36}$$

$$\mathrm{V}[n] = Np(1-p)\,. \tag{1.37}$$

Those formulae can also be obtained directly from Eq. (1.35).

1.11.1 Binomial Distribution and Efficiency Estimate

The *efficiency* ε of a device is defined as the probability that the device gives a positive signal upon the occurrence of a process of interest. Particle detectors are examples of such devices. The distribution of the number of positive signals n upon

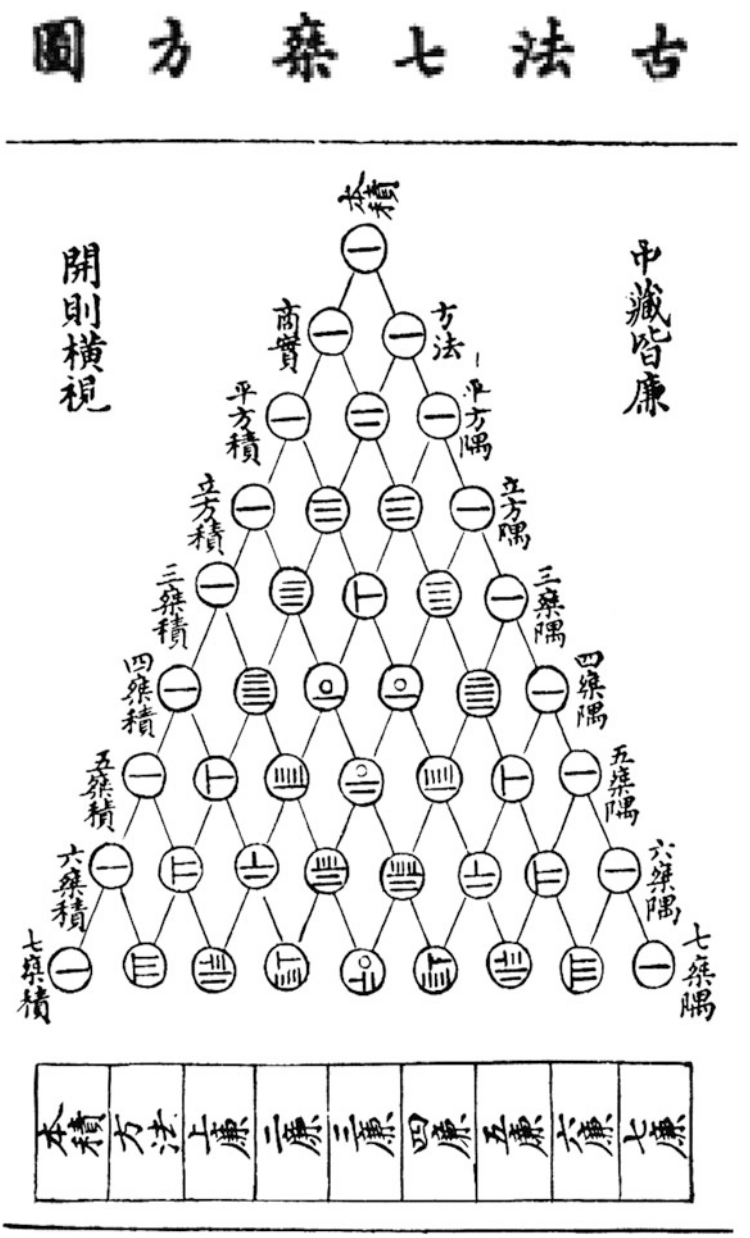

Fig. 1.8 The Yang Hui triangle published in 1303 by Zhu Shijie (1260–1320)

the occurrence of N processes of interest is given by a binomial distribution with probability $p = \varepsilon$. Given a measurement $\hat{n}$ (i.e.: the result of a real experiment, or a particular random extraction of n), a typical problem is the *estimate* of the

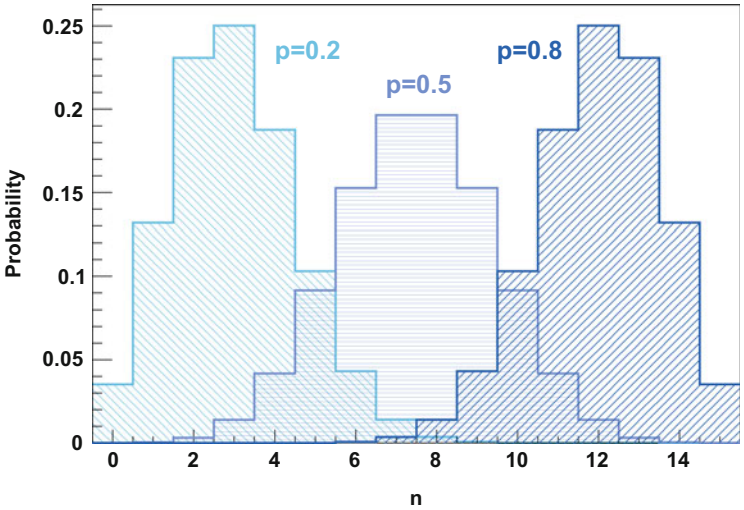

Fig. 1.9 Binomial distribution for $N = 15$ and for $p = 0.2, 0.5$ and 0.8

efficiency ε of the device. The problem of parameter estimates will be discussed more in general in Chaps. 3 and 5 for what concerns the Bayesian and frequentist approaches, respectively.

For the moment, a pragmatic way to estimate the efficiency can be considered by performing a large number N of sampling of our process of interest, and by counting the number of times $\hat{n}$ the device gives a positive signal (i.e.: it has been efficient). This leads to the estimate of the true efficiency ε given by:

$$\hat{\varepsilon} = \frac{\hat{n}}{N} . \tag{1.38}$$

One may argue that if N is sufficiently large, $\hat{\varepsilon}$ will be very close to true efficiency ε. This will be more formally discussed as a consequence of the *law of large numbers*, which will be introduced in Sect. 1.12. The variance of n, $\mathrm{V}[n]$, from Eq. (1.37), is:

$$\sigma_\varepsilon = \sqrt{\mathrm{Var}\left[\frac{n}{N}\right]} = \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}} . \tag{1.39}$$

By simply replacing ε with $\hat{\varepsilon}$ one would have the following *approximated* expression for the standard deviation (interpreted as *error* or *uncertainty*) of $\hat{\varepsilon}$:

$$\delta\varepsilon = \sqrt{\frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{N}} . \tag{1.40}$$

The above formula is just an approximation, and in particular it leads to an error $\delta\varepsilon = 0$ in the cases where $\hat{\varepsilon} = 0$ or $\hat{\varepsilon} = 1$, i.e. when $\hat{n} = 0$ or $\hat{n} = N$ respectively.

Chapter 5 will introduce more rigorous definitions of *error estimates* and will discuss how to evaluate errors on a proper statistical basis. Later on, in Sect. 6.2, we will see how to overcome the problems arising with Eq. (1.40).

1.12 The Law of Large Numbers

When an experiment that produces a random outcome x is repeated N times, the average of the outcomes, $\bar{x} = (x_1 + \cdots + x_N)/N$, is a random variable whose probability distribution has a much smaller fluctuation than each of the original single experiment outcome, x_i. This can be demonstrated, using classical probability and combinatorial analysis, in the simplest cases, and one can show that most of the possible outcomes of $\bar{x}$ have very small probabilities, except the ones very close to the expected average $\langle x \rangle$ (Eq. (1.13)). This is already partially evident in Fig. 1.1, where the distribution of the sum of two dice rolls is peaked around 7 $(= 2 \times 3.5 = 2 \times \langle x \rangle)$ even for $N = 2$. This convergence of the probability distribution of $\bar{x}$ to the single value $\langle x \rangle$, for $N \to \infty$, is called *law of large numbers*, and can be illustrated in a simulated experiment consisting of repeatedly rolling a dice, whose result ranges from 1 to 6 as usual, and plotting the average $\bar{x}$ obtained after the first N extractions, as a function of N. This is shown in Fig. 1.10, where the

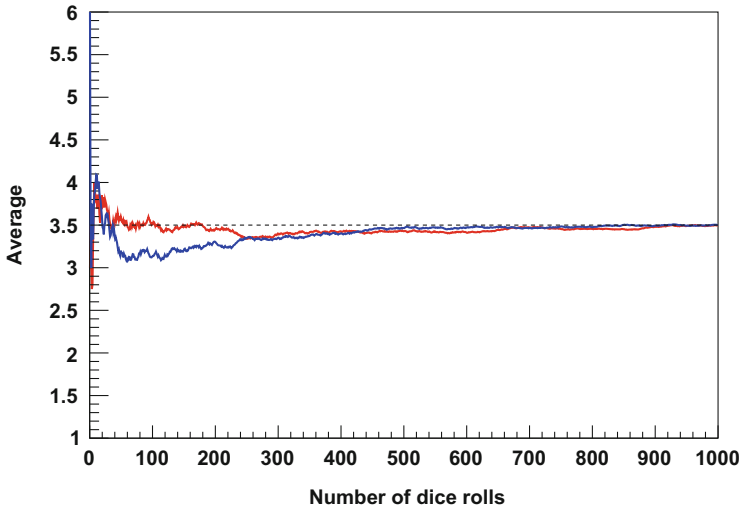

Fig. 1.10 An illustration of the law of large numbers using a simulation of die rolls. The average of the first N out of 1000 random extraction is reported as a function of N. The 1000 extractions have been repeated twice (*red and blue lines*)

plot of $\bar{x}$ as a function of N "converges" to:

$$\langle x \rangle = \frac{1+2+3+4+5+6}{6} = 3.5\,. \tag{1.41}$$

Increasing the number of trials, the probability distribution of the possible average values acquires a narrower shape, and the interval of the values that correspond to a large fraction of the total probability (we could chose—say—90 or 95 %) gets smaller and smaller as the number of trials N increases. If we would ideally increase *to infinity* the total number of trials N, the average value $\bar{x}$ would no longer be a random variable, but would take a single possible value equal to $\langle x \rangle = 3.5$.

The *law of large numbers* has many empirical verifications for the vast majority of random experiments. This broad generality of the law of large numbers even in realistic and more complex cases than the simple ones that can be described by simplified classical models gives raise to the *frequentist* definition of the probability $P(E)$ of an event E according to the following limit:

$$P(E) = p \;\text{ if }\; \forall \varepsilon \lim_{N\to\infty} P\left(\left|\frac{N(E)}{N} - p\right| < \varepsilon\right). \tag{1.42}$$

The limit is intended, in this case, as *convergence in probability*, that means that, by the law of large numbers, the limit only rigorously holds in the non-realizable case of an infinite number of experiments. Rigorously speaking, the definition of frequentist probability in Eq. (1.42) is defined itself in terms of another probability, which could introduce some conceptual problems. F. James et al. report the following sentence [3]: *"this definition is not very appealing to a mathematician, since it is based on experimentation, and, in fact, implies unrealizable experiments ($N \to \infty$)"*.

In practice, anyway, we know that experiments are reproducible on a finite range of time (on this planet, for instance, until the sun and the solar system will continue to exist), so, for the practical purposes of physics application, we can consider the law of large numbers and the frequentist definition of probability, beyond their exact mathematical meaning, as pragmatic definitions that describe to a very good approximation level the concrete situations of the vast majority of the cases we are interested in experimental physics.

References

1. P. Laplace, *Essai Philosophique Sur les Probabilités*, 3rd edn. (Courcier Imprimeur, Paris, 1816)
2. A. Kolmogorov, *Foundations of the Theory of Probability*. (Chelsea Publishing, New York, 1956)
3. W. Eadie, D. Drijard, F. James, M. Roos, B. Saudolet, *Statistical Methods in Experimental Physics*. (North Holland, Amsterdam, 1971)

Chapter 2
Probability Distribution Functions

The problem introduced with Bertrand's paradox, seen in Sect. 1.3.1, occurs because the decomposition of the range of possible values of a random variable x into *equally probable* elementary cases is not possible without ambiguity. The ambiguity arises because of the continuum nature of the problem. We considered in Sect. 1.3 a continuum random variable x whose outcome may take any possible value in a continuum interval $[x_1, x_2]$, and we saw that if x is *uniformly distributed* in $[x_1, x_2]$, a transformed variable $y = Y(x)$ is not in general uniformly distributed in $[y_1, y_2] = [Y(x_1), Y(x_2)]$ (Y is taken as a monotonic function of x). This makes the choice of a continuum variable on which to find *equally probable* intervals of the same size an arbitrary choice.

2.1 Definition of Probability Distribution Function

The concept of probability distribution seen in Sect. 1.5 can be generalized to the continuum case introducing the *probability distribution function* defined as follows. Let's consider a *sample space*, $\{\vec{x} = (x_1, \cdots, x_n) \in \Omega \subseteq R^n\}$. Each random extraction (an *experiment*, in the cases of interest to a physicist) will lead to an outcome (*measurement*) of one point $\vec{x}$ in the sample space Ω. We can associate to any point $\vec{x}$ a *probability density* $f(\vec{x}) = f(x_1, \cdots, x_n)$ which is a real value greater or equal to zero. The probability of an *event* A, where $A \subseteq \Omega$, i.e.: the probability that $\vec{x} \in A$, is given by:

$$P(A) = \int_A f(x_1, \cdots, x_n)\, \mathrm{d}^n x\,. \tag{2.1}$$

The function f is called *probability distribution function* (PDF). The function $f(\vec{x})$ can be interpreted as *differential probability*, i.e.: the probability corresponding to an

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_2

infinitesimal hypercube $\mathrm{d}x_1 \cdots \mathrm{d}x_n$, divided by the infinitesimal hypercube volume:

$$\frac{\mathrm{d}^n P}{\mathrm{d}x_1 \cdots \mathrm{d}x_n} = f(x_1, \cdots, x_n) \,. \tag{2.2}$$

The normalization condition for discrete probability distributions (Eq. (1.3)) can be generalized to continuous probability distribution functions as follows:

$$\int_\Omega f(x_1, \cdots, x_n)\, \mathrm{d}^n x = 1 \,. \tag{2.3}$$

In one dimension, one can write:

$$\int_{-\infty}^{+\infty} f(x)\, \mathrm{d}x = 1 \,. \tag{2.4}$$

Note that the probability of each single point is rigorously zero if f is a real function, i.e.: $P(\{x_0\}) = 0$ for any x_o, since the set $\{x_0\}$ has null measure. The treatment of discrete variables in one dimension can be done using the same formalism of PDFs, extending the definition of PDF to Dirac's delta functions ($\delta(x - x_0)$, with $\int_{-\infty}^{+\infty} \delta(x - x_0)\, \mathrm{d}x = 1$). Dirac's delta functions can be linearly combined with proper weights equal to the probabilities of the discrete values: the PDF corresponding to the case of a discrete random variable x that can take only the values $x_1, \cdots, x_N$ with probabilities $P_1, \cdots, P_N$ respectively can be written as:

$$f(x) = \sum_{i=1}^{N} P_i \delta(x - x_i) \,. \tag{2.5}$$

The normalization condition in this case can be written as:

$$\int_{-\infty}^{+\infty} f(x)\, \mathrm{d}x = \sum_{i=1}^{N} P_i \int_{-\infty}^{+\infty} \delta(x - x_i)\, \mathrm{d}x = \sum_{i=1}^{N} P_i = 1 \,, \tag{2.6}$$

which gives again Eq. (1.3). Discrete and continuous values can be combined introducing linear combinations of continuous PDFs and Dirac's delta functions.

2.2 Average and Variance in the Continuous Case

The definitions of average and variance introduced in Sect. 1.8 are generalized for continuous variables as follows:

$$\boxed{\langle x \rangle = \int x f(x)\, \mathrm{d}x \,,} \tag{2.7}$$

$$\boxed{\mathrm{V}[x] = \int (x - \langle x \rangle)^2 f(x)\,\mathrm{d}x = \langle x^2 \rangle - \langle x \rangle^2 \,.} \tag{2.8}$$

Integrals should be extended over $[-\infty, +\infty]$, or the whole validity range of the variable x. The standard deviation is defined as:

$$\boxed{\sigma_x = \sqrt{V[x]}\,.} \tag{2.9}$$

Other interesting parameters of a continuous PDF are the *median*, which is the value m such that:

$$P(x \leq m) = P(x > m) = \frac{1}{2}\,, \tag{2.10}$$

or equivalently:

$$\int_{-\infty}^{m} f(x)\,\mathrm{d}x = \int_{m}^{+\infty} f(x)\,\mathrm{d}x = \frac{1}{2}\,, \tag{2.11}$$

and the *mode*, which is the value M that corresponds to the maximum of the probability density:

$$f(M) = \max_x f(x)\,. \tag{2.12}$$

2.3 Cumulative Distribution

Given a PDF $f(x)$, its *cumulative distribution* is defined as:

$$\boxed{F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'\,.} \tag{2.13}$$

The cumulative distribution, $F(x)$, is a monotonous increasing function of x, and from the normalization of $f(x)$ (Eq. (2.4)), its value ranges from 0 to 1. In particular:

$$\lim_{x \to -\infty} F(x) = 0\,, \tag{2.14}$$

$$\lim_{x \to +\infty} F(x) = 1\,. \tag{2.15}$$

If the variable x follows the PDF $f(x)$, the PDF of the transformed variable $y = F(x)$ is uniform between 0 and 1, as can be demonstrated in the following:

$$\frac{\mathrm{d}P}{\mathrm{d}y} = \frac{\mathrm{d}P}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}y} = f(x)\frac{\mathrm{d}x}{\mathrm{d}F(x)} = \frac{f(x)}{f(x)} = 1\,. \tag{2.16}$$

This property will be very useful to generate with computers pseudorandom numbers with the desired PDF, as we will see in Chap. 4.

2.4 Continuous Variables Transformation

The transformation of probability distributions upon transformation of variables, presented in Sect. 1.9, can be generalized in the continuous case. If we have a variable transformation $y = Y(x)$, the PDF of the transformed variable y can be generalized from Eq. (1.29) as:

$$f(z) = \int \delta(y - Y(x)) f(x)\, \mathrm{d}x\,. \tag{2.17}$$

Similarly, if we have a variable transformation $z = Z(x, y)$, the PDF of the transformed variable z can be generalized from Eq. (1.30) as:

$$f(z) = \int \delta(z - Z(x, y)) f(x, y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{2.18}$$

In the case of transformations into more than one variable, the generalization is straightforward. If we have for instance: $x' = X'(x, y)$, $y' = Y'(x, y)$, the transformed two-dimensional PDF can be written as:

$$f'(x', y') = \int \delta(x' - X'(x, y)) \delta(y' - Y'(x, y)) f(x, y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{2.19}$$

If the transformation is invertible, the PDF transforms according to the determinant of the Jacobean of the transformation. This factor appears in fact in the transformation of the n-dimensional volume element $\mathrm{d}^n x = \mathrm{d}x_1 \cdots \mathrm{d}x_n$:

$$\frac{\mathrm{d}^n P}{\mathrm{d}^n x} = \frac{\mathrm{d}^n P'}{\mathrm{d}^n x'} \det \left| \frac{\partial x'_i}{\partial x_j} \right|\,. \tag{2.20}$$

2.5 Uniform Distribution

A variable x is *uniformly distributed* in the interval $[a, b[$ if the PDF is constant in the range $x \in [a, b[$. This condition is identical to the condition that was discussed in Sect. 1.3.1. Using the normalization condition, a uniform PDF can be written as:

$$\boxed{u(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \le x < b \\ 0 & \text{if } x < a \text{ or } x \ge b \end{cases}\,.} \tag{2.21}$$

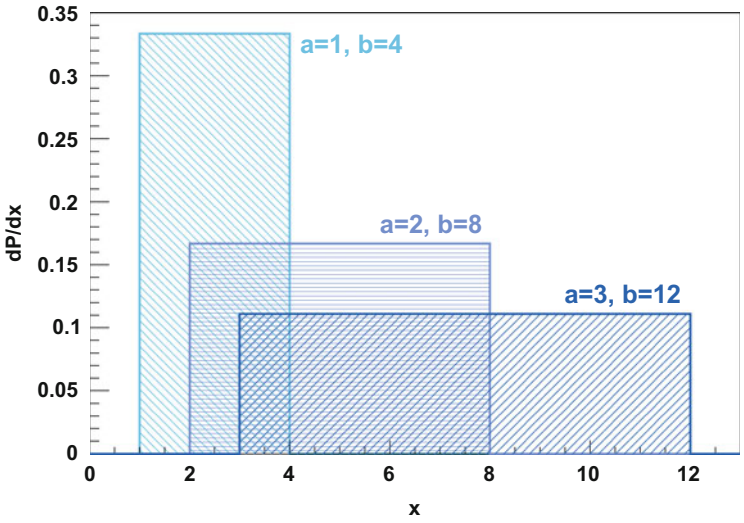

Fig. 2.1 Uniform distributions with different extreme values

Examples of uniform distributions are shown in Fig. 2.1 for different extreme values a and b.

The average of a variable x which is uniformly distributed is:

$$\langle x \rangle = \frac{a+b}{2}\,, \tag{2.22}$$

and the standard deviation of x is:

$$\sigma = \frac{b-a}{\sqrt{12}}\,. \tag{2.23}$$

Example 2.3 – Strip Detectors

A detector instrumented with strips of a given pitch l receives particles uniformly distributed along each strip. The standard deviation of the distribution of the position of the particles' impact point on the strip along the direction transverse to the strips is given by $l/\sqrt{12}$, according to Eq. (2.23).

2.6 Gaussian Distribution

A variable x follows a Gaussian or *normal* distribution if it obeys the following PDF, with μ and σ being fixed parameters:

$$\boxed{g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).} \tag{2.24}$$

The average value and standard deviation of the variable x are μ and σ respectively. Examples of Gaussian distributions are shown in Fig. 2.2 for different values of μ and σ. A random variable obeying a normal distribution is called *normal random variable*.

If $\mu = 0$ and $\sigma = 1$, a normal variable is called *standard normal* following the standard normal distribution $\phi(x)$:

$$\boxed{\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} \tag{2.25}$$

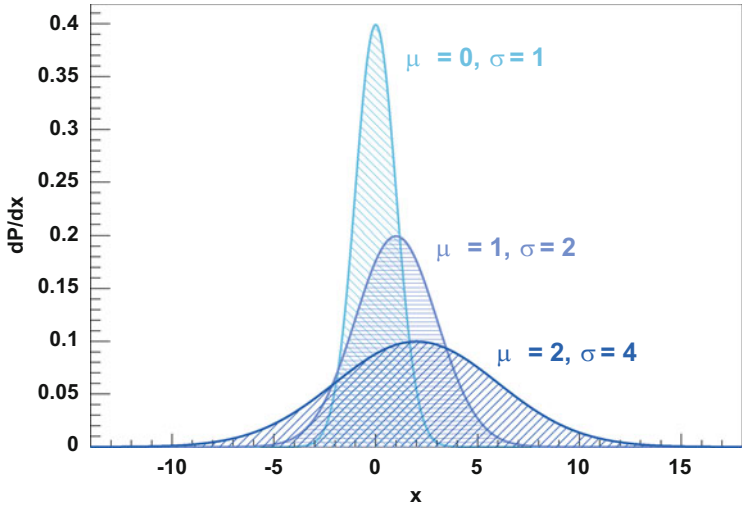

Fig. 2.2 Gaussian distributions with different values of average and standard deviation parameters

and its cumulative distribution is:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x'^2}{2}} \, \mathrm{d}x' = \frac{1}{2} \left[\mathrm{erf}\left(\frac{x}{\sqrt{2}} \right) + 1 \right] . \tag{2.26}$$

The probability for a Gaussian distribution corresponding to the interval $[\mu - Z\sigma, \mu + Z\sigma]$, which is symmetric around μ, is frequently used in many application. It can be computed as:

$$P(Z\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-Z}^{Z} e^{-\frac{x^2}{2}} \, \mathrm{d}x = \Phi(Z) - \Phi(-Z) = \mathrm{erf}\left(\frac{Z}{\sqrt{2}} \right) . \tag{2.27}$$

The most frequently used values are the ones corresponding to 1σ, 2σ and 3σ, and correspond to probabilities of 68.72, 95.45 and 99.73 % respectively.

The importance of the Gaussian distribution is due to the central limit theorem (see Sect. 2.11), which allows to approximate to Gaussian distributions may realistic cases resulting from the superposition of more random effects each having a finite and possibly unknown PDF.

2.7 Log-Normal Distribution

If a random variable y is distributed is distributed according to a normal PDF with average μ and standard deviation σ, the variable $x = e^y$ is distributed according to the following PDF, called *log-normal* distribution:

$$\boxed{f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2} \right) .} \tag{2.28}$$

The PDF in Eq. (2.28) can be derived from Eq. (2.17) from Sect. 2.4 in the case of a normal PDF. A log-normal variable has the following average and standard deviation:

$$\langle x \rangle = e^{\mu + \sigma^2/2} , \tag{2.29}$$

$$\sigma_x = e^{\mu + \sigma^2/2} \sqrt{e^{\sigma^2} - 1} . \tag{2.30}$$

Note that Eq. (2.29) implies that $\langle e^y \rangle > e^{\langle y \rangle}$ for a normal random variable y. In fact, since $e^{\sigma^2/2} > 1$:

$$\langle e^y \rangle = \langle x \rangle = e^{\mu} e^{\sigma^2/2} > e^{\mu} = e^{\langle y \rangle} .$$

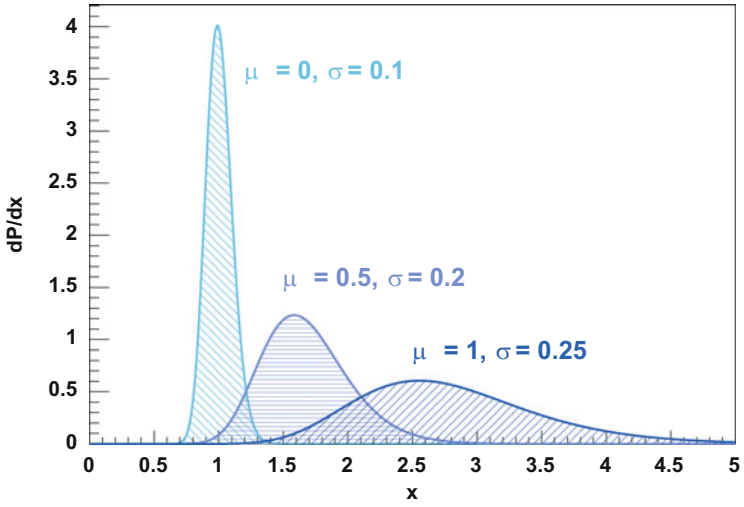

Fig. 2.3 Lognormal distributions with different parameters μ and σ

Examples of lognormal distributions are shown in Fig. 2.3 for different values of μ and σ.

2.8 Exponential Distribution

An *exponential distribution* of a variable $x \geq 0$ is characterized by a PDF proportional to $e^{-\lambda x}$, where λ is a constant. Considering the normalization condition from Eq. (2.4), the complete expression of an exponential PDF is given by:

$$\boxed{f(x) = \lambda e^{-\lambda x}.} \tag{2.31}$$

Examples of exponential distributions are shown in Fig. 2.4 for different values of the parameter λ.

Exponential distributions are widely used in physics.

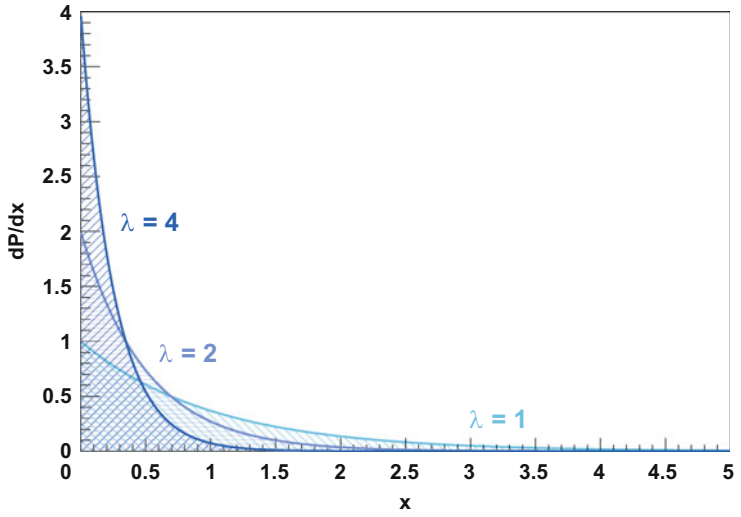

Fig. 2.4 Exponential distributions with different values of the parameter λ

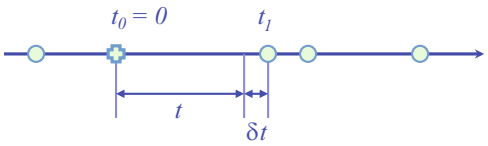

Fig. 2.5 An horizontal axis representing occurrence times (*dots*) of some events. The time origin ($t_0 = 0$) is marked as a *cross*

Example 2.4 – Relation Between Uniform and Exponential Distributions

One interesting problem where an exponential distribution occurs is the distribution of the occurrence time of the first in a sequence of events, or also the time difference between two consecutive events, whose occurrence time is uniformly distributed over a indefinitely wide time range. An example of such a case is sketched in Fig. 2.5. The distribution of the occurrences time t_1 of the first events with respect to a reference time $t_0 = 0$[1] can be determined as follows.

[1]This instant could also coincide with the occurrence of one of the considered events in order to apply the same derivation the time difference of two consecutive events.

Let's consider a time t and another time $t+\delta t$, where $\delta t \ll t$. We will compute first the probability $P(t_1 > t+\delta t)$ that t_1 is greater than $t+\delta t$, which is equal to the probability $P(0, [0, t+\delta t])$ that no event occurs in the time interval $[0, t+\delta t]$. The probability $P(0, [0, t+\delta t])$ is equal to the probability that no event occurs in the interval $[0, t]$ *and* no event occurs in the interval $[t, t+\delta t]$. Since the occurrences of events in two disjoint time intervals are independent events, the combined probability is the product of the two probabilities $P(0, [0, t])$ and $P(0, [t, t+\delta t])$:

$$P(0, [0, t+\delta t]) = P(0, [0, t])P(0, [t, t+\delta t])\,. \tag{2.32}$$

Given the *rate* r of event occurrences, i.e. the expected number of events per unit of time, the probability to have one occurrence in a time interval δt is given by:

$$P(1, [t, t+\delta t]) = r\delta t\,, \tag{2.33}$$

and the probability to have more than one occurrence can be can be neglected with respect to $P(1[t, t+\delta t]) = r\delta t$ at $O(\delta t^2)$. So, the probability to have zero events in δt is equal to:

$$P(0, [t, t+\delta t]) \simeq 1 - P(1, [t, t+\delta t]) = 1 - r\delta t\,, \tag{2.34}$$

and we can write:

$$P(0, [0, t+\delta t]) = P(0, [0, t])(1 - r\delta t)\,, \tag{2.35}$$

or, equivalently:

$$P(t_1 > t+\delta t) = P(t_1 > t)(1 - r\delta t)\,. \tag{2.36}$$

From Eq. (2.36), we can also write:

$$\frac{P(t_1 > t+\delta t) - P(t_1 > t)}{\delta t} = -rP(t_1 > t)\,. \tag{2.37}$$

In the limit $\delta t \to 0$, this leads to the following differential equation:

$$\frac{\mathrm{d}P(t_1 > t)}{\mathrm{d}t} = -rP(t_1 > t)\,. \tag{2.38}$$

Considering the initial condition $P(t_1 > 0) = 1$, Eq. (2.38) has the following solution:

$$P(t_1 > t) = e^{-rt}\,. \tag{2.39}$$

The probability distribution function of the occurrence of the first event can be written as:

$$P(t) = \frac{P(t < t_1 < t + \delta t)}{\delta t} = \frac{\mathrm{d}P(t_1 < t)}{\mathrm{d}t}\,, \tag{2.40}$$

where:

$$P(t_1 < t) = 1 - P(t_1 > t) = 1 - e^{-rt}\,. \tag{2.41}$$

Performing the first derivative with respect to t we have:

$$P(t) = \frac{\mathrm{d}(P(t_1 < t))}{\mathrm{d}t} = \frac{\mathrm{d}(1 - e^{-rt})}{\mathrm{d}t}\,, \tag{2.42}$$

hence:

$$P(t) = re^{-rt}\,. \tag{2.43}$$

The exponential distribution is characteristic of particles lifetimes. The possibility to measure an exponential distribution with the same (logarithmic) slope λ independently on the initial time t_0 allows to measure particle lifetimes without the need to measure the particle's creation time, as it is the case with cosmic-ray muons whose lifetime can be measured at sea level, while the muon has been created in the high atmosphere.

2.9 Poisson Distribution

A discrete variable n is said to be a *Poissonian* variable if it obeys the *Poisson distribution* defined below for a given value of the parameter ν:

$$\boxed{P(n;\nu) = \frac{\nu^n e^{-\nu}}{n!}\,.} \tag{2.44}$$

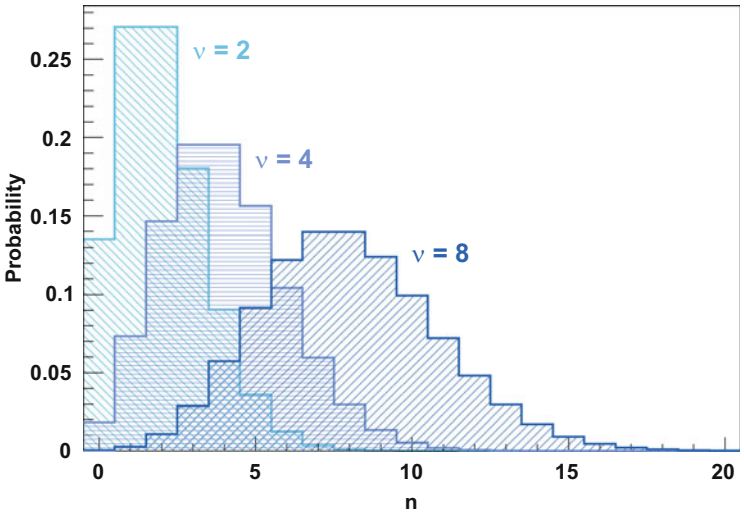

Fig. 2.6 Poisson distributions with different rate parameter

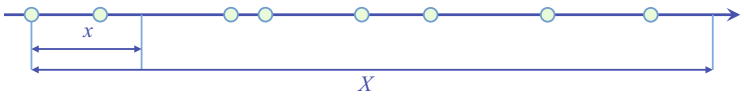

Fig. 2.7 A uniform distribution of events in a variable x. Two intervals are shown of sizes x and X, where $x \ll X$

Examples of Poisson distributions are shown in Fig. 2.6 for different values of the rate parameter ν. It's easy to show that the average and variance of a Poisson distribution are both equal to ν.

Example 2.5 – Relation Between Uniform, Binomial and Poisson Distributions

In the previous Example 2.4, a uniform distribution of a variable t over an indefinitely wide range gave raise to an exponential distribution of the first occurrence time t_1 with respect to a given reference time t_0.

We will now consider again a uniformly distribution of a variable, this time we will call it x, over an indefinitely wide range (x could be either a time or space variable, in many concrete examples) whose *rate* r is known in order to determine the distribution of the number of events present in a finite interval of x. The rate r can be written as N/X, where N is the number of events that occur in an interval X that is much larger than the range of the variable x we are considering (Fig. 2.7). We will eventually take the limit $N \to \infty$, $X \to \infty$, keeping constant the ratio $N/X = r$.

We can consider N and X as constants, i.e. quantities not subject to random fluctuation (consider, for instance, the total number of cosmic rays in a year time, in the case you want to measure the number of cosmic rays passing in a second). If N is fixed and n is the number of events that occur in an interval $x \ll X$, we *expect* a number of events:

$$\nu = \frac{Nx}{X} = rx\,. \tag{2.45}$$

The variable n is subject to the binomial distribution (see Sect. 1.11):

$$P(n;\nu,N) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n}\,, \tag{2.46}$$

that can also be written as:

$$P(n;\nu,N) = \frac{\nu^n}{n!}\frac{N(N-1)\cdots(N-n+1)}{N^n}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n}\,. \tag{2.47}$$

In the limit of $N \to \infty$, the first term, $\frac{\nu^n}{n!}$, remains unchanged, while the remaining three terms tend to 1, $e^{-\nu}$ and 1 respectively. We can write the distribution of n in this limit, which is equal to the Poisson distribution:

$$P(n;\nu) = \frac{\nu^n e^{-\nu}}{n!}\,. \tag{2.48}$$

Poisson distributions have several interesting properties, some of which are listed in the following.

- For large ν, the Poisson distribution can be approximated with a Gaussian having average ν and standard deviation $\sqrt{\nu}$.
- A Binomial distribution characterized by a number of extractions N and probability $p \ll 1$ can be approximated with a Poisson distribution with average $\nu = pN$ (this was seen in the Example 2.5 above).
- If we consider two variables n_1 and n_2 that follow Poisson distributions with averages ν_1 and ν_2 respectively, it is easy to demonstrate, using Eq. (1.30) and a bit of mathematics, that the sum $n = n_1 + n_2$ follows again a Poisson distribution with average $\nu_1 + \nu_2$. In other words:

$$P(n;\nu_1,\nu_2) = \sum_{n_1=0}^{n}\sum_{n_2=0}^{n-n_1}\mathrm{Poiss}(n_1;\nu_1)\,\mathrm{Poiss}(n_2;\nu_2) = \mathrm{Poiss}(n;\nu_1+\nu_2) \tag{2.49}$$

This property descends from the fact that adding two uniform processes similar to what was considered in the example above, leads again to a uniform process, where the total rate is the sum of the two rates.

- Similarly, if we take a random fraction ε of the initial uniform sample, again we have a uniform sample that obeys the Poisson distribution. It can be demonstrated, in fact, that if we have a Poisson variable n_0 with an expected rate ν_0 and we consider the variable n distributed according to a binomial distribution with probability ε and size of the sample n_0, the variable n is again distributed according to a Poisson distribution with average $\nu = \varepsilon\nu_0$. In other words:

$$P(n;\nu_0,\varepsilon) = \sum_{n_0=0}^{\infty} \text{Poiss}(n_0;\nu_0)\,\text{Binom}(n;n_0,\varepsilon) = \text{Poiss}(n;\varepsilon\nu_0) \tag{2.50}$$

This is the case in which, for instance, we count the number of events from a Poissonian process recorded by a detector whose efficiency ε is not ideal ($\varepsilon < 1$).

2.10 Other Distributions Useful in Physics

Other PDF that are used in literature are presented in the following subsections. Though the list is not exhaustive, it covers the most commonly used PDF models.

2.10.1 Argus Function

The Argus collaboration introduced [1] a function that models many cases of combinatorial backgrounds where kinematical bounds produce a sharp edge. The Argus PDF is given by:

$$\boxed{A(x;\theta;\xi) = Nx\sqrt{1-\left(\frac{x}{\theta}\right)^2}\,e^{-\frac{1}{2}\xi^2\left[1-\left(\frac{x}{\theta}\right)^2\right]}}\,, \tag{2.51}$$

where N is a normalization coefficient which depends on the parameters θ and ξ. Examples of Argus distributions are shown in Fig. 2.8 for different values of the parameters θ and ξ. One advantage of this function is that its primitive function can be computed analytically, saving some computer time in the evaluation of the normalization coefficient N in Eq. (2.51). Assuming $\xi^2 \geq 0$, the normalization

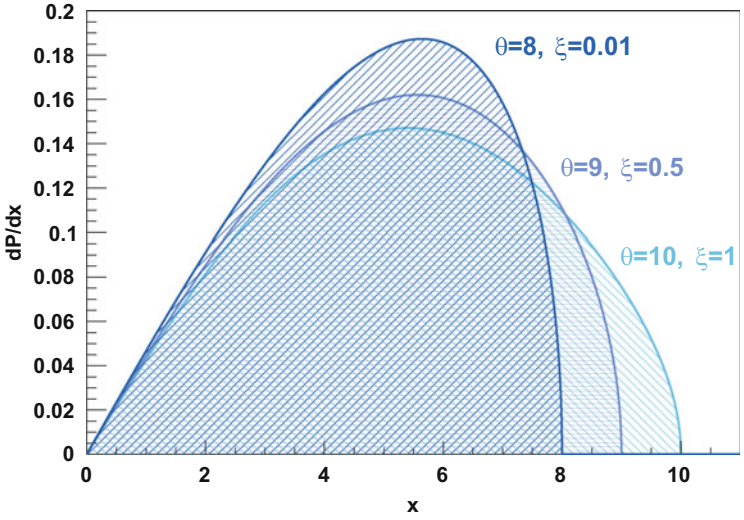

Fig. 2.8 Argus distributions with different parameter values

condition for the Argus PDF can be written as follows:

$$\frac{1}{N}\int A(x;\theta,\xi)\,\mathrm{d}x$$
$$= \frac{\theta^2}{\xi^2}\left\{e^{-\frac{1}{2}\xi^2\left[1-\left(1-\frac{x}{\theta}\right)^2\right]}\sqrt{1-\frac{x^2}{\theta^2}} - \sqrt{\frac{\pi}{2\xi^2}}\mathrm{erf}\sqrt{\frac{1}{2}\xi^2\left(1-\frac{x^2}{\theta^2}\right)}\right\}, \tag{2.52}$$

and the normalized expression of the Argus function becomes:

$$\boxed{A(x;\theta;\xi) = \frac{\xi^3}{\sqrt{2\pi}\Psi(\xi)}\frac{x}{\theta^2}\sqrt{1-\left(\frac{x}{\theta}\right)^2}e^{-\frac{1}{2}\xi^2\left[1-\left(\frac{x}{\theta}\right)^2\right]}}, \tag{2.53}$$

where $\Psi(\xi) = \Phi(\xi) - \xi\phi(\xi) - \frac{1}{2}$, $\phi(\xi)$ being a standard normal distribution (Eq. (2.25)) and $\Phi(\xi)$ its cumulative distribution (Eq. (2.26)).

2.10.2 Crystal Ball Function

Some random variables only approximately follow a Gaussian distribution, but exhibit an asymmetric tail on one of the two sides. In order to provide a description to such distributions, the collaboration working on the Crystal ball experiment at SLAC defined the following PDF [2] where a power-law distribution is used in

place of one of the two Gaussian tail, ensuring the continuity of the function and its first derivative. The Crystal ball PDF is defined as:

$$CB(x;\alpha,n,\bar{x},\sigma) = N \cdot \begin{cases} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right) & \text{for } \frac{x-\bar{x}}{\sigma} > -\alpha \\ A\left(B - \frac{x-\bar{x}}{\sigma}\right)^{-n} & \text{for } \frac{x-\bar{x}}{\sigma} \leq -\alpha \end{cases}, \tag{2.54}$$

where N is a normalization coefficient while A and B can be determined imposing the continuity of the function and its first derivative, which give:

$$A = \left(\frac{n}{|\alpha|}\right)^n e^{-\frac{\alpha^2}{2}}, \; B = \frac{n}{|\alpha|} - |\alpha|\,. \tag{2.55}$$

The parameter α determines the starting point of the power-law tail, measured in units of σ (the Gaussian "core" standard deviation).

Examples of Crystal ball distributions are shown in Fig. 2.9 where the parameter α has been varied, while the parameters of the Gaussian core have been fixed at $\mu = 0, \sigma = 1$ and the power-law exponent was fixed at $n = 2$.

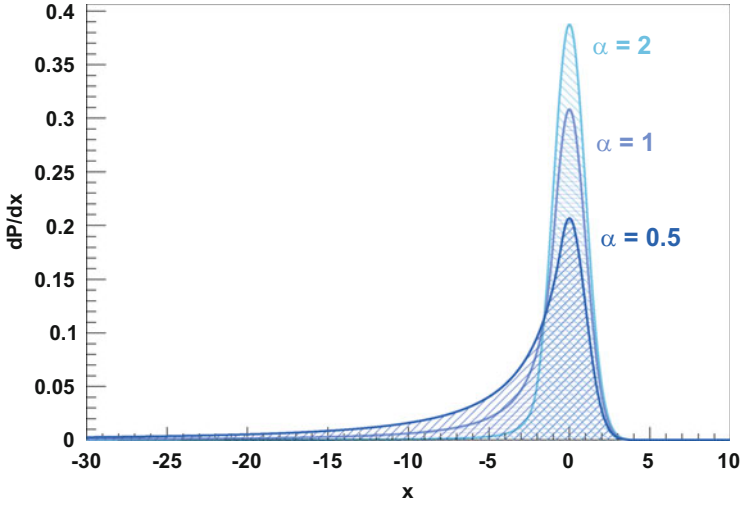

Fig. 2.9 Crystal ball distributions with $\mu = 0$, $\sigma = 1$, $n = 2$ and different values of the tail parameter α

2.10.3 Landau Distribution

A model that describes the fluctuations in energy loss of particles in a thin layers of matter is due to Landau [3, 4]. The distribution of the energy loss x is given by the following integral expression:

$$\boxed{L(x) = \frac{1}{\pi}\int_0^\infty e^{-t\log t - xt}\sin(\pi t)\,\mathrm{d}t\,.} \tag{2.56}$$

More frequently, the distribution is shifted by a constant μ and scaled by a constant σ, assuming the following expression:

$$L(x;\mu,\sigma) = L\left(\frac{x-\mu}{\sigma}\right)\,. \tag{2.57}$$

Examples of Landau distributions are shown in Fig. 2.10 for different values of σ fixing $\mu = 0$. This distribution is also used as empiric model for asymmetric distributions.

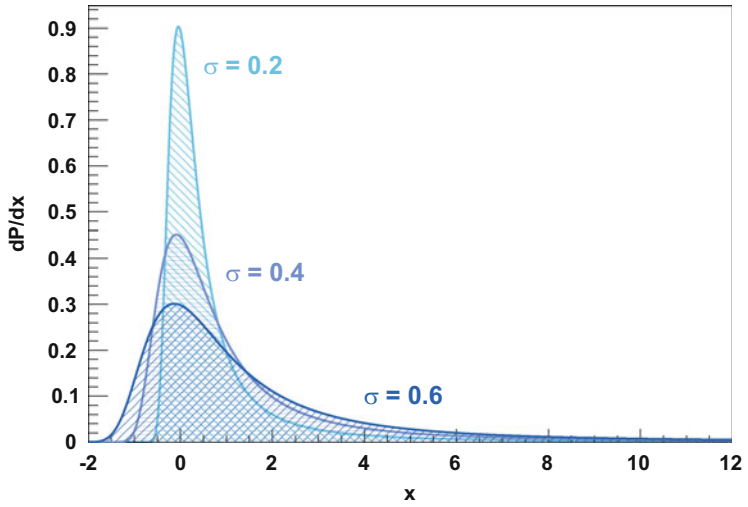

Fig. 2.10 Landau distributions with $\mu = 0$, and different values of σ

2.11 Central Limit Theorem

Given N independent random variables, $x_1, \cdots, x_N$, each distributed according to a PDF having finite variance, the average of those N variables can be approximated, in the limit of $N \to \infty$, to a Gaussian distribution, regardless of the underlying PDFs.

The demonstration of this theorem is not reported here, but quantitative approximate demonstrations of the central limit theorem for specific cases are easy to perform using numerical simulations (Monte Carlo techniques, see Chap. 4).

Two examples of such numerical exercises are shown in Figs. 2.11 and 2.12 where multiple random extractions from two different PDFs are summed and divided by the square root of the number of generated variables, so that this combination has the same variance of the original distribution. The distributions obtained with a large randomly-extracted sample are plotted, superimposed to a Gaussian distribution.

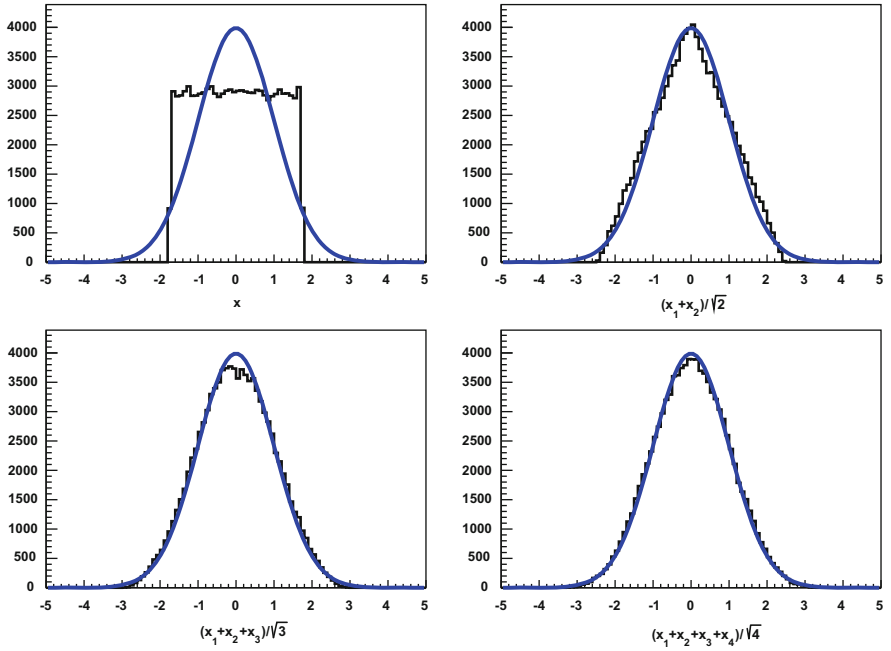

Fig. 2.11 Approximate visual demonstration of the central limit theorem using a Monte Carlo technique. A random variable x_1 is generated uniformly in the interval $[-\sqrt{3}, \sqrt{3}[$, in order to have average value $\mu = 0$ and standard deviation $\sigma = 1$. The *top-left* plot shows the distribution of 10^5 random extractions of x_1; the other plots show 10^5 random extractions of $(x_1 + x_2)/\sqrt{2}$, $(x_1 + x_2 + x_3)/\sqrt{3}$ and $(x_1 + x_2 + x_3 + x_4)/\sqrt{4}$ respectively, where all x_i obey the same uniform distribution as x_1. A Gaussian curve with $\mu = 0$ and $\sigma = 1$, with proper normalization in order to match the sample size, is superimposed to the extracted sample in the four cases. The Gaussian approximation is more and more stringent as a larger number of variables is added

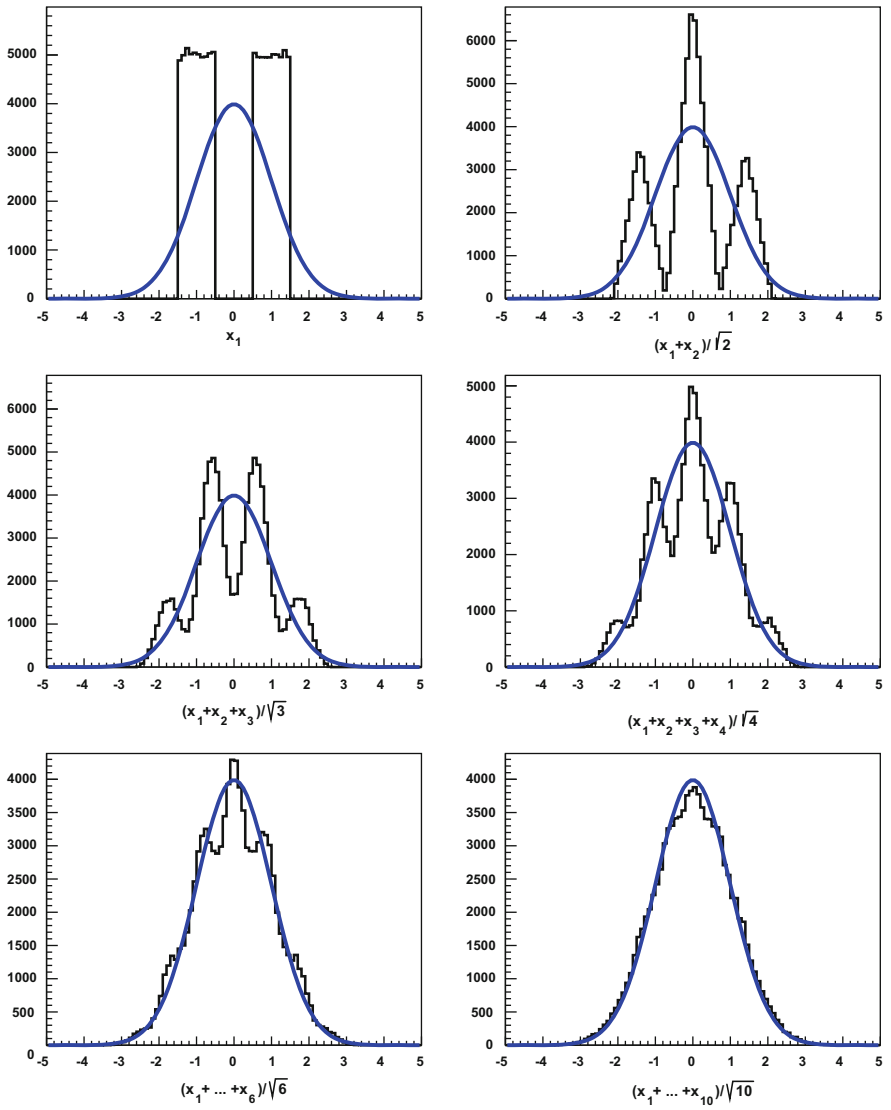

Fig. 2.12 Same as Fig. 2.11, using a PDF that is uniformly distributed in two disjoint intervals, $[-\frac{3}{2}, -\frac{1}{2}[$ and $[\frac{1}{2}, \frac{3}{2}[$, in order to have average value $\mu = 0$ and standard deviation $\sigma = 1$. The sum of 1, 2, 3, 4, 6 and 10 independent random extractions of such a variable, divided by $\sqrt{n}$, $n = 1, 2, 3, 4, 6, 10$ respectively, are shown with a Gaussian distribution having $\mu =$ and $\sigma = 1$ superimposed

2.12 Convolution of Probability Distribution Functions

A typical problem in data analysis is to combine the effect of the detector response, e.g.: due to finite resolution, with the theoretical distribution of some observable quantity x. If the original theoretical distribution is given by $f(x)$ and, for a measured value x' of the observable x, the detector response distribution is given by $r(x;x')$, the PDF that describes the distribution of the measured value of x, taking into account both the original theoretical distribution and the detector response, is give by the *convolution* of the two pdf f and r, defined as:

$$\boxed{g(x) = \int f(x')r(x;x')\,\mathrm{d}x'\,.} \tag{2.58}$$

in many cases r only depends on the difference $x - x'$, and can be described by a one-dimensional pdf: $r(x;x') = r(x-x')$. Sometimes the notation $g = f \otimes r$ is also used in those cases.

Convolutions have interesting properties under Fourier transform. In particular, let's define as usual the Fourier transform of f as:

$$\hat{f}(k) = \int_{-\infty}^{+\infty} f(x)e^{-ikx}\,\mathrm{d}x\,, \tag{2.59}$$

and conversely the inverse transform as:

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{f}(k)e^{ikx}\,\mathrm{d}k\,. \tag{2.60}$$

It is possible to demonstrate that the Fourier transform of the convolution of two PDF is given by the product of the Fourier transforms of the two PDF, i.e.:

$$\widehat{f \otimes g} = \hat{f}\hat{g}\,. \tag{2.61}$$

Conversely, also the Fourier transform of the product of two PDF is equal to the convolution of the two Fourier transforms, i.e.:

$$\widehat{fg} = \hat{f} \otimes \hat{g}\,. \tag{2.62}$$

This property allows in several applications to easily perform convolutions using numerical algorithms that implement the fast Fourier transform (FFT) [5].

A widely used model for detector resolution is a Gaussian PDF:

$$r(x;x') = g(x-x') = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-x')^2}{2\sigma^2}}\,. \tag{2.63}$$

The Fourier transform of a Gaussian PDF, like in Eq. (2.63), can be computed analytically and is given by:

$$\hat{g}(k) = e^{-ik\mu} e^{-\frac{\sigma^2 k^2}{2}} . \tag{2.64}$$

The above expression can be used to simplify the implementation of numerical algorithms that compute FFTs.

2.13 Probability Distribution Functions in More than One Dimension

Probability densities can be defined in spaces with more than one dimension. In the simplest case of two-dimensions, the PDF $f(x, y)$ measures the probability density per unit area, i.e. the ratio of the probability $\mathrm{d}P$ corresponding to an infinitesimal interval around a point (x, y) and the area of the interval $\mathrm{d}x\,\mathrm{d}y$:

$$\frac{\mathrm{d}P}{\mathrm{d}x\,\mathrm{d}y} = f(x, y) . \tag{2.65}$$

In three dimensions the PDF measures the probability density per volume area:

$$\frac{\mathrm{d}P}{\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z} = f(x, y, z) , \tag{2.66}$$

and so on for more dimensions. A multidimensional PDF is also called *joint probability distribution*.

2.13.1 Marginal Distributions

Given a two-dimensional PDF $f(x, y)$, the *marginal distributions* are the probability distributions of the two variables x and y and can be determined by integrating $f(x, y)$ over the other coordinate, y and x, respectively:

$$\boxed{f_x(x) = \int f(x, y)\,\mathrm{d}y ,} \tag{2.67}$$

$$\boxed{f_y(y) = \int f(x, y)\,\mathrm{d}x .} \tag{2.68}$$

The above expressions can also be seen as a special case of continuous variable transformation, as described in Sect. 2.4, where the applied transformations maps the two variables into one of the two: $(x, y) \to x$ or $(x, y) \to y$.

More in general, if we have a PDF in $n = h + k$ variables $(\vec{x}, \vec{y}) = (x_1, \cdots, x_h, y_1, \cdots, y_k)$, the marginal PDF of the subset of h variables $(x_1, \cdots, x_h)$ can be determined by integrating the PDF $f(\vec{x}, \vec{y})$ over the remaining set of variables $(y_1, \cdots, y_k)$:

$$\boxed{f_{\vec{x}}(\vec{x}) = \int f(\vec{x}, \vec{y})\, \mathrm{d}^k y\,.} \tag{2.69}$$

A pictorial view that illustrates the interplay between the joint distribution $f(x, y)$ and the marginal distributions $f_x(x)$ and $f_y(y)$ is shown in Fig. 2.13. Let's remember that, according to Eq. (1.6), two events A and B are independent if $P(A \cap B) = P(A)P(B)$. In the case shown in Fig. 2.13 we can consider as events A and B the cases where the two variables $\hat{x}$ and $\hat{y}$ are extracted in the intervals $[x, x + \delta x[$ and $[y, y + \delta y[$, respectively:

$$A = \text{“}x \le \hat{x} < x + \delta x\text{”} \tag{2.70}$$

and

$$B = \text{“}y \le \hat{y} < y + \delta y\text{”}\,. \tag{2.71}$$

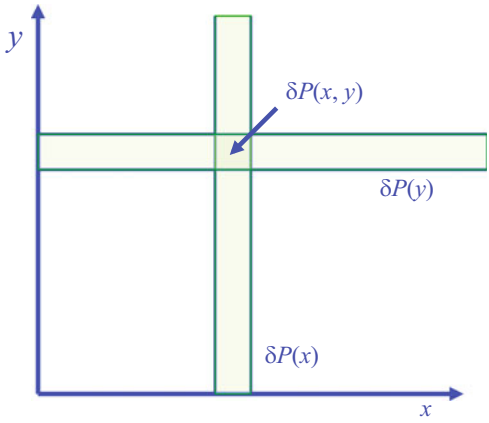

Fig. 2.13 In the two-dimensional plane (x, y), a slice in x corresponds to a probability $\delta P(x) = f_x(x)\delta x$, a slice in y corresponds to a probability $\delta P(y) = f_y(y)\delta y$, and their intersection to a probability $\delta P(x, y) = f(x, y)\delta x \delta y$

So, we have:

$$P(A \cap B) = \text{“}x \leq \hat{x} < x + \text{x and } y \leq \hat{y} < y + \delta y\text{”} = f(x,y)\delta x \delta y\,. \tag{2.72}$$

Since:

$$P(A) = \delta P(x) = f_x(x)\delta x \tag{2.73}$$

and:

$$P(B) = \delta P(y) = f_y(y)\delta y\,, \tag{2.74}$$

we have:

$$P(A)P(B) = \delta P(x,y) = f_x(x)f_y(y)\delta x \delta y\,. \tag{2.75}$$

The equality $P(A \cap B) = P(A)P(B)$ holds if and only if $f(x,y)$ can be factorized into the product of the two marginal PDFs:

$$\boxed{f(x,y) = f_x(x)f_y(y)\,.} \tag{2.76}$$

From this result, we can say that x and y are *independent variables* if their joint PDF can be written as the product of a PDF of the variable x times a PDF of the variable y.

More in general, n variables $x_1, \cdots, x_n$ are said to be independent if their n-dimensional PDF can be factorized into the product of n one-dimensional PDF in each of the variables:

$$\boxed{f(x_1, \cdots, x_n) = f_1(x_1) \cdots f_n(x_n)\,.} \tag{2.77}$$

In a less strict sense, the variables sets $\vec{x} = (x_1, \cdots, x_n)$ and $\vec{y} = (y_1, \cdots, y_m)$ are independent if:

$$\boxed{f(\vec{x},\vec{y}) = f_x(\vec{x})f_y(\vec{y})\,.} \tag{2.78}$$

Note that if two variable x and y are independent, it can be easily demonstrated that they are also uncorrelated, in the sense that their covariance (Eq. (1.22)) is null. Conversely, if two variables are uncorrelated, they are not necessarily independent, as shown in the Example 2.6 below.

Example 2.6 – Uncorrelated Variables that Are Not Independent

An example of PDF that describes uncorrelated variables that are not independent is given by the sum of four two-dimensional Gaussian PDFs as specified below:

$$f(x,y) = \frac{1}{4}(g(x;\mu,\sigma)g(y;0,\sigma) + g(x;-\mu,\sigma)g(y;0,\sigma) \\ g(x;0,\sigma)g(y;\mu,\sigma) + g(x;0,\sigma)g(y;-\mu,\sigma)) \,, \tag{2.79}$$

where g is a one-dimensional Gaussian distribution.

This example is illustrated in Fig. 2.14 which plots the PDF from Eq. (2.79) with numerical values $\mu = 2.5$ and $\sigma = 0.7$.

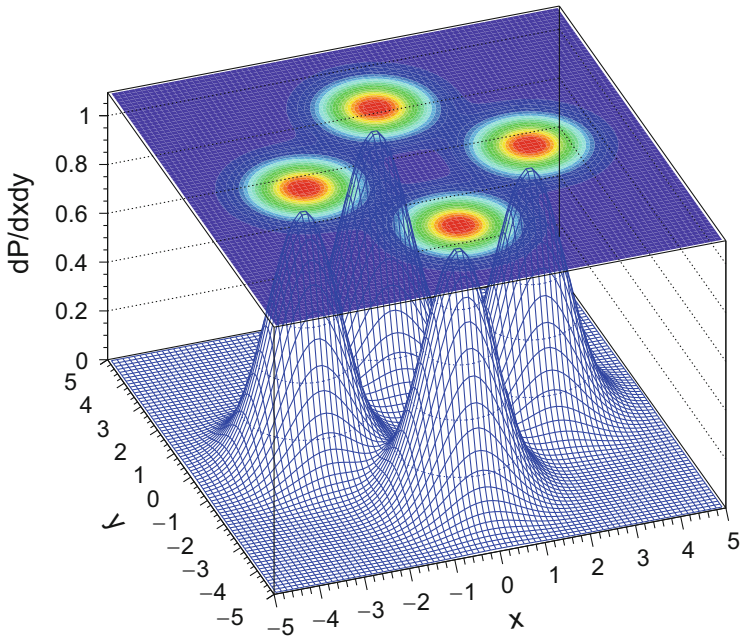

Fig. 2.14 Example of a PDF of two variables x and y that are uncorrelated but are not independent

Considering that, for a variable z distributed according to $g(z;\mu,\sigma)$, the following relations hold:

$$\begin{aligned}\langle z\rangle &= \mu \ ,\\ \langle z^2\rangle &= \mu^2+\sigma^2 \ ,\end{aligned}$$

it is easy to demonstrate that for x and y obeying the PDF in Eq. (2.79), the following relations also hold:

$$\begin{aligned}\langle x\rangle &= \langle y\rangle = 0 \ ,\\ \langle x^2\rangle &= \langle y^2\rangle = \sigma^2+\tfrac{\mu^2}{2} \ ,\\ \langle xy\rangle &= 0 \ .\end{aligned}$$

Hence, according to Eq. (1.22), $\mathrm{cov}(x,y)=0$. Hence, x and y are uncorrelated, but are clearly not independent because the PDF in Eq. (2.79) can't be factorized into the product of two PDF: there is no pair of functions $f_x(x)$ and$f_y(y)$ such that $f(x,y)=f_x(x)f_y(y)$.

2.13.2 Conditional Distributions

Given a two-dimensional PDF $f(x,y)$ and a fixed value x_0 of the variable x, the conditional PDF of y given x_0 is defined as:

$$\boxed{f(y|x_0)=\frac{f(x_0,y)}{\int f(x_0,y')\,\mathrm{d}y'}\,.} \tag{2.80}$$

It can be interpreted as being obtained by "slicing" $f(x,y)$ at $x=x_0$ and normalizing the "sliced" one-dimensional PDF. Reminding Eq. (1.4), and considering again the example in Fig. 2.13, this definition of conditional distribution is consistent with the definition of conditional probability: $P(A|B)=P(A\cap B)/P(B)$, where A ="$y\le\hat{y}<y+\delta y$", $\hat{y}$ being the extracted value of y, and $B=$"$x\le x_0<x+\delta x$",

An illustration of conditional PDF is shown in Fig. 2.15.

In more than two dimensions, we can generalize Eq. (1.4) for a PDF of n $n=h+k$ variables $(\vec{x},\vec{y})=(x_1,\cdots,x_h,y_1,\cdots,y_k)$ as:

$$\boxed{f(\vec{y}|\vec{x}_0)=\frac{f(\vec{x}_0,\vec{y})}{\int f(\vec{x}_0,\vec{y})\,\mathrm{d}y'_1\cdots\mathrm{d}y'_k}\,.} \tag{2.81}$$

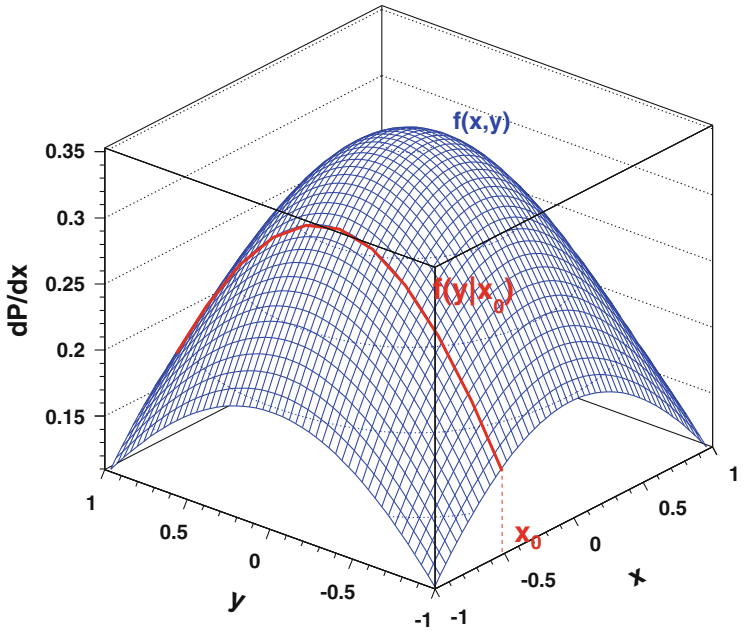

Fig. 2.15 Illustration of conditional PDF in two dimensions

2.14 Gaussian Distributions in Two or More Dimensions

Let's consider in a two-dimensional variable space (x', y') the product of two Gaussian distributions for the variables x' and y' having standard deviations $\sigma_{x'}$ and $\sigma_{y'}$ respectively and for simplicity having both averages $\mu_x = \mu_y = 0$ (a translation can always be applied to generalize to the case $\mu_x, \mu_y \neq 0$):

$$g'(x', y') = \frac{1}{2\pi\sigma_{x'}\sigma_{y'}} \exp\left[-\frac{1}{2}\left(\frac{x'^2}{\sigma_{x'}^2} + \frac{y'^2}{\sigma_{y'}^2}\right)\right]. \tag{2.82}$$

Let's apply a rotation from (x', y') to (x, y) with an angle ϕ defined by:

$$\begin{cases} x' = x\cos\phi - y\sin\phi \\ y' = x\sin\phi + y\cos\phi \end{cases}. \tag{2.83}$$

The transformed PDF $g(x, y)$ can be obtained using Eq. (2.20) considering that $\det\left|\partial x_i'/\partial x_j\right| = 1$, which leads to $g'(x', y') = g(x, y)$. $g(x, y)$ has the form:

$$g(x, y) = \frac{1}{2\pi|C|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left((x, y)C^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right)\right], \tag{2.84}$$

where the matrix C^{-1} is the inverse of the covariance matrix of the variables (x, y). C^{-1} can be obtained by comparing Eqs. (2.82) and (2.84), from which one can substitute the rotated coordinates defined Eq. (2.83) in the following equation:

$$\frac{x'^2}{\sigma_{x'}^2} + \frac{y'^2}{\sigma_{y'}^2} = (x, y)C^{-1}\begin{pmatrix} x \\ y \end{pmatrix}, \tag{2.85}$$

leading to:

$$C^{-1} = \begin{pmatrix} \frac{\cos^2\phi}{\sigma_{x'}^2} + \frac{\sin^2\phi}{\sigma_{y'}^2} & \sin\phi\cos\phi\left(\frac{1}{\sigma_{y'^2}} - \frac{1}{\sigma_{x'}^2}\right) \\ \sin\phi\cos\phi\left(\frac{1}{\sigma_{y'^2}} - \frac{1}{\sigma_{x'}^2}\right) & \frac{\sin^2\phi}{\sigma_{x'}^2} + \frac{\cos^2\phi}{\sigma_{y'}^2} \end{pmatrix}. \tag{2.86}$$

The determinant of C^{-1} that appears in the denominator of Eq. (2.84) under a square root is:

$$|C^{-1}| = \frac{1}{\sigma_{x'}^2\sigma_{y'}^2} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho_{xy}^2)}, \tag{2.87}$$

where ρ_{xy} is the correlation coefficient defined in Eq. (1.23). Inverting the matrix C^{-1}, we can get the covariance matrix of the rotated variables (x, y):

$$C = \begin{pmatrix} \cos^2\phi\sigma_{x'}^2 + \sin^2\phi\sigma_{y'}^2 & \sin\phi\cos\phi(\sigma_{y'}^2 - \sigma_{x'}^2) \\ \sin\phi\cos\phi(\sigma_{y'}^2 - \sigma_{x'}^2) & \sin^2\phi\sigma_{x'}^2 + \cos^2\phi\sigma_{y'}^2 \end{pmatrix}. \tag{2.88}$$

Considering that a covariance matrix should have the form:

$$C = \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}, \tag{2.89}$$

the variances and correlation coefficient of x and y can be determined as:

$$\sigma_x^2 = \cos^2\phi\sigma_{x'}^2 + \sin^2\phi\sigma_{y'}^2, \tag{2.90}$$

$$\sigma_y^2 = \sin^2\phi\sigma_{x'}^2 + \cos^2\phi\sigma_{y'}^2, \tag{2.91}$$

$$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x\sigma_y} = \frac{\sin 2\phi(\sigma_{y'}^2 - \sigma_{x'}^2)}{\sqrt{\sin 2\phi(\sigma_{x'}^4 + \sigma_{y'}^4) + 2\sigma_{x'}^2\sigma_{y'}^2}}. \tag{2.92}$$

The last Eq. (2.92), implies that the correlation coefficient is equal to zero if either $\sigma_{y'} = \sigma_{x'}$ or if ϕ is a multiple of $\frac{\pi}{2}$. We also have the following relation that gives

$\tan 2\phi$ in terms of the elements of the covariance matrix:

$$\tan 2\phi = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_y^2 - \sigma_x^2} \,. \tag{2.93}$$

The transformed PDF can be finally written as:

$$\boxed{g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left[-\frac{1}{2(1-\rho_{xy}^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2xy\rho_{xy}}{\sigma_x\sigma_y}\right)\right] .} \tag{2.94}$$

The geometrical interpretation of σ_x and σ_y in the rotated coordinate system is shown in Fig. 2.16, where the one-sigma ellipse obtained from the following equation is drawn:

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2xy\rho_{xy}}{\sigma_x\sigma_y} = 1 \,. \tag{2.95}$$

It is possible to demonstrate that the distance of the horizontal and vertical tangent lines to the ellipse have a distance from their respective axes equal to σ_y and σ_x respectively.

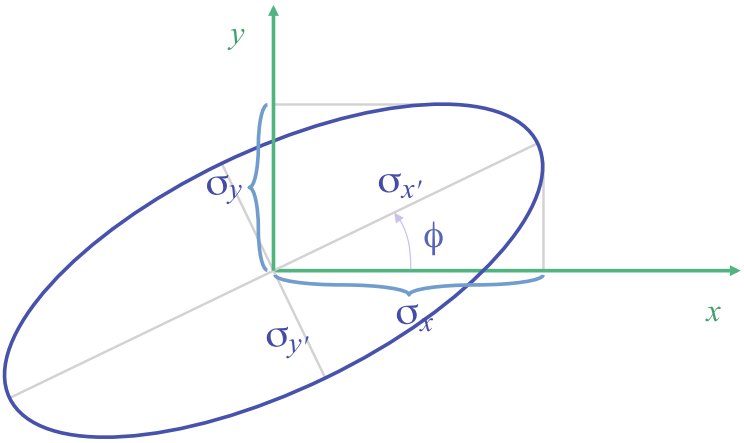

Fig. 2.16 Plot of the one-sigma contour for a two-dimensional Gaussian PDF. The two ellipse axes are equal to $\sigma_{x'}$ and $\sigma_{y'}$; the x' axis is rotated of an angle ϕ with respect to the x axis and the lines tangent to the ellipse parallel to the x and y axes have a distance with respect to the respective axes equal to σ_y and σ_x respectively

Projecting the two-dimensional Gaussian on one of the two coordinates from Eq. (2.94), one obtains the marginal PDFs:

$$g_x(x) = \int_{-\infty}^{+\infty} g(x,y)\,\mathrm{d}y = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{x^2}{2\sigma_x^2}}, \tag{2.96}$$

$$g_y(y) = \int_{-\infty}^{+\infty} g(x,y)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{y^2}{2\sigma_y^2}}. \tag{2.97}$$

In general, projecting a two-dimensional Gaussian PDF in any direction leads to a one-dimensional Gaussian whose standard deviation is equal to the distance of the tangent line to the ellipse perpendicular to the axis along which the two-dimensional Gaussian is projected. This is visually shown in Fig. 2.17.

The probability corresponding to a $Z\sigma$ one-dimensional interval for a Gaussian distribution was reported in Eq. (2.27). In order to extend this result for the two-dimensional case, the integration of $g(x,y)$ should be performed in two dimension over the ellipse that corresponding to $Z\sigma$:

$$P_{2D}(Z\sigma) = \int_{E_Z} g(x,y)\,\mathrm{d}x\,\mathrm{d}y\,, \tag{2.98}$$

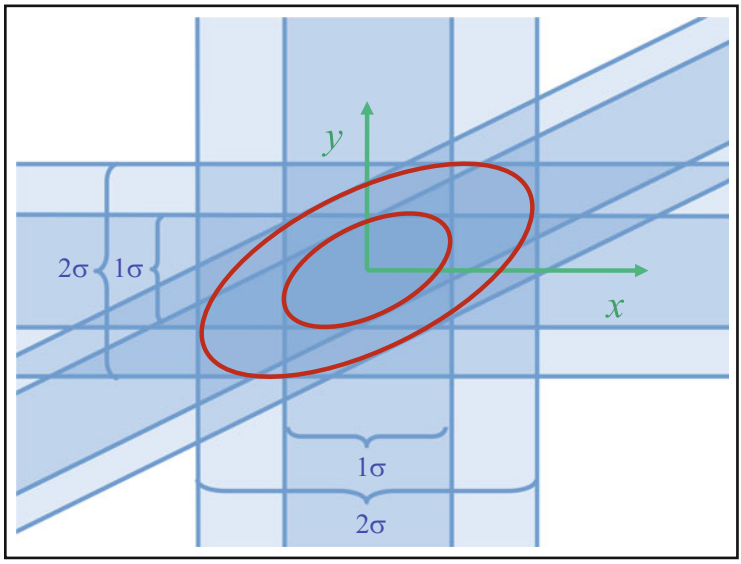

Fig. 2.17 Plot of the two-dimensional one- and two-sigma Gaussian contours

where

$$E_Z = \left\{ (x, y) : \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2xy\rho_{xy}}{\sigma_x \sigma_y} \leq Z \right\} . \tag{2.99}$$

The integral $P_{2D}(Z\sigma)$ simplifies to:

$$P_{2D}(Z\sigma) = \int_0^Z e^{-\frac{r^2}{2}} r \, \mathrm{d}r = 1 - e^{-\frac{Z^2}{2}} , \tag{2.100}$$

that can be compared to the one-dimensional case:

$$P_{1D}(Z\sigma) = \sqrt{\frac{2}{\pi}} \int_0^Z e^{-\frac{x^2}{2}} \, \mathrm{d}x = \mathrm{erf}(\frac{Z}{\sqrt{2}}) . \tag{2.101}$$

The probabilities corresponding to 1σ, 2σ and 3σ are reported for the one- and two-dimensional cases in Table 2.1. The two-dimensional integrals are in all case smaller than the one-dimensional case for a given Z. In particular, to recover the same probability as the one-dimensional interval, one has to enlarge the two-dimensional ellipse from 1σ to 1.515σ, from 2σ to 2.486σ, and from 3σ to 3.439σ.

Figure 2.17 shows the 1σ and 2σ contours for a two-dimensional Gaussian, and three possible choices of 1σ and 2σ one-dimensional bands, two along the x and y axes and one along a third generic oblique direction.

The generalization to n dimensions of the two-dimensional Gaussian described in Eq. (2.94) is:

$$\boxed{g(x_1, \cdots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |C|^{-\frac{1}{2}}} \exp\left[-\frac{1}{2} \left((x_i - \mu_i) C_{ij}^{-1} (x_j - \mu_j) \right) \right] ,} \tag{2.102}$$

where μ_i is the average of the variable x_i and C_{ij} is the $n \times n$ covariance matrix of the variables $x_1, \cdots, x_n$.

Table 2.1 Probabilities corresponding to $Z\sigma$ one-dimensional and two-dimensional contours for different values of Z

	P_{1D}	P_{2D}
1σ	0.6827	0.3934
2σ	0.9545	0.8647
3σ	0.9973	0.9889
1.515σ	0.8702	0.6827
2.486σ	0.9871	0.9545
3.439σ	0.9994	0.9973

References

1. ARGUS collaboration, H. Albrecht et al., Search for hadronic b→u decays. Phys. Lett. **B241**, 278–282 (1990)
2. J. Gaiser, *Charmonium Spectroscopy from Radiative Decays of the J/ψ and ψ′*. Ph.D, thesis, Stanford University, 1982. Appendix F
3. L. Landau, On the energy loss of fast particles by ionization. J. Phys. (USSR) **8**, 201 (1944)
4. W. Allison, J. Cobb, Relativistic charged particle identification by energy loss. Annu. Rev. Nucl. Part. Sci. **30**, 253–298 (1980)
5. R. Bracewell, *The Fourier Transform and Its Applications*, 3rd. edn. (McGraw-Hill, New York, 1999)

Chapter 3
Bayesian Approach to Probability

The Bayesian approach to probability allows to define in a quantitative way probabilities associated to statement whose truth or falsity is not known. This also applies to the knowledge of unknown parameters in physics, allowing to assign a probability to the possibility that a parameter's value lies within a certain interval.

The mathematical tools needed to define a quantitative treatment of such cases start from an extension of the Bayes' theorem. Bayes' theorem, presented in the following section, has general validity for any probability approach, including frequentist probability.

3.1 Bayes' Theorem

According to the definition of conditional probability in Eq. (1.4), the probability of an event A given the condition that the event B has occurred is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \,. \tag{3.1}$$

We can conversely write the probability of the event B given the event A as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \,. \tag{3.2}$$

This situation is visualized in Fig. 3.1. Extracting from Eqs. (3.1) and (3.2) the common term $P(A \cap B)$, we obtain:

$$P(A|B)P(B) = P(B|A)P(A) \,, \tag{3.3}$$

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_3

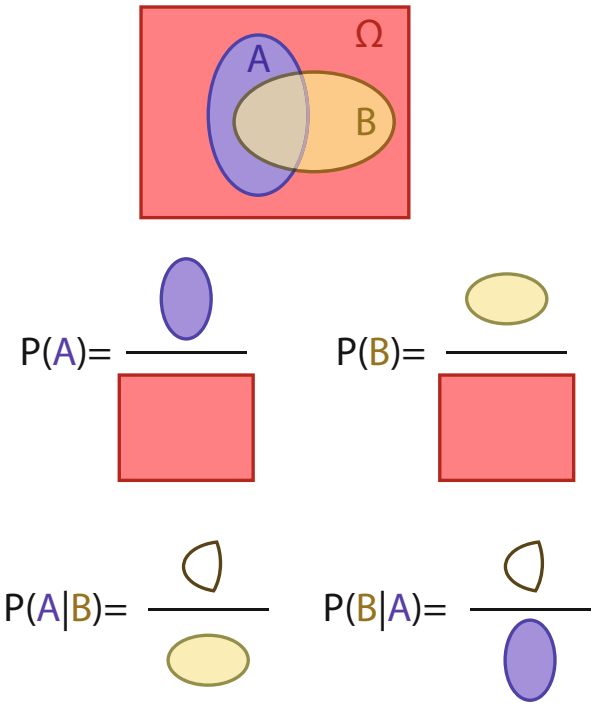

Fig. 3.1 Visualization of the conditional probabilities, $P(A|B)$ and $P(B|A)$ due to Robert Cousins. The events A and B are represented as subsets of a sample space Ω

from which the Bayes' theorem can be derived in the following form:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}.} \tag{3.4}$$

The probabilities $P(A)$ and $P(A|B)$ can be interpreted as probability of the event A *before* the knowledge that the event B has occurred (*prior* probability) and as the probability of the same event A having as further information the knowledge that the event B has occurred (*posterior* probability).

A visual derivation of Bayes' theorem, due to Robert Cousins, is presented in Fig. 3.2.

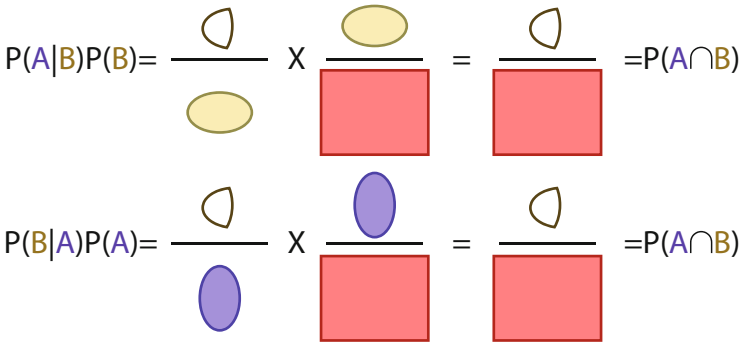

Fig. 3.2 Visualization of the Bayes' theorem due to Robert Cousins. The areas of events A an B, $P(A)$ and $P(B)$ respectively, simplify as we multiply $P(A|B)P(B)$ and $P(B|A)P(A)$

Example 3.7 – An Epidemiology Example

One of the most popular concrete applications of Bayes' theorem occurs in the "inversion" of conditional probability. The most popular case consists of finding the probability that a person who received a positive diagnosis of some a illness is really ill, knowing the probability that the test may give a false positive outcome. This example has been reported in several lecture series and books, for instance, in [1, 2].

Imagine a person has received a diagnosis of a illness through a test. We know that, if a person is really ill, the probability that the test gives a positive result is 100 %. But the test has also a small probability, say 0.2 %, to give a false positive result on a healthy person.

If a random person does the test, and receives a positive diagnosis, what is the probability that he/she is really ill? The naïve answer that this probability is $100\,\% - 0.2\,\% = 99.8\,\%$ is clearly wrong, as will be more clear in the following. But if this question is asked to a person who has a limited statistical background, it is a very likely answer to receive.

The problem can be formalized as follows:

$$P(+|\text{ill}) \simeq 100\%\,, \tag{3.5}$$

$$P(-|\text{ill}) \simeq 0\%\,, \tag{3.6}$$

$$P(+|\text{healthy}) = 0.2\%\,, \tag{3.7}$$

$$P(-|\text{healthy}) = 99.8\%\,. \tag{3.8}$$

What we want to know is the probability that a person is really ill after having received a positive diagnosis, i.e.: $P(\text{ill}|+)$. Using Bayes' theorem, we can "invert" the conditional probability as follows:

$$P(\text{ill}|+) = \frac{P(+|\text{ill})P(\text{ill})}{P(+)}\,, \tag{3.9}$$

which, given that $P(+|\text{ill}) \simeq 1$, gives approximately:

$$P(\text{ill}|+) \simeq \frac{P(\text{ill})}{P(+)}\,. \tag{3.10}$$

We can now identify a missing ingredient in the problem: we need to know $P(\text{ill})$, i.e.: the probability that a random person in the population under consideration is really ill (regardless of any possibly performed test). In a normal situation of a generally healthy population, we will have $P(\text{ill}) \ll P(\text{healthy})$. Using:

$$P(\text{ill}) + P(\text{healthy}) = 1\,, \tag{3.11}$$

and:

$$P(\text{ill } \textit{and } \text{healthy}) = 0\,, \tag{3.12}$$

we can decompose $P(+)$ as follows, according to the law of total probability, in Eq. (1.12) (from Sect. 1.7):

$$P(+) = P(+|\text{ill})P(\text{ill}) + P(+|\text{healthy})P(\text{healthy}) \simeq P(\text{ill}) + P(+|\text{healthy})\,. \tag{3.13}$$

The probability $P(\text{ill}|+)$ can then be written as:

$$P(\text{ill}|+) = \frac{P(\text{ill})}{P(+)} \simeq \frac{P(\text{ill})}{P(\text{ill}) + P(+|\text{healthy})}\,. \tag{3.14}$$

If we assume that $P(\text{ill})$ is smaller than $P(+|\text{healthy})$ then $P(\text{ill}|+)$ will result smaller than 50 %. For instance, if $P(\text{ill}) = 0.15\,\%$, compared with the assumption that $P(+|\text{healthy}) = 0.2\,\%$, then

$$P(\text{ill}|+) = \frac{0.15}{0.15 + 0.20} = 43\%\,. \tag{3.15}$$

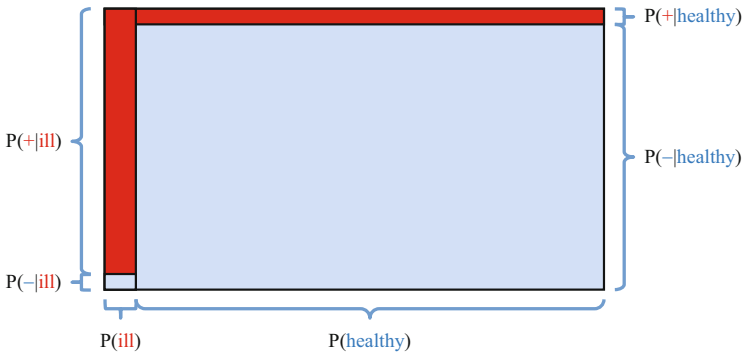

Fig. 3.3 Visualization of the ill/healthy case considered in the Example 3.7. The *red areas* correspond to the cases of a positive diagnosis for a ill person ($P(+|\text{ill})$, *vertical red area*) and a positive diagnosis for a healthy person ($P(+|\text{healthy})$, *horizontal red area*). The probability of being really ill in the case of a positive diagnosis, $P(\text{ill}|+)$, is equal to the ratio of the *vertical red area* and the *total red area*. In the example it was assumed that $P(-|\text{ill}) \simeq 0$

The probability to be really ill, given the positive diagnosis, is very different from the naïve conclusion that one is most likely really hill. The situation can be visualized, changing a bit the proportions for a better presentation, in Fig. 3.3.

Having a high probability of a positive diagnosis in case of illness does not imply that a positive diagnosis turns into a high probability of being really ill. The correct answer, in this case, also depend on the *prior* probability for a random person in the population to be ill, ($P(\text{ill})$). Bayes' theorem allows to compute the *posterior* probability $P(\text{ill}|+)$ in terms of the prior probability and the probability of a positive diagnosis for a ill person, ($P(+|\text{ill})$).

Example 3.8 – Purity of a Sample with Particle Identification

A derivation similar to the previous Example 3.7 can be applied to the case of a detector that performs particle identification, where the conclusions will be less counterintuitive than in the previous case, since this situation is very familiar with a physicist's experience.

Consider, for instance, a muon detector that gives a positive signal for real muons with an efficiency $\varepsilon = P(+|\mu)$ and gives a false positive signal for pions with a probability $\delta = P(+|\pi)$. Given a collection of particles that can be either muons or pions, what is the probability that a selected particle is really a muon, i.e.: $P(\mu|+)$? As before, we can't give an answer, unless we also give the *prior* probability, i.e.: the probability that a random particle from the sample is really a muon or pion. In other words, we should know $P(\mu)$ and $P(\pi) = 1 - P(\mu)$. Using Bayes' theorem, together with Eq. (1.12), we can write, as in the example from the previous section:

$$P(\mu|+) = \frac{P(+|\mu)P(\mu)}{P(+)} = \frac{P(+|\mu)P(\mu)}{P(+|\mu)P(\mu) + P(+|\pi)P(\pi)}\,, \tag{3.16}$$

or, in terms of the fraction of muons $f_\mu = P(\mu)$ and the fraction of pions $f_\pi = P(\pi)$ of the original sample, we can write *purity* of the sample, defined as the fraction of muons in a sample of selected particles, as:

$$\Pi_\mu = f_{\mu,+} = \frac{\varepsilon f_\mu}{\varepsilon f_\mu + \delta f_\pi}\,. \tag{3.17}$$

Another consequence of Bayes' theorem is the relation between ratios of posterior probability and ratios of prior probability. The posteriors' ratio can in fact be written as:

$$\frac{P(\mu|+)}{P(\pi|+)} = \frac{P(+|\mu)}{P(+|\pi)} \cdot \frac{P(\mu)}{P(\pi)}\,. \tag{3.18}$$

The above expression also holds if more than two possible particle types are present in the sample, and does not require to compute the denominator that is present in Eq. (3.16), which would be needed, instead, in order to compute probabilities related to all possible particle cases.

3.2 Bayesian Probability Definition

So far, Bayes' theorem has been applied in cases that also fall under the frequentist domain. Let's now consider again the following formulation of Bayes' theorem, as from Eq. (3.4):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}\,. \tag{3.19}$$

We can interpret it as follows: *before* the occurrence of the event B (or knowledge that B is true), our *degree of belief* of the event A is equal to the *prior* probability, $P(A)$. *After* the occurrence of the event B (or after we know that B is true) our degree of belief of the event A changes and becomes equal to the *posterior* probability $P(A|B)$.

Using this interpretation, we can extend the scope of the definition of probability, in this new Bayesian sense, to events that are not associated to random variables, but represent statements about unknown facts, like *"my football team will win next match"*, or *"the mass of a dark-matter candidate particle is between 1000 and 1500 GeV"*. We can in fact consider a prior probability $P(A)$ of such an unknown statement, representing a measurement of our prejudice about that statement, before the occurrence of any event that could modify our knowledge. After the event B has occurred (and we know that B as occurred), our knowledge of A must change in order to take into account the fact that we know B has occurred, and our degree of belief should be modified becoming equal to the posterior probability $P(A|B)$. In other words, Bayes theorem tells us how we should change our subjective degree of belief from an initial prejudice considering newly available information, according to a quantitative and rational method. Anyway, starting from different priors (i.e.: different prejudices), different posteriors will be determined.

The term $P(B)$ that appears in the denominator of Eq. (3.4) can be considered a normalization factor. If we can decompose the sample space Ω in a partition $A_1, \cdots, A_n$, where $\cup_{i=1}^{n} A_i = \Omega$ and $A_i \cap A_j = 0 \ \forall i, j$, we can write, according to the law of total probability in Eq. (1.12):

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)\,. \tag{3.20}$$

This decomposition was already used in the examples discussed in the previous sections.

The Bayesian definition of probability obeys Kolmogorov's axioms of probability, as defined in Sect. 1.4, hence all properties of classical probability discussed in Chap. 1 also apply to Bayesian probability.

An intrinsic unavoidable feature of Bayesian probability is that the probability associated to an event A can't be defined without having a prior probability of that event, which make Bayesian probability intrinsically subjective.

Example 3.9 – Extreme Cases of Prior Beliefs

Consider a set of possible events $\{A_i\}$ that constitute a non-intersecting partition of Ω. Imagine that we have as *prior* probabilities:

$$P(A_i) = \begin{cases} 1 \text{ if } i = 0 \\ 0 \text{ if } i \neq 0 \end{cases}. \tag{3.21}$$

This corresponds to a person that believes that A_0 is absolutely true, all other alternatives A_i are absolutely false for $i \neq 0$. Whatever event B occurs, that person's posterior probability on any A_i will not be different from the prior probability:

$$P(A_i|B) = P(A_i), \quad \forall B\,. \tag{3.22}$$

In fact, from Bayes' theorem:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}\,. \tag{3.23}$$

But, if $i \neq 0$, clearly:

$$P(A_i|B) = \frac{P(B|A_i) \times 0}{P(B)} = 0 = P(A_i)\,. \tag{3.24}$$

If $i = 0$, instead, we have, assuming $P(B|A_0) \neq 0$:

$$P(A_0|B) = \frac{P(B|A_0) \times 1}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} = \frac{P(B|A_0)}{P(B|A_0) \times 1} = 1 = P(A_0)\,. \tag{3.25}$$

This situation reflects the case that we may call *dogma*, or *religious belief*, i.e.: the case in which someone has such strong prejudices on A_i that no event B, i.e.: no new knowledge, can change his/her degree of belief.

The scientific method allowed to evolve mankind's knowledge of Nature during history by progressively adding more knowledge based on the observation of reality. The history of science is full of examples in which theories known to be true have been falsified by more precise observations, and new better theories have replaced the old ones.

According to Eq. (3.22), instead, scientific progress is not possible in the presence of religious beliefs about observable facts.

3.3 Bayesian Probability and Likelihood Functions

Given a sample $(x_1, \cdots, x_n)$ of n random variables whose PDF is known and depends on m parameters, $\theta_1, \cdots, \theta_m$, the *likelihood function* is defined as the probability density at the point $(x_1, \cdots, x_n)$ given a fixed set of values of the

parameters $\theta_1, \cdots, \theta_m$:

$$L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m) = \left. \frac{\mathrm{d}P(x_1, \cdots, x_n)}{\mathrm{d}x_1 \cdots \mathrm{d}x_n} \right|_{\theta_1, \cdots, \theta_m} . \tag{3.26}$$

As alternative to the notation $L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)$, sometimes the notation $L(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m)$ is also used equivalently, similarly to the notation used in the definition of conditional probability. The likelihood function will be more extensively discussed in Sect. 5.5.1.

From Eq. (3.26), we can define the *posterior* Bayesian probability distribution function for the parameters $\theta_1, \cdots, \theta_m$, given the observation of $(x_1, \cdots, x_n)$, to be:

$$P(\theta_1, \cdots, \theta_m | x_1, \cdots, x_n) = \frac{L(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m) \pi(\theta_1, \cdots, \theta_m)}{\int L(x_1, \cdots, x_n | \theta_1, \cdots, \theta_m) \pi(\theta_1, \cdots, \theta_m) \, \mathrm{d}^m \theta} , \tag{3.27}$$

where the probability distribution function $\pi(\theta_1, \cdots, \theta_m)$ is the *prior* PDF of the parameters $\theta_1, \cdots, \theta_m$, i.e., our *degree of belief* about the m unknown parameters *before* the observation of $(x_1, \cdots, x_n)$.

Fred James et al. gave the following interpretation of the *posterior* probability density from Eq. (3.27): *"The difference between $\pi(\theta)$ and $P(\theta|x)$ shows how one's knowledge (degree of belief) about θ has been modified by the observation x. The distribution $P(\theta|x)$ summarizes all one's knowledge of θ and can be used accordingly"* [3].

3.3.1 Repeated Use of Bayes' Theorem and Learning Process

If we start from a prior probability distribution function $P_0(\theta) = \pi(\theta)$ of an unknown parameter θ, we can apply Bayes' theorem after an observation x_1, and obtain a posterior probability:

$$P_1(\theta) \propto P_0(\theta) L(x_1|\theta) , \tag{3.28}$$

where we have omitted the normalization factor: $\int P_0(\theta) L(x_1|\theta) \, \mathrm{d}\theta$. After a second *independent* observation x_2, we can again apply Bayes' theorem, considering that the combined likelihood corresponding to the two measurements x_1 and x_2 is given by the product of the corresponding likelihoods:

$$L(x_1, x_2|\theta) = L(x_1|\theta) L(x_2|\theta) . \tag{3.29}$$

From Bayes theorem we now have:

$$P_2(\theta) \propto P_0(\theta) L(x_1, x_2|\theta) = P_0(\theta) L(x_1|\theta) L_(x_2|\theta) , \tag{3.30}$$

where we have again omitted a normalization factor. Equation (3.30) can be interpreted as the application of Bayes' theorem to the observation of x_2 having a prior probability $P_1(\theta)$, i.e.: the posterior probability after the observation of x_1, $P_1(\theta)$ (Eq. (3.28)). Considering a third independent observation x_3 we can again apply Bayes' theorem and obtain:

$$P_3(\theta) \propto P_2(\theta)L(x_3|\theta) = P_0(\theta)L(x_1|\theta)L(x_2|\theta)L(x_3|\theta)\,. \tag{3.31}$$

Adding more measurements would allow to apply repeatedly Bayes' theorem. This possibility allows to interpret the application of Bayes' theorem as *learning process*, where one's knowledge about an unknown parameter is influenced and improved by the subsequent observations x_1, x_2, etc.

The more measurement we add, the more the final posterior probability $P_n(\theta)$ will be insensitive to the choice of the prior probability $P_0(\theta) = \pi(\theta)$, because the θ range in which $L(x_1, \cdots, x_n|\theta)$ will be significantly different from zero will be smaller and smaller, hence the prior $\pi(\theta)$ can be approximated to a constant value within this small range, assuming we take a reasonably smooth function.

In this sense, a sufficiently large number of observation may remove, asymptotically, any arbitrariness in posterior Bayesian probability, assuming that the prior is a sufficiently smooth and regular function (e.g.: not in the extreme case mentioned in the Example 3.9).

3.4 Bayesian Inference

Chapter 5 will discuss the problem of performing estimates of unknown parameters using a frequentist approach. Given an observation $\vec{x}$, one can use an algorithm having good statistical properties, and in particular the maximum-likelihood estimator (see Sect. 5.5) is one of the most adopted one, to determine an estimate $\hat{\vec{\theta}}$ of the unknown parameter, or parameter set, $\vec{\theta}$ as the values that maximize the likelihood function. The limited knowledge about the unknown parameter turns into an error, or uncertainty, of the parameter estimate. Section 5.6, and more extensively Chap. 6, will discuss how to determine under the frequentist approach errors and confidence intervals corresponding to a parameter estimate $\hat{\vec{\theta}}$.

This section instead will briefly discuss how the estimate of unknown parameters can be addressed in a straightforward way using the determination of posterior Bayesian PDFs of the unknown parameters of interest $\vec{\theta}$ using Eq. (3.27). First, the *most likely* value of those parameters, $\hat{\vec{\theta}}$, can be determined from the posterior PDF:

$$P(\vec{\theta}|\vec{x}) = \frac{L(\vec{x};\vec{\theta})\pi(\vec{\theta})}{\int L(\vec{x};\vec{\theta}')\pi(\vec{\theta}')\,\mathrm{d}^h\theta'}\,. \tag{3.32}$$

Similarly the *average* value $\bar{\vec{\theta}}$ can also be determined from the same posterior PDF. In particular, if we assume a uniform prior distribution (this may not necessarily the most natural choice), the most likely value, in the Bayesian approach, coincides with the maximum-likelihood estimate, since the posterior PDF is equal, up to a normalization constant, to the product of the likelihood function times the prior PDF, which is assumed to be a constant:

$$P(\vec{\theta}|\vec{x})\Big|_{\pi(\vec{\theta})=\text{const.}} = \frac{L(\vec{x};\vec{\theta})}{\int L(\vec{x};\vec{\theta}')\,\mathrm{d}^h\theta'}\,. \tag{3.33}$$

This result of course does not necessarily hold in the case a non-uniform prior PDF.

If we are interested only in a subset of the parameters, let's say we have two subset of parameters, $\vec{\theta} = (\theta_1, \cdots, \theta_h)$ which are *parameters of interest* and the remaining parameters, $\vec{\nu} = (\nu_1, \cdots, \nu_l)$ which are needed to model our PDF, but should not appear among our final results, the posterior PDF can be written as:

$$P(\vec{\theta}, \vec{\nu}|\vec{x}) = \frac{L(\vec{x};\vec{\theta},\vec{\nu})\pi(\vec{\theta},\vec{\nu})}{\int L(\vec{x};\vec{\theta}',\vec{\nu}')\pi(\vec{\theta}',\vec{\nu}')\,\mathrm{d}^h\theta'\,\mathrm{d}^l\nu'}\,, \tag{3.34}$$

and the posterior PDF for the parameters $\vec{\theta}$ can be obtained as marginal PDF, integrating Eq. (3.34) over all the remaining parameters $\vec{\nu}$:

$$P(\vec{\theta}|\vec{x}) = \int P(\vec{\theta}, \vec{\nu}|\vec{x})\,\mathrm{d}^l\nu = \frac{\int L(\vec{x};\vec{\theta},\vec{\nu})\pi(\vec{\theta},\vec{\nu})\,\mathrm{d}^l\nu}{\int L(\vec{x};\vec{\theta}',\vec{\nu}')\pi(\vec{\theta}',\vec{\nu}')\,\mathrm{d}^h\theta'\mathrm{d}^l\nu'}\,. \tag{3.35}$$

Given a posterior PDF for an unknown parameter of interest θ, it is possible to find intervals $[\theta_{\text{lo}}, \theta_{\text{hi}}]$ corresponding to—say—a 68 % probability (i.e.: equal to a $\pm$ one σ interval for a normal distribution) such that the integral of the posterior PDF from θ_{lo} to θ_{hi} corresponds to 68 %. The choice of the interval for a fixed probability level, usually indicated as CL$= 1 - \alpha$, anyway has still some degree of arbitrariness, since one can choose different possible intervals all having the same probability level $1 - \alpha$ (i.e.: equal area under the PDF curve, in the given interval). Below some examples:

- a central interval $[\theta_{\text{lo}}, \theta_{\text{hi}}]$ such that the two complementary intervals $]-\infty, \theta_{\text{lo}}[$ and $]\theta_{\text{hi}}, +\infty[$ both correspond to probabilities of $\alpha/2$;
- a fully asymmetric interval $]-\infty, \theta_{\text{hi}}]$ with corresponding probability $1 - \alpha$;
- a fully asymmetric interval $]\theta_{\text{lo}}, +\infty]$ with corresponding probability $1 - \alpha$;
- the interval $[\theta_{\text{lo}}, \theta_{\text{hi}}]$ with the smallest length corresponding to the specified probability $1 - \alpha$;
- a symmetric interval around the value with maximum probability $\hat{\theta}$: $[\theta_{\text{lo}} = \hat{\theta} - \delta, \theta_{\text{hi}} = \hat{\theta} + \delta]$, corresponding to the specified probability $1 - \alpha$;
- ... etc.

Example 3.10 – Posterior for a Poisson Distribution

Let's consider a simple case of a Poisson distribution $P(n|s)$ of an integer variable with an expectation value s. Assume we observe the value n. Using Bayes posterior definition from Eq. (3.27), and assuming a prior PDF $\pi(s)$ for s, we have:

$$P(s|n) = \frac{\frac{s^n n^{-s}}{n!}\pi(s)}{\int_0^\infty \frac{s^n n^{-s}}{n!}\pi(s)\,\mathrm{d}s} \,. \tag{3.36}$$

If we take $\pi(s)$ to be a constant,[1] the normalization factor in the denominator becomes:

$$\frac{1}{n!}\int_0^\infty s^n e^s \,\mathrm{d}s = -\left.\frac{\Gamma(n+1,s)}{n!}\right|_0^\infty = 1 \,, \tag{3.37}$$

hence we have:

$$P(s|n) = \frac{s^n e^{-s}}{n!} \,, \tag{3.38}$$

which has the same expression of the original Poisson distribution, but this time it is interpreted as the posterior PDF of the unknown parameter s, given the observation n. Using Eq. (3.38), we can determine that the most *probable* value of s (according to the posterior in Eq. (3.38)) as:

$$\hat{s} = n \,. \tag{3.39}$$

We also have:

$$\langle s \rangle = n + 1 \,, \tag{3.40}$$

$$\mathrm{Var}[s] = n + 1 \,. \tag{3.41}$$

Note that the most probable value $\hat{s}$ is different from the average value $\langle s \rangle$. Those results depend of course on the choice of the prior $\pi(s)$. We took a uniform distribution for $\pi(s)$, but this is of course one of many possible choices.

[1] Considering the prior choice due to Jeffreys, discussed in Sect. 3.7, a constant prior could not be the most "natural" choice. Using Jeffreys' prior in the case of a Poisson distribution would give $\pi(s) \propto \sqrt{s}$.

3.5 Bayes Factors

As seen at the end the Example 3.8, there is a convenient way to compare the probability of two hypotheses using Bayes theorem which does not require the knowledge of all possible hypotheses that can be considered. In the Bayesian approach to probability, this can constitute an alternative to the hypothesis test (that will be introduced in Chap. 7) adopted under the frequentist approach.

Using the Bayes theorem, one can write the ratio of posterior probabilities evaluated under two hypotheses H_0 and H_1, given our observation $\vec{x}$, also called *Bayes factor* [4], as:

$$B_{1/0} = \frac{P(H_1|\vec{x})}{P(H_0|\vec{x})} = \frac{P(\vec{x}|H_1)}{P(\vec{x}|H_0)} \times \frac{\pi(H_1)}{\pi(H_0)}\,, \tag{3.42}$$

where $\pi(H_1)$ and $\pi(H_0)$ are the priors for the two hypotheses.

The Bayes factor can be used as alternative to significance level (see Sect. 8.2) adopted under the frequentist approach in order to determine the evidence of one hypothesis H_1 (e.g.: a signal due to a new particle is present in our data sample) against a null hypothesis H_0 (e.g.: no signal due to a new particle is present in our data sample). The proposed scale for Bayes factor to assess the evidence of H_1 against H_0, as reported in [4], are presented in Table 3.1.

Introducing the likelihood function, the Bayes factor can also be written as:

$$B_{1/0} = \frac{\int L(\vec{x}|H_1, \vec{\theta}_1)\pi(\vec{\theta}_1)\,\mathrm{d}\vec{\theta}_1}{\int L(\vec{x}|H_0, \vec{\theta}_0)\pi(\vec{\theta}_0)\,\mathrm{d}\vec{\theta}_0}\,, \tag{3.43}$$

where the parameters $\vec{\theta}_1$ and $\vec{\theta}_0$ required under the two hypotheses have been introduced. As usual in Bayesian applications, the computation of the Bayes factors requires integrations that in most of the realistic cases can be performed using computer algorithms.

Table 3.1 Assessing evidence with Bayes factors according to the scale proposed in [4]

$B_{1/0}$	Evidence against H_0
1–3	Not worth more than a bare mention
3–20	Positive
20–150	Strong
>150	Very strong

3.6 Arbitrariness of the Prior Choice

The main concern expressed by frequentist statisticians regarding the use of Bayesian probability is its intrinsic dependence on a prior probability that could be chosen by the observer in an arbitrary way. This arbitrariness makes Bayesian probability to some extent *subjective*, in the sense that it depends on one's choice of the prior probability which may change from subject to subject. We saw in the Example 3.9 that, in extreme cases, extreme choices of prior PDFs may lead to insensitiveness on the actual measurements.

It is also true, as remarked in Sect. 3.3.1, that, for reasonable choices of the prior PDF, adding more and more measurements increases one's knowledge about the unknown parameter θ, hence the posterior probability will be less and less sensitive to the choice of the prior probability. In most of those cases, where a large number of measurements is available, the application of Bayesian and frequentist calculations tend to give identical results.

But it is also true that many interesting statistical problems arise in the cases with small number of measurements, where the aim of the adopted statistical method is to extract the maximum possible information from the limited available sample, which is in general precious because it is the outcome of our complex and labour-intensive experiment. In those cases, applying Bayesian or frequentist methods usually leads to numerically different results, which should also be interpreted in very different ways. In those cases, using the Bayesian approach, the choice of the prior probabilities may play a crucial role and it may have relevant influence on the results.

One of the main difficulty arises when choosing a probability distribution to models one's complete ignorance about an unknown parameter θ. A frequently adopted prior distribution in physics that models one's complete lack of knowledge about a parameter θ is a uniform ("flat") PDF in the interval of validity of θ. But it is clear that if we change the parametrization from the original parameter θ to a function of θ (for instance in one dimension one may chose $\exp\theta$, $\log\theta$, $\sqrt{\theta}$ or $1/\theta$, etc.), the resulting transformed parameter will no longer have a uniform prior PDF. This is true, for instance, in the case of the measurement of a particle's lifetime τ. Should one chose a PDF uniform in τ, or in the particle's width, $\Gamma = 1/\tau$? There is no preferred choice provided by any first principle.

An approach to find a prior distribution that is *invariant* under reparametrization is to use the prior's choice proposed by Harold Jeffreys (see next Sect. 3.7). This approach leads, except in a few cases, to prior PDFs that are not uniform. For instance, for a Poissonian counting experiment, like the one considered in the Example 3.10, Jeffreys' prior is proportional to $1/\sqrt{s}$, not uniform as assumed to determine Eq. (3.38).

This *subjectiveness* in the choice of the prior PDF is intrinsic to the Bayesian approach and raises criticism by supporters of the frequentist approach, which object that results obtained under the Bayesian approach are to some extent *subjective*. Supporters of the Bayesian approach reply that Bayesian result are *intersubjective* [5], in the sense that common prior choices lead to common results. The debate is in some cases still open, and literature still contains opposite opinions about this issue.

3.7 Jeffreys' Prior

One possible approach to the choice of prior PDFs has been proposed by Harold Jeffreys [6]. He proposed to adopt a choice of prior PDF in a form which results invariant under parameter transformation. Jeffreys' choice is, up to a normalization factor, given by:

$$p(\vec{\theta}) \propto \sqrt{I(\vec{\theta})}\,, \tag{3.44}$$

where $I(\vec{\theta})$ is the determinant of the Fisher information matrix defined below:

$$I(\vec{\theta}) = \det\left[\left\langle \frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial\theta_i}\,\frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial\theta_j}\right\rangle\right]. \tag{3.45}$$

It is not difficult to demonstrate that Jeffreys' prior does not change when changing parametrization, i.e.: transforming $\vec{\theta} \to \vec{\theta}' = \vec{\theta}'(\vec{\theta})$.

Jeffreys' prior corresponding to the parameters of some of the most frequently used PDFs are given in Table 3.2. Note that only the case of the mean of a Gaussian gives a uniform Jeffreys' prior.

Table 3.2 Jeffreys' priors corresponding to the parameters of some of the most frequently used PDFs

PDF parameter	Jeffreys' prior
Poissonian mean μ	$p(\mu) \propto 1\sqrt{\mu}$
Poissonian signal mean μ with a background b	$p(\mu) \propto 1\sqrt{\mu + b}$
Gaussian mean μ	$p(\mu) \propto 1$
Gaussian standard deviation σ	$p(\sigma) \propto 1/\sigma$
Binomial success fraction ε	$p(\varepsilon) \propto 1/\sqrt{\varepsilon(1-\varepsilon)}$

3.8 Error Propagation with Bayesian Probability

In the case of Bayesian inference, error propagation proceeds naturally. The outcome of a Bayesian inference process is a posterior PDF for the unknown parameter(s). In order to obtain the PDF for a set of transformed parameters it is sufficient to apply the results we saw in Sect. 2.4 concerning how PDF transform under variables transformation. We remind here that, in the case of a two-variable transformation, $(x, y) \rightarrow (x', y') = (X'(x, y), Y'(x, y))$, a PDF $f(x, y)$ transforms according to:

$$f'(x', y') = \int \delta(x' - X'(x, y))\delta(y' - Y'(x, y))\, f(x, y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{3.46}$$

Given $f'(x', y')$, one can determine again, for the transformed parameters, the most likely values, the average, and probability intervals. The generalization to more variables is straightforward.

Something to be noted is that the most probable values $(\hat{x}, \hat{y})$, i.e: the values that maximize $f(x, y)$, do not necessarily map into values $(X(\hat{x}), Y(\hat{y}))$ that maximize $f'(x', y')$. This issue is also present in the frequentist approach when dealing with the propagation of asymmetric uncertainties, and will be discussed in Sect. 5.9. Section 5.8 will discuss how to propagate errors in the case of parameter transformation using a linear approximation, and under this simplified assumption, which is not always a sufficient approximation, one may assume that values that maximize f map into values that maximize f', i.e.: $(\hat{x}', \hat{y}') = (X(\hat{x}), Y(\hat{y}))$.

References

1. G. Cowan, *Statistical Data Analysis* (Clarendon Press, Oxford, 1998)
2. G. D'Agostini, Telling the truth with statistics. CERN Academic Training, 2005
3. W. Eadie, D. Drijard, F. James, M. Roos, B. Saudolet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam, 1971)
4. R. Kass, E. Raftery, Bayes factors. J. Am. Stat. Assoc. **90**, 773 (1995)
5. G. D'Agostini, *Bayesian Reasoning in Data Analysis: A Critical Introduction* (World Scientific, Hackensack, 2003)
6. H. Jeffreys, An invariant form for the prior probability in estimation problems. Proc. R. Soc. Lond. A Math. Phys. Sci. **186**, 453–461 (1946)

Chapter 4
Random Numbers and Monte Carlo Methods

4.1 Pseudorandom Numbers

Many computer application, ranging from simulations to video games and 3D-graphics application, take advantage of computer-generated numeric sequences having properties very similar to truly random variables. Sequences generated by computer algorithms through mathematical operations are not really random, having no intrinsic unpredictability, and are necessarily deterministic and reproducible. Indeed, it is often a good feature for many application the possibility to reproduce exactly the same sequence of computer-generated numbers given by a computer algorithm.

Good algorithms that generate "random" numbers, or better, *pseudorandom* numbers, given their reproducibility, must obey, in the limit of large numbers, to the desired statistical properties of real random variables.

Numerical methods involving the repeated use of computer-generated pseudo-random numbers are often referred to as *Monte Carlo* methods, from the name of the city hosting the famous casino, which exploits the properties of (truly) random numbers to generate profit.

4.2 Pseudorandom Generators Properties

Good (pseudo)random number generators must be able to generate sequences of numbers that are statistically independent on previous extractions, though unavoidably each number will be determined mathematically, through the generator's algorithm, from the previous numbers.

All numbers in the sequence should be independent and distributed according to the same PDF $P(x)$ (*independent and identically distributed* random variables, or i.i.d.). Those properties can be written as follows:

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_4

$$P(x_i) = P(x_j),\ \forall\, i, j\,, \tag{4.1}$$

$$P(x_n|x_{n-m}) = P(x_n),\ \forall\, n, m\,. \tag{4.2}$$

Example 4.11 – Transition From Regular to "Unpredictable" Sequences

There are several examples of mathematical algorithms that lead to sequences that are poorly predictable. One example of transition from a "regular" to a "chaotic" regime is given by the *logistic map* [1]. The sequence is defined, starting from an initial value x_0, as:

$$x_{n+1} = \lambda x_n(1 - x_n)\,. \tag{4.3}$$

Depending on the value of λ, the sequence may have very different possible behaviors. If the sequence converges to a single asymptotic value x for $n \to \infty$, we would have:

$$\lim_{n\to\infty} x_n = x\,, \tag{4.4}$$

where x must satisfy:

$$x = \lambda x(1 - x)\,. \tag{4.5}$$

Excluding the trivial solution $x = 0$, Eq. (4.5) leads to the solution $x = (1 - \lambda)/\lambda$. This solution is stable for values of λ smaller than 3. Above $\lambda = 3$, the sequence stably approaches a state where it oscillates between two values x_1 and x_2 that satisfy the following system of two equations:

$$x_1 = \lambda x_2(1 - x_2)\,, \tag{4.6}$$

$$x_2 = \lambda x_1(1 - x_1)\,. \tag{4.7}$$

For larger values, up to $1 + \sqrt{6}$, the sequences oscillates between four values, and further *bifurcations* occur for larger values of λ, until it achieves a very complex and poorly predictable behavior. For $\lambda = 4$ the sequence finally densely covers the interval $]0, 1[$. The PDF corresponding to the sequence with $\lambda = 4$ can be demonstrated to be a *beta distribution* with parameters $\alpha = \beta = 0.5$, where the beta distribution is defined as:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1 - x)^{\beta-1}}{\int_0^{\alpha-1}(1 - u)^{\beta-1}\,\mathrm{d}u}\,. \tag{4.8}$$

The behavior of the logistic map for different values of λ is shown in Fig. 4.1.

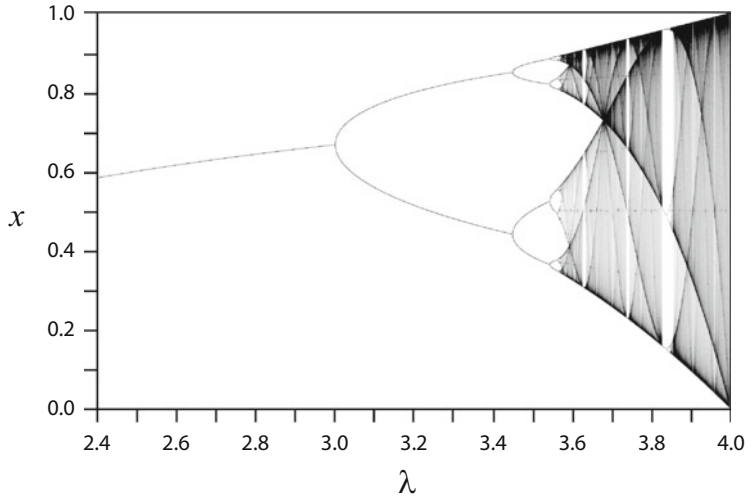

Fig. 4.1 Logistic map [2]

4.3 Uniform Random Number Generators

The most widely used computer-based pseudorandom number generators are conveniently written in such a way to produce sequences of uniformly distributed numbers ranging from zero to one.[1] Starting from uniform random number generators most of the other distribution of interest can be derived using specific algorithms, some of which are described in the following.

The *period* of a random sequence, i.e.: the number of extractions after which the sequence will repeat itself, should be as high as possible, and anyway higher than the number of random numbers that will be used in the specific application.

For instance, the function `lrand48` [3], which is a standard of C programming language, is defined according to the following algorithm:

$$x_{n+1} = (ax_n + c) \bmod m\,, \tag{4.9}$$

where the values of m, a and c are:

$$m = 2^{48} \tag{4.10}$$

$$a = 25214903917 = 5\text{DEECE66D}\,\text{hex} \tag{4.11}$$

$$c = 11 = \text{B}\,\text{hex} \tag{4.12}$$

[1] In realistic cases of finite numeric precision, one of the extreme values is excluded. It would have a corresponding zero probability, in the case of infinite precision, but it is not the case with finite machine precision.

The sequences obtained from Eq. (4.9) for given initial values x_0 are distributed uniformly between 0 and $2^{48} - 1$. This corresponds to sequences of random bits that can be used to return 32-bits integer numbers or can be mapped into sequences of floating-point numbers uniformly distributed in $[0, 1[$, as done by the C function `drand48`. The value x_0 is called *seed* of the random sequence. Choosing different initial seeds produces different sequences. In this way one can repeat a computer-simulated experiment using different random sequences each time changing the initial seed and obtaining different results that simulate the statistical fluctuation of the simulated experiment.

Similarly to `lrand48`, the `gsl_rng_rand` [4] generator of the BSD `rand` function uses the same algorithm but with $a = 41C64E6D$ hex, $c = 3039$ hex and $m = 2^{31}$. The period of `lrand48` is about 2^{48}, while `gsl_rng_rand` has a lower period of about 2^{31}.

A popular random generator that offers good statistical properties is due to Lüscher [5], implemented by James in the RANLUX generator [6], whose period is of the order of 10^{171}. RANLUX is now considered relatively slower than other algorithms, like the L'Ecuyer generator [7], which has a period of 2^{88}, or the Mersenne-Twistor generator [8] which has a period of $2^{19937} - 1$ and is relatively faster than L'Ecuyer's generator.

4.3.1 Remapping Uniform Random Numbers

Given a value x uniformly distributed in $[0, 1[$, it is often convenient to transform it into a variable x' uniformly distributed in another interval $[a, b[$ by performing the following transformation:

$$x' = a + x(b - a)\,. \tag{4.13}$$

With this transformation, $x = 0$ corresponds to $x' = a$ and $x = 1$ corresponds to $x' = b$.

4.4 Non Uniform Random Number Generators

Starting from a uniform pseudorandom number generator, it is possible to define algorithms that generate sequences that have non-uniform distributions.

4.4.1 Gaussian Generators Using the Central Limit Theorem

We already saw in Chap. 1 that, using the central limit theorem, the sum of N random variables, each having a finite variance, is distributed, in the limit $N \to \infty$, according to a Gaussian distribution. For a finite but sufficiently large value of N we can extract N values $x_1, \cdots, x_N$ from a uniform random generator and remap them from $[0, 1[$ to $[-\sqrt{3}, \sqrt{3}[$ using Eq. (4.13), so that the average and variance of each x_i are zero and one respectively. Then, we can compute:

$$x = \frac{x_1 + \cdots + x_N}{\sqrt{N}}\,, \tag{4.14}$$

which has again average equal to zero and variance equal to one. Figure 2.11 shows the distribution of x for N up to four, giving approximately a Gaussian distribution, that is necessarily truncated in the range $[-\sqrt{3N}, \sqrt{3N}]$ by construction. This is anyway not the most effective way to generate approximately Gaussian-distributed pseudorandom numbers. A better algorithm will be described at the end of this section.

4.4.2 Non-uniform Distribution From Inversion of the Cumulative Distribution

If we want to generate a pseudorandom variable x distributed according to a generic function $f(x)$, we can build its cumulative distribution (Eq. (2.13)):

$$F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'\,. \tag{4.15}$$

If we can invert the cumulative distribution $F(x)$, it is possible to demonstrate that, extracting a random number r uniformly distributed in $[0, 1[$, the transformed variable:

$$x = F^{-1}(r) \tag{4.16}$$

is distributed according to $f(x)$. In fact, if we write:

$$r = F(x)\,, \tag{4.17}$$

we have:

$$\mathrm{d}r = \frac{\mathrm{d}F}{\mathrm{d}x}\,\mathrm{d}x = f(x)\,\mathrm{d}x\,. \tag{4.18}$$

Introducing the differential probability $\mathrm{d}P$, we have:

$$\frac{\mathrm{d}P}{\mathrm{d}x} = f(x)\frac{\mathrm{d}P}{\mathrm{d}r}\,. \tag{4.19}$$

Since r is uniformly distributed, we have: $\mathrm{d}P/\mathrm{d}r = 1$, hence:

$$\frac{\mathrm{d}P}{\mathrm{d}x} = f(x)\,, \tag{4.20}$$

which means that x follows the desired PDF. This method only works conveniently if the cumulative $F(x)$ can be easily computed and inverted using either analytical or numerical methods. If not, usually this algorithm may be very slow.

Example 4.12 – Extraction of an Exponential Random Variable

We can use the inversion of the cumulative distribution in order to extract random numbers x distributed according to an exponential PDF:

$$f(x) = \lambda e^{-\lambda x}\,. \tag{4.21}$$

The cumulative PDF is:

$$F(x) = \int_0^x f(x')\,\mathrm{d}x' = 1 - e^{\lambda x}\,. \tag{4.22}$$

Inverting $F(x)$ leads to:

$$1 - e^{\lambda x} = r\,, \tag{4.23}$$

which translates into:

$$x = -\frac{1}{\lambda}\log(1-r)\,. \tag{4.24}$$

If the extraction of r happens in the interval $[0, 1[$, like `drand48`, $r = 1$ will never be extracted, hence the argument of the logarithm will never be equal to zero, ensuring the numerical validity of Eq. (4.24).

Example 4.13 – Extraction of a Uniform Point on a Sphere

Assume we want to generate two variables, θ and ϕ, distributed in such a way that the direction described by them in polar coordinates points to a point uniformly distributed on a sphere. In this case the two-dimensional differential probability must be uniform per unit of solid angle Ω:

$$\frac{\mathrm{d}P}{\mathrm{d}\Omega} = \frac{\mathrm{d}P}{\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi} = k\,, \tag{4.25}$$

where k is a normalization constant such that the PDF integrates to unity over the entire solid angle, 4π. From Eq. (4.25), the combined two-dimensional PDF can be factorized into the product of two PDFs, as functions of θ and ϕ (i.e.: θ and ϕ are independent):

$$\frac{\mathrm{d}P}{\mathrm{d}\theta\,\mathrm{d}\phi} = f(\theta)g(\phi) = k\sin\theta\,, \tag{4.26}$$

where:

$$f(\theta) = \frac{\mathrm{d}P}{\mathrm{d}\theta} = c_1\,, \tag{4.27}$$

$$g(\phi) = \frac{\mathrm{d}P}{\mathrm{d}\phi} = c_2\sin\theta\,. \tag{4.28}$$

The constants c_1 and c_2 ensure the normalization of $f(\theta)$ and $g(\phi)$. θ can be extracted inverting the cumulative distribution of $f(\theta)$ (Eq. (4.16)), and since $g(\phi)$ is uniformly distributed, ϕ can be extracted just remapping the interval $[0, 1[$ into $[0, 2\pi[$ (Eq. (4.13)), leading to:

$$\theta = \arccos(1 - 2r_1) \in]0, \pi]\,, \tag{4.29}$$

$$\phi = 2\pi r_2 \in [0, 2\pi[\,, \tag{4.30}$$

where r_1 and r_2 are extracted as uniformly distributed in $[0, 1[$.

4.4.3 Gaussian Numbers Generation

In order to generate Gaussian random numbers, the inversion of the cumulative distribution would lead to the inversion of an error function, which is not easy to perform numerically. One possibility is to extract, according to the algorithm described in the following, pairs of random numbers Gaussian distributed in two dimensions and use them as independent Gaussian numbers. This is easier to

perform by moving from cartesian coordinates (x, y) to polar coordinates (r, ϕ). In particular, the radial Gaussian cumulative PDF was already introduced in Eq. (2.100), and leads, in the simplest case of a normal Gaussian (with average in $(0, 0)$ and standard deviations in both coordinates equal to one), to:

$$P_{2D}(\rho) = \int_0^\rho e^{-\frac{r^2}{2}} r \, \mathrm{d}r = 1 - e^{-\frac{\rho^2}{2}} . \tag{4.31}$$

This leads to the Box Muller transformation [9] from two variables r_1 and r_2 uniformly distributed in $[0, 1[$ into two variables z_1 and z_2 distributed according to a normal Gaussian:

$$r = \sqrt{-2 \log r_1} \tag{4.32}$$

$$\phi = 2\pi r_2 \tag{4.33}$$

$$z_1 = r \cos \phi \tag{4.34}$$

$$z_2 = r \sin \phi \tag{4.35}$$

A normal Gaussian random number z can be easily transformed into a Gaussian random number x with a given average μ and standard deviation σ using the following transformation:

$$x = \mu + \sigma z . \tag{4.36}$$

More efficient generators for Gaussian random numbers exist. For instance [10] describes the so-called Ziggurat algorithm.

4.5 Monte Carlo Sampling

There are many cases in which the cumulative distribution of a PDF is not easily computable and invertible. In those case, other methods allow to generate random numbers according to a given PDF.

4.5.1 Hit-or-Miss Monte Carlo

A rather general-purpose and simple to implement method to generate random numbers according to a given PDF is the *hit-or-miss Monte Carlo*. It assumes we have a PDF defined in an interval $x \in [a, b[$, that is known as a function $f(x)$, but not necessarily normalized (i.e.: $\int_a^b f(x) \, \mathrm{d}x$ may be different from one). We also assume

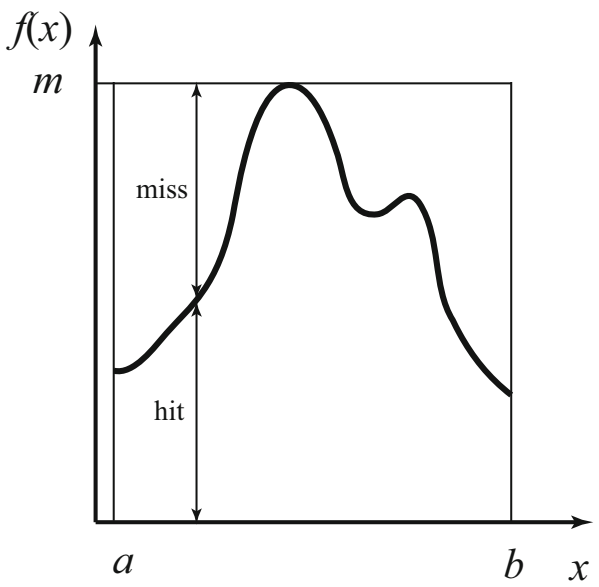

Fig. 4.2 Sketch of hit-or-miss Monte Carlo method

that we know the maximum value m of f, or at least a value m that is greater or equal to the maximum of f. The situation is sketched in Fig. 4.2.

The method consist of first extracting a uniform random number x in the interval $[a, b[$, and then computing $f = f(x)$. Then, a random number r is extracted uniformly in $[0, m[$. If $r > f$ ("miss") we repeat the extraction of x, until $r < f$ ("hit"). In this case, we accept x as the desired extracted value. In this way, the probability distribution of the accepted x is our initial (normalized) PDF by construction. A possible inconvenient of this method is that it rejects a fraction of extractions equal to the ratio of area under the curve $f(x)$, and the area of the rectangle that contains f. The method has an *efficiency* (i.e.: the fraction of accepted values of x) equal to:

$$\varepsilon = \frac{\int_a^b f(x)\,\mathrm{d}x}{(b-a)\times m}\,, \tag{4.37}$$

which may lead to a suboptimal use of the computing power, in particular if the shape of $f(x)$ is very peaked.

Hit-or-miss Monte Carlo can also be applied to multi-dimensional cases. In those cases, one first extracts a multi-dimensional point $\vec{x} = (x_1, \cdots, x_n)$, then accepts or rejects $\vec{x}$ according to a random extraction $r \in [0, m[$, compared with $f(x_1, \cdots, x_n)$.

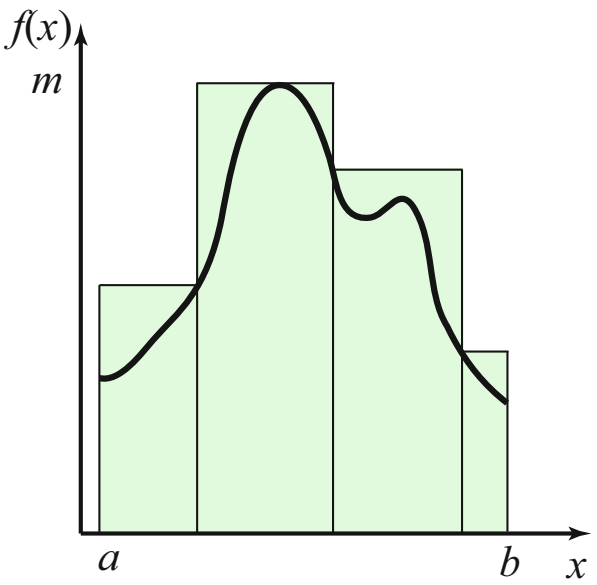

Fig. 4.3 Variation of hit-or-miss Monte Carlo using the importance sampling

4.5.2 Importance Sampling

If the function f is very peaked, the efficiency ε of this method (Eq. (4.37)) may be very low. In those cases the algorithm may be adapted to become more efficient by identifying in a preliminary stage a partition of the interval $[a, b[$ such that in each subinterval the function f has a smaller variation than in the overall range. In each subinterval the maximum of f is estimated. This situation is sketched in Fig. 4.3. First of all, a subinterval of the partition is randomly chosen with a probability proportional to its area. Then, the hit-or-miss approach is followed in the corresponding rectangle. This approach is often referred to as *importance sampling*.

A possible variation of this method is to use, instead of the aforementioned partition, an "envelope" for the function f, i.e.: a function $g(x)$ that is always greater than $f(x)$: $g(x) \geq f(x),\ \forall x \in [a, b[$, and for which a convenient method to extract x according to the normalized distribution $g(x)$ is known.

It's evident that the efficiency of the importance sampling may be significantly higher than the "plain" hit-or-miss Monte Carlo if the partition and the corresponding maxima in each subinterval are properly chosen.

Example 4.14 – Combined Application of Different Monte Carlo Techniques

We want to find an algorithm to generate a random variable x distributed according to a PDF:

$$f(x) = Ce^{-\lambda x}\cos^2 kx\,, \tag{4.38}$$

where C is a normalization constant, and λ and k are two known parameters. $f(x)$ describes an oscillating term ($\cos^2 kx$) dumped by an exponential term ($e^{-\lambda x}$).

We can use as "envelope" the function $Ce^{-\lambda x}$, so we first generate a random number x according to this exponential distribution. Then we apply an hit-or-miss technique, and we accept or reject x according to a probability proportional to $\cos^2 kx$. The probability distribution, given the two independent processes, is the product of the exponential envelope times the cosine-squared term. In summary, the algorithm may proceed as follows:

1. generate r uniformly in $[0, 1[$
2. compute $x = -\frac{1}{\lambda}\log(1-r)$.
3. generate s uniformly in $[0, 1[$
4. if $s > \cos^2 kx$ repeat the extraction at the point 2
5. else return x

4.6 Numerical Integration with Monte Carlo Methods

Monte Carlo methods are often used as numerical methods to compute integrals with the computer. If we use the *hit-or-miss* method described in Sect. 4.5.1, for instance, we can estimate the integral $\int_a^b f(x)\,\mathrm{d}x$ from the fraction of the accepted hits n over the total number of extractions N:

$$I = \int_a^b f(x)\,\mathrm{d}x \simeq \hat{I} = (b-a) \times \frac{n}{N}\,. \tag{4.39}$$

With this approach, n follows a binomial distribution. If we can use the approximate expression in Eq. (1.40) (see Sect. 1.11.1, n should not be too close to either 0 or N), we can estimate the error on $\hat{I}$ as:

$$\sigma_{\hat{I}} = (b-a)\sqrt{\frac{I(1-I)}{N}}\,. \tag{4.40}$$

Equation 4.40 shows that the error on $\hat{I}$ decreases as $\sqrt{N}$. We can apply the same hit-or-miss method to a multi-dimensional integration problem. In that case, we

would have the same expression for the error, which decreases as $\sqrt{N}$, regardless of the number of dimensions d of the problem. Other numerical methods, not based on random number extractions, may suffer from severe computing time penalties as the number of dimensions d increases, and this makes Monte Carlo methods advantageous in cases of high number of dimensions, compared to other non-random-based techniques.

In the case of hit-or-miss Monte Carlo, anyway, numerical problems may arise in the algorithm to find the maximum value of the input function in the multi-dimensional range of interest. Also, partitioning the multi-dimensional integration range in an optimal way, in case the of importance sampling, can be a non-trivial problem.

References

1. R. May, Simple mathematical models with very complicated dynamics. Nature **621**, 459 (1976)
2. Logistic map. public domain image, 2011. https://commons.wikimedia.org/wiki/File:LogisticMap_BifurcationDiagram.png
3. T.O. Group, The single UNIX® specification (1997). http://www.unix.org. Version 2
4. T.G. project, GNU operating system—GSL—GNU scientific library (1996–2011), http://www.gnu.org/software/gsl/
5. M. Lüscher, A portable high-quality random number generator for lattice field theory simulations. Comput. Phys. Commun. **79**, 100–110 (1994)
6. F. James, Ranlux: A fortran implementation of the high-quality pseudorandom number generator of Lüscher. Comput. Phys. Commun. **79**, 111–114 (1994)
7. P. L'Ecuyer, Maximally equidistributed combined Tausworthe generators. Math. Comput. **65**, 203–213 (1996)
8. M. Matsumoto, T. Nishimura, Mersenne twistor: a 623-dimensionally equidistributed unifor pseudorandom number generator. ACM Trans. Model. Comput. Simul. **8**, 3–30 (1998)
9. G.E.P. Box, M. Muller, A note on the generation of random normal deviates. Ann. Math. Stat. **29**, 610–611 (1958)
10. G. Marsglia, W. Tsang, The Ziggurat method for generating random variables. J. Stat. Softw. **5**, 8 (2000)

Chapter 5
Parameter Estimate

This chapter describes how to determine *unknown parameters* of some probability distribution by sampling the values of random variables that obey such distribution. In physics this procedure is applied when measuring *parameters of interest* by running repeated experiments.

A typical case is running a particle collider and recording multiple collision events for further analysis. Theory provides a probability model that predicts the distribution of the observable quantities our detector is sensitive to. Some parameters of the theory are unknown, and the measurement of those parameters of interest is the goal of our experiment.

The outcome of an experiment is related to a probability distribution that results from the combination of the theoretical model and the effect of the experimental detector response, since usually the observed distributions of the measured quantities is affected by detector's finite resolution effects, miscalibrations, presence of background, etc. The detector response can be described by a probability model that depends on unknown parameters. Those additional unknown parameters, referred to as *nuisance parameters*, arise in such a way in the problem and appear together with the unknown parameters from theory. Examples of nuisance parameters are the detector resolution, calibration constants, amount of background, etc. Those parameters can be, in many cases, determined from experimental data samples. In some cases, dedicated data samples may be needed (e.g.: data from test beams in order to determine the calibration constants of a detector, cosmic-ray runs to determine alignment constants, etc.), or dedicated simulation programs.

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_5

5.1 Measurements and Their Uncertainties

We can *determine*, or *estimate* the value of the unknown parameters (either parameters of interest and nuisance parameters) using the data collected by our experiment, which lead to an *approximate* knowledge of those parameters within some *uncertainties*.

The procedure to determine parameters from a data sample is also called *best fit*, because we determine the parameters that *best fit* the theoretical model on the experimental data sample by finding an optimal set of parameter values (see Sect. 5.5).

The data provided by an experiment consist of measured values of the observable quantities that constitute our *data sample* which is a sampling of the PDF determined by the combination of the theoretical and the instrumental effects.

From the observed data sample, the value of one or more parameters θ is inferred, up to some level of uncertainty, and, as result of the measurement of a parameter θ, one quotes a central value and uncertainty, denoted usually as:

$$\theta = \hat{\theta} \pm \delta\theta \, .$$

Above, $\delta\theta$ represents the *error* or *uncertainty* associated to the measurement $\hat{\theta}$. The interval $[\hat{\theta} - \delta\theta, \hat{\theta} + \delta\theta]$ is referred to as *confidence interval* with a probability value is associated to it. Usually 68 % is implicitly assumed, which corresponds to an interval of $\pm 1\sigma$ for a normal distribution. In some cases, asymmetric positive and negative uncertainties are taken, so we quote:

$$\theta = \hat{\theta}^{+\delta\theta_+}_{-\delta\theta_-} \, ,$$

corresponding to the confidence interval $[\hat{\theta} - \delta\theta_+, \hat{\theta} + \delta\theta_-]$.

The meaning of the error $\delta\theta$ (or the errors δ_- and δ_+) and of the confidence interval may be different depending on the statistical approach taken.

As introduced in Sect. 1.1, two main complementary statistical approaches exist in literature. Those correspond to two different interpretation of the confidence interval, or the central value and its uncertainty:

- **Frequentist approach**: for a large fraction, usually by convention 68 % ("one σ", or alternatively 90 or 95 %), of repeated experiments, in the limit of infinitely large number of repetitions of the experiment, the unknown true value of θ is contained in the quoted confidence interval $[\hat{\theta} - \delta\theta, \hat{\theta} + \delta\theta]$, where $\hat{\theta}$ and $\delta\theta$ may vary from one experiment to another, being the result of the measured data sample in each experiment. This property of the estimated interval is referred to as statistical *coverage*. Interval estimates that have a larger or smaller (frequentist) probability to contain the true value, in this sense, are said to *overcover* or *undercover*, respectively.

- **Bayesian approach**: one's *degree of belief* that the unknown parameter is contained in the quoted interval $[\hat{\theta} - \delta\theta, \hat{\theta} + \delta\theta]$ can be quantified with a 68 % (or 90 or 95 %) probability. The mathematical quantitative definition and the procedures to quantify the degree of belief under the Bayesian approach has been discussed in Chap. 3.

Within both the frequentist and the Bayesian approaches there are still some degrees of arbitrariness in the definition of confidence intervals, as already seen in Sect. 3.4 (central interval, extreme intervals, smallest-length interval, etc.).

The process of determining the estimated values ($\hat{\theta}$) and the corresponding uncertainty ($\delta\theta$) of some unknown parameters from experimental data that correspond to the sampling of a probability distribution is also called *inference*. The presence of a finite uncertainty reflects the statistical fluctuations of the data sample due to the intrinsic (theoretical) and/or experimental (due to the effect of the detection instrument) randomness of our observable (see the diagram in Fig. 5.1). The smallest the amount of fluctuation of data (i.e.: a distribution that concentrates in a small interval a large probability), the smallest the uncertainty in the determination of the unknown parameters. Ideally, if the data sample would exhibit no fluctuation and if our detector would have a perfect resolution, a perfect knowledge of the unknown parameters would be possible. This case is never present in real experiments, and every real-life measurement is affected by some level of uncertainty.

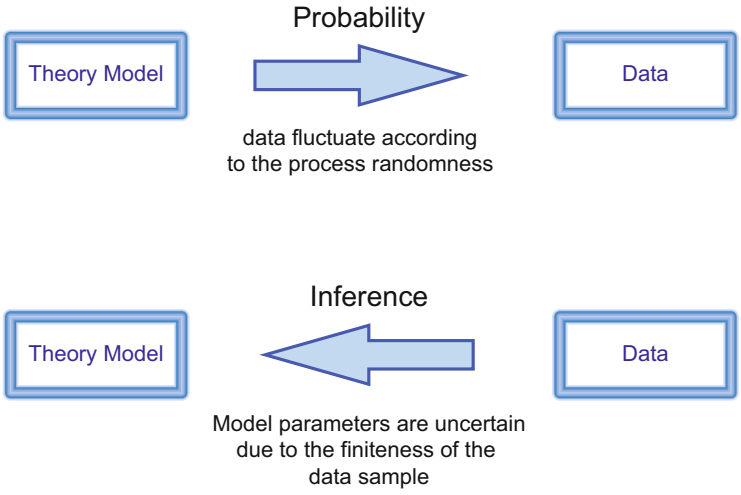

Fig. 5.1 Relation between probability and inference

5.2 Nuisance Parameters and Systematic Uncertainties

In order to define the PDF describing a data model, in many cases it is necessary to introduce parameters that are not of direct interest to our problem. For instance, when determining ("fitting") the yield of a signal peak, it is sometimes needed to determine from data other parameters, like the experimental resolution that gives the peak width, detector efficiencies that are needed to determine the signal production yield from the measured signal yield, parameters to define the shapes and amounts of possible backgrounds, and so on. Those parameters are often referred to as *nuisance* parameters. In some cases, nuisance parameters can't be determined from the same data sample used to measure the parameters of interest and their estimate should be taken form other measurements. The uncertainty on their determination, external to the considered fit problem, will reflect into uncertainties on the estimate of parameters of interest, as we will see in the following (see Sect. 8.11).

Uncertainties due to the propagation of imperfect knowledge of nuisance parameters that can't be constrained from the same data sample used for the main fit of the parameters of interest gives raise to *systematic uncertainties*, while uncertainties purely related to the fit are referred to as *statistical uncertainties*.

5.3 Estimators

Determining the *estimate* of unknown parameters requires the definition of a mathematical procedures to determine the central values as a function of the observed data sample. In general, such a function of the data sample that returns the estimated value of a parameter is called *estimator*. Estimators can be defined in practice by more or less complex mathematical formulae or numerical algorithms. We are interested in estimators that have "good" statistical properties, and such properties that characterize good estimators will be discussed in Sect. 5.4.

Example 5.15 – A Very Simple Estimator in a Gaussian Case

As first and extremely simplified example, let's assume a Gaussian distribution whose standard deviation σ is known (e.g.: the resolution of our apparatus), and whose average μ is the unknown parameter of interest. Let's assume our data sample consists of a single measurement x which is distributed according to the Gaussian distribution under consideration. We can define as estimator of μ the function that returns the single value x that has been extracted, i.e.: we consider as estimate of μ the value $\hat{\mu} = x$.

If we repeat many times (ideally: an infinite number of times) the experiment, we will have different values of $\hat{\mu} = x$, distributed according to the original Gaussian.

In 68.27 % of the experiments, in the limit of an infinite number of experiments, the fixed and unknown true value μ lies in the *confidence interval* $[\hat{\mu} - \sigma, \hat{\mu} + \sigma]$, i.e.: $\mu - \sigma < \hat{\mu} < \mu + \sigma$ or $\mu \in [\hat{\mu} - \sigma, \hat{\mu} + \sigma]$, and in the remaining 31.73 % of the cases μ will lie outside the same interval. This expresses the *coverage* of the interval $[\hat{\mu} - \sigma, \hat{\mu} + \sigma]$ at the 68.27 % *confidence level*.

We can quote our estimate $\mu = \hat{\mu} \pm \sigma$, in this sense. $\pm\sigma$ is the *error* or *uncertainty* we assign to the measurement $\hat{\mu}$, with the frequentist meaning defined above.

In realistic cases we have a data sample that is more rich than a single measurement, and PDF models more complex than a Gaussian, which we considered in the Example 5.15. The definition of an estimator may require in general complex mathematics or in many cases computer algorithms.

5.4 Properties of Estimators

Different estimators may have different statistical properties that makes one or another estimator more suitable for a specific problem. In the following we will present some of the main properties of estimators. Section 5.5 will introduced the maximum-likelihood estimators which have good performances in terms of most of the properties indicators described in the following.

5.4.1 Consistency

An estimator is said to be *consistent* if it converges, in probability, to the true unknown parameter value as the number of measurements n that tends to infinity, i.e., if:

$$\forall \varepsilon \lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1 \,. \tag{5.1}$$

5.4.2 Bias

The bias of an estimator is the average deviation of the estimate from its true value:

$$b(\theta) = \left\langle \hat{\theta} - \theta \right\rangle = \left\langle \hat{\theta} \right\rangle - \theta \,. \tag{5.2}$$

The average is intended as the expected value over an infinitely large number of repeated experiments.

5.4.3 Minimum Variance Bound and Efficiency

The variance of any consistent estimator is subject to a lower bound due to Cramér [1] and Rao [2] which is given by:

$$V[\hat{\theta}] \geq V_{\mathrm{CR}}(\hat{\theta}) = \frac{\left(1 + \frac{\partial b(\hat{\theta})}{\partial \theta}\right)^2}{\left\langle \left(\frac{\partial \ln L(x_1, \cdots, x_n; \theta)}{\partial \theta} \right)^2 \right\rangle} \,. \tag{5.3}$$

The denominator of Eq. (5.3) is called *Fisher information* in statistical literature, and was already introduced in Sect. 3.7 when defining Jeffreys' priors in the Bayesian approach.

The ratio of the Cramér–Rao bound to the estimator's variance is called estimator's *efficiency*:

$$\varepsilon(\hat{\theta}) = \frac{V_{\mathrm{CR}}(\hat{\theta})}{V[\hat{\theta}]} \,. \tag{5.4}$$

Any consistent estimator $\hat{\theta}$ has efficiency $\varepsilon(\hat{\theta})$ lower or equal to one.

Example 5.16 – Estimators with Variance Below the Cramér–Rao Bound are Not Consistent

It is possible to find estimators that have variance lower than the Cramér–Rao bound, but this implies that they are not consistent. Examples can be found by dropping the hypothesis that the estimator is consistent.

For instance, an estimator of an unknown parameter that gives a constant value (say zero) as estimate of the parameter, regardless of the data sample values, has zero variance, but is of course not consistent.

An estimator of this kind is clearly not very useful in practice.

5.4.4 *Robust Estimators*

If a sample's distribution has (hopefully slight) deviations from the hypothetical theoretical PDF model, some properties of estimators that assume the theoretical PDF model correctly describes the data may not hold necessarily. The entries in our data sample that exhibit large deviations from the theoretical PDF are called *outliers*. An important property of an estimator, in presence of outliers, is to have a limited sensitivity to the presence of such data that deviate from the theoretical model. This property, that can be better quantified, is in general defined as *robustness*.

An example of robust estimator of the average value of a sample $x_1, \cdots, x_n$ is the *median* $\bar{x}$, defined in the following way:

$$\bar{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd}, \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{if } n \text{ is even}. \end{cases} \tag{5.5}$$

Clearly, the presence of outliers at the left or right tails of the distribution will not change significantly the value of the median, if it is dominated by measurements in the "core" of the distribution, while the usual average (Eq. (1.13)) could be shifted from the true value more and more as much as the outlier distribution is broader than the "core" part of the distribution. *Trimmed averages*, i.e.: averages computed by removing from the sample a fraction of the rightmost and leftmost data in the sample is also less sensitive to the presence of outliers.

It is convenient to define the *breakdown point* as the maximum fraction of incorrect measurements (i.e. outliers) above which the estimate may grow arbitrarily large in absolute value. In particular, trimmed averages that remove a fraction f of the events can be demonstrated to have a breakdown point f, while the median has a breakdown point of 0.5.

A more detailed discussion of robust estimator is beyond the purpose of this text. Eadie et al. [3] contains a more extensive discussion of robust estimators.

5.5 Maximum-Likelihood Method

The most frequently used estimate method is based on the construction of the combined probability distribution of all measurements in our data sample, called *likelihood function*, which was already introduced in Sect. 3.3. The central values of the parameters we want to estimate are obtained by finding the parameter set that correspond to the maximum value of the likelihood function. This approach gives the name of *maximum-likelihood method* to this technique.

Maximum-likelihood fits are very frequently used because of very good statistical properties according to the indicators discussed in Sect. 5.4.

The simple example of estimator discussed in the Example 5.15, assuming the simple case of a Gaussian PDF with unknown average μ and known standard deviation σ, where the sample consisted of a a single measurement x obeying this given Gaussian PDF, is a very simple example of the application of the maximum-likelihood method.

5.5.1 Likelihood Function

We define the *likelihood function* as the PDF that characterizes our set of experimental observables, evaluated at the values of those observables that corresponds to our data sample, for given values of the unknown parameters. If we measure the values $x_1, \cdots x_n$ of n random variables and our PDF model depends on m unknown parameters $\theta_1, \cdots, \theta_m$, we define the likelihood function L as:

$$L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m) = f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)\,, \tag{5.6}$$

where f is the (joint) PDF of the random variables $x_1, \cdots x_n$. As already anticipated in Sect. 3.3, often, the notation: $L(x_1, \cdots, x_n|\theta_1, \cdots, \theta_m)$ is also used, resembling the notation adopted for conditional probability (see Sect. 1.6).

We can define the *maximum-likelihood estimator* of the unknown parameters $\theta_1, \cdots, \theta_m$ the function that returns the values of the parameters for which the likelihood function, evaluated on the measured sample, is maximum. We implicitly assume that there is a unique maximum value, otherwise the determination of the maximum-likelihood estimate is not unique.

If we have N repeated measurements each consisting of the n values of the random variables $(x_1, \cdots, x_n)$, we can consider the probability density corresponding to the total sample $\mathbf{x} = \{(x_1^1, \cdots, x_n^1), \cdots, (x_1^N, \cdots, x_n^N)\}$. If we assume that the events are independent of each other,[1] the likelihood function of the sample consisting of the N events can be written as the product of the PDFs corresponding to the measurement of each single event, i.e.:

$$\boxed{L(\mathbf{x}; \vec{\theta}) = \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)\,.} \tag{5.7}$$

[1] In this case, we will often use the term *event* with the meaning frequently used in physics, more than in statistics, i.e.: an event is the collection of measurements of the observable quantities $x_1, \cdots, x_n$ which is repeated in one or more experiments. Measurements performed in different events are assumed to be uncorrelated. In this sense, each sequence of event variables $x_i^1, \cdots, x_i^N$, for all $i = 1, \cdots, n$, can be considered a sampling of *independent and identically distributed* random variables, or i.i.d., see Sect. 4.2.

In order to more easily manage the maximization of the likelihood, often the logarithm of the likelihood function is computed, so that when the product of many terms appears in the likelihood definition, this is transformed into the sum of the logarithms of such terms. The logarithm of the likelihood function becomes:

$$-\ln L(\mathbf{x};\vec{\theta}) = -\sum_{i=1}^{N} \ln f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)\,. \tag{5.8}$$

5.5.1.1 Numerical Implementations: MINUIT

The maximization of the likelihood function L, or the equivalent—but often more convenient—minimization of $-\ln L$, can be performed analytically only in the simplest cases. Most of realistic cases require numerical methods implemented as computer algorithms. The computer algorithm MINUIT [4] is one of the most widely used minimization software tool in the field of high-energy and nuclear physics since the years 1970s, and has been reimplemented from the original fortran version in C++, as available in the ROOT software toolkit [5].

5.5.2 Extended Likelihood Function

If the number of events N is also a random variable that obeys the distribution $p(N;\theta_1,\cdots,\theta_m)$, which may also depend on the m unknown parameters, we can define the *extended likelihood function* as:

$$L = p(N;\theta_1,\cdots,\theta_m)\prod_{i=1}^{N} f(x_1^i,\cdots,x_n^i;\theta_1,\cdots,\theta_m)\,. \tag{5.9}$$

In most of the cases of interest in physics, $p(N;\theta_1,\cdots,\theta_m)$ is a Poisson distribution whose average μ may depend on the m unknown parameters. In this case, we can write:

$$L = \frac{e^{-\mu(\theta_1,\cdots,\theta_m)}\mu(\theta_1,\cdots,\theta_m)^N}{N!}\prod_{i=1}^{N} f(x_1^i,\cdots,x_n^i;\theta_1,\cdots,\theta_m)\,. \tag{5.10}$$

In the case for instance where the PDF is a linear combination of two PDFs, one for "signal" (P_s), one for "background" (P_b), whose yields, s and b are unknown, we can write the extended likelihood function as:

$$L(x_i;s,b,\theta) = \frac{(s+b)^n e^{-(s+b)}}{n!}\prod_{i=1}^{N}(f_s P_s(x_i;\theta) + f_b P_b(x_i;\theta))\,, \tag{5.11}$$

where the fractions of signal and background f_s and f_b are:

$$f_s = \frac{s}{s+b}\,, \tag{5.12}$$

$$f_b = \frac{b}{s+b}\,. \tag{5.13}$$

Replacing Eqs. (5.12) and (5.13) in (5.11), we have:

$$L(x_i; s, b, \theta) = \frac{(s+b)^n e^{-(s+b)}}{n!} \prod_{i=1}^{N} \frac{(sP_s(x_i;\theta) + bP_b(x_i;\theta))}{s+b} \tag{5.14}$$

$$= \frac{e^{-(s+b)}}{n!} \prod_{i=1}^{N} (sP_s(x_i;\theta) + bP_b(x_i;\theta))\,. \tag{5.15}$$

Writing the logarithm of the likelihood function we have the more convenient expression:

$$-\ln L(x_i; s, b, \theta) = s + b + \sum_{i=1}^{n} \ln(sP_s(x_i;\theta) + bP_b(x_i;\theta)) - \ln n!\,. \tag{5.16}$$

The last term, $-\ln n!$, is constant with respect to the unknown parameters and can be omitted when minimizing the function $-\ln L(x_i; s, b, \theta)$.

An example of extended maximum-likelihood fit of a two-component PDF like the one illustrated above is shown in Fig. 5.2. The points with the error bars represent

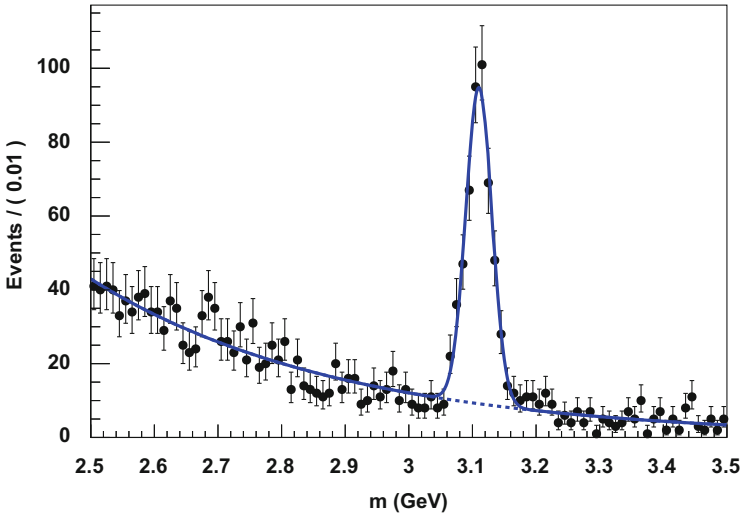

Fig. 5.2 Example of unbinned extended maximum-likelihood fit of a simulated dataset. The fit curve is superimposed on the data points as *solid blue line*

the data sample, shown as binned histogram for convenience, which is randomly extracted according to the assumed PDF. The unknown parameters present in are fit according to the likelihood function in Eq. (5.16) with a PDF modeled as sum of a Gaussian component ("signal") and an exponential component ("background"). The mean and standard deviation of the Gaussian component and the decay parameter of the exponential component are fit simultaneously with the number of signal and background events, s and b.

5.5.3 *Gaussian Likelihood Functions*

Assuming we have n measurements of a variable x whose PDF is a Gaussian with average μ and standard deviation σ, we have:

$$-2\ln L(x_i;\mu,\sigma) = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} + n(\ln 2\pi + 2\ln\sigma)\,. \tag{5.17}$$

The minimization of $-2\ln L(x_i;\mu,\sigma)$ can be performed analytically by finding the zeros of the first derivative with respect to μ and σ^2, and the following maximum-likelihood estimates for μ and σ^2 can be obtained:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \tag{5.18}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\,. \tag{5.19}$$

The maximum-likelihood estimate $\hat{\sigma}^2$ is affected by *bias* (Sect. 5.4.2), in the sense that its average value deviates from the true σ^2. The bias, anyway, decreases as $n \to \infty$. The way to correct the bias of Eq. (5.19) is discussed in the Example 5.18.

5.6 Errors with the Maximum-Likelihood Method

Once we have determined the estimate $\hat{\theta}$ of our parameter of interest, using the maximum-likelihood method, we have to determine a *confidence interval* that corresponds to a *coverage* of 68.27 %, or "1σ", in most of the cases.

In cases of non-Gaussian PDFs, in the presence of a large number of measurements that we combined into a single estimate, we can still use a Gaussian approximation. If we are far from that ideal limit, the result obtained under the Gaussian assumption may be only an approximation and deviation from the exact coverage may occur. If we use as PDF model an n-dimensional Gaussian

(Eq. (2.102)), it is easy to demonstrate that we can obtain the n-dimensional PDF covariance matrix as the inverse of the matrix of the second-order partial derivatives of the negative logarithm of the likelihood function, i.e.[2]:

$$C_{ij}^{-1} = -\frac{\partial \ln L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)}{\partial \theta_i \, \partial \theta_j} \,. \tag{5.20}$$

This covariance matrix gives the n-dimensional confidence contour having the correct coverage if the PDF model is exactly Gaussian.

Let's consider the Gaussian likelihood case of Eq. (5.17), from Sect. 5.5.3. In that case, if we compute the derivative of $-\ln L$ with respect to the average parameter μ, we obtain:

$$\frac{1}{\sigma_\mu^2} = \frac{\partial^2(-\ln L)}{\partial \mu^2} = \frac{n}{\sigma^2} \,, \tag{5.21}$$

which leads to the following error on the estimated average:

$$\sigma_\mu = \frac{\sigma}{\sqrt{n}} \,. \tag{5.22}$$

This expression coincides with the evaluation of the standard deviation of the average from Eq. (5.18) using the general formulae from Eqs. (1.14) and (1.15).

Another method frequently used to determine the uncertainty on maximum-likelihood estimates is to look at the excursion of $-2 \ln L$ around the minimum value (i.e.: the value that maximizes L), and find the interval where $-2 \ln L$ increases by one with respect to its minimum value. This method leads to identical results of Eq. (5.20) in the Gaussian case, but may lead to different results in the case of non-Gaussian PDFs. In particular, this approach in general may lead to asymmetric errors, but better follows the non-local behavior of the likelihood function.[3] This procedure is graphically shown in Fig. 5.3 for the case of a single parameter θ.

In more than one dimension the same method can be applied, obtaining an approximate multidimensional ellipsoid. For instance, the two-dimensional contour plot shown in Fig. 5.4 shows the values of the parameters of s and b that minimize Eq. (5.16) (central point) for the sample shown in Fig. 5.2. The approximate ellipse corresponds to the points for which $-\ln L(x_i; s, b, \theta)$ in Eq. (5.16) increases by two with respect to its minimum, corresponding to a 1σ uncertainty.

[2] For the users of the program MINUIT, this estimate correspond to call the method MIGRAD/HESSE.

[3] Again, for MINUIT users this procedure correspond to the call of the method MINOS.

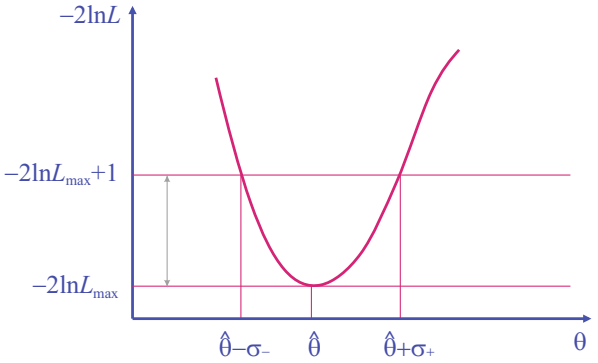

Fig. 5.3 Scan of $-2\ln L$ as a function of a parameter θ. the error interval is determined looking at the excursion of $-2\ln L$ around the minimum, at $\hat{\theta}$, finding the values for which $-2\ln L$ increases by one unit

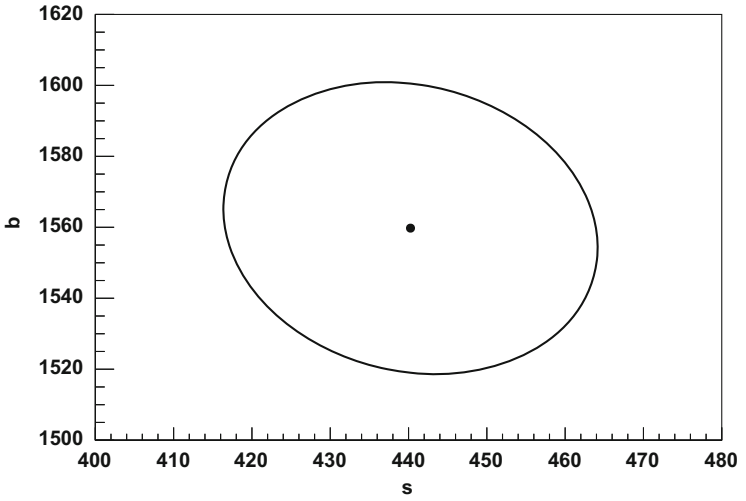

Fig. 5.4 Two-dimensional contour plot showing the 1σ error ellipse for the fit parameters s and b (number of signal and background events) corresponding to the fit in Fig. 5.2

The exact coverage, anyway, is not ensured even in this method where the excursion of the negative log-likelihood function is evaluated, though usually the coverage is improved with respect to the parabolic approximation in Eq. (5.20). Chapter 6 will discuss a more rigorous treatment of uncertainty intervals that ensure the proper coverage.

Example 5.17 – Maximum-Likelihood Estimate for an Exponential Distribution

We want to compute the maximum-likelihood estimate and error for the parameter λ of an exponential PDF:

$$f(t) = \lambda e^{-\lambda t}\,, \tag{5.23}$$

given n measurements of t: $t_1, \cdots, t_n$.

By minimizing analytically the likelihood function, written as the product of $f(t_1) \cdots f(t_n)$, one can easily show that the maximum-likelihood estimate of λ, which is the inverse of the average "lifetime", $\tau = 1/\lambda$, is:

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right)^{-1}\,, \tag{5.24}$$

with an error, using Eq. (5.20), of:

$$\sigma_\lambda = \hat{\lambda}/\sqrt{n}\,. \tag{5.25}$$

The derivation of this result is left as exercise to the reader.

5.6.1 Properties of Maximum-Likelihood Estimators

Maximum-likelihood estimators can be demonstrated to be consistent. Moreover, maximum-likelihood estimators may have a bias, but it is possible to demonstrate that the bias tends to zero as the number of measurements n tends to infinity. Maximum-likelihood estimators have efficiencies (compared to the Cramér–Rao bound, from Eq. (5.4)) that asymptotically, for large number of measurements, tend to one. Hence, maximum-likelihood estimators have, asymptotically, the lowest possible variance compared to any other estimator. Finally, maximum-likelihood estimator are invariant under reparameterization. That is, if we find a maximum of the likelihood in term of some parameters, in the new parameterization, the transformed parameters also maximize the likelihood.

Example 5.18 – Bias of the Maximum-Likelihood Estimate of a Gaussian Variance

The maximum-likelihood estimate of the variance of Gaussian-variables average from Eq. (5.19) is given by:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \,. \tag{5.26}$$

The bias is defined in Eq. (5.2). It can be shown analytically that the expected value of $\hat{\sigma}^2$ is:

$$\left\langle \hat{\sigma}^2 \right\rangle = \frac{n-1}{n}\sigma^2 \,, \tag{5.27}$$

where σ is the true variance. Hence, the maximum-likelihood estimate $\hat{\sigma}^2$ tends to underestimate the variance, and it has a bias given by:

$$b_{\hat{\sigma}^2} = \left\langle \hat{\sigma}^2 \right\rangle - \sigma^2 = \left(\frac{n-1}{n} - 1\right)\sigma^2 = -\frac{\sigma^2}{n} \tag{5.28}$$

which decreases with n, as in general expected for a maximum-likelihood estimate. In particular, an unbiased estimate can be obtained by multiplying the maximum-likelihood estimate by a correction factor $n/(n-1)$, which gives:

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \,. \tag{5.29}$$

5.7 Minimum χ^2 and Least-Squares Methods

Let's consider a case in which a number n of measurements $(y_1 \pm \sigma_1, \cdots y_n \pm \sigma_n)$ is performed, and each measurement $y_i \pm \sigma_i$ corresponds to a value x_i of a variable x. Assume we have a model for the dependence of y on the variable x as follows:

$$y = f(x; \vec{\theta}) \,, \tag{5.30}$$

where $\vec{\theta} = (\theta_1, \cdots, \theta_m)$ is a set of unknown parameters we want to determine. If the measurements y_i are Gaussianly distributed around $f(x_i; \vec{\theta})$ with standard deviation equal to σ_i, we can write the likelihood function for this problem as the product of

n Gaussian PDFs:

$$L(\vec{y};\vec{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(y_i - f(x_i;\vec{\theta}))^2}{2\sigma_i^2}\right], \tag{5.31}$$

where $\vec{y} = (y_1, \cdots, y_n)$.

Maximizing $L(\vec{y};\vec{\theta})$ is equivalent to minimize $-2\ln L(\vec{y};\vec{\theta})$:

$$-2\ln L(\vec{y};\vec{\theta}) = \sum_{i=1}^{n} \frac{(y_i - f(x_i;\vec{\theta}))^2}{\sigma_i^2} + \sum_{i=1}^{n} \ln 2\pi\sigma_i^2 . \tag{5.32}$$

The last term does not depend on the parameters $\vec{\theta}$, if the uncertainties σ_i are known and fixed, hence it is a constant that we can drop when performing the minimization. So, we can minimize the following quantity corresponding to the first term in Eq. (5.32):

$$\boxed{\chi^2(\vec{\theta}) = \sum_{i=1}^{n} \frac{(y_i - f(x_i;\vec{\theta}))^2}{\sigma_i^2} .} \tag{5.33}$$

An example of fit performed with the minimum-χ^2 method is shown in Fig. 5.5.

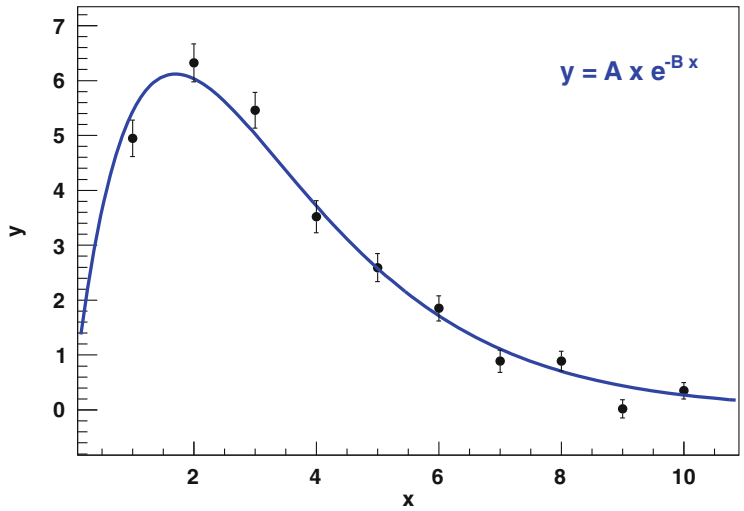

Fig. 5.5 Example of minimum-χ^2 fit of a simulated dataset. The points with the error bars, assumed resulting from Gaussian PDFs, are used to fit a function model of the type $y = f(x) = Axe^{-Bx}$, where A and B are unknown parameters determined form the fit procedure. The fit curve is superimposed as *solid blue line*

In case the uncertainties σ_i are all equal, it is possible to minimize the expression:

$$\boxed{R^2 = \sum_{i=1}^{n} (y_i - f(x_i; \vec{\theta}))^2 = \sum_{i=1}^{n} r_i \,.} \tag{5.34}$$

This minimization is referred to as *least-squares method*.

5.7.1 Linear Regression

In the easiest case of a linear function, the minimum-χ^2 problem can be solved analytically. The function f can be written as:

$$y = f(x; a, b) = a + bx \,, \tag{5.35}$$

a and b being free parameters. The χ^2 can then be written as:

$$\chi^2(a, b) = \sum_{i=1}^{n} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \,. \tag{5.36}$$

The analytical minimization can proceed as usual, imposing $\dfrac{\partial \chi^2(a,b)}{\partial a} = 0$ and $\dfrac{\partial \chi^2(a,b)}{\partial b} = 0$, which gives:

$$\hat{b} = \frac{\mathrm{cov}(x, y)}{\mathrm{V}[x]} \,, \tag{5.37}$$

$$\hat{a} = \langle y \rangle - \hat{b} \, \langle x \rangle \,, \tag{5.38}$$

where, introducing the weights w_i defined as:

$$w_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{n} 1/\sigma_j^2} \,, \tag{5.39}$$

we have:

$$\langle x \rangle = \sum_{i=1}^{n} w_i x_i \,, \tag{5.40}$$

$$\langle y \rangle = \sum_{i=1}^{n} w_i y_i \,, \tag{5.41}$$

$$\mathrm{V}[x] = \langle x \rangle^2 - \langle x^2 \rangle = (\sum_{i=1}^{n} w_i x_i)^2 - \sum_{i=1}^{n} w_i x_i^2 \,, \tag{5.42}$$

$$\mathrm{cov}(x, y) = \langle xy \rangle - \langle x \rangle \, \langle y \rangle = \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i \,. \tag{5.43}$$

The uncertainties on the parameter estimates $\hat{a}$ and $\hat{b}$ can be estimated as described in Sect. 5.6:

$$\frac{1}{\sigma_{\hat{a}}^2} = -\frac{\partial^2 \ln L}{\partial^2 a} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial^2 a} \tag{5.44}$$

and similarly:

$$\frac{1}{\sigma_{\hat{b}}^2} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial^2 b} \,. \tag{5.45}$$

Hence, we have:

$$\sigma_{\hat{a}} = \sqrt{\left(\sum_{i=1}^{n} \frac{1}{\sigma_i} \right)^{-1}} \,, \tag{5.46}$$

$$\sigma_{\hat{b}} = \sqrt{\left(\sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} \right)^{-1}} \,. \tag{5.47}$$

5.7.2 *Goodness of Fit*

One advantage of the minimum-χ^2 method is that the expected distribution of $\hat{\chi}^2$, the minimum χ^2 value, is known and is given by:

$$\boxed{P(\chi^2; n) = \frac{2^{-n/2}}{\Gamma(n/2)} \chi^{n-2} e^{-\chi^2/2} \,,} \tag{5.48}$$

where n is the *number of degrees of freedom* of the problem, i.e.: the number of measurements n minus the number of fit parameters m. We define the p-value as $P(\chi^2 \geq \hat{\chi}^2)$, the probability that χ^2 is greater or equal to the value $\hat{\chi}^2$ obtained at the fit minimum. If the data follow the assumed Gaussian distributions, the p-value is expected to be a random variable uniformly distributed from 0 to 1 (see Sect. 2.3).

Obtaining a small p-value of the fit could be symptom of a poor description of the theoretical model $y = f(x;\vec{\theta})$. For this reason, the minimum χ^2 value can be used as a measurement of the goodness of the fit. Anyway, setting a threshold, say p-value less than 0.05, to determine whether a fit can be considered acceptable or not, will always discard on average 5 % of the cases, even if the PDF model correctly describes the data, due to statistical fluctuations.

Note also that the p-value can not be considered as the *probability of the fit hypothesis to be true*. Such probability would only have a meaning in the Bayesian approach (see Chap. 3), and in that case it would require a different type of evaluation.

Unlike minimum-χ^2 fits, in general, for maximum-likelihood fits the value of $-2\ln L$ for which the likelihood function is maximum does not provide a measurement of the goodness of the fit. It is possible in some cases to obtain a goodness-of-fit measurement by finding the ratio of the likelihood functions evaluated in two hypotheses: the Wilks' theorem (see Sect. 7.6) ensures that such ratios, if the likelihood function obeys some regularity conditions, and if the two hypotheses are *nested* (see Chap. 7), is asymptotically, for a large number of repeated measurements, distributed as a χ^2 distribution (Eq. (5.48)). A more extensive discussion about the relation between likelihood functions, their ratios and χ^2 can be found in [6], and will be also discussed in Sect. 5.10.2 for what concerned binned fits.

5.8 Error Propagation

Once we have measured with our inference procedure the m parameters $\theta_1, \cdots, \theta_m$, in several cases we need to evaluate a new set of parameters, $\eta_1, \cdots, \eta_k$ that can be determined as functions of the measured ones. The uncertainty on the original parameters reflects in the uncertainty on the new parameter set.

The best option to determine the uncertainties on the new parameters would be to reparameterize the likelihood problem using the new set of parameters and to perform again the maximum-likelihood fit in terms of the new parameters. This is not always possible, in particular when the details of the likelihood problem are not available, for instance when retrieving a measurement from a published paper.

In those cases, the simplest procedure may be to perform a local linear approximation of the transformation functions that express the new parameters as a function of the measured ones. If the errors are sufficiently small, projecting them on the new variables using the linear approximation, leads to a sufficiently accurate result. In general, we can express the covariance matrix H_{ij} of the transformed parameters in terms of the covariance matrix Θ_{kl} of the original parameters with the following expression:

$$H_{ij} = \sum_{k,l} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \Theta_{k_l} \,, \tag{5.49}$$

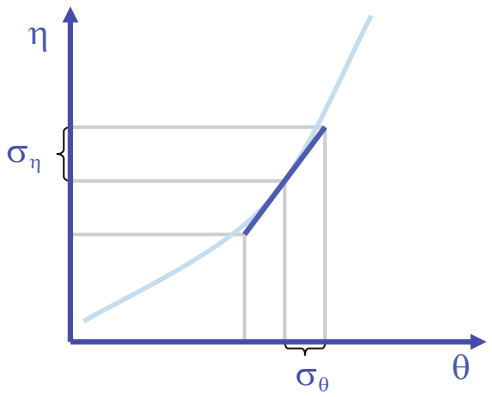

Fig. 5.6 Plot of a variable transformation $\eta = \eta(\theta)$, and visualization of the procedure of error propagation using local linear approximation

or, in matrix form:

$$\boldsymbol{H} = \boldsymbol{A}^T \boldsymbol{\Theta} \boldsymbol{A} \,, \tag{5.50}$$

where:

$$A_{ij} = \frac{\partial \eta_i}{\partial \theta_j} \,. \tag{5.51}$$

This procedure is visualized in the simplest case of a one-dimensional transformation in Fig. 5.6.

5.8.1 Simple Cases of Error Propagation

If we have to rescale a variable x:

$$y = ax \tag{5.52}$$

then the uncertainty squared will become:

$$\sigma_y^2 = \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 \sigma_x^2 = a^2 \sigma_x^2 \,, \tag{5.53}$$

hence:

$$\boxed{\sigma_{ax} = |a| \sigma_x \,.} \tag{5.54}$$

If we sum two uncorrelated variables x and y:

$$z = x + y \tag{5.55}$$

then the uncertainty squared will become:

$$\sigma_z^2 = \left(\frac{\mathrm{d}z}{\mathrm{d}x}\right)^2 \sigma_x^2 + \left(\frac{\mathrm{d}z}{\mathrm{d}y}\right)^2 \sigma_y^2 = \sigma_x^2 + \sigma_y^2 \,, \tag{5.56}$$

hence:

$$\boxed{\sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2}\,.} \tag{5.57}$$

More in general, also considering a possible correlation term, using Eq. (5.49), one may obtain:

$$\boxed{\sigma_{x+y} = \sqrt{\sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y}\,.} \tag{5.58}$$

In case of the product of two variables x and y:

$$z = xy \tag{5.59}$$

the relative uncertainty add in quadrature, plus a possible correlation term:

$$\boxed{\left(\frac{\sigma_{xy}}{xy}\right) = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 + \frac{2\rho\sigma_x\sigma_y}{xy}}\,.} \tag{5.60}$$

For the non-linear case of a power law:

$$y = x^n \,, \tag{5.61}$$

one has:

$$\left(\frac{\sigma_{x^n}}{x^n}\right) = |n| \left(\frac{\sigma_x}{x}\right) \,. \tag{5.62}$$

5.9 Issues with Treatment of Asymmetric Errors

In Sect. 5.6 we observed that maximum-likelihood fits may lead to asymmetric errors. The propagation of asymmetric errors and the combination of more measurements having asymmetric errors may require special care. If we have two

measurements: $x = \hat{x}^{+\sigma_x^+}_{-\sigma_x^-}$ and $y = \hat{y}^{+\sigma_y^+}_{-\sigma_y^-}$, the naïve extension of the sum in quadrature of errors, derived in Eq. (5.57), would lead to the (incorrect!) sum in quadratures of the positive and negative errors:

$$x + y = (\hat{x} + \hat{y})^{+\sqrt{(\sigma_x^+)^2+(\sigma_y^+)^2}}_{-\sqrt{(\sigma_x^-)^2+(\sigma_y^-)^2}}, \tag{5.63}$$

which has no statistical motivation. One reason for which Eq. (5.63) is incorrect, is the central limit theorem. In the case of a very large sample, symmetric errors can be approximated by the standard deviation of the distribution of that sample. In case of an asymmetric (skew) distribution, asymmetric errors may be related to a measurement of the skewness (Eq. (1.25)) of the distribution. Adding more random variables, each characterized by an asymmetric PDF, should lead to a resulting PDF that approaches a Gaussian more than the original PDFs, hence it should have more symmetric errors. From Eq. (5.63), instead, the error asymmetry would never decrease by adding more and more measurements with asymmetric errors.

One statistically correct way to propagate asymmetric errors on quantities (say η) that are expressed as functions of some original parameters (say θ) is to reformulate the fit problem in terms of the new parameters, and perform again the fit and error evaluation for the derived quantities (η). This approach is sometimes not feasible in the case when measurements with asymmetric errors are taken as result of a previous measurement (e.g.: from a publication) that do not specify the complete underlying likelihood model. In those cases, the treatment of asymmetric errors requires some assumptions on the underlying PDF model which is missing in the available documentation of the model's description. Discussions about how to treat asymmetric errors can be found in [7–9] using a frequentist approach. D'Agostini [10] also approaches this subject using the Bayesian approach (see Chap. 3). In the following part of the section the derivation from [9] will be briefly presented as example.

Imagine that the uncertainty on a parameter x' may arise from a non-linear dependence on another parameter x (i.e.: $x' = f(x)$) which has a symmetric uncertainly σ.

Figure 5.7 shows a simple case where a random variable x, distributed according to a Gaussian PDF, is transformed into a variable x' through a piece-wise linear transformation, leading to an asymmetric PDF. The two straight-line sections, with different slopes, join with continuity at the central value of the original PDF. This leads to a resulting PDF of the transformed variable which is piece-wise Gaussian: the two half-Gaussians, each corresponding to a 50 % probability, have different standard deviation parameters, σ'_+ and σ'_-. Such a PDF is also called *bifurcated* Gaussian in some literature.

If we have as original measurement: $x = \hat{x} \pm \sigma$, this will lead to a resulting measurement of the transformed variable: $x' = \hat{x}'^{+\sigma'_+}_{-\sigma'_-}$, where σ'_+ and σ'_- depend on σ through factors equal to the two different slopes: $\sigma'_+ = \sigma \cdot s_+$ and $\sigma'_- = \sigma \cdot s_-$ respectively, as seen from the Fig. 5.7.

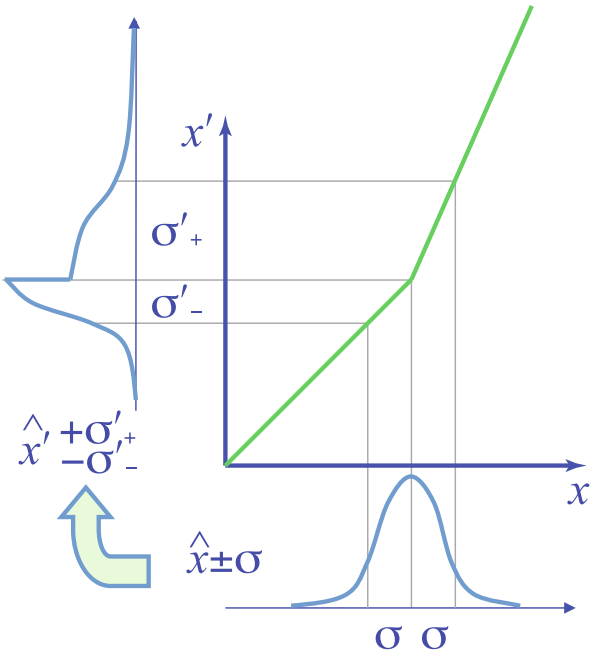

Fig. 5.7 Transformation of a variable x into a variable x' through a piece-wise linear transformation characterized by two different slopes. If x obeys a Gaussian PDF with standard deviation σ, x' obeys a *bifurcated* Gaussian, made of two Gaussian halves having different standard deviation parameters, σ'_+ and σ'_-

One consequence of the transformed PDF shape is that the average value of the transformed variable, $\langle x' \rangle$, is different from the most likely value, $\hat{x'}$. While the average value of the sum of two variables is equal to the sum of the average values of the two variables (Eq. (1.19)), this is not the case for the most likely value of the sum of the two variables.

In fact, using a naïve error treatment, like the one in Eq. (5.63), could even lead to a bias in the estimate of combined values. In particular, the average value of x' is:

$$\langle x' \rangle = \hat{x'} + \frac{1}{\sqrt{2\pi}}(\sigma'_+ - \sigma'_-) \,, \tag{5.64}$$

In that assumed case of a piece-wise linear transformation, we can consider, in addition to Eq. (5.64), also the corresponding expressions for the variance:

$$\mathrm{Var}[x'] = \left(\frac{\sigma'_+ + \sigma'_-}{2}\right)^2 + \left(\frac{\sigma'_+ - \sigma'_-}{2}\right)^2 \left(1 - \frac{2}{\pi}\right) \,, \tag{5.65}$$

and for the *unnormalized skewness*, defined in Eq. (1.26):

$$\gamma[x'] = \left\langle x'^3 \right\rangle - 3\langle x' \rangle \left\langle x'^2 \right\rangle + 2\langle x' \rangle^3 \tag{5.66}$$

$$= \frac{1}{2\pi}\left[2(\sigma_+^3 - \sigma_-^3) - \frac{3}{2}(\sigma_+ - \sigma_-)(\sigma_+^3 + \sigma_-^3) + \frac{1}{\pi}(\sigma_+ - \sigma_-)^3\right] . \tag{5.67}$$

The unnormalized skewness that leads to Eq. (5.67) add linearly, like the average and variance. If we have two variables affected by asymmetric errors, say: $x_{1}{}^{+\sigma_1^+}_{-\sigma_1^-}$ and $x_{2}{}^{+\sigma_2^+}_{-\sigma_2^-}$, and assume an underlying piece-wise linear model, as in the case treated here, we can compute the average, variance and unnormalized skewness for the sum of the two:

$$\langle x_1 + x_2 \rangle = \langle x_1 \rangle + \langle x_2 \rangle \ , \tag{5.68}$$

$$\mathrm{Var}[x_1 + x_2] = \mathrm{Var}[x_1] + \mathrm{Var}[x_2] \, , \tag{5.69}$$

$$\gamma[x_1 + x_2] = \gamma[x_1] + \gamma[x_2] \, , \tag{5.70}$$

where the three above quantities for x_1 and x_2 separately can be computed from Eqs. (5.64), (5.65) and (5.67). Inverting the relation between $\langle x_1 + x_2 \rangle$, $\mathrm{Var}[x_1 + x_2]$ and $\gamma[x_1 + x_2]$ and $\widehat{x_1 + x_2}$, σ^+_{1+2} and σ^-_{1+2} one can obtain, using numerical techniques, the correct estimate for $x_{1+2}{}^{+\sigma^+_{1+2}}_{-\sigma^-_{1+2}}$.

Barlow [9] also considers the case of a parabolic dependence and obtains a procedure to estimate $x_{1+2}{}^{+\sigma^+_{1+2}}_{-\sigma^-_{1+2}}$ with this second model.

Any estimate of the sum of two measurements affected by asymmetric errors requires an assumption of an underlying PDF model, and results may differ for different assumptions, depending case by case on the numerical values.

5.10 Binned Samples

The maximum-likelihood method discussed in Sect. 5.5 is used to perform parameter estimates using the complete set of information present in our measurements sample. In the case of a very large number of measurement, computing the likelihood function may become unpractical from the numerical point of view, and the implementation could require very large computing power.

For this reason, it's frequently preferred to perform the parameter estimate using a *summary* of the sample's information which is obtained by binning the distribution of the random variable (or variables) of interest and using as information the number of entries in each single bin.

In practice, we build an histogram, in one or more variables, of the experimental distribution. If the sample is composed of independent extractions from a given random distribution, the number of entries in each bin obeys a Poisson distribution whose expected number of entries in each bin can be determined from the theoretical distribution and depends on the unknown parameters one wants to estimate.

5.10.1 Minimum-χ^2 Method for Binned Histograms

In the case of a sufficiently large number of entries in each bin, the Poisson distribution describing the number of entries in each bin can be approximated by a Gaussian with variance equal to the average number of entries in that bin, according to what was derived in Sect. 2.9. In this case, the expression for $-2\ln L$ becomes:

$$-2\ln L = \sum_i \frac{(n_i - \int_{x_i^-}^{x_i^+} f(x;\theta_1,\cdots,\theta_m)\,\mathrm{d}x)^2}{n_i} + n\ln 2\pi + \sum_i \ln n_i\,, \tag{5.71}$$

where $[x_i^-, x_i^+]$ is the interval corresponding to the i^{th} bin. If the binning is sufficiently fine, we can write:

$$-2\ln L = \sum_i \frac{(n_i - f(x_i;\theta_1,\cdots,\theta_m)\,\delta x_i)^2}{n_i} + n\ln 2\pi + \sum_i \ln n_i\,, \tag{5.72}$$

where x_i is center of the i^{th} bin and δx_i is the bin's width.

Dropping the terms that are constant with respect to the unknown parameters θ_j, and hence are not relevant for the minimization, the maximum-likelihood problem reduces to the minimization of the following χ^2:

$$\chi^2 = \sum_i \frac{(n_i - \mu_i(\theta_1,\cdots,\theta_m))^2}{n_i}\,, \tag{5.73}$$

where the bin size δx_i has been absorbed in the definition of μ_i, for simplicity of notation:

$$\mu_i(\theta_1,\cdots,\theta_m) = f(x_i;\theta_1,\cdots,\theta_m)\delta x_i\,. \tag{5.74}$$

An alternative approach due to Pearson consists of replacing the denominator with the "theoretical" expected number of entries μ_i.

The value of χ^2 at the minimum can be used, as discussed in Sect. 5.7.2, as measurement of the goodness of the fit, where in this case the number of degrees of freedom is the number of bins n minus the number of fit parameters k.

5.10.2 Binned Poissonian Fits

In binned samples in general the Gaussian hypothesis assumed in Sect. 5.10.1 does not necessarily hold when the number of entries per bin may be small. In the Poissonian case, valid in general, also for small number of events, the negative

log-likelihood function that replaces Eq. (5.72) is:

$$-2\ln L = -2\ln \prod_i \text{Poiss}(n_i; \mu_i(\theta_1, \cdots, \theta_m)) \tag{5.75}$$

$$= -2\ln \prod_i \frac{e^{-\mu_i(\theta_1,\cdots,\theta_m)}\mu_i(\theta_1, \cdots, \theta_m)^{n_i}}{n_i!}\,. \tag{5.76}$$

Using the approach proposed in [6], we can divide the likelihood function by it's maximum value, which we obtain replacing μ_i with n_i. Hence, we obtain:

$$\chi^2_\lambda = -2\ln \frac{L(n_i; \mu_i(\theta_1, \cdots, \theta_m))}{L(n_i; n_i)} = -2\ln \frac{e^{-\mu_i}{\mu_i}^{n_i}}{n_i!}\frac{n_i!}{e^{-n_i}n_i^{n_i}} \tag{5.77}$$

$$= 2\sum_i \left[\mu_i(\theta_1, \cdots, \theta_m) - n_i + n_i \ln\left(\frac{n_i}{\mu_i(\theta_1, \cdots, \theta_m)}\right)\right]. \tag{5.78}$$

From Wilks' theorem (see Sect. 7.6) the distribution of χ^2_λ can be approximated with the χ^2 distribution (Eq. (5.48)), hence χ^2_λ can be used to determine a p-value that provides a measure of the goodness of the fit.

In case the Wilks' theorem hypotheses do not hold, e.g.: insufficiently large number of measurements, the distribution of χ^2_λ for the specific problem can still be determined by generating a large number of Monte Carlo pseudo-experiments that reproduce the theoretical PDF, and the p-value can be computed accordingly. This technique is often called *toy Monte Carlo*.

5.11 Combining Measurements

The problem of combining two or more measurements of the same unknown quantity θ can be addressed in general building a likelihood function that combines the two or more experimental samples and contains the same unknown parameter set. Minimizing the new likelihood gives an estimate of θ that combines the information of the available samples.

This option is not always pursuable, either for intrinsic complexity of the problem, or because the full original samples are not available, and only the final results are known, as in the case of combining results available from different publications.

In the case of Gaussian approximation, as assumed in Sect. 5.10.1, we can use the minimum χ^2 method to perform a combination of measurements having different uncertainties, as seen in the following sections.

5.11.1 Weighted Average

Imagine we have two measurements of the same quantity x, which are $x_1 \pm \sigma_1$ and $x_2 \pm \sigma_2$. Assuming a Gaussian distribution for x_1 and x_2 and no correlation between the two measurements, we can build a χ^2 as:

$$\chi^2 = \frac{(x - x_1)^2}{\sigma_1^2} + \frac{(x - x_2)^2}{\sigma_2^2} \,. \tag{5.79}$$

We can find the value $x = \hat{x}$ that minimizes the χ^2 by imposing:

$$0 = \frac{\partial \chi^2}{\partial x} = 2\frac{(\hat{x} - x_1)}{\sigma_1^2} + 2\frac{(\hat{x} - x_2)}{\sigma_2^2} \tag{5.80}$$

which gives:

$$\hat{x} = \frac{\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \,. \tag{5.81}$$

The variance of the estimate $\hat{x}$ can be computed using Eq. (5.20):

$$\frac{1}{\sigma_{\hat{x}}^2} = -\frac{\partial^2 \ln L}{\partial x^2} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial x^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \,. \tag{5.82}$$

Equation (5.81) is called *weighted average*, and can be generalized for n measurements as:

$$\hat{x} = \frac{\sum_{i_1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \,, \tag{5.83}$$

where $w_i = 1/\sigma_i^2$, and in general:

$$\sigma_{\hat{x}} = \sqrt{\frac{1}{\sum_{i=1}^{n} w_i}} \,. \tag{5.84}$$

5.11.2 χ^2 in n Dimensions

The χ^2 definition from Eq. (5.79) can be generalized in the case of n measurements $\vec{x} = (x_1, \cdots, x_n)$ of the parameter $\vec{\mu} = (\mu_1, \cdots, \mu_n)$, where $\mu_i = \mu_i(\theta_1, \cdots, \theta_m)$ and the correlation matrix of the measurements $(x_1, \cdots, x_n)$ is σ_{ij}, as follows:

$$\chi^2 = \sum_{i,j=1}^{n} (x_i - \mu_i)\sigma_{ij}^{-1}(x_j - \mu_j) = (\vec{x} - \vec{\mu})^T \boldsymbol{\sigma}^{-1} (\vec{x} - \vec{\mu}) \tag{5.85}$$

$$= \left(x_1 - \mu_1, \cdots, x_n - \mu_n\right) \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ \cdots \\ x_n - \mu_n \end{pmatrix}. \tag{5.86}$$

The χ^2 can be minimized in order to have the best-fit estimate of the unknown parameters $\theta_1, \cdots, \theta_m$. Numerical algorithms are needed for the general case, analytical minimizations are as usual possible only in the simplest cases.

5.11.3 The Best Linear Unbiased Estimator

Let's consider the case of two measurements of the same quantity x: $x = x_1 \pm \sigma_1$ and $x = x_2 \pm \sigma_2$, which have a correlation coefficient ρ. In this case, the χ^2 can be written as:

$$\chi^2 = \begin{pmatrix} x - x_1 & x - x_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x - x_1 \\ x - x_2 \end{pmatrix}. \tag{5.87}$$

Following the same procedure used to obtain Eq. (5.81), we can easily obtain:

$$\hat{x} = \frac{x_1(\sigma_2^2 - \rho\sigma_1\sigma_2) + x_2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}, \tag{5.88}$$

with the following estimate for the variance:

$$\sigma_{\hat{x}}^2 = \frac{\sigma_1^2\sigma_2^2(1 - \rho^2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}. \tag{5.89}$$

In the general case of more than two measurement, the minimization of the χ^2 is equivalent to finding the *best linear unbiased estimate* (BLUE), i.e.: the unbiased linear combination of the measurements $\vec{x} = (x_1, \cdots, x_n)$ whose weights

$\vec{w} = (w_1, \cdots, w_n)$ give the lowest variance. If we assume:

$$\hat{x} = \sum_i w_i x_i = \vec{x} \cdot \vec{w}\,, \tag{5.90}$$

the condition of a null bias implies that:

$$\sum_i w_i = 1\,. \tag{5.91}$$

The variance can be expressed as:

$$\sigma^2_{\hat{x}} = \vec{w}^{\mathrm{T}} \boldsymbol{\sigma}\, \vec{w}\,, \tag{5.92}$$

where σ_{ij} is the covariance matrix of the n measurements. It can be shown [11] that the weights can be determined according to the following expression:

$$\vec{w} = \frac{\boldsymbol{\sigma}^{-1}\vec{u}}{\vec{u}^{\mathrm{T}}\boldsymbol{\sigma}^{-1}\vec{u}}\,, \tag{5.93}$$

where $\vec{u}$ is the vector having all elements equal to the unity: $\vec{u} = (1, \cdots, 1)$.

The interpretation of Eq. (5.88) becomes more intuitive [12] if we introduce the *common error*, defined as:

$$\sigma_{\mathrm{C}} = \rho\sigma_1\sigma_2\,. \tag{5.94}$$

Imagine, for instance, that the two measurements are affected by a common uncertainty, like the knowledge of the integrated luminosity in the case of a cross section measurement, while the other uncertainties are uncorrelated. In that case, we could write the two measurements as:

$$x = x_1 \pm \sigma_1' \pm \sigma_{\mathrm{C}} \tag{5.95}$$

$$x = x_2 \pm \sigma_2' \pm \sigma_{\mathrm{C}} \tag{5.96}$$

where $\sigma_1'^2 = \sigma_1^2 - \sigma_{\mathrm{C}}^2$ and $\sigma_2'^2 = \sigma_2^2 - \sigma_{\mathrm{C}}^2$. Equation (5.88) becomes:

$$\hat{x} = \frac{\dfrac{x_1}{\sigma_1^2 - \sigma_{\mathrm{C}}^2} + \dfrac{x_2}{\sigma_2^2 - \sigma_{\mathrm{C}}^2}}{\dfrac{1}{\sigma_1^2 - \sigma_{\mathrm{C}}^2} + \dfrac{1}{\sigma_2^2 - \sigma_{\mathrm{C}}^2}} \tag{5.97}$$

with a variance:

$$\sigma_{\hat{x}}^2 = \frac{1}{\frac{1}{\sigma_1^2 - \sigma_C^2} + \frac{1}{\sigma_2^2 - \sigma_C^2}} + \sigma_C^2 \,. \tag{5.98}$$

Equation (5.97) is equivalent to the weighted average from Eq. (5.81), where the errors $\sigma_1'^2$ and $\sigma_2'^2$ are used. Equation (5.98) shows that the additional term σ_C has to be added in quadrature to the usual error in the case of no correlation.

In general, as can be seen from Eq. (5.88), weights with the BLUE method can become negative. This leads to the counter-intuitive result that the combined measurement lies outside the range delimited by the two central values. Also, in the case when $\rho = \sigma_1/\sigma_2$, the weight of the first measurement is zero, hence the combined central value is not influenced by x_2. Conversely, if $\rho = \sigma_2/\sigma_1$, the central value is not influenced bye x_1.

Example 5.19 – Reusing Multiple Times the Same Measurement Doesn't Improve a Combination

Assume to have a single measurement, $x \pm \sigma$, and we want to use it twice to determine again the same quantity. Without considering the correlation coefficient $\rho = 1$, one would expect to reduce the uncertainty σ "for free" by a factor $\sqrt{2}$, which is clearly a wrong answer.

The correct use of Eqs. (5.88) and (5.89) lead, in the limit when $\rho = 1$, to $\hat{x} = x \pm \sigma$, i.e.: as expected, no precision is gained by using the same measurement twice.

5.11.3.1 Iterative Application of the BLUE Method

The BLUE method is unbiased by construction assuming that the true uncertainties and their correlations are known. Anyway, it can be proven that BLUE combinations may exhibit a bias if uncertainty estimates are used in place of the true ones, and in particular if the uncertainty estimates depend on measured values. For instance, when contributions to the total uncertainty are known as relative uncertainties, the actual uncertainty estimates are obtained as the product of the relative uncertainties times the measured central values. An iterative application of the BLUE method can be implemented in order to mitigate such a bias.

L. Lyons et al. remarked in [13] the limitation of the application of the BLUE method in the combination of lifetime measurements where uncertainty estimates $\hat{\sigma}_i$ of the true unknown uncertainties σ_i were used, and those estimates had a dependency on the measured lifetime.

They also demonstrated that the application of the BLUE method violates, in that case, the "combination principle": if the set of measurements is split into a number of subsets, then the combination is first performed in each subset and finally all subset combinations are combined into a single combination, this result differs from the combination of all individual results of the entire set.

L. Lyons et al. recommended to apply iteratively the BLUE method, rescaling at each iteration the uncertainty estimates according to the central value obtained with the BLUE method in the previous iteration, until the sequence converges to a stable result. Lyons et al. showed that the bias of the BLUE estimate is reduced compared to the standard application of the BLUE method. A more extensive study of the iterative application of the BLUE method is also available in [14].

References

1. H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, 1946)
2. C.R. Rao, Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 8189 (1945)
3. W. Eadie, D. Drijard, F. James, M. Roos, B. Saudolet, *Statistical Methods in Experimental Physics* (North Holland, London, 1971)
4. F. James, M. Roos, MINUIT: Function minimization and error analysis. Cern Computer Centre Program Library, Geneve Long Write-up No. D506, 1989
5. R. Brun, F. Rademakers, ROOT—an object oriented data analysis framework, in *Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. Meth.*, vol. A389 (1997), pp. 81–86. See also http://root.cern.ch/
6. S. Baker, R. Cousins, Clarification of the use of chi-square and likelihood functions in fit to histograms. Nucl. Instrum. Methods **A221**, 437–442 (1984)
7. R. Barlow, Asymmetric errors, in *Proceedings of PHYSTAT2003*, SLAC, Stanford, 8–11 September 2003. http://www.slac.stanford.edu/econf/C030908/
8. R. Barlow, Asymmetric statistical errors. arXiv:physics/0406120v1 (2004)
9. R. Barlow, Asymmetric systematic errors. arXiv:physics/0306138 (2003)
10. G. D'Agostini, Asymmetric uncertainties sources, treatment and potential dangers. arXiv:physics/0403086 (2004)
11. L. Lions, D. Gibaut, P. Clifford, How to combine correlated estimates of a single physical quantity. Nucl. Instrum. Methods **A270**, 110–117 (1988)
12. H. Greenlee, Combining CDF and D0 physics results, in *Fermilab Workshop on Confidence Limits*, March 2000
13. L. Lyons, A.J. Martin, D.H. Saxon, On the determination of the B lifetime by combining the results of different experiments. Phys. Rev. **D41**, 982–985 (1990)
14. L. Lista, The bias of the unbiased estimator: a study of the iterative application of the BLUE method. Nucl. Instrum. Methods **A764**, 82–93 (2014); and corr. ibid. **A773**, 87–96 (2015)

Chapter 6
Confidence Intervals

6.1 Neyman's Confidence Intervals

Section 5.6 presented approximate methods to determine uncertainties of maximum-likelihood estimates. Basically either the negative log-likelihood function was approximated to a parabola at the minimum, corresponding to a Gaussian PDF approximation, or the excursion of the negative log-likelihood around the minimum was considered to possibly give asymmetric uncertainties. None of those methods guarantees an exact coverage of the uncertainty intervals. In many cases the provided level of approximation is sufficient, but this may not always be the cases, in particular for measurements with a small number of events or PDF models that exhibit significant deviation from the Gaussian approximation.

A more rigorous and general treatment of *confidence intervals* under the frequentist approach is due to Neyman [1], which will be discussed in the following in the simplest case of a single parameter.

Let's consider a variable x distributed according to a PDF which depends on an unknown parameter θ. We have in mind that x could be the value of the estimator of the parameter θ. Neyman's procedure to determine confidence intervals proceeds in two steps:

1. the construction of a *confidence belt*;
2. its inversion to determine the confidence interval.

6.1.1 Construction of the Confidence Belt

As first step, the confidence belt is determined by scanning the parameter space, varying θ within its allowed range. For each fixed value of the parameter $\theta = \theta_0$, the corresponding PDF which describes the distribution of x, $f(x|\theta_0)$, is known.

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_6

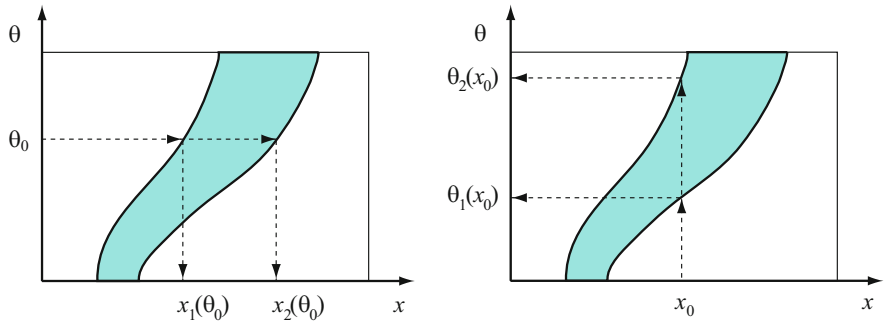

Fig. 6.1 Graphical illustration of Neyman's belt construction (*left*) and inversion (*right*)

According to the PDF $f(x|\theta_0)$, an interval $[x_1(\theta_0), x_2(\theta_0)]$ is determined whose corresponding probability is equal to the specified *confidence level*, defined as CL= $1-\alpha$, usually equal to 68.27 % (1σ), 90 or 95 %:

$$1-\alpha = \int_{x_1(\theta_0)}^{x_2(\theta_0)} f(x|\theta_0)\,\mathrm{d}x\,. \tag{6.1}$$

Neyman's construction of the confidence belt is graphically illustrated in Fig. 6.1, left.

The choice of $x_1(\theta_0)$ and $x_2(\theta_0)$ has still some arbitrariness, since there are different possible intervals having the same probability, according to the condition in Eq. (6.1). The choice of the interval is referred to in literature as *ordering rule*.

For instance, one can chose an interval centered around the average value of x corresponding to θ_0 (denoted as $\langle x\rangle\,|_{\theta_0}$ below), i.e.: an interval:

$$[x_1(\theta_0), x_2(\theta_0)] = [\langle x\rangle\,|_{\theta_0} - \delta, \langle x\rangle\,|_{\theta_0} + \delta]\,, \tag{6.2}$$

where δ is determined in order to ensure the coverage condition in Eq. (6.1).

Alternatively, one can chose the interval with equal areas of the PDF "tails" at the two extreme sides, i.e.: such that:

$$\int_{-\infty}^{x_1(\theta_0)} f(x|\theta_0)\mathrm{d}x = \frac{\alpha}{2} \quad \text{and} \quad \int_{x_2(\theta_0)}^{+\infty} f(x|\theta_0)\mathrm{d}x = \frac{\alpha}{2}\,. \tag{6.3}$$

Other possibilities consist of choosing the interval having the smallest size, or fully asymmetric intervals on either side: $[x_1(\theta_0), +\infty]$ or $[-\infty, x_2(\theta_0)]$. Other possibilities are also considered in literature. Figure 6.2 shows three of the possible cases described above.

A special ordering rule was introduced by Feldman and Cousins based on a likelihood-ratio criterion and will be discussed in Sect. 6.4.

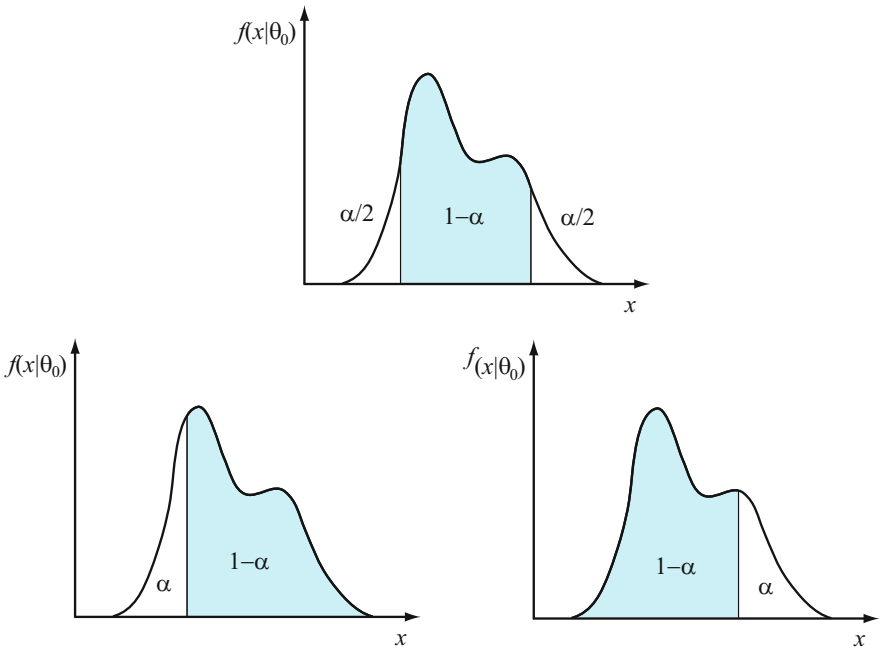

Fig. 6.2 Three possible choices of ordering rule: central interval (*top*) and fully asymmetric intervals (*bottom left, right*)

Given a choice of the ordering rule, the intervals $[x_1(\theta), x_2(\theta)]$, for all possible values of θ, define the Neyman belt in the space (x, θ) as shown in Fig. 6.1.

6.1.2 Inversion of the Confidence Belt

As second step of the Neyman procedure, given a measurement $x = x_0$, the confidence interval for θ is determined by inverting the Neyman belt (Fig. 6.1, right): two extreme values $\theta_1(x_0)$ and $\theta_2(x_0)$ are determined as the intersections of the vertical line at $x = x_0$ with the two boundary curves of the belt.

The interval $[\theta_1(x_0), \theta_2(x_0)]$ has, by construction, a coverage equal to the desired confidence level, $1-\alpha$. This means that, if θ is equal to the true value θ^{true}, extracting $x = x_0$ randomly according to the PDF $f(x|\theta^{\text{true}})$, θ^{true} will be included in the determined confidence interval $[\theta_1(x_0), \theta_2(x_0)]$ in a fraction $1-\alpha$ of the cases, in the limit of a large number of extractions.

6.2 Binomial Intervals

In Sect. 1.11.1 the binomial distribution (Eq. (1.35)) was introduced:

$$P(n;N) = \frac{n!}{N!(N-n)!} p^N (1-p)^{N-n} . \tag{6.4}$$

Given an extraction of n, $n = \hat{n}$, the parameter p can be estimated as:

$$\hat{p} = \frac{\hat{n}}{N} . \tag{6.5}$$

An approximate estimate of the uncertainty on $\hat{p}$ is often obtained by replacing p with its estimate $\hat{p}$ in the expression for the variance from Eq. (1.37):

$$\delta p \simeq \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} . \tag{6.6}$$

This may be justified by the law of large numbers because for $n \to \infty$ $\hat{p}$ and p will coincide. This is clearly not the case in general for finite values of n, and Eq. (6.6) clearly gives a null error in case either $\hat{n} = 0$ or $\hat{n} = N$, i.e. for $\hat{p} = 0$ or 1, respectively.

A general solution to this problem is due to Clopper and Pearson [2], and consists of finding the interval $[p_1, p_2]$ that gives (at least) the correct coverage.

If the confidence level is $1 - \alpha$, we can find p_1 such that:

$$P(n \geq \hat{n}|N, p_1) = \sum_{n=\hat{n}}^{N} \frac{n!}{N!(N-n)!} p_1^n (1-p_1)^{N-n} = \frac{\alpha}{2} , \tag{6.7}$$

and p_2 such that:

$$P(n \leq \hat{n}|N, p_2) = \sum_{n=0}^{\hat{n}} \frac{n!}{N!(N-n)!} p_2^n (1-p_2)^{N-n} = \frac{\alpha}{2} . \tag{6.8}$$

This corresponds, in a discrete case, to the Neyman inversion described in Sect. 6.1.

The presence of a discrete observable quantity, n in this case, does not allow a continuous variation of the observable and the probability associated to possible discrete intervals $[n_1(p_0), n_2(p_0)]$, corresponding to an assumed parameter value $p = p_0$, can have again discrete possible values. In the Neyman's construction, one has to chose the smallest interval $[n_1, n_2]$, consistently with the adopted ordering rule, that has at least the desired coverage. In this way, the confidence interval $[p_1, p_2]$ determined by the inversion procedure for the parameter p could *overcover*, i.e.: may have a corresponding probability larger than the desired $1 - \alpha$. In this sense, the interval is said to be *conservative*.

Another case of a discrete application of the Neyman belt inversion in a Poissonian problem can be found in Sect. 8.6.

Example 6.20 – Application of the Clopper–Pearson Method

As exercise, we compute the 90 % CL interval in case of a measurement $\hat{n} = N$ with the Clopper–Pearson method.

We need to determine the values p_1 and p_2 such that Eqs. (6.7) and (6.8) hold. In this case, those give, considering that $\alpha = 0.10$:

$$P(n \geq N|N, p_1) = \frac{N!}{N!\,0!}\, p_1^N (1-p_1)^0 = p_1^N = 0.05\,,$$
$$P(n \leq N|N, p_2) = 1\,.$$

So, for p_2 we should consider the largest allowed value $p_2 = 1.0$, since the probability $P(n \leq 10|10, p_2)$ is always equal to one. The first equation can be inverted and gives:

$$p_1 = \exp[\ln(0.05)/N]\,.$$

For instance, for $N = 10$ we have $p_1 = 0.741$, and the confidence interval is $[0.74, 1.00]$, which has not a null size, as would have been the case using the approximated expression in Eq. (6.6).

Symmetrically, the Clopper–Pearson evaluation of the confidence interval for $\hat{n} = 0$ gives $[0.00, 0.26]$ in the case $N = 10$.

6.3 The "Flip-Flopping" Problem

In order to determine confidence intervals, a consistent choice of the ordering rule has to be adopted. Feldman and Cousins demonstrated [3] that the ordering rule choice must not depend on the outcome of our measurement, otherwise the quoted confidence interval or upper limit could not correspond to the correct confidence level (i.e.: do not respect the coverage).

In some cases, experiments searching for a rare signal decide to quote their final result in different possible ways, switching from a central interval to an upper limit, depending on the outcome of the measurement. A typical choice is to:

- quote an upper limit if the measured signal yield is not larger than at least three times its uncertainty;
- quote instead the central value with its uncertainty if the measured signal exceeds three times its uncertainty.

This "3σ" significance criterion will be discussed in more details later, see Sect. 8.2 for a more general definition of *significance level*.

This problem is sometimes referred to in literature as *flip-flopping*, and can be illustrated in a simple example [3]. Imagine a model where a random variable x obeys a Gaussian distribution with a fixed and known standard deviation σ, for simplicity we can take $\sigma = 1$, and an unknown average μ which is bound to be greater or equal to zero. This is the case of a signal yield, or cross section measurement.

The quoted central value must always be greater than or equal to zero, given the assumed constraint. Imagine we decide to quote, as measured value for μ, zero if the significance, defined as x/σ, is less than 3σ, otherwise we quote the measured value x:

$$\mu = \begin{cases} x & \text{if } x/\sigma \geq 3 \\ 0 & \text{if } x/\sigma < 3 \end{cases}, \tag{6.9}$$

As confidence interval, given the measurement of x, we could decide to quote a central value if $x/\sigma \geq 3$ with a symmetric error: $\pm\sigma$ at the 68.27 % CL, or $\pm 1.645\sigma$ at 90 % CL, and instead, we may quote an upper limit x^{up}, i.e.: the confidence interval $[0, x^{\text{up}}]$, if $x/\sigma < 3$. The upper limit on μ corresponds to $\mu < x^{\text{up}} = x + 1.282$ at 90 % CL, given the corresponding area under a Gaussian PDF.

In summary, the quoted confidence interval at 90 % CL is:

$$[\mu_1, \mu_2] = \begin{cases} [x - 1.645, x + 1.645] & \text{if } x/\sigma \geq 3 \\ [0, x + 1.282] & \text{if } x/\sigma < 3 \end{cases}, \tag{6.10}$$

The situation is shown in Fig. 6.3.

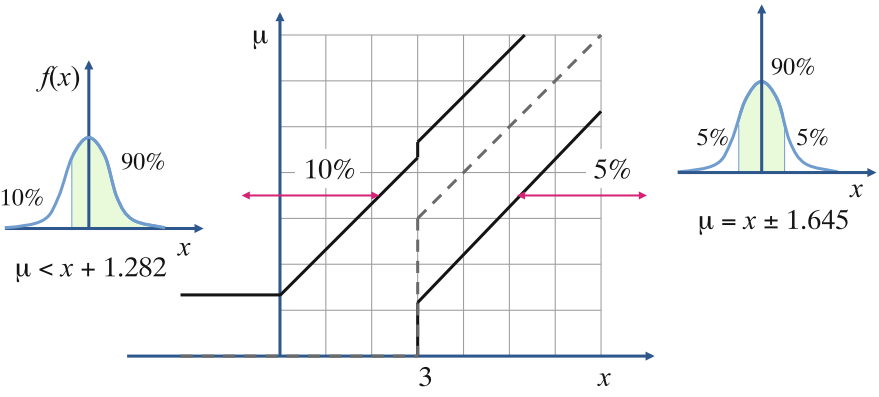

Fig. 6.3 Illustration of the *flip-flopping* problem. The plot shows the quoted central value of μ as a function of the measured x (*dashed line*), and the 90 % confidence interval corresponding to a choice to quote a central interval for $x/\sigma \geq 3$ and an upper limit for $x/\sigma < 3$

The choice to switch from a central interval to a fully asymmetric interval (upper limit) based on the observation of x produces an incorrect coverage: looking at Fig. 6.3, depending on the value of μ, the interval $[x_1, x_2]$ obtained by crossing the confidence belt by an horizontal line, one may have cases where the coverage decreases from 90 to 85 %, which is lower than the desired CL (purple line in Fig. 6.3). Next Sect. 6.4, presents the method due to Feldman and Cousins to preserve consistently the coverage for this example without incurring the flip-flopping problem.

6.4 The Unified Feldman–Cousins Approach

In order to avoid the flip-flopping problem and to ensure the correct coverage, the ordering rule proposed by Feldman and Cousins [3] provides a Neyman confidence belt, as defined in Sect. 6.1, that smoothly changes from a central or quasi-central interval to an upper limit in the case of low observed signal yield.

The ordering rule is based on the likelihood ratio that will be further discussed in Sect. 7.3: given a value θ_0 of the unknown parameter θ, the chosen interval of the variable x used for the Neyman-belt construction is defined by the ratio of two PDFs of x, one under the hypothesis that θ is equal to the considered fixed value θ_0, the other under the hypothesis that θ is equal to the maximum-likelihood estimate value $\theta_{\text{best}}(x)$ corresponding to the given measurement x. The likelihood ratio must be greater than a constant k_α whose value depends on the chosen confidence level $1-\alpha$. In a formula:

$$\lambda(x|\theta_0) = \frac{f(x|\theta_0)}{f(x|\theta_{\text{best}}(x))} > k_\alpha \,. \tag{6.11}$$

The constant k_α should be taken such that the integral of the PDF evaluated in the range corresponding to $\lambda(x|\theta_0) > k_\alpha$ is equal to $1-\alpha$. Hence, the confidence interval R_α for a given value $\theta = \theta_0$ is given by:

$$R_\alpha(\theta_0) = \{x : \lambda(x|\theta_0) > k_\alpha\} \,, \tag{6.12}$$

and the constant k_α must be chosen in such a way that:

$$\int_{R_\alpha} f(x|\theta_0)\, \mathrm{d}x = 1-\alpha \,. \tag{6.13}$$

This case is illustrated in Fig. 6.4.

Feldman and Cousins computed the confidence interval for the simple Gaussian case discussed in Sect. 6.3. The value for $\mu = \mu_{\text{best}}(x)$ that maximizes the likelihood function, given x and under the constraint $\mu \geq 0$, is:

$$\mu_{\text{best}}(x) = \max(x, 0) \,. \tag{6.14}$$

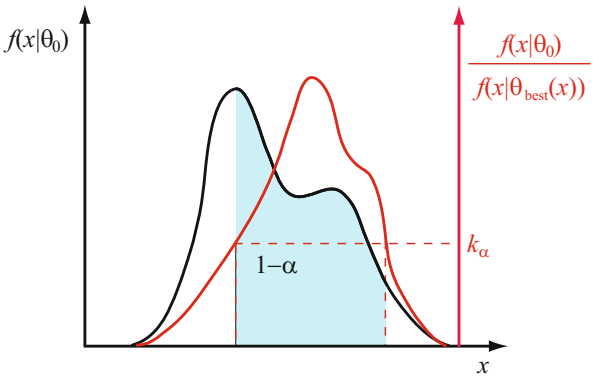

Fig. 6.4 Ordering rule in the Feldman–Cousins approach, based on the likelihood ratio $\lambda(x|\theta_0) = f(x|\theta_0)/f(x|\theta_{\text{best}}(x))$

The PDF for x using the maximum-likelihood estimate for μ becomes:

$$f(x|\mu_{\text{best}}(x)) = \begin{cases} \frac{1}{\sqrt{2\pi}} & \text{if } x \geq 0 \\ \frac{1}{\sqrt{2\pi}} e^{-x^2/2} & \text{if } x < 0 \end{cases} . \tag{6.15}$$

The likelihood ratio in Eq. (6.11) can be written in this case as:

$$\lambda(x|\mu) = \frac{f(x|\mu)}{f(x|\mu_{\text{best}}(x))} = \begin{cases} \exp(-(x-\mu)^2/2 & \text{if } x \geq 0 \\ \exp(x\mu - \mu^2/2) & \text{if } x < 0 \end{cases} . \tag{6.16}$$

The interval $[x_1(\mu_0), x_2(\mu_0)]$, for a given $\mu = \mu_0$, can be found numerically using the equation $\lambda(x|\mu) > k_\alpha$ and imposing the desired confidence level $1 - \alpha$ from Eq. (6.13).

The results are shown in Fig. 6.5, and can be compared with Fig. 6.3. Using the Feldman–Cousins approach, for large values of x one gets the usual symmetric confidence interval. As x moves towards lower values, the interval becomes more and more asymmetric, and at some point it becomes fully asymmetric (i.e.: $[0, \mu^{\text{up}}]$), determining an upper limit μ^{up}. For negative values of x the result is always an upper limit avoiding unphysical values with negative values of μ. As seen, this approach smoothly changes from a central interval to an upper limit, yet ensuring the correct required (90 % in this case) coverage.

More cases treated using the Feldman–Cousins approach will be presented in Chap. 8.

The application of the Feldman–Cousins method requires, in most of the cases, numerical treatment, even for simple PDF models, like the Gaussian case discussed above, because it requires the inversion of the integral in Eq. (6.13). The procedure requires to scan the parameter space, and in case of complex model may be very CPU intensive. For this reason, other methods are often preferred to the Feldman–Cousins for such complex cases.

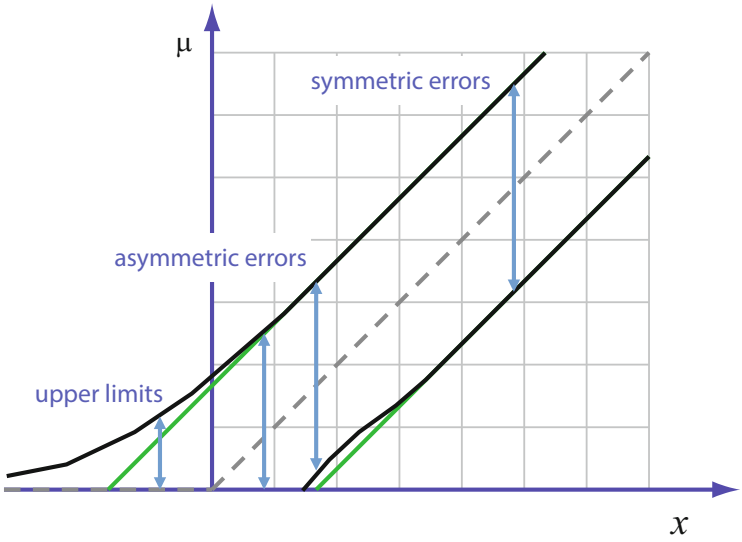

Fig. 6.5 Neyman confidence belt constructed using the Feldman–Cousins ordering

References

1. J. Neyman, Outline of a theory of statistical estimation based on the clasiscal theory of probability. Philos. Trans. R. Soc. Lond. A Math. Phys. Sci. **236**, 333–380 (1937)
2. C.J. Clopper, E. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26**, 404–413 (1934)
3. G. Feldman, R. Cousins, Unified approach to the classical statistical analysis of small signals. Phys. Rev. **D57**, 3873–3889 (1998)

Chapter 7
Hypothesis Tests

7.1 Introduction to Hypothesis Tests

A key task in most of physics measurements is to discriminate between two or more *hypotheses* on the basis of the observed experimental data.

A typical example in physics is the identification of a particle type (e.g.: as a muon vs pion) on the basis of the measurement of a number of discriminating variables provided by a particle-identification detector (e.g.: the depth of penetration in an iron absorber, the energy release in scintillator crystals or measurements from a Cherenkov detector, etc.).

Another example is to determine whether a sample of events is composed of background only or it contains a mixture of background plus signal events. This may allows to ascertain the presence of a new signal, which may lead to a discovery.

This problem in statistics is known as *hypothesis test*, and methods have been developed to assign an observation, which consists of the measurements of specific discriminating variables, to one of two or more hypothetical models, considering the predicted probability distributions of the observed quantities under the different possible assumptions.

In statistical literature when two hypotheses are present, these are called *null hypothesis*, H_0, and *alternative hypothesis*, H_1.

Assume that the observed data sample consists of the measurement of a number n of variables, $\vec{x} = (x_1, \cdots, x_n)$ randomly distributed according to some probability density function, which is in general different under the hypotheses H_0 and H_1. A measurement of whether the observed data sample better agrees with H_0 or rather with H_1 can be given by the value of a function of the measured sample $\vec{x}$, $t = t(\vec{x})$, called *test statistic*, whose PDF under the considered hypotheses can be derived from the (joint) PDF of the observable quantities $\vec{x}$.

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_7

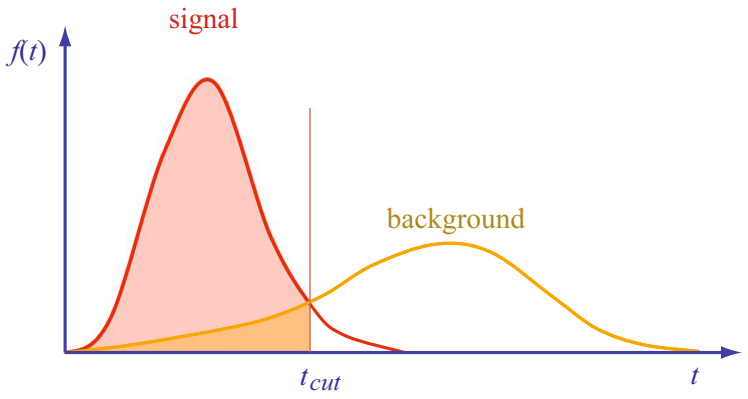

Fig. 7.1 Probability distribution functions for a discriminating variable $t(x) = x$ which has two different PDFs for the signal (*red*) and background (*yellow*) hypotheses under test

One simple example is to use a single variable x which has discriminating power between two hypotheses, say *signal* = "muon" versus *background* = "pion", as shown in Fig. 7.1. A good "separation" of the two cases can be achieved if the PDFs of x under the hypotheses H_1 = *signal* and H_0 = *background* are appreciably different.

On the basis of the observed value $\hat{x}$ of the discriminating variable x, a simple test statistics can be defined as the measured value itself:

$$\hat{t} = t(\hat{x}) = \hat{x}\,. \tag{7.1}$$

A *selection requirement* (in physics jargon sometimes called *cut*) can be defined by identifying a particle as a muon if $\hat{t} \leq t_{cut}$ or as a pion if $\hat{t} > t_{cut}$, where the value t_{cut} is chosen a priori.

Not all real muons will be correctly identified as a muon according to this criterion, as well as not all real pions will be correctly identified as pions. The expected fraction of selected signal particles (muons) is usually called signal *selection efficiency* and the expected fraction of selected background particles (pions) is called *misidentification probability*.

Misidentified particles constitute a background to positively identified signal particles. Applying the required selection (cut), in this case $t \leq t_{\text{cut}}$, on a data sample made of different detected particles, each providing a measurements of x, the selected data sample will be enriched of signal, reducing the fraction of background in the selected data sample with respect to the original unselected sample. The sample will be actually enriched if the selection efficiency is larger than the misidentification probability, which is the case considering the shapes of the PDFs in Fig. 7.1 and the chosen selection cut.

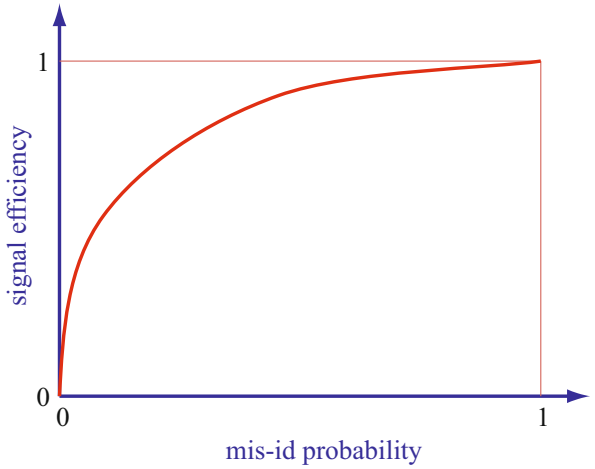

Fig. 7.2 Signal efficiency versus background misidentification probability

Statistical literature defines the *significance level* α as the probability to reject the hypothesis H_1 if it is true. The case of rejecting H_1 if true is called *error of the first kind*. In our example, this means selecting a particle as a pion in case it is a muon. Hence the *selection efficiency* for the signal corresponds to $1 - \alpha$.

The probability β to reject the hypothesis H_0 if it is true (*error of the second kind*) is the *misidentification probability*, i.e.: the probability to incorrectly identify a pion as a muon, in our case.

Varying the value of the selection cut t_{cut} different values of selection efficiency and misidentification probability (and the corresponding values of α and β) are determined. A typical curve representing the signal efficiency versus the misidentification probability obtained by varying the selection requirement is shown in Fig. 7.2. A good selection should have a low misidentification probability corresponding to a high selection efficiency. But clearly the background rejection can't be perfect (i.e.: the misidentification probability can't drop to zero) if the distributions $f(x|H_0)$ and $f(x|H_1)$ overlap, as in Fig. 7.2.

More complex examples of cut-based selections involve multiple variables, where selection requirements in multiple dimensions can be defined as regions in the discriminating variables space. Events are accepted as "signal" or as "background" if they fall inside or outside the selection region. Finding an optimal selection in multiple dimensions is usually not a trivial task. Two simple examples of selections with very different performances in terms of efficiency and misidentification probability are shown on Fig. 7.3.

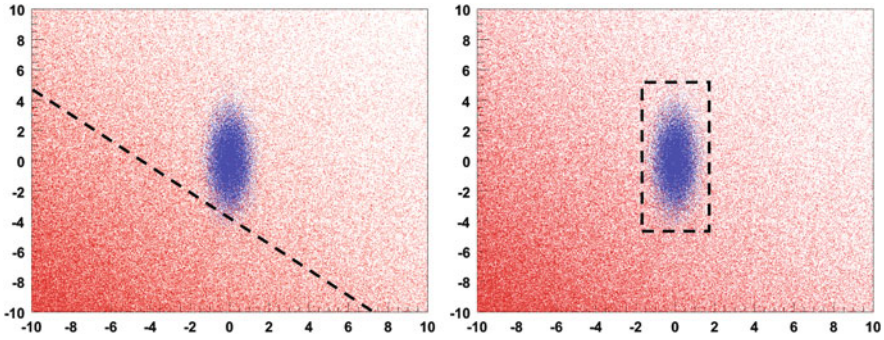

Fig. 7.3 Examples of two-dimensional selections of a signal (*blue dots*) against a background (*red dots*). A linear cut is chosen on the *left plot*, while a box cut is chosen on the *right plot*

7.2 Fisher's Linear Discriminant

A discriminator in more dimensions that uses a linear combination of n discriminating variables is due to Fisher [1]. The method aims at finding an hyperplane in the n-dimensional space which gives the *best* expected separation between two random variables sets whose PDFs in the multivariate space are known.

The criterion consists of maximizing the distance of the means of the two PDFs when projected on the axis perpendicular to the selection hyperplane, while minimizing the variance *within* each class, defined below. More quantitatively, given a direction vector $\vec{w}$, if we project the two samples along the direction $\vec{w}$, being μ_1 and μ_2 the averages and σ_1 and σ_2 the standard deviations of the two PDF projections, the Fisher discriminant is defined as:

$$\boxed{J(\vec{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\,.} \tag{7.2}$$

The numerator of Eq. (7.2) can be written as:

$$\mu_1 - \mu_2 = \vec{w}^T(\vec{m}_1 - \vec{m}_2)\,, \tag{7.3}$$

where $\vec{m}_1$ and $\vec{m}_2$ are the averages in n dimensions of the two PDFs. Moreover, we can write the square of Eq. (7.3) as:

$$(\mu_1 - \mu_2)^2 = \vec{w}^T(\vec{m}_1 - \vec{m}_2)(\vec{m}_1 - \vec{m}_2)^T\vec{w} = \vec{w}^T\boldsymbol{S_B}\vec{w}\,, \tag{7.4}$$

where we have defined the *between classes* scatter matrix $\boldsymbol{S_B}$ as:

$$\boldsymbol{S_B} = (\vec{m}_1 - \vec{m}_2)(\vec{m}_1 - \vec{m}_2)^T . \tag{7.5}$$

The projections of the two $n \times n$ covariance matrices $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ give:

$$\sigma_1^2 = \vec{w}^T \boldsymbol{S_1} \vec{w} , \tag{7.6}$$

$$\sigma_2^2 = \vec{w}^T \boldsymbol{S_2} \vec{w} , \tag{7.7}$$

from which the denominator of Eq. (7.2) becomes:

$$\sigma_1^2 + \sigma_2^2 = \vec{w}^T (\boldsymbol{S_1} + \boldsymbol{S_2}) \vec{w} = \vec{w}^T \boldsymbol{S_W} \vec{w} , \tag{7.8}$$

where we have defined the *within classes* scatter matrix $\boldsymbol{S_W}$ as:

$$\boldsymbol{S_W} = \boldsymbol{S_1} + \boldsymbol{S_2} . \tag{7.9}$$

The Fisher's discriminant from Eq. (7.2) can be rewritten as:

$$J(\vec{w}) = \frac{(\vec{w}^T (\vec{m}_1 - \vec{m}_2))^2}{\vec{w}^T (\boldsymbol{S_1} + \boldsymbol{S_2}) \vec{w}} = \frac{\vec{w}^T \boldsymbol{S_B} \vec{w}}{\vec{w}^T \boldsymbol{S_W} \vec{w}} . \tag{7.10}$$

The problem of finding the vector $\vec{w}$ that maximizes $J(\vec{w})$ can be solved by performing the derivatives of $J(\vec{w})$ with respect to the components w_i of $\vec{w}$. It can also be demonstrated that the problem is equivalent to solve the eigenvalues equation:

$$\boldsymbol{S_B} \vec{w} = \lambda \boldsymbol{S_W} \vec{w} , \tag{7.11}$$

i.e.:

$$\boldsymbol{S_W}^{-1} \boldsymbol{S_B} \vec{w} = \lambda \vec{w} , \tag{7.12}$$

which leads to the solution:

$$\boxed{\vec{w} = \boldsymbol{S_W}^{-1} \boldsymbol{S_B} (\vec{m}_1 - \vec{m}_2) .} \tag{7.13}$$

A practical way to find the Fisher's discriminant is to provide two *training samples* that approximate the two PDFs by generating with a Monte Carlo technique (see Chap. 4) a very large number of random extractions of the n discriminating variables for each of the two PDFs. The averages and covariance matrices can be

determined from the Monte Carlo samples and Eq. (7.13) can be applied numerically to find the direction $\vec{w}$ that maximizes Fisher's discriminant.

7.3 The Neyman–Pearson Lemma

The performance of a selection criterion can be considered optimal if it achieves the smallest misidentification probability for a desired value of the selection efficiency.

A test statistic that ensures the optimal performance in this sense is provided by the Neyman–Pearson lemma [2]. According to this lemma, such test statistic is defined as the ratio of the likelihood functions evaluated for the observed data sample $\vec{x}$ under the two hypotheses H_1 and H_0:

$$\lambda(\vec{x}) = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)} \,. \tag{7.14}$$

For a fixed signal efficiency $\varepsilon = 1 - \alpha$ the selection that corresponds to the lowest possible misidentification probability β is given by:

$$\boxed{\lambda(\vec{x}) = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)} \geq k_\alpha \,,} \tag{7.15}$$

where the value of the "cut" k_α should be set in order to achieve the required value of α. This corresponds to chose a point in the curve shown in Fig. 7.2 such that $\varepsilon = 1 - \alpha$ corresponds to a required value.

If the multidimensional PDFs that characterize our discrimination problem are known, the Neyman–Pearson lemma provides a procedure to implement a selection that achieves the optimal performances.

In many realistic cases it is not easy to determine the correct model for the multidimensional PDFs, and approximated solutions may be adopted. Numerical methods and algorithms exist to find selections in the variable space that have performances in terms of efficiency and misidentification probability close to the optimal limit given by the Neyman–Pearson lemma. Some of those algorithms achieve great complexity.

Among such methods some of the most frequently used ones in High Energy Physics are *Artificial Neural Networks* and *Boosted Decision Trees*, which are covered in this text. An introduction to Artificial Neural Network can be found in [3], while an introduction to Boosted Decision Trees is available in [4].

7.4 Likelihood Ratio Discriminant

If the n variables $x_1, \cdots, x_n$ that characterize our problem are independent, the likelihood function can be factorized into the product of one-dimensional marginal PDFs:

$$\lambda(x_1, \cdots, x_n) = \frac{L(x_1, \cdots, x_n|H_1)}{L(x_1, \cdots, x_k|H_0)} = \frac{\prod_{j=1}^{n} f_j(x_j|H_1)}{\prod_{j=1}^{n} f_j(x_i|H_0)}. \tag{7.16}$$

This allows in many cases to simplify the computation of the likelihood ratio and to easily obtain the optimal selection according to the Neyman–Pearson lemma.

Even if it is not possible to factorize the PDFs into the product of one-dimensional marginal PDFs, i.e.: the variables are not independent, the test statistic inspired by Eq. (7.16) can be used as discriminant using the marginal PDFs f_j for the individual variables:

$$\lambda(x_1, \cdots, x_n) = \frac{\prod_{j=1}^{n} f_j(x_j|H_1)}{\prod_{j=1}^{n} f_j(x_i|H_0)}. \tag{7.17}$$

In case the PDFs cannot be exactly factorized, anyway, the test statistic defined in Eq. (7.17) will differ from the exact likelihood ratio in Eq. (7.14) and hence it will correspond to worse performances in term of α and β compared with the best possible values provided by the Neyman–Pearson lemma.

In some cases, anyway, the simplicity of this method can justify its application in spite of the suboptimal performances. Indeed, the marginal PDFs f_j can be simply obtained using Monte Carlo *training samples* with a large number of entries that allow to produce histograms corresponding to the distribution of the individual variables x_j that, to a good approximation, reproduce the marginal PDFs.

Some numerical applications implement this factorized likelihood-ratio technique after applying proper variable transformation in order to reduce or eliminate the variables' correlation. This mitigates the decrease of performance compared with the Neyman–Pearson limit, but does not necessarily allow to reach the optimal performances, because uncorrelated variables are not necessarily independent (see for instance the Example 2.6).

7.5 Kolmogorov–Smirnov Test

A test due to Kolmogorov [5], Smirnov [6] and Chakravarti [7] can be used to assess the hypothesis that a data sample is compatible with a given distribution.

Assume to have a sample $x_1, \cdots, x_n$, ordered by increasing values of x that we want to compare with a distribution $f(x)$, which is assumed to be a continuous function.

The discrete cumulative distribution of the sample can be defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \theta(x - x_i) \,, \tag{7.18}$$

where the function θ is defined as:

$$\theta(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases} \,. \tag{7.19}$$

The distribution $F_n(x)$ can be compared to the cumulative distribution of $f(x)$ defined as (Eq. 2.13):

$$F(x) = \int_{-\infty}^{x} f(x') \, \mathrm{d}x' \,. \tag{7.20}$$

The maximum distance between the two cumulative distributions $F_n(x)$ and $F(x)$ is used to quantify the agreement of the data sample $x_1, \cdots, x_n$ with $f(x)$:

$$D_n = \sup_x |F_x(x) - F(x)| \,. \tag{7.21}$$

The definition of $F_n(x)$, $F(x)$ and D_n is visualized in Fig. 7.4.

For large n, D_n converges to zero in probability. The distribution of the test statistic $K = \sqrt{n} D_n$, if the null hypothesis (the sample $x_1, \cdots, x_n$ is distributed

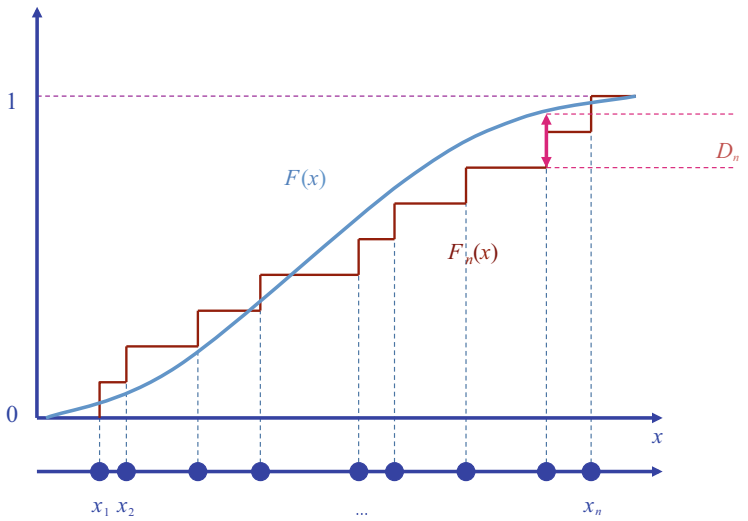

Fig. 7.4 Graphical representation of the Kolmogorov–Smirnov test

according to $f(x)$) is true, does not depend on the distribution $f(x)$, and the probability that K is smaller or equal to a given value k is given by Marsaglia et al. [8]:

$$P(k \leq k) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-i^2k^2} = \frac{\sqrt{2\pi}}{k}\sum_{i=1}^{\infty}e^{-\frac{(2i-1)^2\pi^2}{8k^2}} . \tag{7.22}$$

It's important to notice that Kolmogorov–Smirnov is a non-parametric test, i.e.: if some parameters of the distribution $f(x)$ are determined by the sample $x_1, \cdots, x_n$ than the test cannot be applied.

A pragmatic solution to this problem in case parameters of $f(x)$ have been estimated from the sample $x_1, \cdots, x_n$ is to still use the test statistic k, but not to rely on Eq. (7.22) and determine the distribution of k for the specific problem empirically by using a toy Monte Carlo generation. This is implemented, for instance, as optional method in the package root [9].

The Kolmogorov–Smirnov test can also be used to compare two samples, say $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$, and asses the hypothesis that both come from the same distribution. In this case, the maximum distance:

$$D_{n,m} = \sup_x |F_n(x) - F_m(y)| \tag{7.23}$$

asymptotically converges to zero as n and m are sufficiently large, and the following test statistic asymptotically follows a Kolmogorov distribution in Eq. (7.22):

$$\sqrt{\frac{nm}{n+m}}D_{n,m} . \tag{7.24}$$

Alternative tests to Kolmogorov–Smirnov are due to Stephens [10], Anderson and Darling [11], Cramér [12] and von Mises [13].

7.6 Wilks' Theorem

When a large number of measurements is available, Wilks' theorem allows to find an approximate asymptotic expression for a test statistic based on a likelihood ratio inspired by the Neyman–Pearson lemma (Eq. 7.14).

Let's assume that two hypotheses H_1 and H_0 can be defined in terms of a set of parameters $\vec{\theta} = (\theta_1, \cdots, \theta_m)$ that appear in the definition of the likelihood function: the condition that H_1 is true can be expressed as $\vec{\theta} \in \Theta_1$, while the condition that H_0 is true is equivalent to $\vec{\theta} \in \Theta_0$. Let's also assume that $\Theta_0 \subseteq \Theta_1$ (*nested hypotheses*). Wilks' theorem [14] ensures, assuming some regularity conditions of the likelihood

function, that the quantity:

$$\chi^2_r = -2 \ln \frac{\sup_{\vec{\theta} \in \Theta_1} \prod_{i=1}^{N} L(\vec{x}_i; \vec{\theta})}{\sup_{\vec{\theta} \in \Theta_0} \prod_{i=1}^{N} L(\vec{x}_i; \vec{\theta})}, \tag{7.25}$$

corresponding to the observed data sample $(\vec{x}_1, \cdots, \vec{x}_N)$, has a distribution that can be approximated for $N \to \infty$, if H_0 is true, with a χ^2 distribution having a number of degrees of freedom equal to the difference between the dimensionality of the set Θ_1 and the dimensionality of the set Θ_0.

As a more specific example, let's assume that we separate the parameter set into one parameter of interest μ and the remaining parameters, which are taken as nuisance parameters (see Sect. 5.2), $\vec{\theta} = (\theta_1, \cdots, \theta_m)$. For instance, μ may be the ratio of signal cross section to it's theoretical value (*signal strength*), and $\mu = 1$ corresponds to a signal cross section equal to the theory prediction.

If we take as H_0 the hypothesis $\mu = \mu_0$, while H_1 is the hypothesis that μ may have any possible value greater or equal to zero, Wilks' theorem ensures that:

$$\chi^2_r(\mu_0) = -2 \ln \frac{\sup_{\vec{\theta}} \prod_{i=1}^{N} L(\vec{x}_i; \mu_0, \vec{\theta})}{\sup_{\mu, \vec{\theta}} \prod_{i=1}^{N} L(\vec{x}_i; \mu, \vec{\theta})} \tag{7.26}$$

is asymptotically distributed as a χ^2 with k degrees of freedom.

The denominator, $\sup_{\mu, \vec{\theta}} \prod_{i=1}^{N} L(\vec{x}_i; \mu, \vec{\theta})$, is the likelihood evaluated at the maximum-likelihood point, i.e.: at the value of the parameters $\mu = \hat{\mu}$ and $\vec{\theta} = \hat{\hat{\theta}}$ that maximize the likelihood function:

$$\prod_{i=1}^{N} L(\vec{x}_i; \hat{\mu}, \hat{\hat{\theta}}).$$

In the numerator, only the nuisance parameters $\vec{\theta}$ are fit and μ is fixed to the constant value $\mu = \mu_0$. If the values of $\vec{\theta}$ that maximize the likelihood function for a fixed $\mu = \mu_0$ are $\vec{\theta} = \hat{\hat{\theta}}(\mu_0)$, then Eq. (7.26) can be written as:

$$\chi^2_r(\mu_0) = \frac{L(\vec{x}|\mu, \hat{\hat{\theta}}(\mu_0))}{L(\vec{x}|\hat{\mu}, \hat{\hat{\theta}})}. \tag{7.27}$$

This test statistic is called in literature *profile likelihood* and will be used later in Sect. 8.12 to determine upper limits.

7.7 Likelihood Ratio in the Search for a New Signal

In the previous section we considered the following likelihood function for a set of observations $(\vec{x}_1, \cdots, \vec{x}_N)$ with a set of parameters $\vec{\theta}$:

$$L(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta}) = \prod_{i=1}^{N} f(\vec{x}_i; \vec{\theta}) \,. \tag{7.28}$$

Two hypotheses H_1 and H_0 are represented as two possible sets of values Θ_1 and Θ_0 of the parameters $\vec{\theta} = (\theta_1, \cdots, \theta_m)$ that characterize the PDFs.

Usually we want to use the number of events N as information in the likelihood definition, hence we use the *extended likelihood function* (see Sect. 5.5.2) multiplying Eq. (7.28) by a Poissonian factor corresponding to the probability to observe a number of events N:

$$L(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta}) = \frac{e^{-\nu(\vec{\theta})} \nu(\vec{\theta})^N}{N!} \prod_{i=1}^{N} f(\vec{x}_i; \vec{\theta}) \,. \tag{7.29}$$

In the Poissonian term the expected number of event ν may also depend on the parameters $\vec{\theta}$: $\nu = \nu(\vec{\theta})$. Typically, we want to discriminate between two hypotheses, which are the presence of only background events in our sample, i.e.: $\nu = b$, against the presence of both signal and background, i.e.: $\nu = \mu s + b$. Here we have introduced the multiplier μ, called *signal strength*, assuming that the expected signal yield from theory is s and we consider all possible values of the expected signal yield, given by μs, by varying μ while keeping s constant at the value predicted from theory. The signal strength μ is widely used in data analyses performed at the Large Hadron Collider.

The hypothesis H_0 corresponding to the presence of background only is equivalent to $\mu = 0$, while the hypothesis H_1 corresponding to the presence of signal plus background allows any non-null positive value of μ.

The PDF $f(\vec{x}_i; \vec{\theta})$ can be written as superposition of two components, one PDF for the signal and another for the background, weighted by the expected signal and background fractions, respectively:

$$f(\vec{x}; \vec{\theta}) = \frac{\mu s}{\mu s + b} f_s(\vec{x}; \vec{\theta}) + \frac{b}{\mu s + b} f_b(\vec{x}; \vec{\theta}) \,. \tag{7.30}$$

In this case the extended likelihood function from Eq. (7.29) becomes:

$$L_{s+b}(\vec{x}_1, \cdots, \vec{x}_M; \mu, \vec{\theta}) = \frac{e^{-(\mu s(\vec{\theta}) + b(\vec{\theta}))}}{N!} \prod_{i=1}^{N} \left(\mu s f_s(\vec{x}_i; \vec{\theta}) + b f_b(\vec{x}_i; \vec{\theta}) \right) \,. \tag{7.31}$$

Note that also s an b may depend on the unknown parameters $\vec{\theta}$:

$$s = s(\vec{\theta}) \,, \tag{7.32}$$

$$b = b(\vec{\theta}) \,. \tag{7.33}$$

For instance in a search for the Higgs boson the theoretical cross section may depend on the Higgs boson's mass, as well as the PDF for the signal f_s, which represents a peak centered around the Higgs boson's mass.

Under the hypothesis H_0, i.e.: $\mu = 0$ (background only), the likelihood function can be written as:

$$L_b(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta}) = \frac{e^{-b(\vec{\theta})}}{N!} \prod_{i=1}^{N} \left(b f_b(\vec{x}_i; \vec{\theta}) \right) \,. \tag{7.34}$$

The term $1/N!$ disappears when performing the likelihood ratio in Eq. (7.14), which becomes:

$$\begin{aligned} \lambda(\mu, \vec{\theta}) &= \frac{L_{s+b}(\vec{x}_1, \cdots, \vec{x}_N; \mu, \vec{\theta})}{L_b(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta})} \\ &= \frac{e^{-(\mu s(\vec{\theta}) + b(\vec{\theta}))}}{e^{-b(\vec{\theta})}} \prod_{i=1}^{N} \frac{\mu s f_s(\vec{x}_i; \vec{\theta}) + b f_b(\vec{x}_i; \vec{\theta})}{b f_b(\vec{x}_i; \vec{\theta})} \\ &= e^{-\mu s(\vec{\theta})} \prod_{i=1}^{N} \left(\frac{\mu s f_s(\vec{x}_i; \vec{\theta})}{b f_b(\vec{x}_i; \vec{\theta})} + 1 \right) \,. \end{aligned} \tag{7.35}$$

Moving to the negative logarithm of the likelihood function, we have:

$$-\ln \lambda(\mu, \theta) = \mu s(\vec{\theta}) - \sum_{i=1}^{N} \ln \left(\frac{\mu s f_s(\vec{x}_i; \vec{\theta})}{b f_b(\vec{x}_i; \vec{\theta})} + 1 \right) \,. \tag{7.36}$$

In the case of a simple event counting, where the likelihood function only accounts for the Poissonian probability term, assuming that the expected signal and background yields depend on the unknown parameters $\vec{\theta}$, the likelihood function only depends on the number of observed event N, and the likelihood ratio which defines the test statistic is:

$$\lambda(\vec{\theta}) = \frac{L_{s+b}(N; \vec{\theta})}{L_b(N; \vec{\theta})} \,, \tag{7.37}$$

where L_{s+b} and L_b are Poissonian probabilities for N corresponding to expected average of $\mu s + b$ and b respectively. More explicitly $\lambda(\vec{\theta})$ can be written as:

$$\lambda(\vec{\theta}) = \frac{e^{-(\mu s(\vec{\theta})+b(\vec{\theta}))}\left(\mu s(\vec{\theta}) + b(\vec{\theta})\right)^N}{N!}\frac{N!}{e^{-b(\vec{\theta})}b(\vec{\theta})^N} = \\ = e^{-\mu s(\vec{\theta})}\left(\frac{\mu s(\vec{\theta})}{b(\vec{\theta})} + 1\right)^N . \tag{7.38}$$

Moving to the negative logarithm the above expression becomes:

$$-\ln\lambda(\vec{\theta}) = \mu s(\vec{\theta}) - N\ln\left(\frac{\mu s(\vec{\theta})}{b(\vec{\theta})} + 1\right), \tag{7.39}$$

which is a simplified version of Eq. (7.36) where the terms f_s and f_b have been dropped.

These results will be used to determine upper limits to s in the search for a new signal, as will be seen in Chap. 8. In particular, this derivation will be very useful for the determination of the upper limits using the *modified frequentist approach*, or CL_s, as discussed in Sect. 8.10.

References

1. R.A. Fisher, The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
2. J. Neyman, E. Pearson, On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. Ser. A **231**, 289–337 (1933)
3. C. Peterson, T.S. Rgnvaldsson, An introduction to artificial neural networks. *LU-TP-91-23. LUTP-91-23*, 1991. 14th CERN School of Computing, Ystad, Sweden, 23 Aug–2 Sep 1991
4. B.P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, G. McGregor, Boosted decision trees as an alternative to artificial neural networks for particle identification. Nucl. Instrum. Methods **A543**, 577–584 (2005)
5. A. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari **4**, 83–91 (1933)
6. N. Smirnov, Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. **19**, 279–281 (1948)
7. I.M. Chakravarti, R.G. Laha, J. Roy, *Handbook of Methods of Applied Statistics*, vol. I (Wiley, New York, 1967)
8. G. Marsaglia, W.W. Tsang, J. Wang, Evaluating Kolmogorov's distribution. J. Stat. Softw. **8**, 1–4 (2003)
9. R. Brun, F. Rademakers, ROOT—an object oriented data analysis framework, in *Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996, Nuclear Instruments and Methods*, vol. A389 (1997), pp. 81–86. See also http://root.cern.ch/
10. M.A. Stephens, EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. **69**, 730–737 (1974)

11. T.W. Anderson, D.A. Darling, Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Ann. Math. Stat. **23**, 193–212 (1952)
12. H. Cramér, On the composition of elementary errors. Scand. Actuar. J. **1928**(1), 13–74 (1928). doi:10.1080/03461238.1928.10416862
13. R.E. von Mises, *Wahrscheinlichkeit, Statistik und Wahrheit* (Julius Springer, Vienna, 1928)
14. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938)

Chapter 8
Upper Limits

8.1 Searches for New Phenomena: Discovery and Upper Limits

The goal of many experiments is to search for new physics phenomena. If an experiment provides a convincing measurement of a new signal, the result should be published claiming a discovery. If the outcome is not sufficiently convincing, the publication can anyway quote an upper limit to the "intensity" of the new signal, which usually allow to exclude parameter sets of a new theory.

In order to give a quantitative measure of how "convincing" the result of an experiment is, there are different possible approached. Using Bayesian statistics (Chap. 3), the posterior probability, given the experiment's measurement and a subjective prior, can quantify the degree of belief that the new signal hypothesis is true. In particular, when comparing two hypotheses, in this case the presence or absence of signal, the Bayes factor (see Sect. 3.5) can be used to quantify how strong the evidence for a new signal is against the background-only hypothesis.

With the frequentist approach, prior probabilities, that introduce a degree of subjectiveness, are not allowed, and the achieved *significance level*, defined quantitatively below in Sect. 8.2, is used to claim a discovery. The significance measures the probability that, in case of presence of background only, a statistical fluctuation in data might have produced by chance the observed features that are interpreted as a new signal.

As in other cases analyzed above, anyway, the interpretation of a discovery in the Bayesian and frequentist approaches are very different, as will be discussed in the following.

In case no convincing new signal is observed, in many cases it is nonetheless interesting to quote as result of the search for the new phenomena the upper limit on the expected yield of the hypothetical new signal. From upper limits to the signal yield often it is possible to indirectly derive limits on the properties of the new signal

L. Lista, *Statistical Methods for Data Analysis in Particle Physics*,
Lecture Notes in Physics 909, DOI 10.1007/978-3-319-20176-4_8

that influence the signal yield, such as the mass of a new particle, if it is related to the theoretical cross section, etc.

The determination of upper limits is in many cases a complex task and the computation frequently requires numerical algorithms. Several methods are adopted in High Energy Physics and are documented in literature to determine upper limits. The interpretation of the obtained limits can be, even conceptually, very different, depending on the adopted method.

This chapter will introduce the concept of significance and will present the most popular methods to set upper limits in High Energy Physics discussing their main benefits and limitations. The interpretation of upper limits under the frequentist and Bayesian approaches will be discussed, devoting special attention to the so-called *modified frequentist approach*, which is a popular method in High Energy Physics that is neither a purely frequentist nor a Bayesian method.

8.2 Claiming a Discovery

8.2.1 The p-Value

Given an observed data sample, claiming the discovery of a new signal requires to determine that the sample is sufficiently *inconsistent* with the hypothesis that only background is present in the data. A test statistic can be used to measure how consistent or inconsistent the observation is with the hypothesis of the presence of background only.

A quantitative measurement of the inconsistency with the background-only hypothesis is given by the *significance*, defined from the probability p (p-value) that the considered test statistics t assumes a value greater or equal to the observed one in the case of pure background fluctuation. We have implicitly assumed that large values of t corresponds to a more signal-like sample. The p-value has by construction (see the end of Sect. 2.3) a uniform distribution between 0 and 1 for the background-only hypothesis and tends to have small values in the presence of a signal.

If the number of observed events is adopted as test statistics (*event counting experiment*), the p-value can be determined as the probability to count a number of events equal or greater than the observed one assuming the presence of no signal and the expected background level. In this cases the test statistics and the p-value may assume discrete values and the distribution of the p-value is only approximately uniform.

Example 8.21 – p-Value for a Poissonian Counting

Figure 8.1 shows a Poisson distribution corresponding to an expected number of events (background) of 4.5. In case the observed number of events is 8, the p-value is equal to the probability to observe 8 or more events, i.e.: it is given by:

$$p = \sum_{n=8}^{\infty} \text{Pois}(n; 4.4) = 1 - e^{-4.5} \sum_{n=0}^{7} \frac{4.5^n}{n!} .$$

Performing the computation explicitly, a p-value of 0.087 can be determined.

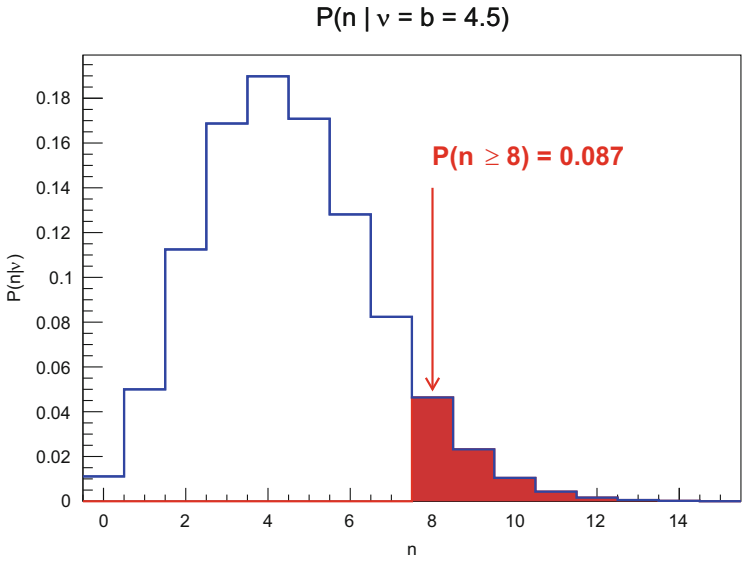

Fig. 8.1 Poisson distribution in the case of a null signal and an expected background of $b = 4.5$. The probability corresponding to $n \geq 8$ (*red area*) is 0.087, and gives a p-value assuming the event counting as test statistics

8.2.2 *Significance*

Instead of quoting the p-value, publications often preferred to quote the equivalent number of standard deviations that correspond to an area p under the rightmost tail

Table 8.1 Significances expressed as "$Z\sigma$" and corresponding p-values in a number of typical cases

$Z\,(\sigma)$	p
1.00	1.59×10^{-1}
1.28	1.00×10^{-1}
1.64	5.00×10^{-2}
2.00	2.28×10^{-2}
2.32	1.00×10^{-2}
3.00	1.35×10^{-3}
3.09	1.00×10^{-3}
3.71	1.00×10^{-4}
4.00	3.17×10^{-5}
5.00	2.87×10^{-7}
6.00	9.87×10^{-10}

of a normal distribution. So, one quotes a "$Z\sigma$" significance corresponding to a given p-value by using the following transformation (see Eq. (2.26)):

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = 1 - \Phi(Z) = \Phi(-Z) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{Z}{\sqrt{2}}\right)\right]. \qquad (8.1)$$

By convention in literature one claims the *"observation"* of the signal under investigation if the significance is at least 3σ ($Z = 3$), which corresponds to a probability of background fluctuation (p-value) of 1.35×10^{-3}. One claims the *"evidence of"* the signal (*discovery*) in case the significance is at least 5σ ($Z = 5$), corresponding to a p-value of 2.87×10^{-7}. Table 8.1 shows a number of typical significance values expressed as "$Z\sigma$" and their corresponding p-values.

8.2.3 Significance and Discovery

Determining the significance, anyway, is only part of the process that leads to a discovery, in the scientific method. Quoting from [1]:

> It should be emphasized that in an actual scientific context, rejecting the background-only hypothesis in a statistical sense is only part of discovering a new phenomenon. One's degree of belief that a new process is present will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data. Here, however, we only consider the task of determining the p-value of the background-only hypothesis; if it is found below a specified threshold, we regard this as "discovery".

In order to evaluate the "plausibility of a new signal" and other factors that give confidence in a discovery, the physicist's judgement cannot, of course, be replaced by the statistical evaluation only. In this sense, we can say that a Bayesian interpretation of the final result is somehow implicitly assumed.

8.3 Excluding a Signal Hypothesis

For the purpose of excluding a signal hypothesis, usually the requirement applied in terms of p-value is much milder than for a discovery. Instead of requiring a p-value of 2.87×10^{-7} or less (the "5σ" criterion), upper limits for a signal exclusion are set requiring $p < 0.05$, corresponding to a 95 % confidence level (CL) or $p < 0.10$, corresponding to a 90 % CL.

In the case of signal exclusion, p indicates the probability of a signal *underfluctuation*, i.e.: the null hypothesis and alternative hypothesis are inverted with respect to the case of a discovery.

8.4 Significance and Parameter Estimates Using Likelihood Ratio

In Sect. 7.7 the following test statistic was introduced in order to achieve the benefits of the Wilks' theorem (see Sect. 7.6):

$$\lambda(\mu, \vec{\theta}) = \frac{L_{s+b}(\vec{x}_1, \cdots, \vec{x}_N; \mu, \vec{\theta})}{L_b(\vec{x}_1, \cdots, \vec{x}_N; \vec{\theta})} \,. \tag{8.2}$$

In the case of a single parameter of interest μ (i.e.: no extra parameter θ is relevant, or equivalently we may set the number of extra parameters to zero: $m = 0$), it is possible to plot $-\ln \lambda(\mu)$ as a function of μ, and the presence of a minimum at $\mu = \hat{\mu}$ is an indication of the possible presence of a signal having a signal strength μ equal to $\hat{\mu}$ within some uncertainty.

In order to determine the significance of the measured signal yield, we can apply Wilks' theorem, since we have nested conditions, if the likelihood function is sufficiently regular. In this case, the distribution of $-2 \ln \lambda(\hat{\mu})$ can be approximated by a χ^2 distribution with one degrees of freedom, and the square root of its value at the minimum gives an approximate estimate of the significance Z:

$$Z = \sqrt{-2 \ln \lambda(\hat{\mu})} \,. \tag{8.3}$$

More in general, in case of a parameter of interest θ, e.g.: the mass of the new particle that produced the signal yield $\mu s(\theta)$, the same procedure may be applied and the value $-\ln \lambda(\theta)$ can be plotted as a function of θ. The same consideration may be done for what concerns the approximate significance evaluation as $Z = \sqrt{-2 \ln \lambda_{\max}}$ as a function of the unknown parameter θ.

This significance is called *local significance*, in the sense that it corresponds to a fixed value of the parameter θ. In Sect. 8.14 the interpretation of significance in the case of parameter estimates from data will be further discussed, and it will be more clear that the estimate of the local significance at a fixed value of a measured

parameter may suffer from a systematic overestimate (so called: *look-elsewhere effect*).

If the background PDF does not depend on θ (for instance, if θ is the mass of an unknown particle), $L_b(\vec{x};\theta)$ also does not depend on θ and the likelihood ratio $\lambda(\theta)$ is equal, up to a multiplicative factor, to the likelihood $L_{s+b}(\vec{x};\theta)$. Hence, the maximum-likelihood estimate of θ, $\theta = \hat{\theta}$, can also be determined by minimizing $-2\ln\lambda(\theta)$ and the error on θ can be determined as usual for maximum-likelihood estimates from the shape of $-2\ln\lambda(\theta)$ around its minimum, finding its intersection with an horizontal line at $-2\ln\lambda(\hat{\theta}) + 1$.

In addition, unlike the likelihood estimate, where the value of the likelihood function at the minimum does not provide any relevant information, the value of $-2\ln\lambda(\theta)$ at the minimum gives an approximate estimate of the significance Z thanks to the Wilks' theorem, as remarked above. This makes the use of likelihood ratio estimators interesting also for parameter estimate, not only to determine upper limits.

Combining the likelihood ratios of several measurements can be performed by multiplying the likelihood functions of individual channels to produce a combined likelihood function. In this way, combining a first measurement that has strong sensitivity to the signal with a second measurement that has low sensitivity gives, as combined test statistic, the product of likelihood ratios λ from both measurements. Since for the second measurement the $s + b$ and b hypothesis give similar values of the likelihood functions, given that the sensitivity to the signal of the second measurement is low, the likelihood ratio of the second additional measurement is close to one. Hence, the combined test statistics (the product of the two) does not differ much from one given by the first measurement only, and the corresponding sensitivity will not be worsened by the presence of the second measurement.

8.4.1 Significance Evaluation with Toy Monte Carlo

A better estimate of the significance may be achieved adopting the test statistics $-2\ln\lambda$ generating a large number of pseudoexperiment (*toy Monte Carlo*) representing randomly extracted samples that assume the presence of no signal ($\mu = 0$) in order to obtain the expected distribution of $-2\ln\lambda$.

The distribution of the generated $-2\ln\lambda$ can be used together with the observed value of $\lambda = \hat{\lambda}$ to determine the p-value which is the probability that λ is less or equal to the observed $\lambda = \hat{\lambda}$:

$$p = P_{s+b}(\lambda(\theta) \leq \hat{\lambda})\,, \tag{8.4}$$

equal to the fraction of generated pseudoexperiments for which $\lambda(\theta) \leq \hat{\lambda}$.

Note that, in order to assess large values of the significance, the number of generated toy Monte Carlo samples required to achieve a sufficient precision may

be very large because the required p-value is very small. Remember that a p-value for a 5σ evidence is (Table 8.1) 2.87×10^{-7}, hence, in order to determine it with sufficient precision the number of toy samples may be as large as $\sim 10^9$.

8.5 Definitions of Upper Limits

In the frequentist approach the procedure to set an upper limit is a special case of determination of the confidence interval (see Sect. 6.1) for the unknown signal yield s, or alternatively the signal strength μ.

In order to determine an upper limit instead of a central interval, the choice of the interval with the desired confidence level (90 % or 95 %, usually) should be fully asymmetric, becoming $s \in [0, s^{\text{up}}[$. When the outcome of an experiment is an upper limit, one usually quotes:

$$s < s^{\text{up}} \text{ at } 95\,\% \text{ C.L (or } 90\,\% \text{ CL)}\,.$$

If the Bayesian approach is adopted, the interpretation of an upper limit s^{up} is different and the interval $s \in [0, s^{\text{up}}[$ has to be interpreted as *credible interval*, meaning that its corresponding *posterior probability* is equal to the CL which is equal to $1 - \alpha$.

8.6 Poissonian Counting Experiments

A simple case, which is anyway realistic for may application, is a counting experiment where the number of observed event is the only information considered. The selected event sample contains a mixture of events due to signal and background processes.

The expected total number of events is $s + b$ where s and b are the expected number of signal and background events, respectively. Assuming the expected background b is known (e.g.: it could be estimated from theory or from a control data sample), the main unknown parameter of the problem is s, which could be equal to zero in case the signal is not present (null hypothesis). The likelihood function in the case of such a counting experiment is:

$$L(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}\,, \tag{8.5}$$

where n is the observed number of events.

8.6.1 Simplified Significance Evaluation for Counting Experiments

In case we use as single information the number of observed events n that we want to compare with an expected number of background events b, if b is sufficiently large and known with negligible uncertainty, the distribution of the number of events, assuming that background only is present in the data ($s = 0$), can be approximated to a Gaussian with average b and standard deviation equal to $\sqrt{b}$. The observed number of events n will be on average equal to b. An excess, quantified as $s = n - b$, should be compared with the expected standard deviation $\sqrt{b}$, and the significance can be approximately evaluated as:

$$Z = \frac{s}{\sqrt{b}} . \tag{8.6}$$

In case the expected background yield b comes from an estimate which has a non-negligible uncertainty σ_b, Eq. (8.6) can be modified as follows:

$$Z = \frac{s}{\sqrt{b + \sigma_b^2}} . \tag{8.7}$$

The above expressions clearly can only be applied in the particular case presented above, and have no general validity.

8.7 Bayesian Approach

The easiest treatment of a counting experiment, at least from the technical point of view, can be done under the Bayesian approach. The Bayesian posterior PDF for s is given by:

$$P(s|n) = \frac{L(n;s)\pi(s)}{\int_0^\infty L(n;s')\pi(s')\,\mathrm{d}s'} . \tag{8.8}$$

The upper limit s^{up} can be computed requiring that the posterior probability corresponding to the interval $[0, s^{\mathrm{up}}[$ is equal to CL, or equivalently that the probability corresponding to $[s^{\mathrm{up}}, \infty[$ is $\alpha = 1 - \mathrm{CL}$:

$$\boxed{\alpha = 1 - \mathrm{CL} = \int_{s^{\mathrm{up}}}^\infty P(s|n)\pi(s)\,\mathrm{d}s = \frac{\int_{s^{\mathrm{up}}}^\infty L(n;s)\pi(s)\,\mathrm{d}s}{\int_0^\infty L(n;s)\pi(s)\,\mathrm{d}s} .} \tag{8.9}$$

Apart from the technical aspects related to the computation of the integrals and the already mentioned arbitrariness in the choice of the prior $\pi(s)$ (see Sect. 3.6), the above expression poses no particular problem.

8.7.1 Bayesian Upper Limits for Poissonian Counting

In the simplest case of negligible background, $b = 0$ and assuming a uniform prior for s, the posterior PDF for s has the same expression as the Poissonian probability itself, as it was demonstrated in the Example 3.10:

$$P(s|n) = \frac{s^n e^{-s}}{n!}\,. \tag{8.10}$$

In the case of no observed events, i.e.: $n = 0$, we have:

$$P(s|0) = e^{-s}\,, \tag{8.11}$$

and:

$$\alpha = 1 - \mathrm{CL} = \int_{s^{\mathrm{up}}}^{\infty} e^{-s}\,\mathrm{d}s = e^{-s^{\mathrm{up}}}\,, \tag{8.12}$$

which gives:

$$s^{\mathrm{up}} = -\ln(\alpha)\,. \tag{8.13}$$

For $\alpha = 0.05$ (95 % CL) and $\alpha = 0.10$ (90 % CL), we can set the following upper limits:

$$s^{\mathrm{up}} = \ 3.00 \text{ at } \ 95\,\%\,\mathrm{CL}\,, \tag{8.14}$$

$$s^{\mathrm{up}} = \ 2.30 \text{ at } \ 90\,\%\,\mathrm{CL}\,. \tag{8.15}$$

The general case of a possible expected background $b \neq 0$ was treated by Helene [2], and Eq. (8.9) becomes:

$$\alpha = e^{-s^{\mathrm{up}}} \frac{\displaystyle\sum_{m=0}^{n} \frac{(s^{\mathrm{up}} + b)^m}{m!}}{\displaystyle\sum_{m=0}^{n} \frac{b^m}{m!}}\,. \tag{8.16}$$

Table 8.2 Upper limits in presence of negligible background evaluated under the Bayesian approach for different number of observed events n

n	$1-\alpha = 90\,\%$ s^{up}	$1-\alpha = 95\,\%$ s^{up}
0	2.30	3.00
1	3.89	4.74
2	5.32	6.30
3	6.68	7.75
4	7.99	9.15
5	9.27	10.51
6	10.53	11.84
7	11.77	13.15
8	12.99	14.43
9	14.21	15.71
10	15.41	19.96

The above expression can be inverted numerically to determine s^{up} for given α, n and b. In the case of no background ($b = 0$) Eq. (8.16) becomes:

$$\alpha = e^{-s^{\text{up}}} \sum_{m=0}^{n} \frac{s^{\text{up}\,m}}{m!}\,, \tag{8.17}$$

which gives Eq. (8.12) for $n = 0$.

The corresponding upper limits in case of no background for different number of observed events n are reported in Table 8.2. For different number of observed events n and different expected background b, the upper limits derived in [2] at 90 and 95 % CL are shown in Fig. 8.2.

8.7.2 Limitations of the Bayesian Approach

The derivation of Bayesian upper limits presented above assumes a uniform prior on the expected signal yield. Assuming a different prior distribution would result in different upper limits. In general, there is no unique criterion to chose a specific prior PDF to model the complete lack of knowledge about a variable, like in this case the signal yield, as it was already discussed in Sect. 3.6. In case of search for new signals, sometimes the signal yield is related to other parameters of the theory (e.g.: the mass of unknown particles, or specific coupling constants). In that case, should one chose a uniform prior for the signal yield or a uniform prior for the theory parameters? The choice is not obvious and no prescription may be provided from first principles.

A possible approach is to chose more prior PDFs that reasonably model one's ignorance about the unknown parameters and to verify that the obtained upper limit is not too sensitive to the choice of the prior.

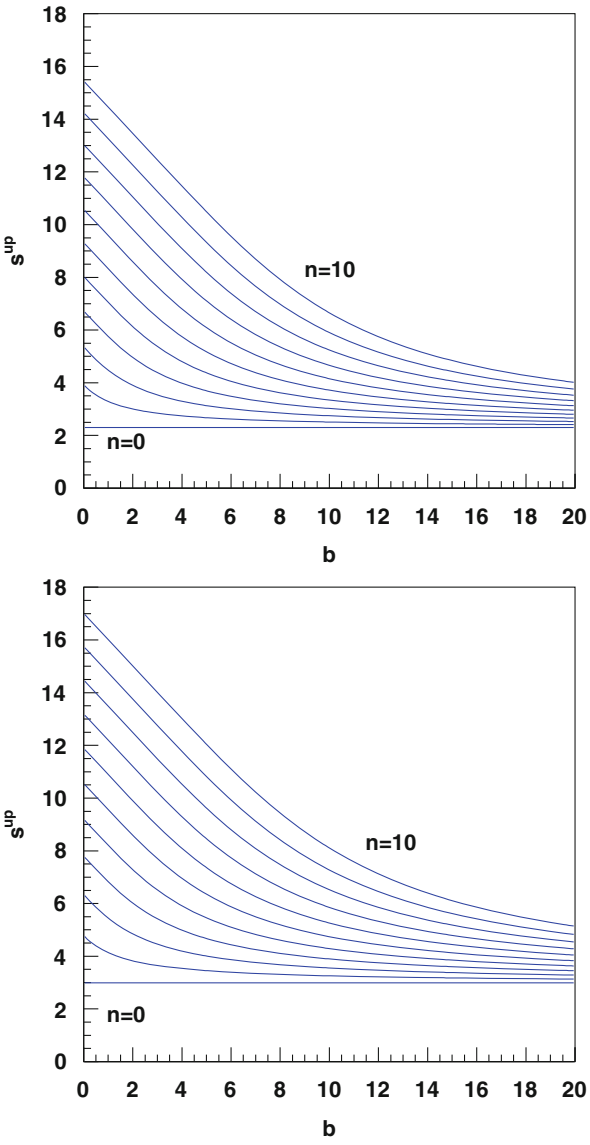

Fig. 8.2 Upper limits at the 90 % CL (*top*) and 95 % CL (*bottom*) for Poissonian process using the Bayesian approach as a function of the expected background b and for number of observed events n from $n = 0$ to $n = 10$

8.8 Frequentist Upper Limits

Frequentist upper limits can be computed rigorously by inverting the Neyman belt (see Sect. 6.1). In cases where the confidence interval is fully asymmetric, i.e.: it has the form $[0, s^{\mathrm{up}}[$, the result may be quoted as:

$$s < s^{\mathrm{up}} ,$$

at the given confidence level. In order to better interpret the upper limits obtained from the Neyman belt inversion in the general case, a simple example of a counting experiment will be analyzed first, similarly to what was considered in Sect. 8.7.1 under the Bayesian approach.

8.8.1 The Counting Experiment Case

Let's assume an experiment counts a number n of selected events when expecting b events from background and s from signal. Let's also assume for simplicity, as in Sect. 8.7.1, that the expected background is negligible: $b \simeq 0$. The probability to observe n events when we expect s events is given by a Poisson distribution:

$$P(n;s) = \frac{e^{-s}s^n}{n!}\,. \tag{8.18}$$

We can set an upper limit on the expected signal yield s by *excluding* the values of s for which the probability to observe n events or less (p-value) is below the value $\alpha = 1 - \mathrm{CL}$. For $n = 0$ we have:

$$P(0;s) = e^{-s}\,, \tag{8.19}$$

and the p-value is equal to e^{-s}. Setting $p > \alpha = 1 - \mathrm{CL}$ allows values of the expected signal yield s that satisfy:

$$p = s^{-s} > \alpha = 1 - \mathrm{CL}\,. \tag{8.20}$$

The above relation can be inverted, and gives:

$$s < -\ln\alpha = s^{\mathrm{up}}\,, \tag{8.21}$$

which, for $\alpha = 5\,\%$ or $\alpha = 10\,\%$ gives:

$$s < 3.00 \;\text{ at }\; 95\,\%\,\mathrm{CL}, \tag{8.22}$$

$$s < 2.30 \;\text{ at }\; 90\,\%\,\mathrm{CL}. \tag{8.23}$$

Those results coincide accidentally with the results obtained under the Bayesian approach.

The coincidence of the values of limits computed under the Bayesian and frequentist approaches for this simplified but commonly used case may lead to confusion. There is no intrinsic reason for which limits evaluated under the two approaches should coincide, and in general, with very few exceptions, like in this case, Bayesian and frequentist limits don't coincide numerically. Moreover,

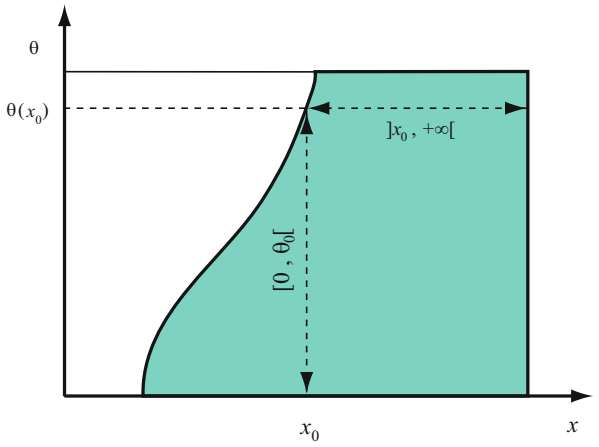

Fig. 8.3 Graphical illustration of Neyman's belt construction for upper limits determination

regardless of their numerical coincidence like in this case, the interpretation of Bayesian and frequentist limits is very different, as already discussed several times.

8.8.2 *Upper Limits from Neyman's Confidence Intervals*

In Sect. 6.1 Neyman's confidence intervals construction was presented. Upper or lower limits on an unknown parameter θ can be determined using fully asymmetric intervals for the observed quantity x. In particular, assuming that the Neyman belt is monotonically increasing, the choice of intervals $]x_1(\theta_0), +\infty[$ for x leads to a confidence interval $[0, \theta(x_0)[$ for θ which corresponds to the upper limit:

$$\theta < \theta^{\text{up}} = \theta(x_0)\,. \tag{8.24}$$

This case is illustrated in Fig. 8.3, which is the equivalent of Fig. 6.1 when adopting a fully asymmetric interval.

8.8.3 *Frequentist Upper Limits on Discrete Variables*

In the case of a discrete observed variable n, like for a Poissonian counting experiment, when constructing the Neyman's belt it's not always possible to find an interval $\{n_1, \ldots, n_k\}$ that has the exact desired coverage because of the intrinsic discreteness of the problem. The only possibility in such cases is to take the smallest interval having a probability greater or equal to the desired CL. The upper limit determined in those cases is *conservative*, i.e. the procedure ensures that the

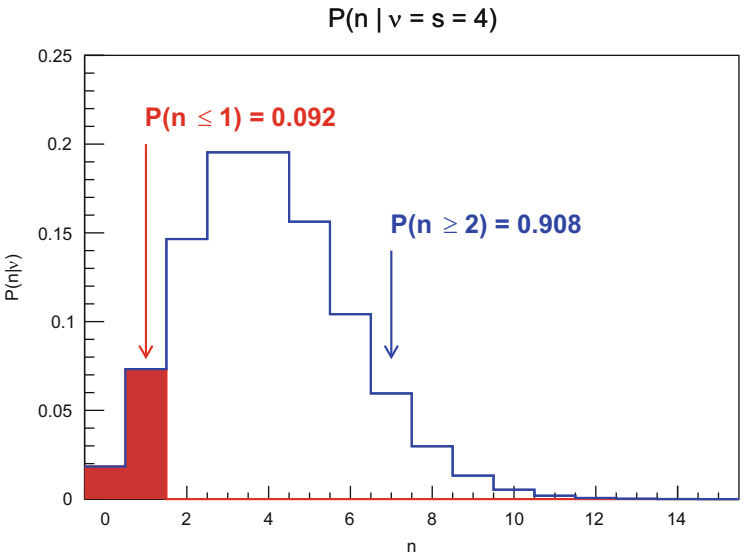

Fig. 8.4 Poisson distribution in the case of a signal $s = 4$ and $b = 0$. The *white bins* show the smallest possible fully asymmetric confidence interval ($\{2, 3, 4, \cdots\}$ in this case) that gives at least the coverage of $1 - \alpha = 90\,\%$

probability that the true value s lies within the determined confidence interval $[s_1, s_2]$ is *greater or equal* to CL $= 1 - \alpha$ (*overcoverage*).

Figure 8.4 shows an example of Poisson distribution corresponding to the case with $s = 4$ and $b = 0$. Using a fully asymmetric interval as ordering rule, the interval $\{2, 3, \cdots\}$ of the discrete variable n corresponds to a probability $P(n \geq 2) = 1 - P(0) - P(1) = 0.9084$, and is the smallest interval which has a probability greeter or equal to a desired CL of 0.90: the interval $\{3, 4, \cdots\}$ would have a probability $P(n \geq 3)$ less than 90 %, while enlarging the interval to $\{1, 2, \cdots\}$ would produce a probability $P(n \geq 1)$ larger than $P(n \geq 2)$.

Given an observation of n events, we could set the upper limit s^{up} such that:

$$s^{\text{up}} = \min_{\sum_{m=0}^{n} P(m;s) < \alpha} (s)\,. \tag{8.25}$$

In the simplest case where $n = 0$, we have:

$$s^{\text{up}} = \min_{P(0;s) < \alpha} (s) = \min_{e^{-s} < \alpha} (s) = -\ln \alpha\,, \tag{8.26}$$

hence, again we find the same result that was derived in Sect. 8.8.1 with a simplified approach:

$$s^{\text{up}} = -\ln(\alpha)\,. \tag{8.27}$$

This leads, for $\alpha = 0.05$ (95 % CL) or $\alpha = 0.1$ (90 % CL),

$$s^{\text{up}} = 2.00 \text{ at } 95\,\%\text{CL}\,, \tag{8.28}$$

$$s^{\text{up}} = 2.30 \text{ at } 90\,\%\text{CL}\,. \tag{8.29}$$

From the purely frequentist point of view, anyway, this result suffers from the flip-flopping problem discussed in Sect. 6.3: if we decide a priori to quote an upper limit as our final result the procedure to chose a fully asymmetric interval in the Neyman's construction leads to the correct overage. But if we choose to switch from fully asymmetric to central interval in case we observe a significant signal, this would lead to an incorrect coverage.

8.8.4 *Feldman–Cousins Upper Limits for Counting Experiments*

The Feldman–Cousins (FC) approach (see Sect. 6.4) may be a solution for this case.

In the Poissonian counting case, the 90 % confidence belt obtained with the FC approach is shown in Fig. 8.5 for the case in which $b = 3$. The results in the case of no background ($b = 0$) are reported in Table 8.3 for different values of the number of observed events n. Figure 8.6 shows the value of the 90 % CL upper limit computed using the FC approach as a function of the expected background b for different values of n.

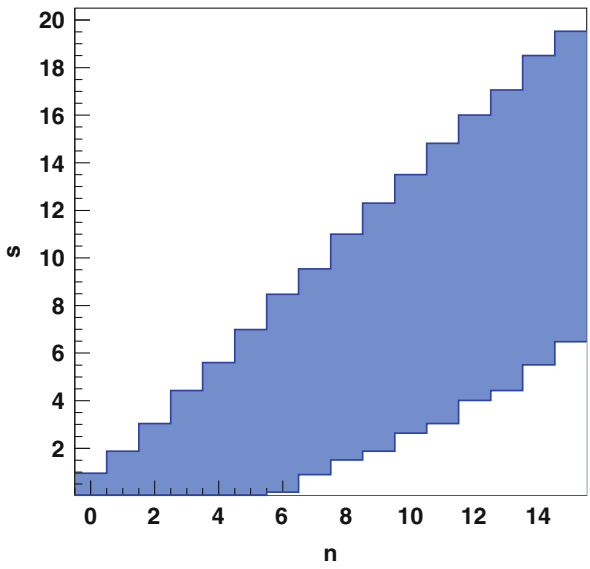

Fig. 8.5 90 % Confidence belt for a Poissonian process using Feldman–Cousins ordering, in the case of $b = 3$

Table 8.3 Upper and lower limits in presence of negligible background ($b = 0$) obtained using the Feldman–Cousins approach

	$1-\alpha = 90\,\%$		$1-\alpha = 95\,\%$	
n	s^{lo}	s^{up}	s^{up}	s^{lo}
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.99	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

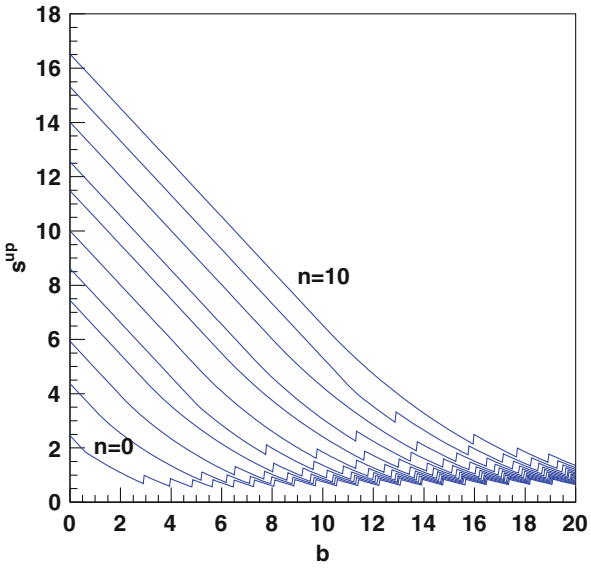

Fig. 8.6 Upper limits at 90 % confidence belt for Poissonian process using Feldman–Cousins ordering as a function of the expected background b and for number of observed events n from 0 to 10

Comparing Table 8.3 with Table 8.2, which reports the Bayesian results, FC upper limits are in general numerically larger than Bayesian limits. In particular, for the case $n = 0$, the upper limit increases from 2.30 to 2.44 for a 90 % CL and from 3.00 to 3.09 for a 95 % CL. Anyway, as remarked before, the interpretation of frequentist and Bayesian limits is very different, and the numerical comparison of those upper limits should not lead to any specific interpretation.

A peculiar feature of FC upper limits is that, for $n = 0$, a larger expected background b corresponds to a more stringent, i.e.: lower, upper limit, as can be seen in Fig. 8.6 (lowest curve). This feature is absent in Bayesian limits that do not depend on the expected background b for $n = 0$ (see Fig. 8.2).

This dependence of upper limits on the expected amount of background is somewhat counterintuitive: imagine two experiments (say A and B) performing a search for a rare signal designed to achieve a very low background level (say 0.01 and 0.1 events for A and B, respectively). If both measure zero counts, which is for both the most likely outcome, the experiment that achieves the most stringent limit is the one which has the highest expected background level (B in this case), i.e.: the one which has the worse expected performances (0.1 expected events for B, vs 0.01 for A).

The Particle Data Group published in their review [3] the following sentence about the interpretation of frequentist upper limits, in particular for what concerns the difficulty to interpret a more stringent limit for an experiment with worse expected background, in the case of no event observed ($n = 0$):

> The intervals constructed according to the unified [Feldman Cousins] procedure for a Poisson variable n consisting of signal and background have the property that for $n = 0$ observed events, the upper limit decreases for increasing expected background. This is counter-intuitive, since it is known that if $n = 0$ for the experiment in question, then no background was observed, and therefore one may argue that the expected background should not be relevant. The extent to which one should regard this feature as a drawback is a subject of some controversy.

This feature of frequentist limits is often considered unpleasant by physicists. The need to agree on a common procedure to determine upper limits, mainly triggered by the need to combine the results of the four LEP experiments on Higgs boson search [4], led to the proposal of a new method that modifies the purely frequentist approach, as will be discussed in the following Section. This comes to some extent from a generalization of the attempt performed by G. Zech discussed in Sect. 8.9 below.

8.9 Can Frequentist and Bayesian Upper Limits Be "Unified"?

The coincidence of Bayesian and frequentist upper limits in the simplest event counting case motivated an effort attempted by Zech [5] to conciliate the two approaches, namely the limits obtained by Helene in [2] and the frequentist approach.

In order to determine the probability distribution of the number of events from the sum of two Poissonian processes with s and b expected number of events from signal and background, respectively, one can write, using Eq. (2.49), the probability distribution for the total observed number of events n as:

$$P(n; s, b) = \sum_{n_b=0}^{n} \sum_{n_s=0}^{n-n_b} P(n_b; b) P(n_s; s) \,. \tag{8.30}$$

Zech proposed to modify the first term of Eq. (8.30), $P(n_b; b)$, to take into account that the observation of n events should put a constraint on the possible values of n_b, which can only range from 0 to n. In this way, Zech attempted to replace $P(n_b; b)$ with:

$$P'(n_b; b) = P(n_b; b) / \sum_{n'_b=0}^{n} P(n'_b; b) \,. \tag{8.31}$$

This modification leads to the same result obtained by Helene in Eq. (8.16), which apparently indicates a possible convergence of Bayesian and frequentist approaches.

This approach was later criticized by Highland and Cousins [6] who demonstrated that the modification introduced by Eq. (8.31) produced an incorrect coverage, and Zech himself admitted the non rigorous application of the frequentist approach [7].

This attempt could not demonstrate a way to conciliate the Bayesian and frequentist approaches, which, as said, have completely different interpretations. Anyway, Zech's intuition anticipated the formulation of the *modified frequentist approach* that will be discussed in Sect. 8.10, which is nowadays widely used in High Energy Physics.

8.10 Modified Frequentist Approach: The CL_s Method

The concerns about frequentist limits discussed at the end of Sect. 8.8.4 have been addressed with the definition of a procedure that was adopted for the first time for the combination of the results of the search for the Higgs boson [4] of the four LEP experiments, Aleph, Delphi, Opal and L3.

The modification of the purely frequentist confidence level with the introduction of a conservative corrective factor can cure, as will be presented in the following, the aforementioned counterintuitive peculiarities of the frequentist limit procedure in case of no observed signal events.

The so-called *modified frequentist approach* will be illustrated using the test statistic adopted in the original proposal, which is the ratio of the likelihood functions evaluated under two different hypotheses: the presence of signal plus background (H_1, corresponding to the likelihood function L_{s+b}), and the presence of background only (H_0, corresponding to the likelihood function L_b):

$$\lambda(\vec{\theta}) = \frac{L_{s+b}(\vec{x}; \vec{\theta})}{L_b(\vec{x}; \vec{\theta})} \,. \tag{8.32}$$

Different test statistics have been applied after the original definition of the LEP procedure, but the method described in the following remains unchanged for all different kinds of test statistics.

If we consider, in addition to the pure counting of the number of events N, a set of n variables $\vec{x} = (x_1, \cdots, x_n)$ that characterize each event which are measured for N events $(\vec{x}_1, \cdots \vec{x}_N)$, as it was already introduced in Sect. 7.7, the ratio of the extended likelihood functions can be written as (Eq. 7.35):

$$\lambda(\mu, \vec{\theta}) = e^{-\mu s(\vec{\theta})} \prod_{i=1}^{N} \left(\frac{\mu s(\vec{\theta}) f_s(\vec{x}_i; \vec{\theta})}{b(\vec{\theta}) f_b(\vec{x}_i; \vec{\theta})} + 1 \right), \tag{8.33}$$

where the functions f_s and f_b are the PDFs for signal and background of the variables $\vec{x}$. The negative logarithm of the test statistic is (Eq. 7.36):

$$-\ln \lambda(\mu, \vec{\theta}) = \mu s(\vec{\theta}) - \sum_{i=1}^{N} \ln \left(\frac{\mu s(\vec{\theta}) f_s(\vec{x}_i; \vec{\theta})}{b(\vec{\theta}) f_b(\vec{x}_i; \vec{\theta})} + 1 \right). \tag{8.34}$$

In order to quote an upper limit using the frequentist approach, the distribution of the test statistics λ (or equivalently $-2 \ln \lambda$) in the hypothesis of signal plus background $(s + b)$ has to be known, and the p-value corresponding to the observed value $\lambda = \hat{\lambda}$ (denoted below as CL_{s+b}) has to be determined.

The proposed modification to the purely frequentist approach consists of finding two p-values corresponding to both the $s+b$ and b hypotheses (below, for simplicity of notation, the set of parameters $\vec{\theta}$ also includes μ):

$$CL_{s+b}(\vec{\theta}) = P_{s+b}(\lambda(\vec{\theta}) \leq \hat{\lambda}), \tag{8.35}$$

$$CL_b(\vec{\theta}) = P_b(\lambda(\vec{\theta}) \leq \hat{\lambda}). \tag{8.36}$$

From those two probabilities, the following quantity can be derived:

$$\boxed{CL_s(\vec{\theta}) = \frac{CL_{s+b}(\vec{\theta})}{CL_b(\vec{\theta})}.} \tag{8.37}$$

Upper limits are determined excluding the range of the parameters of interest (e.g.: the signal strength μ or a new particle's mass) for which $CL_s(\vec{\theta})$ is lower than the conventional exclusion confidence level, typically 95 % or 90 %. For this reason, the modified frequentist approach is often referred to as the *CL_s method*.

In most of the cases, the probabilities P_{s+b} and P_b in Eqs. (8.35) and (8.36) are not trivial to obtain analytically and are determined numerically using randomly generated pseudoexperiments, or *toy Monte Carlo*. In this way, CL_{s+b} and CL_b can be estimated as the fraction of generated pseudoexperiments for which $\lambda(\vec{\theta}) \leq \hat{\lambda}$, assuming the presence of both signal and background, or background only, respectively. An example of the outcome of this numerical approach is shown in Fig. 8.7.

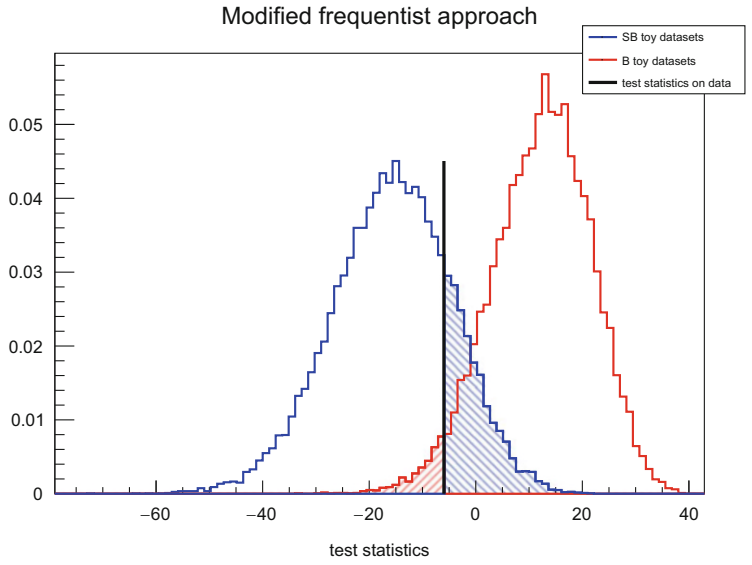

Fig. 8.7 Example of determination of CL_s from pseudoexperiments. The distribution of the test statistics $-2\ln\lambda$ is shown in *blue* assuming the signal-plus-background hypothesis and in red assuming the background-only hypothesis. The *black line* shows the value of the test statistics measured in data, and the hatched areas represent CL_{s+b} (*blue*) and $1-\mathrm{CL}_b$ (*red*)

This method does not produce the desired (90 % or 95 %, usually) coverage from the frequentist point of view, but does not suffer from the problematic features of frequentist upper limits that were observed at the end of Sect. 6.4.

The CL_s method has convenient statistical properties:

- It is conservative from the frequentist point of view. In fact, since $\mathrm{CL}_b \leq 1$, we have that $\mathrm{CL}_s(\vec{\theta}) \geq \mathrm{CL}_{s+b}(\vec{\theta})$. So, it *overcovers*; this means that a CL_s upper limit is less stringent than a purely frequentist limit.
- Unlike the upper limits obtained using the Feldman–Cousins approach, if no signal event is observed ($n = 0$), the CL_s upper limit does not depend on the expected amount of background.

For a simple Poissonian counting experiment with expected signal s and a background b, using the likelihood ratio from Eq. (7.39), it is possible to demonstrate analytically that the CL_s approach leads to a result identical to the Bayesian one (Eq. 8.16) and to what was derived by Zech (see Sect. 8.9). In general, it turns out that in many realistic applications, the CL_s upper limits are numerically very similar to Bayesian upper limits computed assuming a uniform prior. But of course the Bayesian interpretation of upper limits cannot be applied to limits obtained using the CL_s approach.

On the other hand, the interpretation of limits obtained using the CL_s method is not obvious, and it does not match neither the frequentist nor the Bayesian approaches. CL_s limits have been defined as [8]:

> approximation to the confidence in the signal hypothesis one might have obtained if the experiment had been performed in the complete absence of background.

8.11 Incorporating Nuisance Parameters and Systematic Uncertainties

Some of the parameters in the set considered so far, $\vec{\theta} = (\theta_1, \cdots, \theta_m)$, are not of direct interest for our measurement, but are needed to model unknown characteristics of our data sample. Those parameters are defined *nuisance parameters* (Sect. 5.2). Nuisance parameters may appear, for instance, when the yield and/or the distribution of the observed background is estimated with some uncertainty from simulation or from control samples in data, or when the modeling of distributions for signal and/or background events are not (perfectly) known, e.g.: to take into account the effect of detector response (resolution, calibration, etc.).

If we are only interested in the measurement of the signal strength μ only, all other parameters θ_i can be considered as nuisance parameters. In case, for instance, we are also interested in the measurement of the mass of a new particle, like the Higgs boson's mass, the parameter corresponding to the particle mass, say $m = \theta_1$, is, like μ, a *parameter of interest* and all other remaining parameters $\theta_2, \cdots, \theta_m$ can be considered nuisance parameters.

More in general, let's divide the parameter set in two subsets: the parameters of interest, $\vec{\theta} = (\theta_1, \cdots, \theta_h)$, and the nuisance parameters, $\vec{\nu} = (\nu_1, \cdots, \nu_l)$.

8.11.1 Nuisance Parameters with the Bayesian Approach

The treatment of nuisance parameters is a well defined task under the Bayesian approach and was already discussed in Sect. 3.4. The posterior joint probability distribution for all the unknown parameters can be defined as follows (Eq. 3.32):

$$P(\vec{\theta}, \vec{\nu}|\vec{x}) = \frac{L(\vec{x}; \vec{\theta}, \vec{\nu})\pi(\vec{\theta}, \vec{\nu})}{\int L(\vec{x}; \vec{\theta}', \vec{\nu}')\pi(\vec{\theta}', \vec{\nu}')\, \mathrm{d}^h\theta'\, \mathrm{d}^l\nu'}\,, \tag{8.38}$$

where $L(\vec{x}; \vec{\theta}, \vec{\nu})$ is the as usual the likelihood function and $\pi(\vec{\theta}, \vec{\nu})$ is the prior distribution of the unknown parameters.

The probability distribution of $\vec{\theta}$ can be obtained as marginal PDF, integrating the joint PDF over all nuisance parameters:

$$P(\vec{\theta}|\vec{x}) = \int P(\vec{\theta}, \vec{\nu}|\vec{x})\, \mathrm{d}^l\nu = \frac{\int L(\vec{x}; \vec{\theta}, \vec{\nu})\pi(\vec{\theta}, \vec{\nu})\, \mathrm{d}^l\nu}{\int L(\vec{x}; \vec{\theta}', \vec{\nu}')\pi(\vec{\theta}', \vec{\nu}')\, \mathrm{d}^h\theta' \mathrm{d}^l\nu'} \,. \tag{8.39}$$

The problem is well defined, and the only difficulty is the numerical integration in multiple dimensions. Several algorithms can be adopted for the evaluation of Bayesian integrals. A particularly performant algorithm in those cases is the Markov-chain Monte Carlo [9].

8.11.2 Hybrid Treatment of Nuisance Parameters

The treatment of nuisance parameters under the frequentist approach is more difficult to perform rigorously. Cousins and Highlands [10] proposed to adopt the same approach used for the Bayesian treatment to determine approximate likelihood functions for the signal-plus-background and the background-only hypotheses. The hybrid likelihood functions can be written, integrating Eqs. (7.31) and (7.34), as:

$$\begin{aligned} &L_{s+b}(\vec{x}_1, \cdots, \vec{x}_k|\mu, \vec{\theta}) \\ &\quad = \frac{1}{n!} \int e^{-\left(\mu s(\vec{\theta},\vec{\nu}) + b(\vec{\theta},\vec{\nu})\right)} \prod_{i=1}^{n} \Big(\mu s(\vec{\theta}, \vec{\nu}) f_s(\vec{x}_i; \vec{\theta}, \vec{\nu}) + b(\vec{\theta}, \vec{\nu}) f_b(\vec{x}_i; \vec{\theta}, \vec{\nu})\Big)\, \mathrm{d}^l\nu \,, \end{aligned} \tag{8.40}$$

$$L_b(\vec{x}_1, \cdots, \vec{x}_k|\vec{\theta}) = \frac{1}{n!} \int e^{-b(\vec{\theta},\vec{\nu})} b(\vec{\theta}, \vec{\nu})^n \prod_{i=1}^{n} f_b(\vec{x}_i; \vec{\theta}, \vec{\nu})\, \mathrm{d}^l\nu \,. \tag{8.41}$$

In order to include detector resolution effects, for instance to model the width of a signal peak, the hybrid approach requires the convolution of the likelihood function with the experimental resolution function.

The above likelihood functions can be used to compute CL_s limits, as it was done in the combined Higgs limit at LEP [4].

This *hybrid* Bayesian-frequentist approach does not ensure an exact frequentist coverage, and it may tend to undercover in some cases [11]. It has been proven on simple models to provide results that are numerically close to Bayesian evaluation performed assuming a uniform prior [12].

8.11.2.1 Event Counting Uncertainties

In the case of an event counting problem, if the number of background events is known with some uncertainty, the PDF of the background estimate b' can be modeled as a function of the true unknown expected background b, $P(b';b)$. The likelihood functions, which depend on the parameter of interest s (we don't use the signal strength μ here for simplicity) and the unknown nuisance parameter b, can be written as:

$$L_{s+b}(n,b';s,b) = \frac{(s+b)^n}{n!}e^{-(s+b)}P(b';b)\,, \tag{8.42}$$

$$L_b(n,b';b) = \frac{b^n}{n!}e^{-b}P(b';b)\,. \tag{8.43}$$

In order to eliminate the dependence on the nuisance parameter b, the hybrid likelihoods, using the Cousins–Highlands can be written as:

$$L_{s+b}(n,b';s) = \int_0^\infty \frac{(s+b)^n}{n!}e^{-(s+b)}P(b';b)\,\mathrm{d}b\,, \tag{8.44}$$

$$L_b(n,b') = \int_0^\infty \frac{b^n}{n!}e^{-b}P(b';b)\,\mathrm{d}b\,. \tag{8.45}$$

In the simplified case, for instance when $P(b';b)$ is a Gaussian function, the integration can be performed analytically [13]. In this case, when the standard deviation of the distribution is not much smaller than b', $P(b';b)$ extends to negative values of b, and the integration includes unphysical regions with negative signal yields. For this reason, the above equations can be safely used only when the probability to have negative values of b are negligible.

In order to avoid such cases, the use of distributions whose range is limited to positive values is preferred. For instance, a log-normal distribution (see Sect. 2.7) is usually preferred to a plain Gaussian.

8.12 Upper Limits Using the Profile Likelihood

A procedure that accounts for the treatment of nuisance parameters avoiding the hybrid Bayesian approach adopts as test statistic the *profile likelihood* defined in Eq. (7.27):

$$\lambda(\mu) = \frac{L(\vec{x}|\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\vec{x}|\hat{\mu},\hat{\hat{\vec{\theta}}})}\,. \tag{8.46}$$

Above, in Eq. (8.46), $\hat{\mu}$ and $\hat{\vec{\theta}}$ are the best fit values of μ and $\vec{\theta}$ corresponding to the observed data sample, and $\hat{\hat{\vec{\theta}}}(\mu)$ is the best fit value for $\vec{\theta}$ obtained for a fixed value of μ. Above we assumed that all parameters are treated as nuisance parameter and μ is the only parameter of interest.

The profile likelihood is introduced in order to satisfy the conditions required to apply Wilks' theorem (see Sect. 7.6).

A scan of $-2\ln\lambda(\mu)$ as a function of μ reveals a minimum at the value $\mu = \hat{\mu}$, where $-2\ln\lambda(\hat{\mu}) = 0$ by construction. As for the usual likelihood function, the excursion of $-2\ln\lambda(\mu)$ around the minimum and the intersection of the corresponding curve with a straight line corresponding to $-2\ln\lambda(\mu) = 1$ allow to determine an uncertainty interval for μ similarly to what was discussed in Sect. 5.6. According to Wilks' theorem (see Sect. 7.6) if μ corresponds to the true value, then the distribution of $-2\ln\lambda(\mu)$ follow a χ^2 distribution with one degree of freedom.

Usually the addition of nuisance parameters $\vec{\theta}$ has the effect of broadening the shape of the profile likelihood function as a function of the parameter of interest μ compared with cases where the nuisance parameters are kept fixed. This effect increases the uncertainty on μ when systematic uncertainties are included in the test statistic.

Compared with the Cousins–Highland hybrid treatment of nuisance parameters, the profile likelihood offers several advantages, including a CPU-faster numerical implementation because it requires no numerical integration, which may be more time-consuming than the minimizations required to compute the profile likelihood.

Given that the profile likelihood is based on a likelihood ratio, according to the Neyman–Pearson lemma (see Sect. 7.3) it has optimal performances for what concerns the test of the two hypotheses assumed in the numerator and in the denominator of Eq. (8.46).

The test statistics $t_\mu = -2\ln\lambda(\mu)$ can be used to determine p-values p_μ corresponding to the various hypotheses on μ. Those p-values can be computed in general by generating sufficiently large toy Monte Carlo samples.

8.13 Variations of Profile-Likelihood Test Statistics

Different variation in the definition of the profile likelihood have been proposed and adopted for various data analysis cases. A review of most of the adopted test statistics is presented in [1] where approximate formulae have been provided to compute significances in asymptotic limits. Some examples are reported below.

8.13.1 Test Statistic for Positive Signal Strength

In order to enforce the condition $\mu \geq 0$, since we can't have negative yields from new signals, the test statistic $t_\mu = -2 \ln \lambda(\mu)$ can be modified as follows:

$$\tilde{t}_\mu = -2 \ln \tilde{\lambda}(\mu) = \begin{cases} -2 \ln \frac{L(\vec{x}|\mu, \hat{\hat{\theta}}(\mu))}{L(\vec{x}|\hat{\mu}, \hat{\hat{\theta}})} & \hat{\mu} \geq 0\,, \\ -2 \ln \frac{L(\vec{x}|\mu, \hat{\hat{\theta}}(\mu))}{L(\vec{x}|0, \hat{\hat{\theta}}(0))} & \hat{\mu} < 0\,. \end{cases} \tag{8.47}$$

In practice, one limits the best-fit value for μ to values greater or equal to zero.

8.13.2 Test Statistics for Discovery

In order to asses the presence of a new signal, one wants to test the case of a positive signal strength μ against the case $\mu = 0$. This can be done using the test statistic t_0 which is equal to $t_\mu = -2 \ln \lambda(\mu)$ evaluated for $\mu = 0$. The test statistic t_μ, anyway, may reject the hypothesis $\mu = 0$ also in case of a downward fluctuations in the data that would yield a negative best-fit value μ. A modification of t_0 has been proposed in order to be only sensitive to an excess in data that yields a significantly high value of $\hat{\mu}$ [1]:

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0\,, \\ 0 & \hat{\mu} < 0\,. \end{cases} \tag{8.48}$$

The p-value p_0 corresponding to the test statistic q_0 can be evaluated using toy Monte Carlo samples generated assuming background-only events. The distribution of q_0 will have a spike at $q_0 = 0$ (a Dirac's delta $\delta(q_0)$ component) corresponding to all cases which give a negative $\hat{\mu}$.

8.13.3 Test Statistics for Upper Limits

Similarly to the definition of q_0, in order to set an upper limit on μ one doesn't want to exclude a given value of μ if an upward fluctuation in data occurred, i.e.: if $\hat{\mu} > \mu$. In order to avoid those cases, the following modification of t_μ has been proposed:

$$q_\mu = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu\,, \\ 0 & \hat{\mu} > \mu\,. \end{cases} \tag{8.49}$$

Here the distribution of q_μ will present a spike at $q_\mu = 0$ corresponding to those cases where $\hat{\mu} > \mu$.

8.13.4 Higgs Test Statistic

A modified version of q_μ defined above has been adopted for Higgs search at LHC. The modified test statistics is:

$$\tilde{q}_\mu = \begin{cases} -2\ln \frac{L(\vec{x}|\mu,\hat{\hat{\theta}}(\mu))}{L(\vec{x}|0,\hat{\hat{0}})} & \hat{\mu} < 0\,, \\ -2\ln \frac{L(\vec{x}|\mu,\hat{\hat{\theta}}(\mu))}{L(\vec{x}|\hat{\mu},\hat{\hat{\theta}})} & 0 \le \hat{\mu} \le \mu\,, \\ 0 & \hat{\mu} > \mu\,. \end{cases} \tag{8.50}$$

Above, the constant introduced for $\hat{\mu} < 0$, as for q_0, protects against unphysical values of the signal strength, while the cases which have an upward fluctuations of the data, such to give $\hat{\mu} > \mu$, are not considered as evidence against the signal hypothesis with signal strength equal to a considered value of μ, and the test statistics is set to zero in those cases, in order not to spoilt the upper limit performances of this test statistic.

8.13.5 Asymptotic Approximations

For the definition of the test statistics $\tilde{q}_\mu$, as well for the most adopted variations of the profile likelihood, asymptotic approximations which use Wilks' theorem and approximate formulae by Wald [14] have been computed and are treated extensively in [1].

Using the test statistic for discovery q_0, the asymptotic approximation gives the following simplified expression for the significance:

$$Z_0 \simeq \sqrt{q_0}\,. \tag{8.51}$$

The asymptotic approximation for the distribution of $\tilde{q}_\mu$ is:

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2\sqrt{2\pi}} \frac{1}{\sqrt{\tilde{q}_\mu}} e^{-\tilde{q}_\mu/2} & 0 < \tilde{q}_\mu \le \mu^2/\sigma^2\,, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)} \exp\left[-\frac{1}{2}\frac{(\tilde{q}_\mu+\mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}\right] & \tilde{q}_\mu > \mu^2/\sigma^2\,, \end{cases} \tag{8.52}$$

where $\delta(\tilde{q}_\mu)$ is a Dirac delta function, to model the cases in which the test statistics is set to zero, and where $\sigma^2 = \mu^2/q_{\mu,A}$, in which $q_{\mu,A}$ is the value of the test statistics $-2\ln\lambda$ evaluated on the so-called *Asimov set* [15]. The Asimov set is a *representative* data set in which the yields of all data samples are set to their expected values and nuisance parameter are set at their nominal value.

Asimov sets can also be used to compute approximate estimates of expected experimental sensitivity, which would require the extraction of a large number of

pseudoexperiments. The square roots of the test statistics evaluated at the Asimov data sets corresponding to the assumed signal strength μ can be used to approximate the median significance, assuming a data sample distributing according to the background-only hypothesis:

$$\text{med}[Z_\mu|0] = \sqrt{\tilde{q}_{\mu,A}}\,. \tag{8.53}$$

For a comprehensive treatment of asymptotic approximations, again, use [1].

Example 8.22 – A Realistic Example of Bump Hunting

The following example will present a classic "bump-hunting" case. In particular, we will try to estimate a signal's significance using a variation of the profile-likelihood technique. Eventually, in the following Example 8.23, the systematic uncertainty on the expected background yield will be added.

A randomly generated sample has been used to mimic a realistic data sample and it will be compared with two hypotheses:

1. background only, assuming an exponential model;
2. background plus signal, adding on top of the exponential background a Gaussian signal.

The two cases are shown in Fig. 8.8, superimposed to the data histogram.

The likelihood function for the binned case can be built using a Poisson distribution for the number of entries in each bin $n_1, \cdots, n_{n_{\text{bins}}}$:

$$L(\vec{n}|\vec{s},\vec{b},\mu,\beta) = \prod_{i=1}^{n_{\text{bins}}} \text{Pois}(n_i|\beta b_i + \mu s_i)\,. \tag{8.54}$$

In Eq. (8.54) the distribution for the signal is modeled as $\vec{s} = (s_i, \cdots, s_{\text{bins}})$ and for the background as $\vec{b} = (b_i, \cdots, b_{\text{bins}})$. The normalization of $\vec{s}$ and $\vec{b}$ are given by theory expectations, and variations of the normalization scale can be modeled by the extra parameter β and μ. μ is the *signal strength*, and has been introduced in several cases before. β has the same role as μ for the background yield instead of the signal.

In order to measure the signal yield, the signal strength μ should be determined, which is the parameter of interest of the problem. β, instead, can be considered as a nuisance parameter. Let's assume for the moment $\beta = 1$, i.e.: the uncertainty on the expected background yield from theory can be considered negligible. We can use as test statistics:

$$q = -2\ln\frac{L(\vec{n}|\vec{s},\vec{b},\mu,\beta=1)}{L(\vec{n}|\vec{s},\vec{b},\mu=0,\beta=1)}\,. \tag{8.55}$$

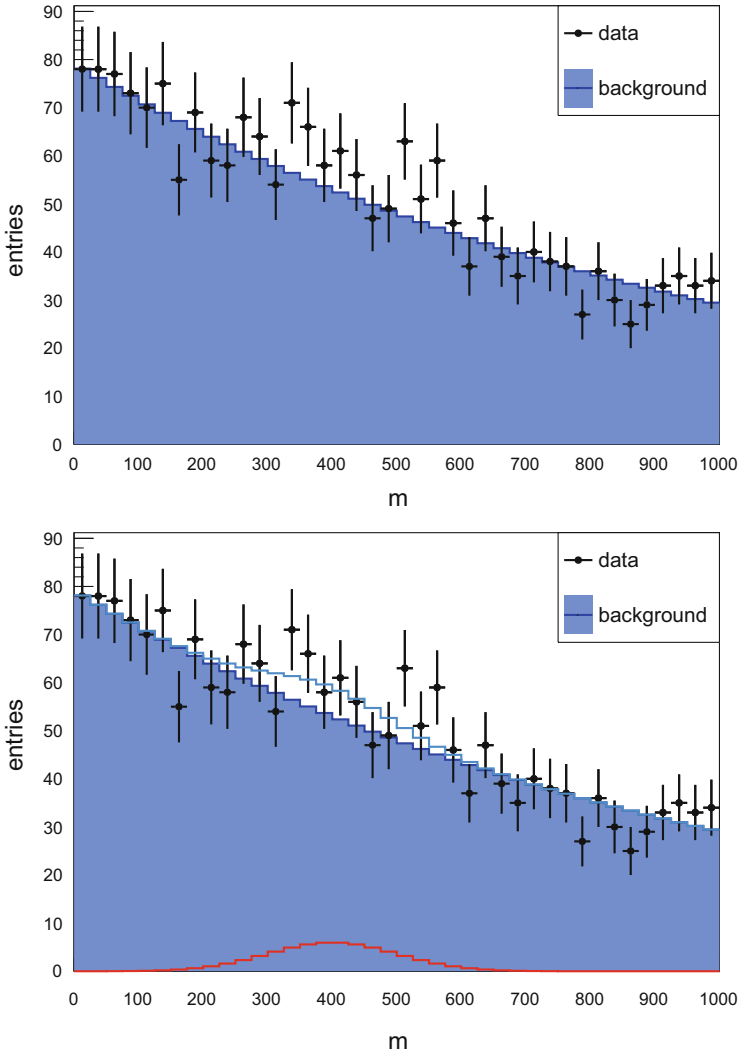

Fig. 8.8 A toy Monte Carlo data sample superimposed to an exponential background model (*left*) and to an exponential background model plus a Gaussian signal (*right*)

Note that this is somewhat different from the profile likelihood in Eq. (8.46), and we used here the ratio of the likelihood functions L_{s+b}/L_b.

This test statistic was used at Tevatron for several searches for new physics. Note that compared with the ordinary profile likelihood (Eq. 8.46) this test statistic can be written as:

$$q = -2\ln\lambda(1) + 2\ln\lambda(0)\,. \tag{8.56}$$

The distributions of the test statistic t for the background-only and for the signal-plus-background hypotheses are shown in Fig. 8.9. The figure has been produced generating 100,000 randomly-extracted toy samples in the two hypotheses.

The p-value corresponding to the background-only hypothesis can be evaluated as the fraction of randomly-extracted toy samples having a value of the test statistic lower than the one observed in data.

From the distributions shown in Fig. 8.9, 375 out of 100,000 toy samples have a value of q below the one in our data sample, hence the p-value is 3.7 %. Assuming a binomial uncertainty, the p-value can be determined with an uncertainty of 0.2 %.

Using Eq. (8.1) to convert the p-value into a significance level, this corresponds to a significance $Z = 2.7$.

An alternative way to determine the significance is to look at the scan of the test statistics as a function of the parameter of interest μ. This can be seen in Fig. 8.10 which shows that the minimum value of q is reached for $\mu = 1.24$.

Considering the range of μ where q exceeds the minimum value by not more than 1, the asymmetric uncertainty interval for μ can be determined as $\hat{\mu} = 1.24^{+0.49}_{-0.48}$. Indeed, the interval exhibits a very small asymmetry.

The minimum value of q can be used to determine the significance in the asymptotic approximation. If the null hypothesis ($\mu = 0$, assumed in the denominator of Eq. (8.55)) is true, than the Wilks' theorem holds, and we have the approximated expression $Z \simeq \sqrt{-q_{\min}} = 2.7$, in agreement with the estimate obtained with the toy generation.

Note that in Fig. 8.10 the test statistic is zero for $\mu = 0$ and reaches a negative minimum for $\mu = \hat{\mu}$, while the profile likelihood defined in Eq. (8.46) has a minimum value of zero.

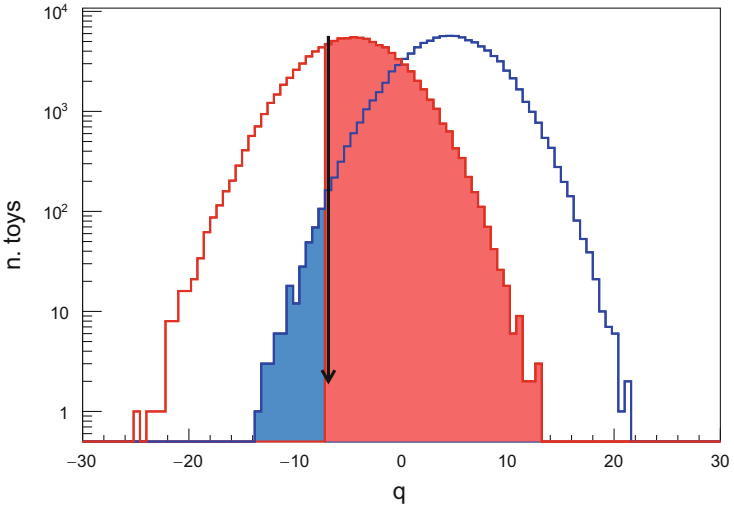

Fig. 8.9 Distribution of the test statistic q for the background-only hypothesis (*blue*) and for the signal-plus-background hypothesis (*red*). Superimposed is the value determined with the presented data sample (*black arrow*). p-values can be determined from the *shaded areas* of the two PDFs

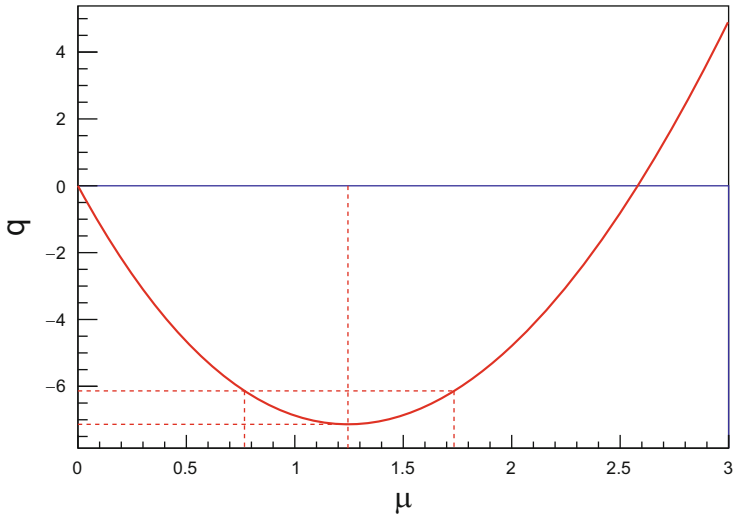

Fig. 8.10 Scan of the test statistic q as a function of the parameter of interest μ

Example 8.23 – Adding a Systematic Uncertainty

Let's assume, continuing with the Example 8.22, that we know the background normalization with a finite uncertainty. A 10 % uncertainty corresponds to an estimate of the nuisance parameter $\beta = \beta' \pm \delta\beta = 1.0 \pm 0.1$. The extreme cases where $\beta = 0.9$ or $\beta = 1.1$ are shown in Fig. 8.11.

The test statistic can be modified as follows:

$$q = -2 \ln \frac{\sup\limits_{0.9 \le \beta \le 1.1} L(\vec{n}|\vec{s}, \vec{b}, \mu, \beta)}{\sup\limits_{0.9 \le \beta \le 1.1} L(\vec{n}|\vec{s}, \vec{b}, \mu = 0, \beta)} \,. \tag{8.57}$$

The scan of the new test statistic is shown in Fig. 8.12. Compared with the case where no uncertainty was included, the shape of the test-statistic scan is now broader and the minimum is less deep. This results in a larger uncertainty: $\mu = 1.40^{+0.61}_{-0.60}$ and a smaller significance: $Z = 2.3$.

Note that this approaches considers implicitly a uniform distribution of the estimate β' of the parameter β, while a different PDF could be adopted. In that case, the test statistic can be modified as follows:

$$q = -2 \ln \frac{\sup\limits_{-\infty < \beta < +\infty} L(\vec{n}|\vec{s}, \vec{b}, \mu, \beta, \beta')}{\sup\limits_{-\infty < \beta < +\infty} L(\vec{n}|\vec{s}, \vec{b}, \mu = 0, \beta, \beta')} \,, \tag{8.58}$$

where the likelihood function should now be written as:

$$L(\vec{n}|\vec{s}, \vec{b}, \mu, \beta, \beta') = \prod_{i=1}^{n_{\text{bins}}} \text{Pois}(n_i|\beta b_i + \mu s_i) P(\beta'|\beta) \,, \tag{8.59}$$

and the term $P(\beta'|\beta)$ is the PDF for the estimated background yield β', given the true value β. Typical options for $P(\beta'|\beta)$ are a Gaussian distribution, with the problem that it could also lead to negative values of β', or a lognormal distribution (see Sect. 2.7), which constrains β' to be positive.

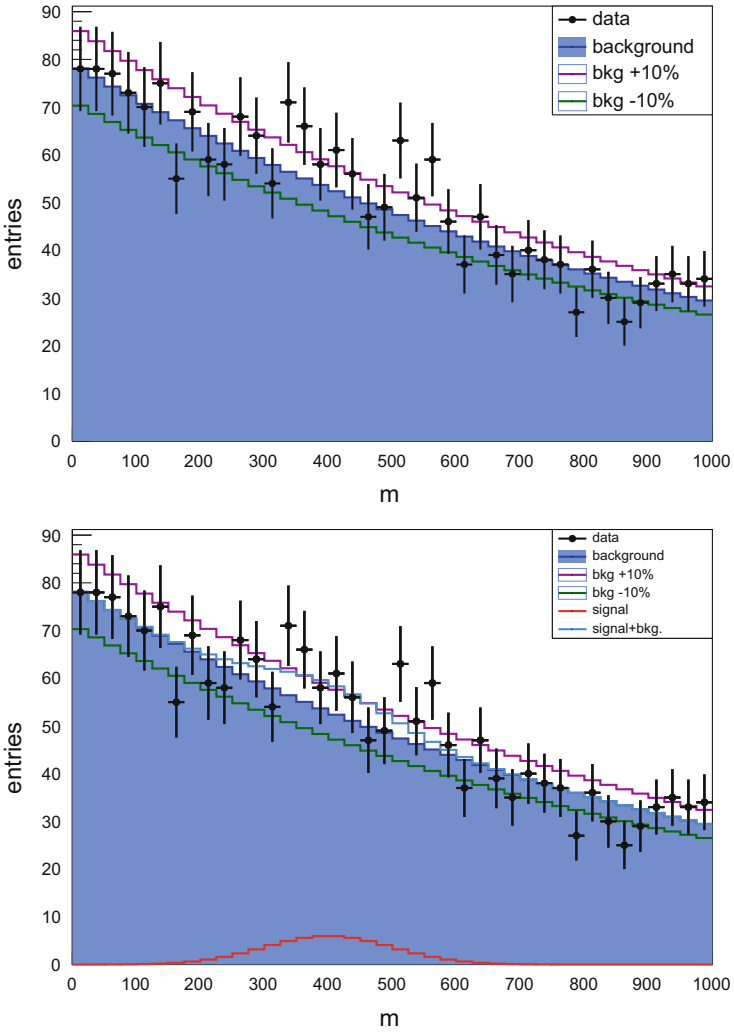

Fig. 8.11 Toy data sample superimposed to an exponential background model (*top*) and to an exponential background model plus a Gaussian signal (*bottom*) adding a 10 % uncertainty to the background normalization

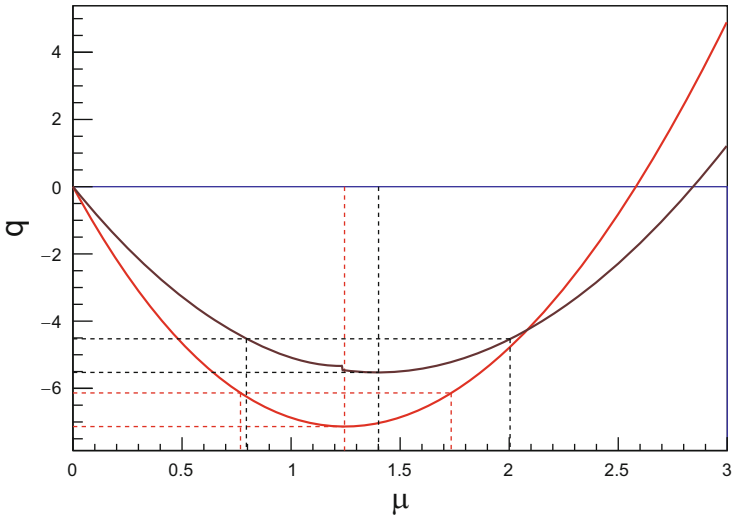

Fig. 8.12 Scan of the test statistic q as a function of the parameter of interest μ including systematic uncertainty on β (*dark brown*) compared with the case with no uncertainty (*red*)

8.14 The Look-Elsewhere Effect

In several cases experiments look for a peak in the distribution of an observable variable, typically a reconstructed mass of a particle, but the location of the peak is not known a priori. This was the case, for instance, for the search for the Higgs boson at LHC, or any search for new resonances. This is not the case, instead, for the search for a rare decay of a known particle (imagine $B_s \to \mu^+\mu^-$ at LHC).

If an excess in data compared with the background expectation is found at *any* mass value, the excess could be interpreted as a possible signal of a new resonance at the observed mass. Anyway, the peak could be produced either by the presence of a real signal or by a background fluctuation.

The computation of the signal significance can be done using the p-value of the measured test statistics q assuming a fixed value m_0 of the resonance mass. This is called *local significance*, and can be written as:

$$p(m_0) = \int_{q_{\text{obs}(m_0)}}^{\infty} f(q(m_0)|\mu = 0)\, \mathrm{d}q\,, \tag{8.60}$$

where $f(q|\mu)$ is the PDF of the adopted test statistics q for a given value of the signal strength μ.

The local significance gives the probability corresponding to a background fluctuation at a fixed value of the mass m_0. The probability to have a background overfluctuation at *any* mass value, called *global* p-value, is larger than the local p-

value, which would be an underestimate if taken as the probability of a background fluctuation at *any* mass value in the range of interest.

The magnitude of the effect is larger as the mass resolution gets worse. In fact, assuming a small intrinsic width of the new resonance, a very good mass resolution implies that a peak can appear from a background fluctuation only if values of the reconstructed mass for different background events are by chance all close each other within the experimental resolution, which is less likely as the resolution is smaller.

More in general, when an experiment is looking for a signal where one or more parameters $\vec{\theta}$ are unknown (e.g.: could be both the mass and the width or other properties of a new signal), in the presence of an excess in data with respect to the background expectation, the unknown parameter (or parameters) can be determined from the data sample itself. In those cases, the local significance, expressed in terms of a p-value computed at fixed values of the unknown parameter set $\vec{\theta}_0$, is an underestimate of the global significance.

The global p-value can be computed using as test statistics the largest value of the estimator over the entire parameter range:

$$q(\hat{\vec{\theta}}) = \max_{\substack{\theta_i^{\min} < \theta_i < \theta_i^{\max}, \\ i=1,\cdots,m}} q(\vec{\theta})\,. \tag{8.61}$$

The distribution of $q(\hat{\vec{\theta}})$ from Eq. (8.61) is not easy to determine, and usually CPU-intensive random extraction of pseudo experiments are needed. Note that in this case the Wilks' theorem cannot be applied because the values of the parameters $\vec{\theta}$ determined from data is not defined for the case where $\mu = 0$, hence the hypotheses assumed at the numerator and the denominator are not "nested".

In order to determine a significance close to the discovery level of 5σ, p-values smaller than the 3×10^{-7} need to be evaluated, hence tens of millions of pseudo experiment representing the background-only case need to be generated.

8.14.1 Trial Factors

An approximate way to determine the global significance taking into account the look-elsewhere effect is reported in [16], relying on the asymptotic behavior of likelihood-ratio estimators. It is possible to demonstrate [17] that the probability that the test statistic $q(\hat{m})$ is larger than a given value u, i.e.: the value of the global p-value, is bound by:

$$p^{\text{glob}} = P(q(\hat{m}) > u) \leq \langle N_u \rangle + P(\chi^2 > u)\,, \tag{8.62}$$

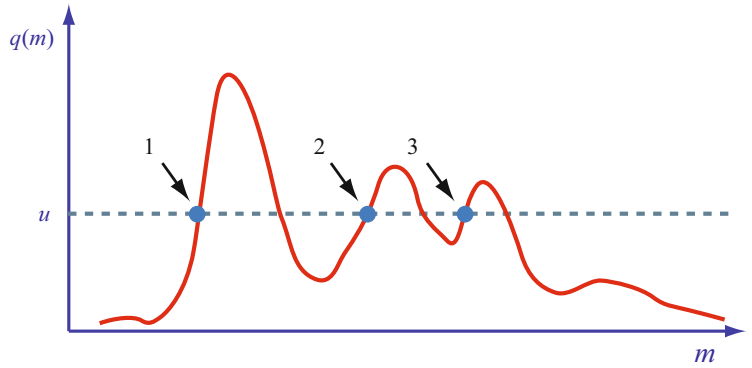

Fig. 8.13 Visual illustration of upcrossing, computed to determine $\langle N_{u_0} \rangle$. In this example, we have $N_u = 3$

where $P(\chi^2 > u)$ comes from an asymptotic approximation of the distribution of the local test statistics $q(m)$ as a χ^2 distribution with one degree of freedom. The term $\langle N_u \rangle$ in Eq. (8.62), instead, is the average number of *upcrossings*, i.e. the average number of times the curve $q = q(m)$ crosses an horizontal line at a given level $q = u$ with a positive derivative, which can be evaluated using toy Monte Carlo. This is visualized in an example in Fig. 8.13.

The value of $\langle N_u \rangle$ could be very small, depending on the level u. Fortunately, a scaling law exists, so, starting from a different level u_0 one can extrapolate $\langle N_{u_0} \rangle$ as:

$$\langle N_u \rangle = \langle N_{u_0} \rangle \, e^{-(u-u_0)/2} \, . \tag{8.63}$$

This allows to evaluate $\langle N_{u_0} \rangle$ generating a number of pseudo experiment much smaller than what would be needed to determine $\langle N_u \rangle$ with comparable precision.

In the case of more than one parameter the number of upcrossing is replaced by the *Euler characteristic* given by the number of disconnected components minus the number of "holes" in the multidimensional sets of the parameter space defined by $q(m) > u$ [18].

References

1. G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics. Eur. Phys. J. **C71**, 1554 (2011)
2. O. Helene, Upper limit of peak area. Nucl. Instrum. Methods **A212**, 319 (1983)
3. C. Amsler et al., The review of particle physics. Phys. Lett. **B667**, 1 (2008)
4. G. Abbiendi et al., Search for the standard model Higgs boson at LEP. Phys. Lett. **B565**, 61–75 (2003)
5. G. Zech, Upper limits in experiments with background or measurement errors. Nucl. Instrum. Methods **A277**, 608 (1989)

6. V. Highland, R. Cousins, Comment on "upper limits in experiments with background or measurement errors" [Nucl. Instrum. Methods **A277**, 608–610 (1989)]. Nucl. Instrum. Methods **A398**, 429 (1989)
7. G. Zech, Reply to "comment on "upper limits in experiments with background or measurement errors" [Nucl. Instrum. Methods **A277**, 608–610 (1989)]". Nucl. Instrum. Methods, **A398**, 431 (1989)
8. A. Read, Modified frequentist analysis of search results (the CL_s method), in *1st Workshop on Confidence Limits*, (CERN) (2000)
9. B. Berg, *Markov Chain Monte Carlo Simulations and Their Statistical Analysis* (World Scientific, Singapore, 2004)
10. R. Cousins, V. Highland, Incorporating systematic uncertainties into an upper limit. Nucl. Instrum. Methods, **A320**, 331–335 (1992)
11. V. Zhukov, M. Bonsch, Multichannel number counting experiments, in *Proceedings of PHYSTAT2011* (2011)
12. C. Blocker, Interval estimation in the presence of nuisance parameters: 2. Cousins and highland method. CDF/MEMO/STATISTICS/PUBLIC/7539 (2006)
13. L. Lista, Including gaussian uncertainty on the background estimate for upper limit calculations using Poissonian sampling. Nucl. Instrum. Methods **A517**, 360 (2004)
14. A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. **54**, 426–482 (1943)
15. I. Asimov, Franchise, in *The Complete Stories*, ed. by I. Asimov, vol. 1 (Broadway Books, New York, 1990)
16. E. Gross, O. Vitells, Trial factors for the look elsewhere effect in high energy physics. Eur. Phys. J. **C70**, 525 (2010)
17. R. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **74**, 33 (1987)
18. O. Vitells, Look elsewhere effect, in *Proceedings of PHYSTAT2011* (2011)